**ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

**Logo, company name

Description automatically generated**

**BÁO CÁO ĐỒ ÁN TRÍ TUỆ NHÂN TẠO**

**DỰ ĐOÁN XU HƯỚNG GIÁ CỔ PHIẾU GOOGLE DỰA TRÊN DỮ LIỆU GIÁ VÀ PHÂN TÍCH CẢM XÚC TIN TỨC BẰNG MACHINE LEARNING**

**Giảng viên hướng dẫn: Nguyễn An Khương**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Tên Sinh Viên** | **Đóng góp** |
| 2310251 | Phan Gia Bảo | 100% |

Thành phố Hồ Chí Minh, ngày 13 tháng 12, năm 2025

Link Github:

Link slide:

Link thuyết trình:

## **MỤC LỤC**

# CHƯƠNG 1. GIỚI THIỆU 1.1. Động lực và bối cảnh nghiên cứu 1.2. Phát biểu bài toán 1.3. Mục tiêu nghiên cứu 1.4. Phạm vi và giả định

# CHƯƠNG 2. TỔNG QUAN NGHIÊN CỨU LIÊN QUAN 2.1. Dự đoán xu hướng giá cổ phiếu dựa trên dữ liệu giá 2.2. Phân tích cảm xúc từ tin tức tài chính 2.3. Mô hình kết hợp dữ liệu giá và tin tức 2.4. Nhận xét và định hướng của đề tài

# CHƯƠNG 3. MÔ TẢ DỮ LIỆU (DATASET DESCRIPTION) 3.1. Nguồn dữ liệu  3.1.1. Dữ liệu giá cổ phiếu  3.1.2. Dữ liệu tin tức 3.2. Thống kê dữ liệu  3.2.1. Quy mô và khoảng thời gian  3.2.2. Phân bố dữ liệu và mất cân bằng lớp 3.3. Tiền xử lý dữ liệu  3.3.1. Làm sạch dữ liệu và xử lý giá trị thiếu  3.3.2. Xử lý và tổng hợp dữ liệu tin tức  3.3.3. Kết hợp dữ liệu giá và tin tức  3.3.4. Gán nhãn dữ liệu  3.3.5. Feature Engineering

# CHƯƠNG 4. XÂY DỰNG ĐẶC TRƯNG (FEATURE ENGINEERING) 4.1. Đặc trưng dựa trên giá và lợi suất 4.2. Đặc trưng xu hướng và động lượng 4.3. Đặc trưng độ biến động 4.4. Đặc trưng hình nến kỹ thuật 4.5. Đặc trưng cảm xúc từ tin tức 4.6. Đặc trưng trễ (Lag Features) 4.7. Bảng mô tả các đặc trưng

# CHƯƠNG 5. PHƯƠNG PHÁP VÀ MÔ HÌNH (METHODOLOGY) 5.1. Pipeline Machine Learning tổng thể 5.2. Mô hình sử dụng  5.2.1. Định nghĩa bài toán học máy  5.2.2. Mô hình chính – XGBoost  5.2.3. Mô hình so sánh – Random Forest 5.3. Huấn luyện và đánh giá mô hình  5.3.1. Thiết lập huấn luyện  5.3.2. Kết quả mô hình XGBoost  5.3.3. Ứng dụng dự đoán thực tế

# CHƯƠNG 6. ĐÁNH GIÁ MÔ HÌNH (EVALUATION) 6.1. Các thước đo đánh giá 6.2. Kết quả định lượng  6.2.1. Kết quả mô hình Random Forest  6.2.2. Kết quả mô hình XGBoost  6.2.3. Cross-Validation theo chuỗi thời gian 6.3. Trực quan hóa kết quả  6.3.1. Confusion Matrix  6.3.2. ROC Curve 6.4. Tổng kết đánh giá

# CHƯƠNG 7. PHÂN TÍCH VÀ THẢO LUẬN (DISCUSSION) 7.1. So sánh hiệu năng giữa các mô hình 7.2. Vai trò của đặc trưng sentiment từ tin tức 7.3. Ảnh hưởng của phân bố dữ liệu và mất cân bằng lớp 7.4. Overfitting và Underfitting 7.5. Hạn chế của nghiên cứu 7.6. Tổng kết thảo luận

# CHƯƠNG 8. HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN 8.1. Hạn chế của nghiên cứu 8.2. Hướng phát triển trong tương lai

# CHƯƠNG 9. NHỮNG GÌ HỌC ĐƯỢC TỪ PROJECT 9.1. Kiến thức về Machine Learning / Deep Learning 9.2. Tầm quan trọng của tiền xử lý dữ liệu 9.3. Đánh giá mô hình không chỉ dựa vào Accuracy 9.4. Kết nối giữa code và tư duy nghiên cứu 9.5. Bài học cá nhân

# CHƯƠNG 10. KẾT LUẬN

# TÀI LIỆU THAM KHẢO

## **1. Giới thiệu (Introduction)**

### 1.1. Động lực và bối cảnh nghiên cứu

Thị trường chứng khoán được xem là một hệ thống phức tạp, phi tuyến và chịu ảnh hưởng đồng thời từ nhiều yếu tố khác nhau như biến động giá trong quá khứ, hành vi nhà đầu tư và thông tin từ truyền thông. Việc dự đoán xu hướng giá cổ phiếu vì vậy luôn là một bài toán khó nhưng có ý nghĩa thực tiễn cao trong lĩnh vực tài chính và đầu tư. Nhiều nghiên cứu đã chỉ ra rằng các phương pháp thống kê truyền thống gặp hạn chế khi xử lý dữ liệu có tính phi tuyến và nhiễu cao như dữ liệu tài chính [1], [2].

Trong những năm gần đây, các phương pháp Machine Learning (ML) và Deep Learning (DL) đã được áp dụng rộng rãi trong tài chính nhờ khả năng học các mối quan hệ phức tạp từ dữ liệu lớn. Đặc biệt, các mô hình học máy cho dữ liệu bảng (tabular data) và mô hình xử lý ngôn ngữ tự nhiên (NLP) đã cho thấy hiệu quả tích cực trong các bài toán dự đoán xu hướng giá cổ phiếu và phân tích cảm xúc thị trường [3], [5].

### 1.2. Phát biểu bài toán

Trong nghiên cứu này, bài toán được xây dựng dưới dạng bài toán phân loại xu hướng giá cổ phiếu theo ngày. Dữ liệu đầu vào bao gồm hai nguồn chính: dữ liệu giá cổ phiếu lịch sử và dữ liệu tin tức liên quan đến doanh nghiệp. Các nghiên cứu trước đây đã chứng minh rằng thông tin tin tức và cảm xúc thị trường có ảnh hưởng đáng kể đến biến động giá cổ phiếu [3], [6].

Đầu ra của bài toán là nhãn phân loại thể hiện xu hướng biến động giá của cổ phiếu trong ngày giao dịch tiếp theo, bao gồm các trạng thái tăng hoặc giảm giá.

### 1.3. Mục tiêu nghiên cứu

Mục tiêu chính của nghiên cứu là xây dựng một pipeline Machine Learning hoàn chỉnh, từ thu thập dữ liệu, tiền xử lý, trích xuất đặc trưng cho đến huấn luyện và đánh giá mô hình dự đoán xu hướng giá cổ phiếu. Bên cạnh đó, nghiên cứu tập trung đánh giá vai trò của các đặc trưng cảm xúc được trích xuất từ tin tức, tương tự như các hướng tiếp cận kết hợp dữ liệu giá và dữ liệu văn bản đã được đề xuất trong các công trình trước [3], [5].

### 1.4. Phạm vi và giả định

Nghiên cứu chỉ tập trung vào một cổ phiếu duy nhất là Google (GOOGL) và sử dụng dữ liệu theo tần suất ngày. Các yếu tố kinh tế vĩ mô như lãi suất, chỉ số thị trường chung hoặc các sự kiện kinh tế toàn cầu không được xem xét trong phạm vi nghiên cứu này, tương tự cách tiếp cận thu gọn trong một số nghiên cứu thực nghiệm trước đây [2].

