**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

# BÁO CÁO HỌC PHẦN SEMINAR

Đề tài

# HIỆU QUẢ CỦA TRÍ TUỆ NHÂN TẠO TRONG DỰ ĐOÁN THỊ TRƯỜNG CHỨNG KHOÁN DỰA TRÊN HỌC MÁY


Sinh viên thực hiện: Hồ Tú Minh – 22022674

Đỗ Quang Dũng – 22022561

Phạm Long Nhật – 22022520


**HÀ NỘI – 2025**

Ý kiến đánh giá:……...

……

……

……

……

Điểm số: ……  Điểm chữ: ……

*Hà Nội, ngày 10 tháng 05 năm 2025*

Giảng viên đánh giá

*(Ký, ghi rõ họ tên)*

# MỤC LỤC

# DANH MỤC HÌNH ẢNH



# DANH MỤC BẢNG BIỂU



# MÃ NGUỒN DỰ ÁN

https://github.com/logthatisnatural/btl-seminar

# CHƯƠNG 1: GIỚI THIỆU

## 1.1. Giới thiệu đề tài

Thị trường chứng khoán luôn là một kênh đầu tư hấp dẫn để gia tăng vốn. Với sự phát triển của công nghệ truyền thông, thị trường chứng khoán ngày càng trở nên phổ biến đối với các nhà đầu tư cá nhân trong những thập kỷ gần đây. Khi số lượng cổ đông và công ty niêm yết tăng lên hàng năm, nhiều người đang tìm kiếm giải pháp để dự đoán xu hướng tương lai của thị trường chứng khoán, nhằm tối đa hóa lợi nhuận.

Việc dự báo thị trường chứng khoán là một vấn đề đầy thách thức với vô số yếu tố phức tạp tác động đến biến động giá cả. Khối lượng giao dịch hàng ngày trên thị trường chứng khoán là rất lớn, điều này tạo động lực mạnh mẽ cho các nhà nghiên cứu quan tâm đến việc nghiên cứu bài toán dự báo thị trường chứng khoán.

## 1.2. Mục đích, ý nghĩa của đề tài

Khả năng dự báo chính xác thị trường chứng khoán có ý nghĩa kinh tế to lớn, ảnh hưởng trực tiếp đến quyết định đầu tư và quản lý rủi ro của hàng triệu nhà đầu tư, từ cá nhân nhỏ lẻ đến các tổ chức lớn. Trong bối cảnh thị trường ngày càng biến động và phức tạp, việc đưa ra các dự báo đáng tin cậy không chỉ giúp tối ưu hóa lợi nhuận mà còn giảm thiểu rủi ro thua lỗ, bảo vệ tài sản của nhà đầu tư.

Tuy nhiên, như  việc dự báo thị trường chứng khoán luôn là một thách thức lớn do tính phi tuyến, ngẫu nhiên và chịu ảnh hưởng của vô số yếu tố vĩ mô, vi mô, và tâm lý. Các phương pháp phân tích truyền thống, dù là phân tích kỹ thuật hay phân tích cơ bản, thường đòi hỏi nhiều thời gian, công sức và đặc biệt là dễ bị chi phối bởi các thiên lệch hành vi của con người. Nhà đầu tư cá nhân, với nguồn lực hạn chế và thiếu kinh nghiệm, càng dễ rơi vào bẫy tâm lý đám đông, dẫn đến những quyết định sai lầm và thua lỗ đáng tiếc.

Mặc dù trí tuệ nhân tạo (AI) và học máy (ML) đã chứng minh được tiềm năng vượt trội trong nhiều lĩnh vực dự báo khác, vẫn tồn tại một khoảng trống nghiên cứu đáng kể về hiệu quả thực tế của chúng trong điều kiện thị trường chứng khoán cụ thể. Thị trường chứng khoán có những đặc thù riêng biệt như:

- Tính chất biến động của thị trường: Thường có biến động cao hơn, tính thanh khoản chưa đồng đều, và dễ bị ảnh hưởng bởi các yếu tố tin đồn hoặc tâm lý.
- Hệ thống thông tin chưa hoàn thiện: Thông tin đôi khi không minh bạch hoặc không được cập nhật kịp thời, gây khó khăn cho việc phân tích cơ bản.
- Khung pháp lý và chính sách: Có thể thay đổi nhanh chóng, tạo ra những yếu tố bất định.

Nghiên cứu này tập trung vào việc đánh giá khả năng của AI/ML trong việc xử lý lượng lớn dữ liệu phức tạp và đưa ra dự đoán về thị trường chứng khoán.

## 1.3. Phạm vi đề tài

Nghiên cứu tập trung vào việc đánh giá hiệu quả của trí tuệ nhân tạo (AI) và học máy (ML) trong dự báo thị trường chứng khoán, thông qua tích hợp phân tích kỹ thuật và phân tích cơ bản để cung cấp dự đoán chính xác và tín hiệu đầu tư. Phạm vi bao gồm:

- Kết hợp phân tích kỹ thuật và cơ bản: Sử dụng dữ liệu giá lịch sử và cảm xúc công chúng để xây dựng mô hình dự báo toàn diện, cải thiện khả năng dự đoán xu hướng giá và hành vi thị trường.
- Đánh giá hiệu suất mô hình học máy: So sánh độ chính xác, tin cậy và khả thi của các thuật toán học máy, dựa trên các chỉ số đánh giá chuẩn, để dự đoán giá cổ phiếu và tín hiệu mua/bán/nắm giữ.
- Ứng dụng và giới hạn: Cung cấp tín hiệu đầu tư thực tiễn, đồng thời xem xét hạn chế của AI trong dự báo thị trường và đề xuất hướng nghiên cứu mô hình kết hợp để nâng cao hiệu quả.

## 1.4. Phương pháp nghiên cứu

Nghiên cứu được tiến hành qua ba giai đoạn chính: thu thập và xử lý dữ liệu, huấn luyện mô hình học máy, và đánh giá hiệu suất mô hình. Phương pháp này dựa trên hai cách tiếp cận phân tích thị trường chứng khoán là phân tích kỹ thuật và phân tích cơ bản, tận dụng các kỹ thuật trí tuệ nhân tạo (AI) để dự báo xu hướng giá cổ phiếu. Quá trình được thiết kế cẩn thận để đảm bảo dữ liệu đầy đủ, mô hình tối ưu và đánh giá chính xác, nhằm cung cấp cái nhìn tổng quan về khả năng dự báo thị trường bằng AI.

Giai đoạn đầu tiên tập trung vào thu thập và xử lý dữ liệu từ hai nguồn chính. Dữ liệu giá cổ phiếu lịch sử, bao gồm giá mở, đóng, cao, thấp và khối lượng giao dịch, được lấy từ một nền tảng tài chính trực tuyến. Dữ liệu này được xử lý để tính toán các chỉ báo kỹ thuật như trung bình động, chỉ số sức mạnh tương đối và các chỉ báo xu hướng, tạo thành tập đặc trưng cho phân tích kỹ thuật. Đồng thời, các bài đăng công khai trên mạng xã hội liên quan đến cổ phiếu được thu thập, phân tích cảm xúc và giảm chiều đặc trưng bằng các kỹ thuật thống kê, hình thành dữ liệu cho phân tích cơ bản. Quá trình tiền xử lý đảm bảo dữ liệu được làm sạch và chuẩn hóa để phù hợp với mục tiêu dự báo.

Ở giai đoạn thứ hai, các mô hình học máy và học sâu được huấn luyện để dự báo dựa trên hai loại phân tích. Với phân tích kỹ thuật, các thuật toán hồi quy được sử dụng để dự đoán giá cổ phiếu dựa trên dữ liệu lịch sử. Với phân tích cơ bản, các thuật toán phân loại được áp dụng để dự đoán tín hiệu mua, bán hoặc nắm giữ, dựa trên cảm xúc công chúng từ

mạng xã hội. Các mô hình được huấn luyện trên tập dữ liệu đã tiền xử lý, với các kỹ thuật tối ưu hóa như chia tập huấn luyện và kiểm tra, cũng như điều chỉnh siêu tham số để cải thiện hiệu suất. Quá trình này nhằm khai thác các mẫu dữ liệu và đánh giá mức độ rủi ro đầu tư.

Giai đoạn cuối cùng là đánh giá hiệu suất mô hình bằng các chỉ số phù hợp. Đối với phân tích kỹ thuật, các chỉ số như độ chính xác dự đoán, sai số bình phương trung bình và sai số tuyệt đối trung bình được sử dụng để đo lường khả năng dự báo giá. Đối với phân tích cơ bản, các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu và diện tích dưới đường cong được áp dụng để đánh giá khả năng dự đoán tín hiệu. Kết quả từ các mô hình được so sánh để xác định điểm mạnh, điểm yếu và hạn chế. Nghiên cứu cũng đề xuất các hướng cải tiến, đặc biệt là việc kết hợp phân tích kỹ thuật và cơ bản trong các mô hình lai để nâng cao độ chính xác dự báo trong tương lai..