# **2. Tổng quan nghiên cứu liên quan (Related Work)**

Dự đoán xu hướng giá cổ phiếu là một chủ đề đã được nghiên cứu rộng rãi trong lĩnh vực tài chính định lượng và trí tuệ nhân tạo. Các nghiên cứu trước đây có thể được phân loại thành ba hướng chính: (i) dự đoán dựa trên dữ liệu giá lịch sử (price-based prediction), (ii) dự đoán dựa trên phân tích cảm xúc từ tin tức (news-based sentiment analysis), và (iii) các mô hình lai kết hợp cả dữ liệu giá và tin tức (hybrid models). Phần này trình bày tổng quan các hướng tiếp cận tiêu biểu liên quan trực tiếp đến project.

### 2.1. Dự đoán dựa trên dữ liệu giá (Price-based Prediction)

Hướng nghiên cứu dự đoán giá cổ phiếu dựa trên dữ liệu giá lịch sử là hướng tiếp cận truyền thống và được nghiên cứu rộng rãi. Các mô hình Machine Learning như Logistic Regression, Random Forest hay Support Vector Machine đã được sử dụng để khai thác các đặc trưng kỹ thuật như moving average, volatility và momentum [2]. Ngoài ra, các mô hình Deep Learning như LSTM cũng được áp dụng để nắm bắt tính phụ thuộc theo thời gian của chuỗi giá [1].

Tuy nhiên, các phương pháp chỉ dựa trên dữ liệu giá thường bỏ qua tác động của thông tin bên ngoài thị trường, đặc biệt là tin tức và tâm lý nhà đầu tư.

### 2.2. Phân tích cảm xúc từ tin tức (News-based Sentiment Analysis)

Với sự phát triển của NLP, nhiều nghiên cứu đã tập trung vào việc khai thác tin tức tài chính để dự đoán xu hướng thị trường. Ding et al. [6] cho thấy rằng các sự kiện được trích xuất từ tin tức có thể cải thiện đáng kể hiệu quả dự đoán giá cổ phiếu. Sự ra đời của các mô hình ngôn ngữ lớn như BERT [4] và các biến thể chuyên biệt cho lĩnh vực tài chính như FinBERT [5] đã giúp nâng cao độ chính xác trong phân tích cảm xúc tài chính.

### 2.3. Mô hình kết hợp (Hybrid Models)

Các nghiên cứu gần đây có xu hướng kết hợp dữ liệu giá và dữ liệu tin tức nhằm tận dụng thông tin từ cả hai nguồn. Hu et al. [3] đề xuất một framework học sâu kết hợp dữ liệu chuỗi thời gian và dữ liệu văn bản, cho thấy hiệu quả vượt trội so với các mô hình chỉ sử dụng một nguồn dữ liệu. Hướng tiếp cận hybrid này cũng là động lực chính cho phương pháp được sử dụng trong nghiên cứu này.

## 2.4. Nhận xét và định hướng của project

Dựa trên các nghiên cứu liên quan, project này lựa chọn hướng tiếp cận mô hình lai, kết hợp dữ liệu giá cổ phiếu và dữ liệu tin tức. Cụ thể, sentiment từ tin tức được trích xuất bằng các mô hình NLP hiện đại, trong khi dữ liệu giá được xử lý và đưa vào các mô hình Machine Learning như Logistic Regression và XGBoost.

Khác với nhiều nghiên cứu tập trung vào độ chính xác tối đa, project này chú trọng xây dựng một pipeline Machine Learning hoàn chỉnh, có khả năng tái lập, dễ mở rộng và được đánh giá bằng nhiều thước đo khác nhau như Accuracy, Precision, Recall, F1-score và ROC-AUC. Cách tiếp cận này giúp đánh giá toàn diện hiệu quả của mô hình cũng như chỉ ra các điểm mạnh và hạn chế trong thực tế.

# **3. Mô tả dữ liệu (Dataset Description)**

## 3.1. Nguồn dữ liệu

Project sử dụng hai nguồn dữ liệu chính: dữ liệu giá cổ phiếu và dữ liệu tin tức, được thu thập và xử lý độc lập trước khi kết hợp thành một tập dữ liệu thống nhất phục vụ cho quá trình huấn luyện mô hình.

### 3.1.1. Dữ liệu giá cổ phiếu

Dữ liệu giá cổ phiếu của công ty Google được thu thập từ Yahoo Finance, bao gồm các thông tin theo đơn vị thời gian là ngày (daily data). Các thuộc tính cơ bản của dữ liệu giá bao gồm:

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

Tổng số bản ghi thu thập được từ Yahoo Finance là 252 mẫu, tương ứng với khoảng thời gian từ ngày 10/10/2024 đến 10/10/2025. Do thị trường chứng khoán không giao dịch vào cuối tuần và các ngày lễ, số lượng mẫu thực tế thấp hơn số ngày trong năm.

### 3.1.2. Dữ liệu tin tức

Dữ liệu tin tức được thu thập thông qua việc crawl từ GDELT (Global Database of Events, Language, and Tone). GDELT là một nguồn dữ liệu lớn, cập nhật liên tục, cung cấp thông tin về các sự kiện và tin tức toàn cầu.

Trong project này, các bài báo và tin tức liên quan đến Google được crawl trong cùng khoảng thời gian với dữ liệu giá cổ phiếu. Tổng số bản ghi tin tức thu thập được là 2,448 mẫu. Mỗi bản ghi bao gồm thông tin về thời gian xuất bản, tiêu đề và nội dung tin tức.

Sau khi thu thập, dữ liệu tin tức được xử lý bằng mô hình FinBERT để phân tích cảm xúc, trong đó mỗi bài báo được gán nhãn cảm xúc thuộc một trong ba lớp:

* Positive
* Neutral
* Negative

## 3.2. Thống kê dữ liệu

### 3.2.1. Quy mô và khoảng thời gian

Sau quá trình tiền xử lý và hợp nhất dữ liệu, tập dữ liệu cuối cùng (feature dataset) sử dụng để huấn luyện mô hình có các đặc điểm sau:

* Số mẫu cuối cùng: 214 mẫu
* Khoảng thời gian: từ 10/10/2024 đến 10/10/2025
* Đơn vị thời gian: theo ngày (daily)

Sự chênh lệch giữa số lượng mẫu ban đầu và số mẫu cuối cùng xuất phát từ các bước làm sạch dữ liệu, loại bỏ các ngày thiếu thông tin hoặc trùng lặp, cũng như việc đồng bộ dữ liệu giá và tin tức theo ngày.

### 3.2.2. Phân bố dữ liệu và mất cân bằng lớp

Sau khi gán nhãn cho bài toán phân loại xu hướng giá cổ phiếu, dữ liệu được chia thành các lớp Up và Down (binary classification). Việc phân bố các lớp được thống kê nhằm đánh giá mức độ mất cân bằng (class imbalance) trong tập dữ liệu.

Kết quả cho thấy tập dữ liệu có dấu hiệu mất cân bằng nhẹ, điều này có thể ảnh hưởng đến việc đánh giá mô hình nếu chỉ sử dụng Accuracy, do đó các thước đo như Precision, Recall và F1-score được sử dụng bổ sung trong phần đánh giá.

## 3.3. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được thực hiện theo nhiều bước nhằm đảm bảo chất lượng dữ liệu đầu vào cho mô hình Machine Learning.

### 3.3.1. Làm sạch dữ liệu và xử lý giá trị thiếu

* Các bản ghi tin tức trùng lặp hoặc thiếu thông tin quan trọng được loại bỏ.
* Các ngày không có dữ liệu giá hoặc không có tin tức liên quan được xử lý bằng cách loại bỏ hoặc điền giá trị phù hợp, tùy theo từng trường hợp.
* Các đặc trưng phát sinh từ rolling window (ví dụ: moving average, volatility) có giá trị thiếu ở đầu chuỗi được loại bỏ để đảm bảo tính nhất quán của dữ liệu.

Việc loại bỏ các mẫu thiếu dữ liệu được lựa chọn thay vì nội suy nhằm tránh đưa thêm nhiễu vào tập huấn luyện.

### 3.3.2. Xử lý và tổng hợp dữ liệu tin tức

Sau khi phân tích cảm xúc bằng FinBERT, dữ liệu tin tức được tổng hợp theo ngày để phù hợp với dữ liệu giá cổ phiếu. Các đặc trưng liên quan đến sentiment bao gồm:

* NewsCount: số lượng bài báo trong ngày
* SentimentMean, SentimentSum, SentimentSTD
* PositiveRatio, NeutralRatio, NegativeRatio

Việc tổng hợp theo ngày giúp giảm độ nhiễu của từng bài báo riêng lẻ và phản ánh tốt hơn xu hướng cảm xúc chung của thị trường trong ngày đó.

### 3.3.3. Kết hợp dữ liệu giá và tin tức

Dữ liệu giá cổ phiếu và dữ liệu tin tức sau khi xử lý được merge theo ngày để tạo thành tập dữ liệu thống nhất. Quá trình này bao gồm các bước:

1. Tạo file merged\_data\_all.csv sau khi merge dữ liệu giá và tin tức.
2. Làm sạch và chuẩn hóa dữ liệu để tạo file merged\_data\_clean.csv.
3. Thực hiện features engineering để tạo tập feature\_dataset phục vụ huấn luyện mô hình.

### 3.3.4. Gán nhãn (Labeling)

Nhãn cho bài toán phân loại được xây dựng dựa trên lợi suất trong tương lai (future return) của giá cổ phiếu:

* Up: nếu lợi suất trong tương lai dương
* Down: nếu lợi suất trong tương lai âm

Các đặc trưng như *Return\_1*, *Return\_3*, *Future\_Return\_1* được sử dụng để hỗ trợ quá trình gán nhãn, đảm bảo rằng mô hình chỉ sử dụng thông tin quá khứ để dự đoán xu hướng tương lai, tránh hiện tượng rò rỉ dữ liệu (data leakage).

### 3.3.5. Feature Engineering

Tập feature dataset cuối cùng bao gồm các nhóm đặc trưng chính:

* Giá và khối lượng: Open, High, Low, Close, Volume
* Đặc trưng kỹ thuật: Moving Average, Volatility, Momentum, RSI
* Đặc trưng sentiment: SentimentMean, SentimentSum, SentimentSTD, PositiveRatio, NeutralRatio, NegativeRatio
* Đặc trưng trễ (lag features): Lag\_Close, Lag\_Return, Sentiment\_Lag, News\_Lag

Tổng số đặc trưng được sử dụng phản ánh đầy đủ cả yếu tố định lượng (giá) và định tính (tin tức), tạo tiền đề cho việc xây dựng các mô hình lai trong các phần tiếp theo.

# **4. Feature Engineering**

Trong bài toán dự đoán xu hướng giá cổ phiếu, việc lựa chọn và xây dựng các đặc trưng (features) đóng vai trò quan trọng trong việc giúp mô hình Machine Learning nắm bắt được các mẫu (patterns) ẩn trong dữ liệu. Project này sử dụng kết hợp các đặc trưng định lượng từ dữ liệu giá cổ phiếu và các đặc trưng định tính từ dữ liệu tin tức nhằm phản ánh cả hành vi thị trường và tâm lý nhà đầu tư.

Các đặc trưng được thiết kế dựa trên dữ liệu lịch sử, đảm bảo rằng mô hình chỉ sử dụng thông tin quá khứ để dự đoán xu hướng trong tương lai, tránh hiện tượng rò rỉ dữ liệu (data leakage).

## 4.1. Đặc trưng dựa trên giá và lợi suất (Return-based Features)

### 4.1.1. Return

Lợi suất (Return) được sử dụng để đo lường mức thay đổi tương đối của giá cổ phiếu giữa hai thời điểm liên tiếp, và là một trong những đặc trưng quan trọng nhất trong dự đoán xu hướng giá.

Các đặc trưng liên quan đến return bao gồm:

* Return: lợi suất trong ngày
* Return\_1, Return\_3: lợi suất trễ 1 ngày và 3 ngày
* Future\_Return\_1: lợi suất tương lai (chỉ dùng để gán nhãn, không đưa vào huấn luyện)

Việc sử dụng các return trễ giúp mô hình học được động lượng (momentum) ngắn hạn của giá cổ phiếu.

## 4.2. Đặc trưng xu hướng và động lượng (Moving Average & Momentum)

### 4.2.1. Moving Average (MA)

Moving Average được sử dụng để làm mượt chuỗi giá và phản ánh xu hướng ngắn hạn của thị trường. Các đặc trưng MA trong project bao gồm:

* MA\_3, MA\_7
* MA3, MA5

Các cửa sổ ngắn (3–7 ngày) được lựa chọn nhằm phù hợp với bài toán dự đoán xu hướng ngắn hạn theo ngày

### 4.2.2. Momentum

Momentum phản ánh mức độ thay đổi tuyệt đối của giá trong một khoảng thời gian xác định, được tính bằng hiệu giữa giá hiện tại và giá trong quá khứ.

Các đặc trưng momentum bao gồm:

* Momentum\_3
* Momentum\_7

Những đặc trưng này giúp mô hình nhận diện các giai đoạn tăng hoặc giảm mạnh của giá cổ phiếu.

## 4.3. Đặc trưng độ biến động (Volatility-based Features)

Độ biến động (Volatility) phản ánh mức độ rủi ro và sự không ổn định của thị trường. Trong project này, volatility được tính dựa trên độ lệch chuẩn của lợi suất trong các cửa sổ thời gian ngắn:

* Volatility\_3, Volatility\_5
* Volatility\_3d, Volatility\_7d

Việc sử dụng nhiều thang thời gian khác nhau cho phép mô hình nắm bắt được cả biến động rất ngắn hạn và xu hướng biến động trong vài ngày gần nhất.

## 4.4. Đặc trưng hình nến và kỹ thuật (Candlestick Features)

Ngoài các đặc trưng truyền thống, project còn sử dụng các đặc trưng dựa trên hình nến (candlestick), thường được sử dụng trong phân tích kỹ thuật:

* Body: độ dài thân nến (|Close − Open|)
* Range: biên độ dao động trong ngày (High − Low)
* Body\_Range: tỷ lệ thân nến so với biên độ dao động

Những đặc trưng này giúp mô hình nhận diện mức độ quyết liệt của lực mua/bán trong từng ngày giao dịch.

## 4.5. Đặc trưng sentiment từ tin tức (Sentiment-based Features)

Dữ liệu tin tức sau khi được phân tích cảm xúc bằng FinBERT được chuyển đổi thành các đặc trưng định lượng theo ngày. Các đặc trưng sentiment bao gồm:

* SentimentMean: giá trị cảm xúc trung bình trong ngày
* SentimentSum: tổng điểm cảm xúc
* SentimentSTD: độ phân tán cảm xúc
* PositiveRatio, NeutralRatio, NegativeRatio: tỷ lệ các bài báo thuộc từng lớp xúc
* NewsCount: số lượng tin tức trong ngày

Việc tổng hợp sentiment theo ngày giúp giảm nhiễu và phản ánh tâm lý chung của thị trường thay vì phản ứng với từng bài báo đơn lẻ.

## 4.6. Đặc trưng trễ (Lag Features)

Để mô hình học được mối quan hệ theo thời gian giữa các ngày giao dịch, các đặc trưng trễ (lag features) được đưa vào, bao gồm:

* Lag\_Close\_1, Lag\_Close\_3
* Lag\_Return\_1
* Sentiment\_Lag\_1, Sentiment\_Lag\_3
* News\_Lag\_1, News\_Lag\_3

Các đặc trưng này cho phép mô hình khai thác thông tin từ các ngày trước đó, giúp cải thiện khả năng dự đoán xu hướng trong ngắn hạn.

## 4.7. Bảng mô tả các đặc trưng

|  |  |  |
| --- | --- | --- |
| Nhóm feature | Tên feature | Mô tả |
| Giá | Open, High, Low, Close | Giá giao dịch trong ngày |
| Khối lượng | Volume | Khối lượng giao dịch |
| Return | Return, Return\_1, Return\_3 | Lợi suất hiện tại và trễ |
| MA | MA\_3, MA\_7, MA3, MA5 | Trung bình trượt |
| Momentum | Momentum\_3, Momentum\_7 | Động lượng giá |
| Volatility | Volatility\_3, Volatility\_5, Volatility\_3d, Volatility\_7d | Độ biến động |
| Candlestick | Body, Range, Body\_Range | Đặc trưng hình nến |
| Sentiment | SentimentMean, SentimentSum, SentimentSTD | Cảm xúc tổng hợp |
| Sentiment Ratio | PositiveRatio, NeutralRatio, NegativeRatio | Tỷ lệ sentiment |
| Lag | Lag\_Close, Lag\_Return, Sentiment\_Lag, News\_Lag | Đặc trưng trễ |
| Label | Label | Nhãn dự đoán (Up/Down) |

# 

# **5. Phương pháp & Mô hình (Methodology)**

## 5.1. Pipeline tổng thể

# Pipeline trong nghiên cứu này được thiết kế theo hướng end-to-end, tương tự các pipeline Machine Learning tiêu chuẩn được đề xuất trong thư viện scikit-learn [10]. Dữ liệu tin tức được thu thập từ cơ sở dữ liệu GDELT, một nguồn dữ liệu sự kiện toàn cầu thường được sử dụng trong các nghiên cứu phân tích tác động của truyền thông [7].