### 1.5. Nội dung báo cáo

Báo cáo gồm 5 chương, phản ánh quá trình nghiên cứu dự báo giá cổ phiếu bằng trí tuệ nhân tạo, từ lý thuyết đến thực nghiệm và đánh giá:

**Chương 1: Đặt vấn đề -** Trình bày bối cảnh, mục đích, ý nghĩa, phạm vi và phương pháp nghiên cứu dự báo giá cổ phiếu.

**Chương 2: Cơ sở lý thuyết -** Giới thiệu các khái niệm về phân tích kỹ thuật và cơ bản, cùng các lý thuyết nền tảng như học máy và xử lý dữ liệu tài chính.

**Chương 3: Thực nghiệm-** Mô tả quy trình thu thập, xử lý dữ liệu giá cổ phiếu và cảm xúc công chúng, cùng phương pháp xây dựng và đánh giá mô hình dự báo.

**Chương 4: Kết quả -** Tổng hợp kết quả dự báo giá và tín hiệu giao dịch, phân tích hiệu suất mô hình

**Chương 5: Đánh giá và kết luận** – Tổng hợp kết quả, đánh giá thực nghiệm và đưa ra đề xuất hướng cải tiến cho nghiên cứu tương lai..

## CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

### 2.1. Các giả thiết nền tảng về thị trường tài chính

Mô hình ngôn ngữ lớn (LLMs) là các hệ thống trí tuệ nhân tạo tiên tiến, được xây dựng dựa trên kiến trúc mạng nơ-ron sâu, huấn luyện trên các tập dữ liệu văn bản quy mô lớn nhằm xử lý và sinh ra ngôn ngữ tự nhiên. Các mô hình này hoạt động bằng cách dự đoán xác suất của từ hoặc cụm từ tiếp theo trong một chuỗi văn bản, từ đó thực hiện các

nhiệm vụ phức tạp như trả lời câu hỏi, giải thích khái niệm, dịch ngôn ngữ, và sinh văn bản. Trong lĩnh vực giáo dục, LLMs đóng vai trò quan trọng trong việc hỗ trợ học tập cá nhân hóa, cung cấp các giải pháp linh hoạt để giải đáp thắc mắc và hướng dẫn học sinh qua các bài tập học thuật. Đặc biệt, khả năng phân tích ngữ cảnh và tạo ra phản hồi có tính ngữ nghĩa cao giúp LLMs trở thành công cụ hiệu quả trong việc nâng cao chất lượng tự học, đặc biệt trong các môn học như Toán, Lịch sử, và Địa lý ở cấp độ phổ thông.

2.1.1. Giả thuyết Thị trường Hiệu quả (Efficient Market Hypothesis - EMH)

EMH do Eugene Fama phát triển, cho rằng giá tài sản đã phản ánh đầy đủ mọi thông tin có sẵn, khiến việc dự báo giá để kiếm lợi nhuận vượt trội là không thể. Giá cả biến động ngẫu nhiên và không thể dự đoán dựa trên thông tin cũ hay công khai.

Trong đó có ba dạng hiệu quả

- Dạng yếu: Giá phản ánh thông tin lịch sử giá và khối lượng. Phân tích kỹ thuật vô ích.
- Dạng bán mạnh: Giá phản ánh tất cả thông tin công khai. Cả phân tích kỹ thuật và cơ bản đều không hiệu quả.
- Dạng mạnh: Giá phản ánh mọi thông tin (công khai và nội bộ). Ngay cả người biết thông tin trong nội bộ cũng không thể kiếm lợi nhuận bất thường.

Ý nghĩa: EMH đặt ra thách thức cơ bản cho mọi nỗ lực dự báo, khuyến nghị chiến lược đầu tư thụ động

2.1.2. Giả thuyết Thị trường Thích ứng (Adaptive Market Hypothesis - AMH)

Quan điểm cải tiến: AMH của Andrew Lo cho rằng thị trường không hoàn toàn hiệu quả mà thích ứng và tiến hóa theo thời gian. Hiệu quả thị trường dao động tùy thuộc vào điều kiện và khả năng học hỏi của người tham gia.

Cơ hội dự báo: AMH chỉ ra rằng vẫn tồn tại cơ hội kiếm lợi nhuận bất thường trong những giai đoạn thị trường kém hiệu quả hoặc khi có thay đổi lớn. Tuy nhiên, các chiến lược cần được liên tục đổi mới và thích ứng.

Phù hợp với AI/ML: AMH cung cấp cơ sở lý thuyết vững chắc cho AI/ML vì các mô hình này có khả năng:

- Học hỏi liên tục: Phân tích lượng lớn dữ liệu và nhận diện mẫu hình phức tạp.
- Thích ứng: Tự điều chỉnh theo điều kiện thị trường thay đổi.
- Giảm thiểu thiên lệch cảm xúc: Đưa ra quyết định khách quan dựa trên dữ liệu.

Khả năng này giúp AI/ML khai thác những bất thường và cơ hội sinh lời tạm thời mà AMH đề cập trong một thị trường liên tục biến đổi.

## 2.2. Các phương pháp phân tích truyền thống

### 2.2.1. Phân tích Kỹ thuật (Technical Analysis)

**Nguyên lý cơ bản**: Lịch sử giá và khối lượng giao dịch chứa đựng thông tin về xu hướng tương lai, các mô hình dự đoán có thể sử dụng thông tin này để đưa ra dự đoán về giá đóng của cổ phiếu, từ đó cung cấp thông tin để nhà đầu tư ra quyết định.

**Bài toán dự đoán giá đóng trong phân tích kỹ thuật:** Bài toán được định nghĩa là một bài toán hồi quy (regression) nhằm dự đoán giá đóng cửa của cổ phiếu trong phiên giao dịch tiếp theo dựa trên dữ liệu lịch sử và các chỉ báo kỹ thuật.

- **Đầu vào:** Dữ liệu lịch sử giá cổ phiếu (giá mở, cao, thấp, đóng, khối lượng giao dịch) và các chỉ báo kỹ thuật như trung bình động, chỉ số sức mạnh tương đối (RSI), hoặc dải Bollinger.
- **Đầu ra:** Một giá trị số thực biểu thị giá đóng cửa dự đoán của cổ phiếu trong phiên giao dịch tiếp theo.

Mục tiêu là xây dựng một mô hình học máy có khả năng dự đoán chính xác giá đóng, hỗ trợ nhà đầu tư đưa ra quyết định giao dịch hiệu quả.

Các chỉ báo chính được sử dụng để xây dựng mô hình dự đoán bao gồm:

1. **SMA.** Chỉ báo này là giá đóng cửa trung bình của các giai đoạn gần nhất của một cổ phiếu cụ thể. Công thức toán học tính SMA được thể hiện như sau

$$\text{SMA}\ (t, N) = \frac{\sum_{k=1}^{N} CP(t\text{-}k)}{N}$$

*CP là giá đóng, N cho biết số ngày được đánh giá, và k thể hiện các ngày liên quan đến một CP cụ thể.*

2. **EMA.** Chỉ báo này theo dõi giá cổ phiếu tương tự như SMA, nhưng nó chú trọng hơn vào các giá đóng cửa gần đây nhất bằng cách gán trọng số cho chúng. Công thức sau mô tả quá trình tính toán của chỉ báo này:

$$EMA(t, \Delta) = \big(CP(t) - EMA(t-1)\big) \times \Gamma + EMA(t-1);\ \Gamma\ = \frac{2}{\Delta\ +\ 1}$$

*t là thời điểm hiện tại, Δ là số ngày, và Γ là hệ số làm mịn.*

3. **MACD.** Chỉ báo này cố gắng so sánh xu hướng ngắn hạn và dài hạn của giá cổ phiếu. Công thức sau mô tả chỉ báo này:

$$MACD = EMA(t,k) - EMA(t,d)$$

*k và d là các chu kỳ của xu hướng ngắn hạn và dài hạn. Thông thường, các giá trị này được coi là k=12 và d=26 ngày.*

4. **OBV.** Chỉ báo này sử dụng dòng chảy khối lượng cổ phiếu để thể hiện xu hướng giá và cho biết liệu khối lượng này đang chảy vào hay chảy ra. Công thức sau giải thích khái niệm OBV:

$$\mathrm{OBV} = \mathrm{OBV_{pr}} + \begin{cases} volume, if\ \mathrm{CP} > \mathrm{CP_{pr}} \\ 0, \mathrm{if\ CP} = \mathrm{CP_{pr}} \\ -\mathrm{volume}, \mathrm{if\ CP} = \mathrm{CP_{pr}} \end{cases}$$

*Trong đó OBVpr là OBV trước đó, volume là khối lượng giao dịch mới nhất, và CPpr là giá đóng cửa trước đó.*

5. **RSI.** Chỉ báo này đo lường tình trạng quá mua hoặc quá bán đặc trưng của một cổ phiếu. Thực tế, nó cho thấy xu hướng mua/bán một cổ phiếu. RSI được mô tả như sau:

$$\mathrm{RSI} = \frac{100}{1 + \mathrm{RS(t)}}, \mathrm{RS(t)} = \frac{\mathrm{AvgGain(t)}}{\mathrm{AvgLoss(t)}}$$

2.2.2. Phân tích Cơ bản (Fundamental Analysis)

**Phân tích cơ bản** là phương pháp truyền thống nhằm đánh giá giá trị nội tại của một doanh nghiệp thông qua việc kiểm tra báo cáo tài chính (doanh thu, lợi nhuận, P/E, P/B) và đánh giá triển vọng kinh doanh cùng ngành (mô hình kinh doanh, vị thế cạnh tranh, yếu tố vĩ mô). Mục tiêu chính là đưa ra các quyết định đầu tư dài hạn dựa trên sức khỏe và tiềm năng của công ty.