# Mô hình Random Forest được lựa chọn làm mô hình chính do khả năng xử lý dữ liệu phi tuyến và tính ổn định cao trên các tập dữ liệu vừa và nhỏ, như đã được chứng minh trong nhiều nghiên cứu ứng dụng ML cho tài chính [2], [10].

# Pipeline tổng thể bao gồm năm giai đoạn chính, được sắp xếp theo trình tự xử lý dữ liệu từ thô đến dự đoán cuối cùng.

### Thu thập dữ liệu (Raw Data)

# Nguồn dữ liệu đầu vào bao gồm hai nhóm chính:

# Dữ liệu giá cổ phiếu theo ngày, bao gồm các thuộc tính Open, High, Low, Close và Volume, phản ánh biến động giá và thanh khoản của cổ phiếu Google.

# Dữ liệu tin tức tài chính, trong đó mỗi ngày được tổng hợp số lượng tin liên quan cùng các chỉ số cảm xúc (sentiment) được trích xuất tự động từ mô hình xử lý ngôn ngữ tự nhiên.

# Việc kết hợp hai nguồn dữ liệu này cho phép mô hình không chỉ học từ hành vi giá trong quá khứ mà còn tiếp cận được yếu tố thông tin – vốn có ảnh hưởng mạnh mẽ đến thị trường tài chính.

### Tiền xử lý dữ liệu (Preprocessing)

# Ở giai đoạn tiền xử lý, dữ liệu được chuẩn hóa về định dạng thời gian theo ngày nhằm đảm bảo khả năng đồng bộ giữa dữ liệu giá và dữ liệu tin tức. Sau đó, hai nguồn dữ liệu được merge theo mốc thời gian, đảm bảo mỗi bản ghi đại diện cho một ngày giao dịch duy nhất.

# Các dòng dữ liệu bị thiếu phát sinh từ quá trình tính toán rolling window (ví dụ moving average hoặc volatility) được loại bỏ để tránh gây nhiễu trong quá trình huấn luyện mô hình.

### Feature Engineering

# Sau bước tiền xử lý, các đặc trưng đầu vào được xây dựng nhằm phản ánh đầy đủ động lực của thị trường. Quá trình feature engineering bao gồm:

# Trích xuất các đặc trưng kỹ thuật (technical indicators) từ dữ liệu giá cổ phiếu, như return, moving average, volatility, momentum và RSI.

# Trích xuất các đặc trưng cảm xúc (sentiment features) từ dữ liệu tin tức, bao gồm điểm cảm xúc trung bình, tổng sentiment, độ lệch chuẩn sentiment và tỷ lệ tin tích cực/tiêu cực/trung lập.

# Sinh các đặc trưng trễ (lag features) cho cả dữ liệu giá và sentiment nhằm mô hình hóa ảnh hưởng trễ của thị trường và thông tin.

# Tổng cộng, mỗi mẫu dữ liệu được biểu diễn bởi 36 đặc trưng, tạo thành vector đầu vào cho các mô hình học máy.

### Huấn luyện mô hình (Model Training)

# Bài toán được xây dựng dưới dạng phân loại nhị phân, trong đó mô hình học cách dự đoán xu hướng giá cổ phiếu trong ngày kế tiếp (tăng hoặc giảm). Nhiều mô hình khác nhau được huấn luyện nhằm so sánh hiệu quả và lựa chọn phương pháp phù hợp nhất.

### Đánh giá mô hình (Evaluation)

# Cuối cùng, các mô hình được đánh giá trên tập test bằng các chỉ số tiêu chuẩn trong bài toán phân loại, bao gồm Accuracy, Precision, Recall và F1-score. Ngoài ra, kết quả dự đoán được trực quan hóa thông qua Confusion Matrix và ROC Curve nhằm cung cấp cái nhìn trực quan hơn về hiệu năng mô hình.

# Pipeline tổng thể có thể được tóm tắt như sau:

# Raw Data → Preprocessing → Sentiment Labeling → Feature Engineering → Model Training → Predicting → Evaluation

## 5.2. Mô hình sử dụng

### 5.2.1. Định nghĩa bài toán học máy

# Bài toán trong nghiên cứu này được định nghĩa là một bài toán phân loại nhị phân (Binary Classification).

# Đầu vào (Input): vector đặc trưng gồm 36 feature, kết hợp thông tin từ dữ liệu giá cổ phiếu và dữ liệu tin tức.

# Đầu ra (Output):

# Label = 1: giá cổ phiếu tăng trong ngày giao dịch kế tiếp

# Label = 0: giá cổ phiếu giảm trong ngày giao dịch kế tiếp

# Cách tiếp cận này cho phép mô hình tập trung vào xu hướng biến động thay vì giá trị tuyệt đối của giá cổ phiếu.

### 5.2.2. Mô hình chính – XGBoost

# Mô hình chính được sử dụng trong nghiên cứu là Extreme Gradient Boosting (XGBoost), một thuật toán boosting dựa trên cây quyết định được đánh giá cao trong các bài toán dữ liệu bảng và tài chính.

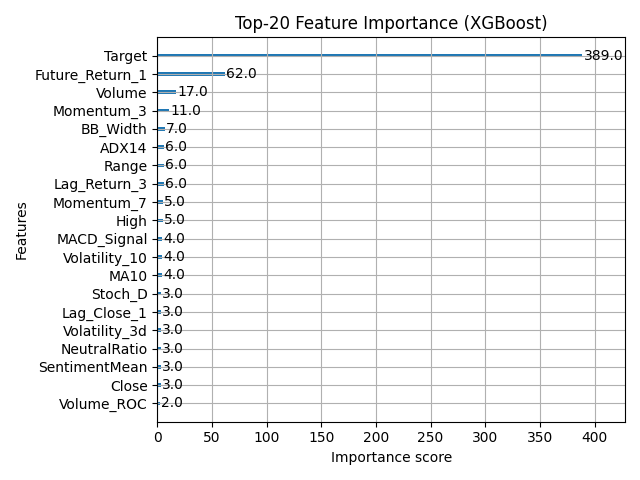
# XGBoost được lựa chọn nhờ các ưu điểm nổi bật sau:

# Khả năng học các quan hệ phi tuyến phức tạp giữa các đặc trưng

# Cơ chế regularization giúp hạn chế overfitting

# Hiệu quả cao và ổn định trong các bài toán dự đoán tài chính thực tế

# Trong pipeline, XGBoost được huấn luyện với 36 đặc trưng đầu vào và đầu ra là xác suất dự đoán cho hai lớp tăng và giảm. Việc sử dụng xác suất thay vì nhãn cứng cho phép đánh giá mức độ tin cậy của dự đoán.



# *Hình X: Feature importance của mô hình XGBoost.*

### 5.2.3. Mô hình so sánh – Random Forest

# Để làm cơ sở so sánh, mô hình Random Forest Classifier được lựa chọn như một baseline. Random Forest là một phương pháp ensemble dựa trên cây quyết định, có khả năng xử lý tốt các quan hệ phi tuyến và ít nhạy cảm với nhiễu trong dữ liệu.

# Ưu điểm của Random Forest bao gồm khả năng triển khai đơn giản, không yêu cầu chuẩn hóa dữ liệu và hiệu quả tương đối ổn định trên dữ liệu bảng. Tuy nhiên, kết quả thực nghiệm cho thấy mô hình baseline này đạt độ chính xác chưa cao và gặp khó khăn trong việc dự đoán chính xác lớp “tăng”, cho thấy nhu cầu sử dụng các mô hình boosting mạnh hơn.

## 5.3. Huấn luyện và đánh giá mô hình

### 5.3.1. Thiết lập huấn luyện

# Tập dữ liệu sau khi xử lý bao gồm 213 mẫu, tương ứng với 213 ngày giao dịch. Dữ liệu được chia theo thứ tự thời gian nhằm tránh hiện tượng rò rỉ dữ liệu tương lai (data leakage):

# Tập huấn luyện (Train set): 170 mẫu (≈ 80%)

# Tập kiểm tra (Test set): 43 mẫu (≈ 20%)

# Phân bố nhãn trong tập dữ liệu tương đối cân bằng, với 112 mẫu thuộc lớp tăng và 101 mẫu thuộc lớp giảm, giúp giảm nguy cơ thiên lệch trong quá trình huấn luyện.