**Bài toán phân tích cảm xúc (Sentiment Analysis)**: Bài toán nghiên cứu trong phân tích cơ bản được định nghĩa là một bài toán phân loại nhị phân (binary classification) nhằm dự đoán tác động của cảm xúc công khai được thể hiện qua các phát ngôn trên mạng xã hội đối với xu hướng giá cổ phiếu. Đầu ra, đầu vào của bài toán như sau:

- Đầu vào: Một đoạn văn bản (một dòng tweet trên Twitter, thông báo trên web hay fanpage chính thức của công ty) liên quan đến thị trường chứng khoán nói chung
- Đầu ra: một nhãn phân loại nhị phân biểu thị tác động của đoạn văn bản đó lên giá cổ phiếu:
    - "Tích cực" (Positive): Nếu nội dung văn bản được dự đoán sẽ có tác động tăng giá cổ phiếu (tương đương với tín hiệu "Mua" hoặc "Giữ").
    - "Tiêu cực" (Negative): Nếu nội dung văn bản được dự đoán sẽ có tác động giảm giá cổ phiếu (tương đương với tín hiệu "Bán").

Mục tiêu sẽ là xây dựng một mô hình học máy có khả năng phân loại chính xác cảm xúc công khai từ dữ liệu văn bản không cấu trúc, từ đó cung cấp tín hiệu hỗ trợ quyết định đầu tư cho nhà giao dịch

## 2.3. Cơ sở lý thuyết Trí tuệ Nhân tạo, Học máy và Học sâu

2.3.1. Học máy có giám sát (Supervised Learning)

**Linear Regression**: Hồi quy tuyến tính là một trong những thuật toán học máy đơn giản nhưng mạnh mẽ và được sử dụng rộng rãi để mô hình hóa mối quan hệ giữa một biến phụ thuộc (biến mục tiêu) và một hoặc nhiều biến độc lập (biến dự báo). Về cơ bản, nó tìm cách vẽ một đường thẳng (hoặc mặt phẳng, siêu mặt phẳng trong không gian nhiều chiều) phù hợp nhất với dữ liệu, sao cho khoảng cách từ các điểm dữ liệu đến đường thẳng đó là nhỏ nhất.

Mục tiêu của hồi quy tuyến tính là dự đoán một giá trị liên tục (ví dụ: giá nhà, doanh số bán hàng, nhiệt độ). Mô hình giả định rằng mối quan hệ giữa các biến có thể được biểu diễn dưới dạng tuyến tính. Công thức tổng quát của hồi quy tuyến tính đơn giản là y = ax + b, trong đó y là biến phụ thuộc, x là biến độc lập, a là hệ số góc và b là hệ số chặn. Đối với hồi quy tuyến tính đa biến, công thức sẽ phức tạp hơn:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Để tìm ra đường thẳng "phù hợp nhất", thuật toán sử dụng phương pháp Bình phương Tối thiểu - Ordinary Least Squares. Phương pháp này tính toán các hệ số (a,b hoặc b0,...,bn) sao cho tổng bình phương của các phần dư (hiệu số giữa giá trị thực tế và giá trị dự đoán) là nhỏ nhất.

**Logistic Regression**: Hồi quy Logistic là một thuật toán học máy có giám sát, mặc dù có từ "Regression" (hồi quy) trong tên nhưng nó chủ yếu được sử dụng cho các bài toán phân loại (classification), đặc biệt là phân loại nhị phân (ví dụ: có/không, đúng/sai,

0/1). Khác với hồi quy tuyến tính dự đoán giá trị liên tục, hồi quy Logistic dự đoán xác suất một trường hợp thuộc về một lớp nhất định.

Về cơ bản, nó hoạt động bằng cách lấy một tổ hợp tuyến tính của các biến đầu vào (giống như hồi quy tuyến tính), sau đó truyền kết quả này qua một hàm Sigmoid (còn gọi là hàm logistic). Hàm Sigmoid có công thức:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Hàm có nhiệm vụ "ép" mọi giá trị đầu vào thực z về một giá trị trong khoảng từ 0 đến 1, có thể được diễn giải như một xác suất. Nếu xác suất này vượt quá một ngưỡng nhất định (thường là 0.5), đối tượng sẽ được phân loại vào một lớp (ví dụ: lớp 1), ngược lại sẽ thuộc lớp còn lại (ví dụ: lớp 0).

Hồi quy Logistic tìm cách tối ưu hóa các hệ số (trọng số) của mô hình để sao cho xác suất dự đoán gần với nhãn thực tế nhất. Điều này thường được thực hiện bằng cách tối thiểu hóa một hàm mất mát (loss function) như Cross-Entropy. Mô hình này tạo ra một đường biên phân chia tuyến tính (linear decision boundary) trong không gian đặc trưng để phân tách các lớp dữ liệu.

**Support Vector Machine (SVM)**: Máy Vector Hỗ trợ (SVM) là một thuật toán học máy có giám sát mạnh mẽ và linh hoạt, được sử dụng rộng rãi cho cả bài toán phân loại (classification) và hồi quy (regression), mặc dù phổ biến hơn trong phân loại.

Ý tưởng cốt lõi của SVM là tìm kiếm một siêu mặt phẳng (hyperplane) tối ưu trong không gian đa chiều để phân chia các lớp dữ dữ liệu. Đối với bài toán phân loại nhị phân, siêu mặt phẳng này có nhiệm vụ phân tách hai lớp dữ liệu một cách rõ ràng nhất có thể.

Điểm đặc biệt của SVM là nó không chỉ tìm một siêu mặt phẳng bất kỳ mà là siêu mặt phẳng có khoảng cách lề (margin) lớn nhất giữa nó và các điểm dữ liệu gần nhất của mỗi lớp. Các điểm dữ liệu gần siêu mặt phẳng nhất, nằm trên lề, được gọi là vector hỗ trợ (support vectors). Chính những vector hỗ trợ này quyết định vị trí và hướng của siêu mặt phẳng, và chúng là những điểm dữ liệu quan trọng nhất trong quá trình huấn luyện mô hình. Bằng cách tối đa hóa lề, SVM cố gắng đảm bảo rằng mô hình có khả năng tổng quát hóa tốt và ít bị ảnh hưởng bởi nhiễu.

**k-Nearest Neighbors (k-NN)** là một thuật toán học máy có giám sát đơn giản nhưng hiệu quả, được sử dụng cho cả bài toán phân và hồi quy. Ý tưởng cốt lõi của k-NN là dựa

trên nguyên tắc rằng các điểm dữ liệu tương tự nhau sẽ nằm gần nhau trong không gian đặc trưng.

Trong k-NN, để dự đoán nhãn (hoặc giá trị) của một điểm dữ liệu mới, thuật toán sẽ: Tìm k điểm dữ liệu gần nhất (hàng xóm) trong tập huấn luyện, dựa trên một thước đo khoảng cách (thường là khoảng cách Euclidean). Đối với bài toán phân loại, nhãn của điểm dữ liệu mới được xác định bằng cách lấy đa số phiếu (majority vote) từ nhãn của k hàng xóm gần nhất

Tham số k (số lượng hàng xóm) là một siêu tham số quan trọng, cần được điều chỉnh để cân bằng giữa độ nhạy với nhiễu (k nhỏ) và độ tổng quát hóa (k lớn). k-NN là một thuật toán "lười" (lazy learning), nghĩa là nó không xây dựng mô hình rõ ràng trong giai đoạn huấn luyện mà chỉ lưu trữ dữ liệu huấn luyện và thực hiện tính toán tại thời điểm dự đoán.

**Cây quyết định (Decision Tree)** là một thuật toán học máy có giám sát, được sử dụng cho cả bài toán phân loại và hồi quy. Nó hoạt động bằng cách biểu diễn các quyết định dưới dạng một cấu trúc cây, trong đó mỗi nút (node) đại diện cho một điều kiện kiểm tra trên một đặc trưng, mỗi nhánh (branch) đại diện cho kết quả của điều kiện đó, và các lá (leaf nodes) đại diện cho nhãn lớp hoặc giá trị dự đoán.

**Random Forest** là một thuật toán học máy dựa trên tập hợp (ensemble), sử dụng nhiều cây quyết định để cải thiện độ chính xác và độ ổn định so với cây quyết định đơn lẻ. Nó được sử dụng cho cả bài toán phân loại và hồi quy. Đối với bài toán phân loại, nhãn dự đoán cuối cùng được xác định bằng cách lấy đa số phiếu (majority vote) từ tất cả các cây quyết định. Đối với bài toán hồi quy, giá trị dự đoán là trung bình của các giá trị dự đoán từ các cây.

**XGBoost** là một thuật toán học máy dựa trên kỹ thuật tăng cường độ dốc (gradient boosting), được tối ưu hóa để đạt hiệu suất cao và tốc độ nhanh. Đây là một trong những thuật toán mạnh mẽ nhất trong các bài toán phân loại và hồi quy. XGBoost nổi bật nhờ khả năng xử lý các tập dữ liệu lớn, tính linh hoạt trong việc tùy chỉnh các siêu tham số, và khả năng cung cấp các dự đoán chính xác cao trong nhiều bài toán thực tế

2.3.2. Học sâu (Deep Learning)

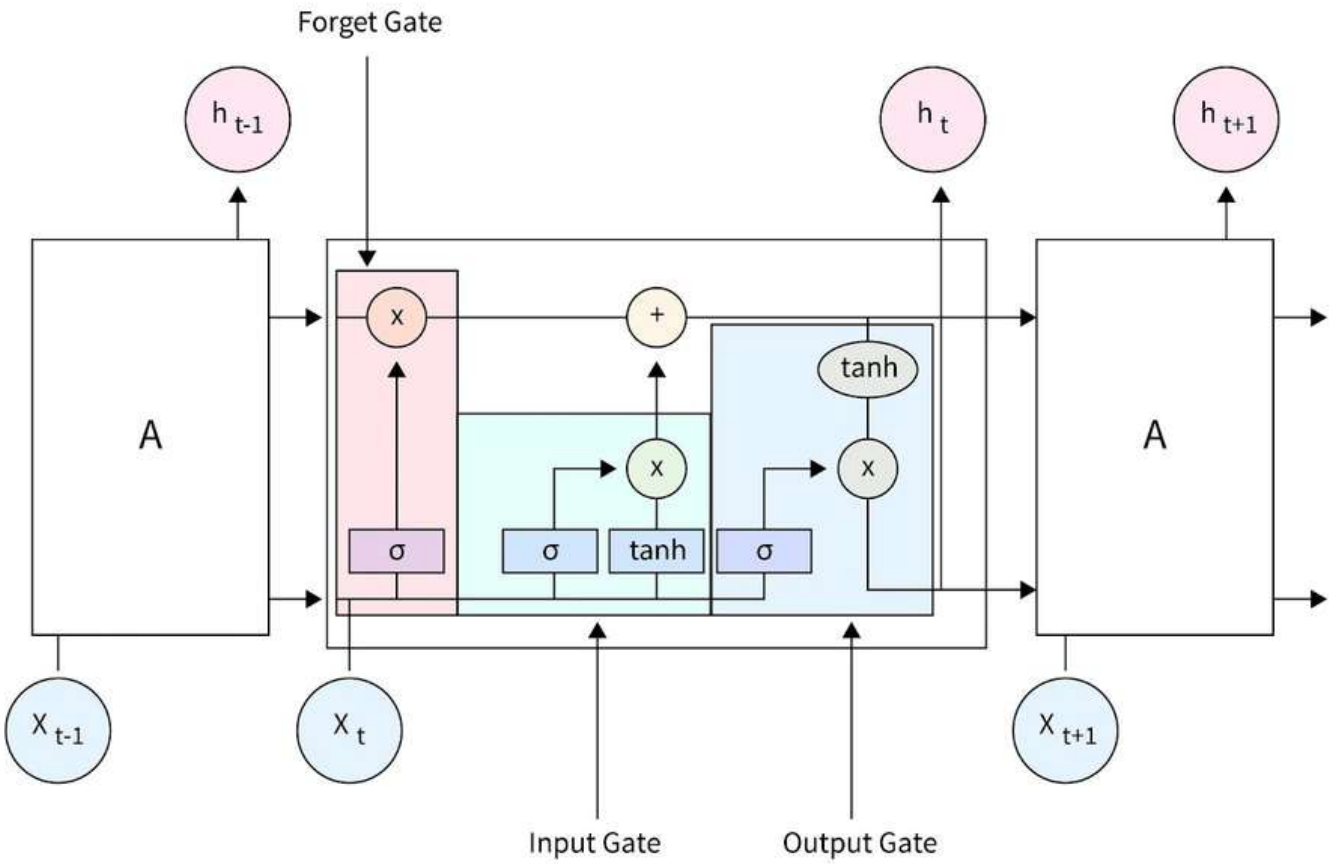
**Mạng nơ-ron hồi quy (LSTM):** Mạng LSTM là một dạng đặc biệt của mạng nơ-ron hồi quy (Recurrent Neural Network - RNN), được thiết kế để giải quyết vấn đề khó khăn trong việc học các phụ thuộc dài hạn (long-term dependencies) mà các RNN truyền thống thường gặp phải. Vấn đề này thường được gọi là "vấn đề gradient biến mất" (vanishing gradient problem) hoặc "vấn đề gradient bùng nổ" (exploding gradient

problem), khiến cho các RNN truyền thống khó có thể lưu giữ thông tin quan trọng từ các bước thời gian xa trong quá khứ khi xử lý chuỗi dữ liệu dài.

Kiến trúc Cốt lõi: Tế bào nhớ (Memory Cell) và Cổng (Gates): Điểm khác biệt cốt lõi và là "trái tim" của kiến trúc LSTM là tế bào nhớ (memory cell), hay còn gọi là trạng thái tế bào (cell state), và một tập hợp các cổng (gates). Tế bào nhớ có khả năng mang thông tin qua các bước thời gian, cho phép nó ghi nhớ hoặc quên đi thông tin một cách có chọn lọc. Các cổng chính là cơ chế điều khiển dòng chảy thông tin vào và ra khỏi tế bào nhớ. Mỗi cổng là một mạng nơ-ron sigmoid layer, có nhiệm vụ xuất ra một giá trị từ 0 đến 1, cho biết mức độ thông tin được "cho phép" đi qua.

Cụ thể, một đơn vị LSTM tại thời điểm t bao gồm:

1. Trạng thái tế bào (Ct - Cell State): Đây là "băng chuyền" chính chạy xuyên suốt chuỗi, mang theo thông tin dài hạn
2. Cổng quên (ft - Forget Gate): Quyết định thông tin nào từ trạng thái tế bào trước đó (Ct−1) nên được "quên đi" hoặc loại bỏ.
3. Cổng đầu vào (it - Input Gate: Quyết định thông tin mới nào từ đầu vào hiện tại nên được "thêm vào" trạng thái tế bào.
4. Cập nhật Trạng thái Tế bào: Kết hợp thông tin từ cổng quên và cổng đầu vào để tạo ra trạng thái tế bào mới (Ct).
5. Cổng đầu ra (ot - Output Gate): Quyết định phần nào của trạng thái tế bào mới (Ct) sẽ được xuất ra làm trạng thái ẩn hiện tại (ht) - đầu ra của đơn vị LSTM cho bước thời gian này.




*Hình 1. Cơ chế hoạt động của mạng LSTM*

# CHƯƠNG 3: THỰC NGHIỆM

Quá trình thực nghiệm của nhóm em bao gồm ba giai đoạn chính: thu thập và xử lý dữ liệu, huấn luyện mô hình, và đánh giá hiệu suất mô hình, áp dụng cho cả hai phương pháp phân tích kỹ thuật và phân tích cơ bản

## 3.1. Phân tích kỹ thuật

3.1.1. Thu thập và xử lý dữ liệu

- **Nguồn dữ liệu**: Dữ liệu lịch sử giá cổ phiếu được thu thập từ trang web "Yahoo Finance", tập trung vào cổ phiếu AAPL (Công ty Apple Inc.) trong khoảng thời gian hơn 10 năm, từ năm 2010 đến năm 2021. Tập dữ liệu bao gồm các thông tin chi tiết như giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất, giá trung bình và khối lượng giao dịch, với đặc điểm không có mẫu dữ liệu bị thiếu. Đây là dữ liệu có cấu trúc, dễ dàng truy cập công khai, phù hợp để áp dụng các thuật toán học máy trong phân tích kỹ thuật. Lượng dữ liệu lớn và liên tục này cung cấp nền tảng vững chắc để các mô hình học máy nhận diện các mẫu hình giá và dự đoán xu hướng thị trường.
- **Xây dựng đặc trưng (Feature Engineering)**: Dữ liệu giá cổ phiếu thô không được sử dụng trực tiếp mà cần được chuyển đổi thành các chỉ số tài chính để làm đầu vào cho các mô hình học máy. Các chỉ số tài chính quan trọng được tính toán bao gồm Đường trung bình động đơn giản (SMA), Đường trung bình động lũy thừa (EMA), Chỉ số sức mạnh tương đối (RSI), Chỉ số MACD, và Khối lượng cân bằng (OBV).
- Các bước xử lý dữ liệu
    - Kiểm tra và xử lý dữ liệu bị thiếu: Dữ liệu từ Yahoo Finance không có mẫu bị thiếu, đảm bảo tính toàn vẹn của tập dữ liệu. Tuy nhiên, để đảm bảo chất lượng, dữ liệu được kiểm tra để loại bỏ bất kỳ giá trị bất thường hoặc lỗi nhập liệu.
    - Chuẩn hóa dữ liệu: Các chỉ số tài chính được chuẩn hóa (ví dụ: chia tỷ lệ hoặc chuẩn hóa min-max) để đảm bảo các đặc trưng có cùng thang đo, giúp cải thiện hiệu suất của các thuật toán học máy như hồi quy tuyến tính (Linear Regression) hoặc mạng nơ-ron (LSTM).
    - Tạo đặc trưng số hóa: Các chỉ số tài chính (SMA, EMA, RSI, MACD, OBV) được tính toán từ dữ liệu giá thô và tổ hợp thành các vector đặc trưng số, sẵn sàng làm đầu vào cho các mô hình học máy.