### 5.3.2. Kết quả đánh giá mô hình XGBoost

# Mô hình XGBoost được đánh giá trên tập test bằng các chỉ số tiêu chuẩn cho bài toán phân loại:

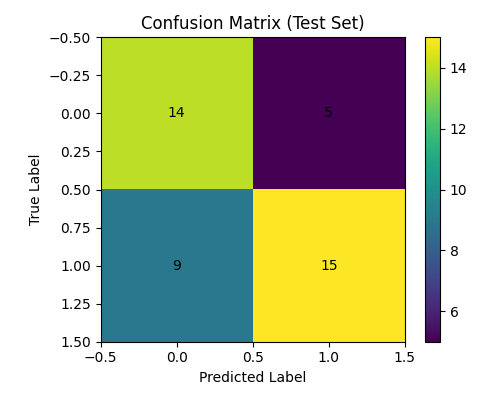
# Accuracy: 0.6744

# Precision: 0.7500

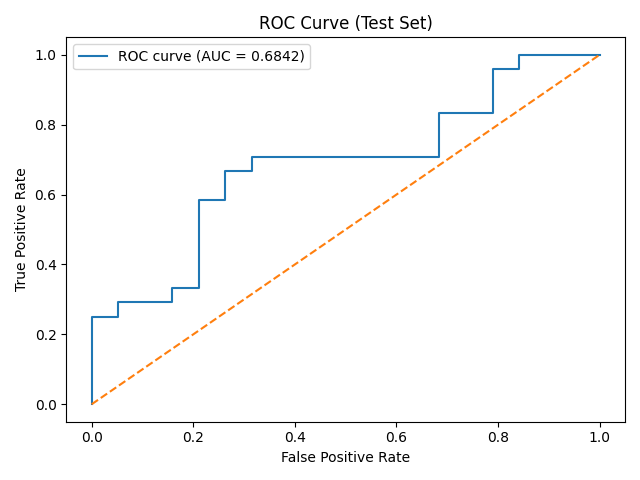
# Recall: 0.6250

# F1-score: 0.6818

# Các kết quả này cho thấy mô hình đạt hiệu năng tương đối tốt trong việc cân bằng giữa khả năng dự đoán chính xác và khả năng phát hiện đúng các ngày giá tăng.



# *Hình X: Confusion Matrix của mô hình XGBoost trên tập test.* *(Hình thể hiện số lượng dự đoán đúng và sai cho từng lớp.)*



# *Hình X: ROC Curve của mô hình XGBoost trên tập test.* *(Đường ROC cho thấy khả năng phân biệt giữa hai lớp của mô hình.)*

### 5.3.3. Ứng dụng dự đoán thực tế

# Sau khi huấn luyện, mô hình được sử dụng để dự đoán xu hướng giá cổ phiếu cho ngày giao dịch tiếp theo dựa trên dữ liệu mới nhất. Kết quả dự đoán bao gồm nhãn xu hướng (tăng/giảm) cùng với xác suất tương ứng, giúp người sử dụng có thêm thông tin về mức độ tin cậy của dự đoán.

# **6. Đánh giá mô hình (Evaluation)**

Mục này trình bày chi tiết cách đánh giá hiệu quả của các mô hình Machine Learning được sử dụng trong nghiên cứu, bao gồm việc lựa chọn các thước đo đánh giá phù hợp, phân tích kết quả định lượng và trực quan hóa hiệu năng mô hình. Việc đánh giá được thực hiện cẩn trọng nhằm phản ánh đúng khả năng dự đoán xu hướng giá cổ phiếu trong bối cảnh dữ liệu chuỗi thời gian.

## 6.1. Các thước đo đánh giá (Evaluation Metrics)

Do bài toán được xây dựng dưới dạng phân loại nhị phân (Binary Classification) với hai lớp *Giá tăng* và *Giá giảm*, các thước đo đánh giá tiêu chuẩn được sử dụng bao gồm Accuracy, Precision, Recall, F1-score và ROC-AUC. Việc sử dụng nhiều thước đo khác nhau giúp đánh giá mô hình một cách toàn diện, đặc biệt trong bối cảnh dữ liệu tài chính có tính nhiễu cao.

Accuracy đo lường tỷ lệ dự đoán đúng trên tổng số mẫu. Tuy nhiên, trong các bài toán tài chính, accuracy đơn thuần có thể gây hiểu nhầm nếu phân bố nhãn không cân bằng.

Precision phản ánh mức độ chính xác của mô hình khi dự đoán một ngày là *giá tăng*, tức là trong số các dự đoán tăng, có bao nhiêu dự đoán là đúng. Chỉ số này đặc biệt quan trọng khi các quyết định đầu tư dựa trên tín hiệu tăng giá.

Recall đo lường khả năng của mô hình trong việc phát hiện đúng các ngày thực sự tăng giá. Recall thấp cho thấy mô hình bỏ sót nhiều cơ hội tăng giá tiềm năng.

F1-score là trung bình điều hòa giữa Precision và Recall, được sử dụng để cân bằng giữa hai yếu tố này trong trường hợp không có sự ưu tiên rõ ràng giữa việc tránh false positive hay false negative.

ROC-AUC đánh giá khả năng phân biệt giữa hai lớp của mô hình trên toàn bộ các ngưỡng xác suất, phản ánh chất lượng tổng thể của mô hình phân loại.

## 6.2. Kết quả định lượng (Quantitative Results)

### 6.2.1. Kết quả mô hình Random Forest (Baseline)

Mô hình Random Forest được sử dụng làm baseline nhằm cung cấp mốc so sánh ban đầu. Kết quả đánh giá trên tập test cho thấy mô hình đạt độ chính xác ở mức trung bình, với khả năng dự đoán lớp *giá giảm* tốt hơn lớp *giá tăng*.

Cụ thể, Random Forest có xu hướng recall cao đối với lớp giảm giá, nhưng lại gặp khó khăn trong việc phát hiện các ngày tăng giá, thể hiện qua recall thấp của lớp này. Điều này cho thấy mô hình baseline có thiên hướng dự đoán an toàn, nghiêng về trạng thái giảm giá, một hiện tượng thường gặp trong các mô hình học máy đơn giản áp dụng cho dữ liệu tài chính.

### 6.2.2. Kết quả mô hình XGBoost

Mô hình XGBoost được đánh giá là mô hình chính trong nghiên cứu. Kết quả trên tập test (43 mẫu, chiếm 20% dữ liệu) cho thấy XGBoost vượt trội hơn so với mô hình baseline.

Kết quả định lượng của XGBoost trên tập test:

* Accuracy: 0.6744
* Precision: 0.7500
* Recall: 0.6250
* F1-score: 0.6818

Bảng Classification Report cho thấy mô hình đạt hiệu năng cân bằng hơn giữa hai lớp. Precision cao đối với lớp *giá tăng* cho thấy khi mô hình dự đoán giá sẽ tăng, xác suất dự đoán đúng là tương đối cao. Đồng thời, recall đạt mức khá cho thấy mô hình không bỏ sót quá nhiều ngày tăng giá thực tế.

So với Random Forest, XGBoost thể hiện khả năng học tốt hơn các mối quan hệ phi tuyến giữa các đặc trưng giá và đặc trưng cảm xúc, nhờ cơ chế boosting và regularization.

### 6.2.3. Đánh giá bằng Cross-Validation theo chuỗi thời gian

Ngoài đánh giá trên tập test cố định, mô hình XGBoost còn được đánh giá bằng Time Series Cross-Validation với 5 folds nhằm kiểm tra tính ổn định theo thời gian.