3.1.2. Huấn luyện mô hình

**Các mô hình học máy được sử dụng**

Các mô hình học máy được sử dụng: Nghiên cứu này đã áp dụng hai thuật toán hồi quy chính để dự đoán giá đóng cửa của cổ phiếu AAPL:

- **Hồi quy tuyến tính (Linear Regression - LR)**: Một thuật toán đơn giản nhưng hiệu quả, phù hợp cho việc dự đoán giá trị liên tục như giá cổ phiếu. Mô hình này tìm cách tối ưu hóa một hàm tuyến tính để dự đoán giá dựa trên các chỉ số tài chính.

- **Mạng nơ-ron hồi quy dài ngắn hạn (Long Short-Term Memory - LSTM)**: Một loại mạng nơ-ron tái phát (RNN) đặc biệt, được thiết kế để xử lý dữ liệu chuỗi thời gian, rất phù hợp với dữ liệu giá cổ phiếu có tính chất tuần tự. LSTM có khả năng ghi nhớ các mẫu dài hạn trong dữ liệu.

**Quy trình huấn luyện (2 bước)**

1. Phân chia dữ liệu
    - Tập dữ liệu lịch sử giá cổ phiếu được chia thành ba tập con:
        1. Tập huấn luyện (khoảng 70-80%): Được sử dụng để huấn luyện các mô hình, giúp chúng học các mẫu hình và mối quan hệ giữa các chỉ số tài chính và giá đóng cửa
        2. Tập đánh giá: (khoảng 10-15%): Được dùng để tinh chỉnh các siêu tham số của mô hình (ví dụ: số lớp ẩn trong LSTM, hệ số học tập) và đánh giá sơ bộ hiệu suất trong quá trình huấn luyện, nhằm tránh hiện tượng overfit.
        3. Tập kiểm thử (khoảng 10-15%): Một tập dữ liệu độc lập, không được sử dụng trong quá trình huấn luyện, để đánh giá hiệu suất cuối cùng của mô hình. Kết quả trên tập kiểm thử phản ánh khả năng tổng quát hóa của mô hình trên dữ liệu mới.
2. Huấn luyện và tinh chỉnh

Mỗi mô hình (LR và LSTM) được huấn luyện độc lập trên tập huấn luyện. Trong quá trình này, các tham số nội bộ của mô hình được điều chỉnh để giảm thiểu lỗi dự đoán giá đóng cửa.

Các siêu tham số của mô hình được tinh chỉnh thông qua các kỹ thuật như tìm kiếm lưới (Grid Search) hoặc tìm kiếm ngẫu nhiên (Random Search), sử dụng tập xác thực để xác định bộ siêu tham số tối ưu, đảm bảo hiệu suất tốt nhất.

3.1.3. Đánh giá hiệu suất của mô hình

Để đánh giá hiệu quả của các mô hình học máy (Hồi quy tuyến tính - LR và Mạng nơ-ron hồi quy dài ngắn hạn - LSTM) trong phân tích kỹ thuật dự đoán giá đóng cửa của cổ phiếu AAPL, nghiên cứu đã sử dụng một tập hợp các chỉ số đánh giá hiệu suất chuyên biệt cho bài toán hồi quy. Các chỉ số này tập trung vào việc đo lường sai số dự đoán giá trị liên tục và mức độ mô hình giải thích được sự biến thiên của dữ liệu, các chỉ số bao gồm:

1. **R-squared ($R^2$)**: đo lường mức độ giải thích của mô hình đối với sự biến thiên của giá đóng. Giá trị $R^2$ gần 1 cho thấy mô hình có khả năng dự đoán tốt và giải thích gần như toàn bộ sự thay đổi trong dữ liệu.

2. **Explained Variation**: tương tự $R^2$, chỉ số đo lường tỷ lệ biến thiên của dữ liệu được mô hình giải thích. Chỉ số này thường được sử dụng để đánh giá mức độ mô hình tái hiện được các mẫu hình trong dữ liệu. Chỉ số này tập trung vào sự biến thiên của dữ đoán so với giá trị trung bình, cung cấp một góc nhìn bổ sung về hiệu suất mô hình
3. **Mean Absolute Percentage Error (MAPE)**: đo lường sai số trung bình dưới dạng phần trăm giữa giá trị dự đoán và giá trị thực tế, giúp đánh giá sai số tương đối của mô hình. MAPE thấp cho thấy mô hình có độ chính xác cao trong việc dự đoán giá
4. **Root Mean Squared Error (RMSE)**: đo lường sai số bình phương trung bình, nhấn mạnh các sai số lớn hơn do bình phương các sai lệch, được sử dụng đặc biệt trong các trường hợp cần giảm thiểu các sai số lớn
5. **Mean Absolute Error (MAE)**: đo lường sai số tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế, cung cấp một cách tiếp cận trực quan để đánh giá sai số mà không nhấn mạnh quá mức vào các sai số lớn

## 3.2 Phân tích cơ bản

3.2.1. Thu thập và xử lý dữ liệu

- **Nguồn dữ liệu**: Dữ liệu cảm xúc công khai được thu thập từ nền tảng Twitter, tập trung vào các tweet liên quan đến công ty Apple Inc. (mã AAPL). Việc lựa chọn Twitter là do đây là một nguồn thông tin nhanh chóng, cập nhật liên tục và phản ánh trực tiếp tâm lý cộng đồng về các sự kiện kinh tế và công ty. Dữ liệu thu thập là các đoạn văn bản (tweet) không cấu trúc, có thể đến từ các hãng tin tức uy tín hoặc từ các cá nhân, nhà đầu tư. Tập dữ liệu bao gồm gần 6000 tweet. Đây là một lượng dữ liệu đủ lớn để các thuật toán học máy có thể học được các mẫu hình cảm xúc.
- **Gán nhãn dữ liệu (Data Labeling)**: Đây là một bước đóng góp quan trọng của tác giả trong việc biến dữ liệu văn bản thô thành định dạng có thể sử dụng cho học máy giám sát. Mỗi tweet được gán một nhãn nhị phân:
    - Nhãn 1 (Tích cực): Nếu nội dung tweet có tác động tích cực hoặc trung tính lên thị trường chứng khoán (ví dụ: tin tốt về lợi nhuận, sản phẩm mới, đối tác). Theo quy ước trong nghiên cứu, điều này tương ứng với tín hiệu "Mua" hoặc "Giữ".
    - Nhãn 0 (Tiêu cực): Nếu nội dung tweet có tác động tiêu cực lên thị trường (ví dụ: tin xấu về kiện tụng, doanh số giảm, bê bối). Điều này tương ứng với tín hiệu "Bán".
- Việc gán nhãn này có thể được thực hiện thông qua chú thích thủ công bởi chuyên gia hoặc thông qua các quy tắc dựa trên từ khóa (rule-based) và sau đó được tinh chỉnh. Sự chính xác trong việc gán nhãn là yếu tố then chốt quyết định chất lượng của mô hình cuối cùng.

- Dữ liệu văn bản thô từ Twitter không thể trực tiếp đưa vào các mô hình học máy. Ta cần thực hiện các bước tiền xử lý nhằm chuyển đổi và tối ưu hoá những dữ liệu này
- Các bước chuẩn hóa định dạng và làm sạch dữ liệu:
    - Mục tiêu chính của bước này là làm sạch, chuẩn hóa và chuyển đổi dữ liệu văn bản thô, nhiễu loạn thành một định dạng gọn gàng, nhất quán và có cấu trúc hơn, sẵn sàng để được mã hóa số hóa (numerical representation) và đưa vào các mô hình học máy. Nếu bỏ qua hoặc thực hiện không kỹ lưỡng, dữ liệu thô sẽ chứa rất nhiều nhiễu, làm giảm đáng kể hiệu suất và độ chính xác của mô hình phân tích cảm xúc. Các tác vụ tiền xử lý bao gồm
    - **Xử lý các mẫu bị thiếu**Bước này có nhiệm vụ đảm bảo tính toàn vẹn của tập dữ liệu. Dữ liệu văn bản thô có thể chứa các dòng trống rỗng, hoặc các tweet chỉ chứa các ký tự không có ý nghĩa (ví dụ: chỉ dấu cách, ký tự đặc biệt không phải chữ). Đầu tiên ta kiểm tra các trường dữ liệu văn bản để tìm các giá trị None, chuỗi rỗng (""), hoặc chuỗi chỉ chứa khoảng trắng/ký tự không liên quan, sau đó chọn các phương pháp xử lý các mẫu bị thiếu. Phương pháp phổ biến nhất trong NLP là loại bỏ các mẫu bị thiếu hoàn toàn hoặc các mẫu rỗng. Việc "điền" dữ liệu văn bản thường không khả thi hoặc có thể gây sai lệch lớn. Ví dụ, nếu một tweet chỉ chứa các ký tự không phải chữ cái sau khi các bước tiền xử lý khác được áp dụng, nó cũng có thể được coi là rỗng và loại bỏ.
    - **Chuyển đổi chữ thường**: Ở bước này ta thực hiện chuẩn hóa các từ, đảm bảo rằng cùng một từ nhưng được viết với cách viết hoa khác nhau không bị coi là hai từ riêng biệt. Ví dụ, "Apple", "apple" và "APPLE" đều được chuyển thành "apple". Điều này giúp giảm đáng kể kích thước từ vựng (vocabulary size) và sự thưa thớt (sparsity) của dữ liệu.
    - **Loại bỏ ký tự đặc biệt và số**: Bước này loại bỏ các yếu tố nhiễu không mang ý nghĩa cảm xúc trực tiếp hoặc không cần thiết cho việc phân tích. Trong ngữ cảnh phân tích cảm xúc tweet, các ký tự này thường làm tăng chiều dữ liệu một cách không cần thiết và có thể gây nhiễu cho mô hình, các loại ký tự thường được loại bỏ bao gồm: dấu câu, biểu tượng cảm xúc, liên kết URL, tên người dùng được nhắc đến, hashtag và số
    - **Loại bỏ từ dừng (Stop-word removal):** Xóa bỏ những từ rất phổ biến trong ngôn ngữ nhưng lại mang rất ít hoặc không có ý nghĩa phân loại độc lập cho bài toán phân tích cảm xúc. Các từ này thường là các giới từ, liên từ, mạo từ, đại từ... Chúng không giúp mô hình phân biệt được cảm xúc tích cực, tiêu cực hay trung tính mà chỉ làm tăng kích thước dữ liệu và gây nhiễu.Ta Sử dụng các danh sách từ dừng chuẩn (stop-word lists) được cung cấp bởi các thư viện NLP như NLTK (đối với tiếng Anh) hoặc các bộ dữ

liệu stop-word tiếng Việt. Mỗi từ trong tweet sẽ được so sánh với danh sách này, nếu trùng khớp sẽ bị loại bỏ, giúp ta giảm chiều dữ liệu, tăng tốc độ xử lý, giúp mô hình tập trung vào các từ khóa mang ý nghĩa thực sự.

    - **Rút gọn từ (Lemmatization):** Đưa các từ về dạng gốc của chúng để giảm sự đa dạng của từ vựng và chuẩn hóa dữ liệu. Nhiều từ có cùng ý nghĩa cơ bản nhưng tồn tại dưới nhiều dạng khác nhau do chia động từ, số ít/số nhiều, hoặc các biến thể ngữ pháp. Việc rút gọn giúp các dạng biến thể này được coi là cùng một thực thể, giảm sự thưa thớt của dữ liệu và cải thiện khả năng tổng quát hóa của mô hình. Đây là quá trình phức tạp, dựa trên từ điển và phân tích ngữ cảnh (Phân tích hình thái học) để đưa từ về dạng gốc có ý nghĩa và đúng ngữ pháp của nó (lemma). Nó chính xác hơn nhưng tốn kém về mặt tính toán hơn.
- Chuyển đổi dữ liệu từ dạng văn bản sang dạng số:
    - Các mô hình học máy yêu cầu đầu vào là dữ liệu số. Dữ liệu văn bản từ các tweet đã được tiền xử lý cần được chuyển đổi thành các vector đặc trưng số. Kỹ thuật được sử dụng trong nghiên cứu là **Continuous Bag of Words (CBOW)**
    - Phương pháp này tạo một vector biểu diễn dựa trên ngữ cảnh của các từ trong tập dữ liệu. CBOW, một biến thể của mô hình Word2Vec, dự đoán một từ dựa trên các từ xung quanh nó, tạo ra các vector dày đặc biểu diễn ngữ nghĩa của từ. Trong ngữ cảnh nghiên cứu, CBOW có thể được sử dụng để chuyển đổi các tweet thành các vector đặc trưng số, phù hợp với các thuật toán học máy như SVM hoặc ANN, cho phép mô hình nắm bắt thông tin ngữ nghĩa từ dữ liệu văn bản.

3.2.2. Huấn luyện mô hình

Nghiên cứu này đã thử nghiệm một loạt các thuật toán phân loại phổ biến và hiệu quả để tìm ra mô hình phù hợp nhất cho bài toán:

- **Logistic Regression (LR)**: Một thuật toán phân loại tuyến tính đơn giản nhưng hiệu quả, phù hợp cho các bài toán phân loại nhị phân.
- **Naive Bayes (Gaussian Naive Bayes - GNB và Bernoulli Naive Bayes BNB)**: Các thuật toán dựa trên định lý Bayes với giả định độc lập có điều kiện giữa các đặc trưng. GNB phù hợp với đặc trưng liên tục (như sau PCA), BNB phù hợp với đặc trưng nhị phân (như BoW).
- **Decision Tree (DT)**: Mô hình phân loại dựa trên cây, dễ giải thích, phân chia dữ liệu dựa trên các quy tắc.

- **Random Forest (RF)**: Một ensemble method (phương pháp kết hợp) xây dựng nhiều cây quyết định và kết hợp các dự đoán của chúng, giúp giảm overfit và cải thiện độ chính xác.
- **k-Nearest Neighbors (KNN)**: Thuật toán phân loại dựa trên khoảng cách, phân loại một điểm dữ liệu mới dựa trên phần lớn nhãn của k hàng xóm gần nhất.
- **Support Vector Machine (SVM)**: Một thuật toán mạnh mẽ tìm một siêu phẳng tối ưu để phân tách dữ liệu thành các lớp khác nhau. SVM nổi tiếng với khả năng hoạt động tốt trên dữ liệu chiều cao và thường là lựa chọn mạnh mẽ cho bài toán phân loại văn bản.
- **XGBoost (eXtreme Gradient Boosting - XGB)**: Một thuật toán boosting dựa trên cây quyết định, nổi tiếng về tốc độ và hiệu suất cao.

**Quy trình huấn luyện (2 bước)**

1. Phân chia dữ liệu

Tập dữ liệu đã được tiền xử lý và biểu diễn dạng số hóa được chia thành ba tập con riêng biệt. Mục đích của việc này là để mô phỏng một cách chân thực nhất việc mô hình sẽ hoạt động như thế nào trên dữ liệu mới trong thế giới thực và để tránh hiện tượng overfit.

4. Tập huấn luyện: Đây là phần lớn nhất của dữ liệu (thường là 70-80%). Tập này được sử dụng để các thuật toán học máy "học" các mẫu hình, các quy tắc và mối quan hệ giữa các đặc trưng văn bản và nhãn cảm xúc tương ứng. Mô hình điều chỉnh các trọng số và tham số nội bộ của nó dựa trên dữ liệu này.
5. Tập đánh giá: Một phần nhỏ hơn của dữ liệu (thường là 10-15%). Tập này đóng vai trò là "đôi mắt" của nhà nghiên cứu trong quá trình huấn luyện. Nó được sử dụng để tinh chỉnh các siêu tham số (hyperparameters) của mô hình (ví dụ: số lượng cây trong Random Forest, tham số C trong SVM, v.v.) và để đánh giá sơ bộ hiệu suất của mô hình trong quá trình huấn luyện. Việc sử dụng tập xác thực giúp ngăn chặn mô hình bị overfit với tập huấn luyện và đảm bảo rằng mô hình có thể tổng quát hóa tốt hơn.
6. Tập kiểm thử: Phần còn lại của dữ liệu (10-15%). Đây là một phần dữ liệu hoàn toàn độc lập và chưa từng được mô hình "nhìn thấy" dưới bất kỳ hình thức nào trong suốt quá trình huấn luyện và tinh chỉnh. Tập này được dành riêng để đánh giá hiệu suất cuối cùng và khách quan của mô hình. Bất kỳ kết quả nào được báo cáo trên tập kiểm thử được coi là chỉ số đáng tin cậy nhất về khả năng tổng quát hóa của mô hình trên dữ liệu mới, chưa biết.

2. Huấn luyện và tinh chỉnh

   Mỗi thuật toán học máy đã chọn được huấn luyện độc lập trên tập huấn luyện. Trong quá trình này, các tham số nội bộ của mô hình được điều chỉnh để tối thiểu hóa lỗi dự đoán.

   Song song với việc huấn luyện, các siêu tham số  của mỗi mô hình cũng được tinh chỉnh. Siêu tham số là các tham số không được học trực tiếp từ dữ liệu mà cần được đặt trước khi quá trình huấn luyện bắt đầu (ví dụ: số lượng n-gram tối đa, loại kernel trong SVM, learning rate, …). Quá trình tinh chỉnh siêu tham số thường sử dụng các kỹ thuật như tìm kiếm lưới (Grid Search) hoặc tìm kiếm ngẫu nhiên (Random Search) kết hợp với đánh giá hiệu suất trên tập xác thực để tìm ra bộ siêu tham số mang lại hiệu suất tốt nhất cho mô hình.