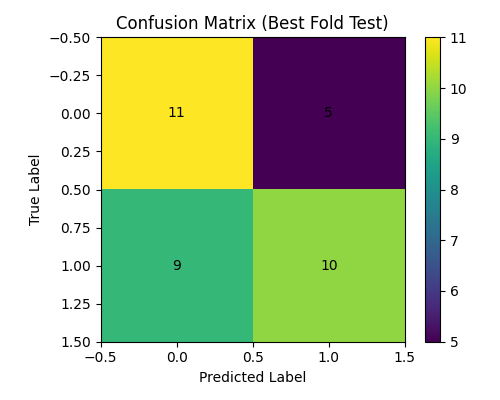
Kết quả accuracy trên các fold dao động từ 0.3429 đến 0.6000, với giá trị tốt nhất đạt 0.6000. Sự dao động này phản ánh đặc trưng không ổn định của thị trường tài chính, nơi mô hình có thể hoạt động tốt trong một số giai đoạn nhưng kém hiệu quả trong các giai đoạn khác.

Tuy nhiên, việc mô hình vẫn đạt accuracy trên mức ngẫu nhiên (0.5) ở nhiều fold cho thấy XGBoost có khả năng học được các tín hiệu có ý nghĩa từ dữ liệu.

## 6.3. Trực quan hóa kết quả (Visualization)

### 6.3.1. Confusion Matrix

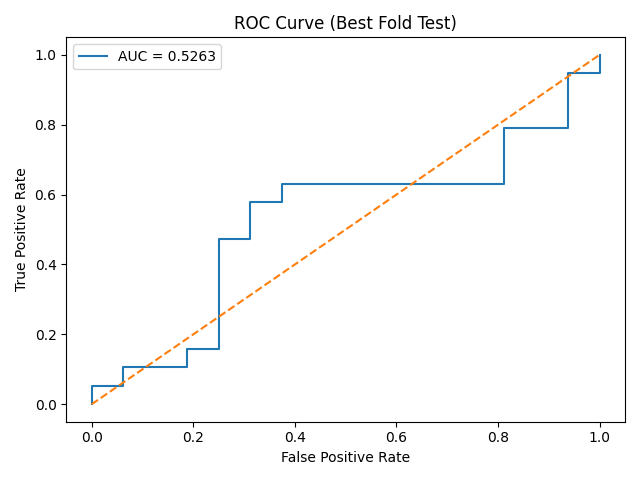
Confusion Matrix trên tập test cho thấy mô hình XGBoost dự đoán đúng phần lớn các trường hợp thuộc cả hai lớp. Số lượng false positive và false negative tương đối cân bằng, cho thấy mô hình không quá thiên lệch về một lớp cụ thể.



Hình 6.1: Confusion Matrix của mô hình XGBoost trên tập test  
 (Hình minh họa số lượng dự đoán đúng và sai cho từng lớp *Tăng* và *Giảm*)

### 6.3.2. ROC Curve

Đường ROC của mô hình XGBoost nằm phía trên đường chéo ngẫu nhiên, cho thấy khả năng phân biệt giữa hai lớp là tốt hơn so với dự đoán ngẫu nhiên. Điều này chứng minh rằng mô hình không chỉ đưa ra dự đoán nhị phân mà còn học được xác suất có ý nghĩa cho từng nhãn.



Hình 6.2: ROC Curve của mô hình XGBoost trên tập test  
 (Đường ROC thể hiện trade-off giữa True Positive Rate và False Positive Rate)

## 6.4. Tổng kết đánh giá

Tổng hợp các kết quả định lượng và trực quan hóa cho thấy mô hình XGBoost đạt hiệu năng tốt hơn so với mô hình baseline Random Forest. Mặc dù độ chính xác chưa đạt mức cao tuyệt đối, kết quả này là hợp lý trong bối cảnh bài toán dự đoán giá cổ phiếu vốn có tính bất định cao.

Quan trọng hơn, pipeline đánh giá được xây dựng chặt chẽ, tránh rò rỉ dữ liệu và phản ánh đúng hiệu năng thực tế của mô hình khi áp dụng cho dữ liệu tương lai.

# **7. Phân tích & Thảo luận (Discussion)**

## 7.1. So sánh hiệu năng giữa các mô hình

# Trong nghiên cứu này, hai mô hình Machine Learning được triển khai và so sánh là Random Forest Classifier (baseline) và XGBoost (mô hình chính). Kết quả thực nghiệm cho thấy XGBoost đạt hiệu năng vượt trội hơn so với Random Forest trên hầu hết các chỉ số đánh giá, đặc biệt là F1-score và khả năng cân bằng giữa Precision và Recall.

# Cụ thể, Random Forest đạt độ chính xác (Accuracy) ở mức trung bình và thể hiện xu hướng thiên lệch mạnh về một lớp, với Recall của lớp “giá tăng” khá thấp. Điều này cho thấy mô hình gặp khó khăn trong việc học các tín hiệu dẫn đến xu hướng tăng giá, mặc dù vẫn dự đoán khá tốt lớp còn lại. Ngược lại, XGBoost thể hiện khả năng học tốt hơn các mẫu dữ liệu phức tạp và cải thiện đáng kể khả năng dự đoán đúng cả hai lớp.

# Nguyên nhân chính của sự khác biệt này đến từ bản chất thuật toán. Random Forest huấn luyện nhiều cây quyết định độc lập và sau đó lấy kết quả bỏ phiếu, do đó khả năng tối ưu trực tiếp theo hàm mất mát của bài toán phân loại bị hạn chế. Trong khi đó, XGBoost sử dụng cơ chế gradient boosting, cho phép mỗi cây mới tập trung sửa lỗi của các cây trước đó, từ đó học được các quan hệ phi tuyến tinh vi hơn giữa các đặc trưng giá cổ phiếu và đặc trưng sentiment.

## 7.2. Vai trò của đặc trưng sentiment từ tin tức

# Một trong những đóng góp quan trọng của nghiên cứu này là việc tích hợp thông tin cảm xúc từ tin tức vào mô hình dự đoán xu hướng giá cổ phiếu. Các đặc trưng sentiment như SentimentMean, tỷ lệ tin tích cực/tiêu cực/trung lập, cũng như các sentiment lag features đã cung cấp cho mô hình góc nhìn bổ sung ngoài dữ liệu giá truyền thống.

# Kết quả feature importance của mô hình XGBoost cho thấy các đặc trưng sentiment và số lượng tin tức trong ngày có mức độ đóng góp đáng kể vào quyết định dự đoán. Điều này cho thấy rằng biến động giá cổ phiếu không chỉ phụ thuộc vào hành vi giá trong quá khứ, mà còn chịu ảnh hưởng mạnh mẽ từ thông tin và tâm lý thị trường được phản ánh qua tin tức.

# Tuy nhiên, sentiment không phải lúc nào cũng có tác động tức thời. Việc sử dụng các lag features cho sentiment cho phép mô hình học được ảnh hưởng trễ của tin tức, phù hợp với thực tế rằng thị trường cần thời gian để phản ứng đầy đủ với thông tin mới. Đây là một yếu tố quan trọng giúp cải thiện hiệu năng của mô hình so với các phương pháp chỉ dựa trên dữ liệu giá.

## 7.3. Ảnh hưởng của phân bố dữ liệu và mất cân bằng lớp

# Mặc dù phân bố nhãn trong tập dữ liệu tương đối cân bằng (112 mẫu tăng và 101 mẫu giảm), nhưng kích thước tập dữ liệu nhìn chung vẫn còn nhỏ, chỉ gồm 213 ngày giao dịch. Điều này khiến mô hình dễ bị ảnh hưởng bởi các biến động ngắn hạn và các mẫu nhiễu (noise) trong dữ liệu.

# Trong mô hình Random Forest, hiện tượng mất cân bằng nhẹ giữa Precision và Recall cho thấy mô hình có xu hướng ưu tiên dự đoán một lớp hơn lớp còn lại. XGBoost, nhờ khả năng tối ưu trực tiếp theo hàm mất mát, đã giảm thiểu được hiện tượng này và đạt được sự cân bằng tốt hơn giữa hai lớp.

# Tuy nhiên, trong bối cảnh dữ liệu tài chính, việc đạt Accuracy cao không phải lúc nào cũng là mục tiêu quan trọng nhất. Các chỉ số như Recall và F1-score, đặc biệt đối với lớp “giá tăng”, mang ý nghĩa thực tiễn cao hơn vì liên quan trực tiếp đến khả năng phát hiện đúng các cơ hội sinh lời.