3.2.3. Đánh giá hiệu suất của mô hình

Để đánh giá một cách toàn diện khả năng phân loại của các mô hình, nhiều metric đánh giá đã được sử dụng:

3. **Confusion Matrix**: Cung cấp cái nhìn tổng quan về số lượng dự đoán đúng và sai của mô hình (True Positive, True Negative, False Positive, False Negative).
4. **Accuracy**: Tỷ lệ các trường hợp được phân loại đúng trên tổng số trường hợp. Đây là một độ đo cơ bản nhưng có thể gây hiểu lầm nếu dữ liệu không cân bằng.
5. **Precision**: Tỷ lệ các trường hợp dự đoán là "tích cực" thực sự là "tích cực". Quan trọng trong các trường hợp mà False Positive gây tốn kém (ví dụ: dự đoán "Mua" sai).
6. **Recall/Sensitivity**: Tỷ lệ các trường hợp "tích cực" thực sự được mô hình dự đoán đúng. Quan trọng khi False Negative gây tốn kém (ví dụ: bỏ lỡ cơ hội "Mua").
7. **F1-score**: Giá trị trung bình điều hòa của Precision và Recall, hữu ích khi cần cân bằng giữa hai độ đo này.
8. **Đường cong ROC** (Receiver Operating Characteristic Curve) và **AUC** (Area Under the Curve): ROC biểu diễn mối quan hệ giữa True Positive Rate và False Positive Rate ở các ngưỡng phân loại khác nhau. AUC là diện tích dưới đường cong ROC, cung cấp một chỉ số tổng thể về khả năng phân loại của mô hình trên tất cả các ngưỡng. Giá trị AUC càng gần 1, mô hình càng tốt

# CHƯƠNG 4: KẾT QUẢ

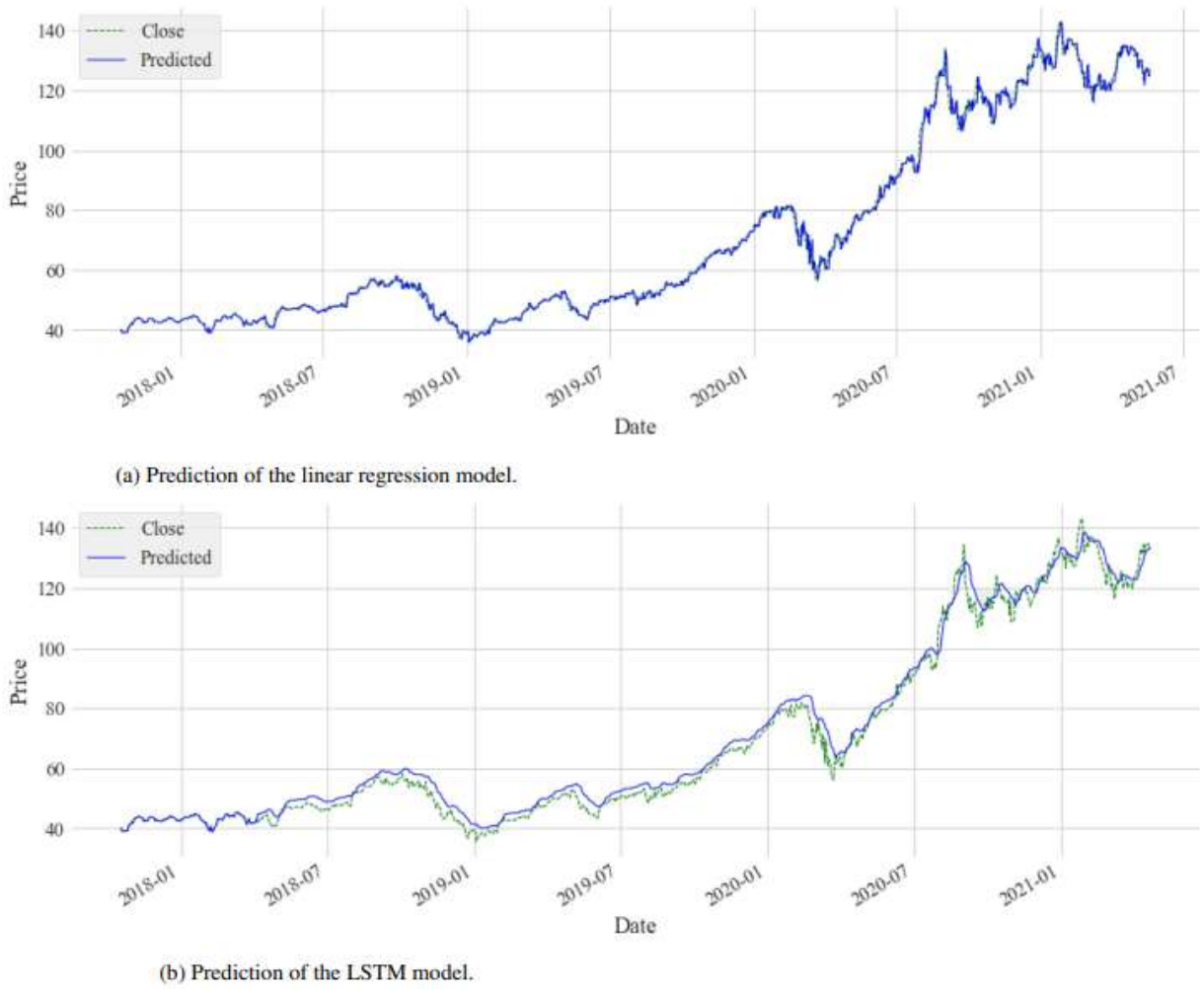
## 4.1. Phân tích kỹ thuật

Phân tích kỹ thuật trong nghiên cứu này tập trung vào việc dự đoán giá cổ phiếu AAPL dựa trên dữ liệu lịch sử từ Yahoo Finance, sử dụng các mô hình hồi quy như Linear Regression (LR) và Long Short-Term Memory (LSTM). Kết quả được đánh giá qua các chỉ số hiệu suất và được minh tổng hợp trong Bảng 1 và Hình 2, cung cấp cái nhìn rõ ràng về khả năng dự đoán của từng mô hình.

| Metric | Linear Regression | LSTM |
|---|---|---|
| $R^2$ | 1.0 | 0.99 |
| **Explained Variation** | 1.0 | 0.99 |
| **MAPE** | 1.56 | 2.99 |
| **RMSE** | 1.82 | 3.42 |
| **MAE** | 1.18 | 2.3 |

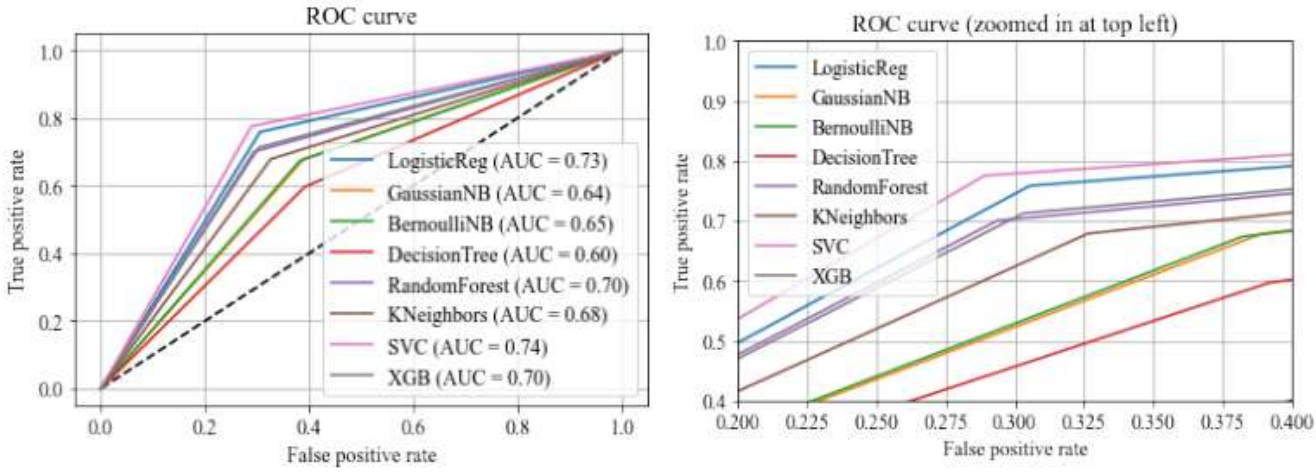
*Bảng 1: So sánh hiệu suất mô hình phân tích kỹ thuật*

Dựa trên Bảng 1, mô hình Linear Regression thể hiện hiệu suất vượt trội so với LSTM. Cụ thể, LR đạt $R^2$ và Explained Variation là 1.0, cho thấy khả năng giải thích hoàn toàn biến thiên của dữ liệu, trong khi LSTM đạt 0.99. Các chỉ số lỗi như MAPE (1.56 so với 2.99), RMSE (1.82 so với 3.42), và MAE (1.18 so với 2.3) của LR đều thấp hơn đáng kể, phản ánh độ chính xác cao hơn và sai số nhỏ hơn khi dự đoán giá đóng cửa.