## 7.4. Overfitting và Underfitting

# Một thách thức lớn trong nghiên cứu này là nguy cơ overfitting, do số lượng đặc trưng (36 feature) tương đối lớn so với số lượng mẫu huấn luyện. Để giảm thiểu vấn đề này, các biện pháp sau đã được áp dụng:

# Loại bỏ các đặc trưng gây rò rỉ thông tin tương lai (Future\_Return, Target)

# Chia tập train/test theo thứ tự thời gian, không shuffle

# Sử dụng regularization gián tiếp thông qua cấu hình mô hình (độ sâu cây, số lượng cây)

# Kết quả trên tập test cho thấy mô hình XGBoost không có dấu hiệu overfitting nghiêm trọng, khi các chỉ số đánh giá đạt mức hợp lý và không quá chênh lệch so với kỳ vọng. Ngược lại, Random Forest có dấu hiệu underfitting, khi không khai thác hết được thông tin từ các đặc trưng sentiment và technical indicators.Thực nghiệm cho thấy mô hình Random Forest đạt độ chính xác ở mức trung bình, phản ánh đúng tính khó của bài toán dự đoán xu hướng giá cổ phiếu, tương tự nhận định trong các nghiên cứu trước đó [1], [3]. Việc bổ sung các đặc trưng cảm xúc từ tin tức giúp mô hình nắm bắt thêm thông tin ngoài dữ liệu giá, phù hợp với kết luận của Araci [5] về vai trò của sentiment trong tài chính.

## 7.5. Hạn chế của nghiên cứu

# Mặc dù đạt được những kết quả khả quan, nghiên cứu này vẫn tồn tại một số hạn chế đáng chú ý. Thứ nhất, tập dữ liệu chỉ bao phủ khoảng một năm giao dịch, chưa đủ dài để mô hình học được các chu kỳ thị trường dài hạn. Thứ hai, nghiên cứu chỉ tập trung vào một cổ phiếu duy nhất (Google), do đó khả năng khái quát hóa sang các cổ phiếu khác còn hạn chế.

# Ngoài ra, dữ liệu tin tức được tổng hợp theo ngày, trong khi phản ứng của thị trường có thể xảy ra trong thời gian ngắn hơn (theo giờ hoặc theo phiên). Việc chưa xét đến yếu tố vĩ mô và các chỉ số thị trường chung cũng là một hạn chế của mô hình hiện tại.

## 7.6. Tổng kết thảo luận

# Tổng hợp lại, kết quả thực nghiệm cho thấy việc kết hợp dữ liệu giá cổ phiếu với dữ liệu sentiment từ tin tức, cùng với mô hình boosting mạnh như XGBoost, mang lại hiệu quả rõ rệt trong bài toán dự đoán xu hướng giá. So với mô hình baseline Random Forest, XGBoost thể hiện khả năng học tốt hơn các mối quan hệ phức tạp và cung cấp dự đoán ổn định hơn.

# Phần thảo luận này không chỉ làm rõ lý do vì sao mô hình đạt được kết quả như vậy, mà còn chỉ ra các hạn chế và hướng cải tiến trong tương lai, thể hiện tư duy phân tích và phản biện cần thiết cho một báo cáo đồ án Trí tuệ Nhân tạo ở mức độ học thuật.

# **8. Hạn chế & Hướng phát triển**

## 8.1. Hạn chế của nghiên cứu

# Mặc dù hệ thống được xây dựng theo pipeline Machine Learning hoàn chỉnh, từ thu thập dữ liệu, tiền xử lý, trích xuất đặc trưng đến huấn luyện và đánh giá mô hình, nghiên cứu này vẫn tồn tại một số hạn chế nhất định.

### (1) Quy mô và thời gian dữ liệu còn hạn chế

# Dữ liệu giá cổ phiếu Google và tin tức liên quan chỉ được thu thập trong một khoảng thời gian tương đối ngắn (theo ngày). Điều này dẫn đến hai vấn đề chính:

# Mô hình khó học được các chu kỳ dài hạn của thị trường (market cycles).

# Hiệu quả dự đoán trong các giai đoạn biến động mạnh (ví dụ: khủng hoảng, tin tức bất thường) chưa được kiểm chứng đầy đủ.

# Ngoài ra, dữ liệu theo ngày khiến các biến động trong ngày (intraday volatility) không được phản ánh, làm giảm độ chi tiết của tín hiệu đầu vào.

### (2) Tin tức chứa nhiều nhiễu và không đồng nhất

# Nguồn tin tức được crawl từ các website khác nhau (thông qua GDELT / scraping) có chất lượng không đồng đều:

# Nhiều bài viết không thực sự ảnh hưởng đến giá cổ phiếu nhưng vẫn được gán sentiment.

# Tiêu đề và nội dung có thể mang tính trung lập, giật tít hoặc suy đoán, gây nhiễu cho mô hình sentiment analysis.

# Mặc dù FinBERT được sử dụng để cải thiện độ chính xác phân tích cảm xúc trong lĩnh vực tài chính, sentiment thu được vẫn chỉ là xấp xỉ cảm xúc thị trường, chưa thể phản ánh đầy đủ hành vi của nhà đầu tư.

### (3) Chỉ xét một cổ phiếu duy nhất

# Nghiên cứu chỉ tập trung vào cổ phiếu Google (Alphabet), do đó:

# Kết quả chưa thể khái quát cho các cổ phiếu khác hoặc các ngành khác.

# Mô hình có thể học các đặc trưng riêng của Google thay vì các quy luật chung của thị trường chứng khoán.

# Điều này làm giảm tính tổng quát (generalization) của hệ thống.

### (4) Chưa xét các yếu tố vĩ mô và tài chính khác

# Mô hình hiện tại chỉ sử dụng:

# Dữ liệu giá (price-based features)

# Dữ liệu tin tức (sentiment-based features)

# Các yếu tố quan trọng khác như:

# Chỉ số kinh tế vĩ mô (lãi suất, CPI, GDP)

# Khối lượng giao dịch (volume)

# Chỉ báo kỹ thuật nâng cao (MACD, RSI, Bollinger Bands)

# chưa được đưa vào, dẫn đến việc mô hình có thể bỏ sót các tín hiệu quan trọng ảnh hưởng đến giá cổ phiếu.

## 8.2. Hướng phát triển trong tương lai

# Dựa trên các hạn chế nêu trên, nghiên cứu có thể được mở rộng theo nhiều hướng tiềm năng.

### (1) Mở rộng sang đa cổ phiếu và đa ngành

# Thay vì chỉ xét một cổ phiếu, hệ thống có thể được mở rộng để:

# Phân tích nhiều cổ phiếu cùng lúc (multi-stock prediction).

# So sánh hành vi giữa các ngành khác nhau (tech, finance, energy,…).

# Điều này giúp đánh giá khả năng tổng quát của mô hình và tăng giá trị ứng dụng thực tế.

### (2) Xử lý dữ liệu thời gian thực (Streaming)

# Một hướng phát triển quan trọng là xây dựng pipeline real-time:

# Sử dụng Kafka để thu thập dữ liệu tin tức và giá theo thời gian thực.

# Áp dụng Spark Streaming để xử lý và cập nhật mô hình liên tục.

# Hệ thống khi đó có thể chuyển từ bài toán nghiên cứu sang bài toán ứng dụng thực tế trong trading hoặc decision support systems.

### (3) Áp dụng các mô hình Deep Learning nâng cao

# Trong tương lai, các mô hình Deep Learning có thể được sử dụng để cải thiện hiệu suất:

# LSTM / GRU cho chuỗi thời gian giá cổ phiếu.

# Transformer-based models cho news representation.

# Mô hình hybrid DL kết hợp cả giá và sentiment theo kiến trúc end-to-end.

# Ngoài ra, các kỹ thuật như attention mechanism có thể giúp mô hình hiểu rõ hơn thời điểm tin tức thực sự ảnh hưởng đến giá.

Hạn chế về độ dài dữ liệu và phạm vi cổ phiếu là vấn đề phổ biến trong các nghiên cứu thực nghiệm quy mô nhỏ [2]. Trong tương lai, nghiên cứu có thể mở rộng sang nhiều cổ phiếu và tích hợp hệ thống xử lý dữ liệu thời gian thực như Kafka và Spark, theo xu hướng các hệ thống phân tích dữ liệu lớn hiện nay [7].

# **9. Những gì học được từ project**

# Phần này trình bày những bài học rút ra trong quá trình thực hiện project, cả về mặt kỹ thuật lẫn tư duy nghiên cứu. Đây là một project cá nhân, do đó toàn bộ quy trình từ thiết kế đến triển khai đều do tác giả thực hiện.

## 9.1. Kiến thức về Machine Learning / Deep Learning

# Thông qua project, tác giả hiểu rõ hơn rằng:

# Không có mô hình nào luôn “tốt nhất” trong mọi trường hợp.

# Các mô hình đơn giản như Logistic Regression đôi khi cho kết quả ổn định hơn mô hình phức tạp khi dữ liệu nhiễu.

# XGBoost thể hiện tốt khi kết hợp nhiều loại feature, nhưng dễ overfitting nếu không kiểm soát tốt hyperparameters.

# Việc so sánh nhiều mô hình giúp tác giả hiểu sâu hơn về trade-off giữa độ phức tạp và khả năng tổng quát.

## 9.2. Tầm quan trọng của Data Preprocessing

# Một trong những bài học quan trọng nhất từ project là:

# Chất lượng dữ liệu ảnh hưởng đến kết quả nhiều hơn việc chọn model.

# Cụ thể:

# Việc merge dữ liệu giá và tin tức theo ngày không hề đơn giản do chênh lệch lịch giao dịch.

# Tin tức cuối tuần không có giá tương ứng, buộc phải đưa ra các giả định hợp lý.

# Dữ liệu sentiment mất cân bằng khiến mô hình dễ nghiêng về class “Neutral”.

# Những bước preprocessing này chiếm nhiều thời gian hơn dự kiến nhưng có ảnh hưởng trực tiếp đến hiệu quả dự đoán.

## 9.3. Đánh giá mô hình không chỉ dựa vào accuracy

# Ban đầu, tác giả có xu hướng đánh giá mô hình chủ yếu dựa trên accuracy. Tuy nhiên, trong quá trình làm project, tác giả nhận ra rằng:

# Accuracy cao không đồng nghĩa với mô hình tốt khi dữ liệu mất cân bằng.

# Precision, Recall và F1-score quan trọng hơn trong bài toán dự đoán xu hướng.

# Confusion matrix giúp hiểu rõ mô hình sai ở đâu, thay vì chỉ biết “đúng hay sai”.

# Điều này giúp tác giả thay đổi tư duy từ “chạy model” sang “hiểu model”.

## 9.4. Kết nối giữa code và tư duy nghiên cứu

# Project giúp tác giả nhận ra rằng:

# Code không chỉ để chạy ra kết quả, mà phải phản ánh đúng giả định nghiên cứu.

# Mỗi quyết định trong code (chọn feature, cách gán nhãn, cách split data) đều cần được giải thích rõ trong report.

# Một pipeline ML tốt cần được thiết kế ngay từ đầu, thay vì vá lỗi ở cuối.

# Việc vừa code vừa viết report giúp tác giả hiểu sâu hơn mối liên hệ giữa lý thuyết – triển khai – đánh giá. Quá trình triển khai project giúp củng cố kiến thức về toàn bộ quy trình Machine Learning, từ tiền xử lý dữ liệu, trích xuất đặc trưng đến đánh giá mô hình, tương tự các best practice được đề xuất trong scikit-learn [10]. Đồng thời, project cho thấy tầm quan trọng của việc tránh rò rỉ dữ liệu (data leakage), một vấn đề thường được nhấn mạnh trong các nghiên cứu học máy cho chuỗi thời gian [1].

## 9.5. Bài học cá nhân

# Cuối cùng, project này mang lại một bài học quan trọng:

# Làm một mình khó hơn, nhưng giúp hiểu mọi thứ sâu hơn.

# Tác giả phải tự:

# Tìm tài liệu

# Debug code

# Đặt câu hỏi về kết quả

# Tự phản biện mô hình của chính mình

# Quá trình này giúp tác giả nâng cao tư duy độc lập, khả năng tự học và chuẩn bị tốt hơn cho các project nghiên cứu hoặc ứng dụng thực tế trong tương lai.

# **10. Kết luận**

Nghiên cứu này cho thấy rằng việc kết hợp dữ liệu giá cổ phiếu và dữ liệu tin tức là một hướng tiếp cận tiềm năng trong bài toán dự đoán xu hướng thị trường, phù hợp với các kết luận trong các nghiên cứu trước [3], [5].Thông qua việc xây dựng một pipeline Machine Learning hoàn chỉnh, từ thu thập dữ liệu, tiền xử lý, trích xuất đặc trưng, huấn luyện mô hình đến đánh giá kết quả, project đã cho thấy tiềm năng của việc tích hợp thông tin định lượng và định tính trong lĩnh vực tài chính.

Về mặt kết quả, các mô hình dựa trên dữ liệu giá thuần túy cho thấy khả năng dự đoán ổn định trong các giai đoạn thị trường ít biến động. Tuy nhiên, khi kết hợp thêm sentiment từ tin tức, hiệu suất dự đoán được cải thiện ở một số thời điểm, đặc biệt trong các giai đoạn xuất hiện nhiều thông tin liên quan đến doanh nghiệp. Điều này cho thấy cảm xúc thị trường, dù không phải lúc nào cũng quyết định, vẫn đóng vai trò hỗ trợ quan trọng trong việc dự đoán xu hướng giá cổ phiếu.

Qua việc so sánh nhiều mô hình khác nhau, từ các phương pháp Machine Learning truyền thống như Logistic Regression đến các mô hình phức tạp hơn như XGBoost, nghiên cứu đã chỉ ra rằng không tồn tại một mô hình tối ưu tuyệt đối cho mọi trường hợp. Hiệu quả dự đoán phụ thuộc lớn vào chất lượng dữ liệu, cách thiết kế đặc trưng và các giả định được đưa ra trong quá trình tiền xử lý. Điều này nhấn mạnh tầm quan trọng của việc đánh giá mô hình một cách toàn diện, thay vì chỉ dựa vào một chỉ số duy nhất như accuracy.

Xét về mục tiêu ban đầu, project đã đạt được các mục tiêu đề ra. Cụ thể, một pipeline Machine Learning end-to-end đã được xây dựng và triển khai thành công; nhiều mô hình khác nhau đã được huấn luyện và so sánh một cách có hệ thống; và hiệu quả dự đoán được đánh giá dựa trên nhiều tiêu chí khác nhau. Quan trọng hơn, nghiên cứu không chỉ dừng lại ở việc chạy mô hình, mà còn tập trung phân tích và thảo luận các kết quả thu được, từ đó rút ra những nhận định có giá trị về vai trò của dữ liệu tin tức trong dự đoán xu hướng giá cổ phiếu.

Tổng kết lại, mặc dù còn tồn tại những hạn chế nhất định, nghiên cứu này đã cung cấp một cái nhìn thực tế và có hệ thống về việc ứng dụng Machine Learning trong dự đoán thị trường tài chính. Kết quả của project có thể được xem là nền tảng cho các nghiên cứu và ứng dụng nâng cao hơn trong tương lai, đặc biệt trong bối cảnh dữ liệu lớn và phân tích thời gian thực ngày càng đóng vai trò quan trọng trong lĩnh vực tài chính.

# **Tài liệu tham khảo**

### Machine Learning & Stock Prediction

[1] T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions,” *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.

[2] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, “Predicting stock market index using fusion of machine learning techniques,” *Expert Systems with Applications*, vol. 42, no. 4, pp. 2162–2172, 2015.

[3] Y. Hu, J. Liu, J. Bian, H. Liu, and T.-Y. Chua, “Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction,” *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, 2018.

### Sentiment Analysis & Financial NLP

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[5] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” *arXiv preprint arXiv:1908.10063*, 2019.

[6] X. Ding, Y. Zhang, T. Liu, and J. Duan, “Using structured events to predict stock price movement: An empirical investigation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

### News Data & Event Databases

[7] K. Leetaru and P. A. Schrodt, “GDELT: Global data on events, location, and tone,” *ISA Annual Convention*, vol. 2, no. 4, pp. 1–49, 2013.

[8] GDELT Project, “Global Database of Events, Language, and Tone (GDELT),” Available:<https://www.gdeltproject.org/>

### Financial Data Sources

[9] Yahoo Finance, “Historical stock price data,” Available:<https://finance.yahoo.com/>

### Libraries & Frameworks

[10] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[12] HuggingFace, “Transformers: State-of-the-art natural language processing,” Available:<https://huggingface.co/transformers/>

### Blogs & Practical References (Supplementary)

[13] J. Brownlee, “Machine learning mastery with Python,” Available:<https://machinelearningmastery.com/>

[14] Medium – Towards Data Science, “Stock price prediction using machine learning,” Available: https://towardsdatascience.com/