*Hình 2: So sánh dự đoán của 2 mô hình phân tích kỹ thuật với giá đóng thực tế*

Hình 2 minh họa trực quan sự khác biệt giữa giá dự đoán (đường xanh đậm) và giá thực tế (đường xanh nhạt) từ năm 2018 đến 2021. Mô hình LR cho thấy đường dự đoán gần khớp hoàn toàn với đường thực tế, trong khi LSTM có xu hướng dao động nhiều hơn, đặc biệt trong các giai đoạn biến động mạnh, như giữa năm 2020, cho thấy hạn chế trong việc xử lý dữ liệu chuỗi thời gian phức tạp. Kết quả này khẳng định LR phù hợp hơn cho dự đoán ngắn hạn dựa trên dữ liệu lịch sử, dù vẫn chưa đủ độ tin cậy cho đầu tư dài hạn, như nghiên cứu kết luận.




(a) ROC curves.

(b) ROC curves from a closer view.

## 4.2. Phân tích cơ bản

Các mô hình học máy đã được huấn luyện và đánh giá trên tập dữ liệu tweet của Apple Inc. Kết quả hiệu suất được tổng hợp trong Bảng 2 và Hình 3:

*Bảng 2: So sánh hiệu suất các mô hình phân tích cơ bản*

Dựa trên Bảng 2, thuật toán SVM đạt hiệu suất cao nhất với độ chính xác 75.5%, Precision 75.7%, Recall 75.5%, F1-score 75.5% và AUC 0.76, cho thấy khả năng phân loại vượt trội. Logistic Regression và Random Forest cùng đạt độ chính xác 72.7%, với các chỉ số Precision, Recall, F1-score ở mức 0.727 và AUC 0.73, thể hiện hiệu suất tốt nhưng kém hơn SVM. XGBoost đạt độ chính xác 70.9%, AUC 0.71, trong khi KNN và ANN chỉ đạt 68.4%, AUC 0.68. Các mô hình Decision Tree, Gaussian Naive Bayes (GNB) và Bernoulli Naive Bayes (BNB) có hiệu suất thấp nhất, lần lượt đạt độ chính xác 62.0% (AUC 0.62), 63.4% (AUC 0.63) và 64.4% (AUC 0.64), cho thấy hạn chế trong việc xử lý dữ liệu tweet không cấu trúc.

*Hình 3: Biểu đồ đường cong ROC so sánh hiệu suất các mô hình phân tích cơ bản*

Hình 3 minh họa đường cong ROC và giá trị AUC của các thuật toán, phản ánh khả năng phân tách giữa cảm xúc tích cực và tiêu cực. SVM đạt AUC cao nhất (0.76), với đường cong ROC vượt trội, khẳng định khả năng phân loại tốt nhất. LR và RF có AUC 0.73, với

đường cong ROC gần nhau, cho thấy hiệu suất ổn nhưng không bằng SVM. XGBoost (AUC 0.71), KNN và ANN (cùng AUC 0.68) có đường cong thấp hơn, phản ánh khả năng phân tách hạn chế hơn. DT, GNB và BNB có AUC thấp nhất (0.62, 0.63 và 0.64), với đường cong ROC gần đường chéo, cho thấy hiệu suất phân loại kém. Theo như kết quả, SVM là mô hình tối ưu nhất trong việc dự đoán cảm xúc công chúng, dù tổng thể vẫn chưa đạt độ chính xác lý tưởng cho dự đoán thị trường chứng khoán.

# CHƯƠNG 5: ĐÁNH GIÁ VÀ KẾT LUẬN

## 5.1. Ưu điểm và hạn chế của nghiên cứu

### 5.1.1 Ưu điểm

- Tận dụng dữ liệu thời gian thực: Phương pháp khai thác dữ liệu phi cấu trúc từ mạng xã hội như Twitter, cho phép nắm bắt nhanh chóng tâm lý thị trường và phản ánh sự thay đổi tức thì trong cảm nhận của nhà đầu tư.
- Quy trình có hệ thống: Nghiên cứu thực hiện quy trình từ thu thập dữ liệu, tiền xử lý, áp dụng kỹ thuật giảm chiều (PCA), đến huấn luyện và đánh giá nhiều mô hình học máy, đảm bảo tính khách quan trong việc chọn mô hình tối ưu.
- Kết hợp phân tích kỹ thuật: Việc tích hợp các chỉ báo kỹ thuật như đường trung bình động, RSI, MACD với phân tích cảm xúc bổ sung khía cạnh định lượng, tăng cường khả năng dự đoán xu hướng giá.
- Tự động hóa hiệu quả: Phương pháp giảm sự phụ thuộc vào phân tích thủ công, cho phép xử lý lượng lớn dữ liệu một cách nhanh chóng và hiệu quả.
- Đa dạng thuật toán: Việc thử nghiệm nhiều thuật toán học máy và chỉ báo kỹ thuật thể hiện sự tìm tòi và đánh giá toàn diện, nâng cao độ tin cậy của nghiên cứu.

### 5.1.2 Hạn chế

- Độ chính xác chưa đủ cao: Mô hình SVM đạt độ chính xác 76%, nhưng vẫn được xem là "trung bình" và chưa đủ tin cậy để áp dụng trong các quyết định đầu tư thực tế, đặc biệt với tỷ lệ lỗi 24% có thể gây thiệt hại lớn.
- Hạn chế của phân loại nhị phân: Phân loại cảm xúc thành "tích cực" hoặc "tiêu cực" quá đơn giản, bỏ qua các sắc thái phức tạp như trung lập hoặc cường độ cảm xúc khác nhau, dẫn đến mất mát thông tin quan trọng.
- Nhiễu từ dữ liệu Twitter: Dữ liệu mạng xã hội dễ bị ảnh hưởng bởi tin đồn, thông tin sai lệch hoặc ngôn ngữ châm biếm, gây khó khăn cho mô hình trong việc phân loại chính xác.

- Thiếu tích hợp dữ liệu tài chính: Nghiên cứu chưa kết hợp đầy đủ các yếu tố định lượng như báo cáo tài chính, chỉ số kinh tế vĩ mô, khiến dự đoán thiếu toàn diện.
- Hạn chế của phân tích kỹ thuật: Các chỉ báo kỹ thuật dựa trên dữ liệu lịch sử, có thể không phản ánh kịp thời các biến động bất ngờ của thị trường.
- Tính dao động của phân tích cảm xúc: Mô hình tĩnh khó phản ánh sự thay đổi nhanh chóng của tâm lý thị trường, làm giảm hiệu quả trong các kịch bản biến động cao.

### 5.2. Các hướng nghiên cứu tiếp theo

- Cải thiện độ chính xác mô hình: Ứng dụng các mô hình học sâu như Mạng Nơ-ron Hồi quy (RNN) hoặc BERT để phân tích cảm xúc, tận dụng khả năng hiểu ngữ cảnh và sắc thái ngôn ngữ nhằm nâng cao chất lượng dự đoán.
- Phân tích cảm xúc đa lớp: Phát triển mô hình phân loại cảm xúc thành nhiều cấp độ (rất tích cực, trung lập, rất tiêu cực) hoặc dự đoán điểm số cường độ cảm xúc để phản ánh chính xác hơn tâm lý thị trường.
- Nâng cấp xử lý ngôn ngữ tự nhiên: Sử dụng kỹ thuật NLP tiên tiến để xử lý tiếng lóng, từ viết tắt và ngữ cảnh tài chính đặc thù, đồng thời giảm thiểu tác động của nhiễu và thông tin sai lệch từ dữ liệu mạng xã hội.
- Mở rộng nguồn dữ liệu: Tích hợp thêm dữ liệu từ diễn đàn đầu tư, báo cáo phân tích, bản tin doanh nghiệp và các chỉ báo tài chính như báo cáo thu nhập, bảng cân đối kế toán để có bức tranh dự đoán toàn diện hơn.
- Phát triển mô hình thích nghi thời gian thực: Xây dựng các mô hình có khả năng học và cập nhật liên tục để thích nghi với sự thay đổi của ngôn ngữ, tâm lý thị trường và xu hướng giá.
- Kết hợp chỉ báo kỹ thuật tiên tiến: Nghiên cứu tích hợp các chỉ báo kỹ thuật như Bollinger Bands, Fibonacci retracement với phân tích cảm xúc để cải thiện khả năng dự đoán trong các kịch

## TÀI LIỆU THAM KHẢO


1. Mokhtari, S., Yen, K. K., & Liu, J. (2021). Effectiveness of artificial intelligence in stock market prediction based on machine learning. International Journal of Computer Applications, 183(7), 1-8. https://doi.org/10.5120/ijca2021921411
2. Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. Expert Systems with Applications, 124, 226-251.

3. Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90, 106181.
4. Kalamı, İ., & Tekin, A. (2021). Stock market prediction using deep learning models with a sentiment analysis approach. Applied Soft Computing, 110, 107605.
5. Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with applications, 38(5), 5311-5319.