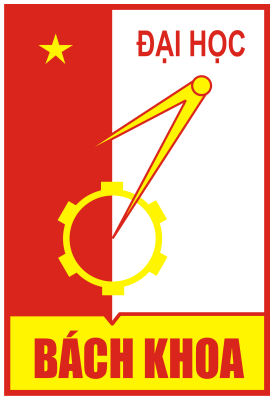
ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN TIN



ĐỒ ÁN II

DỰ BÁO GIÁ CHỨNG KHOÁN DỰA TRÊN MÔ HÌNH RNN VÀ LSTM

LÊ NGỌC HÀ
ha.ln216922@sis.hust.edu.vn

Ngành Hệ thống thông tin quản lý

Giảng viên hướng dẫn: PGS.TS. Nguyễn Đình Hân _____

Chữ ký GVHD

Khoa: Toán Tin

HÀ NỘI, 01/2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. **Mục tiêu và nội dung của đồ án**:
   (a) Mục tiêu: Tìm hiểu về mô hình RNN và LSTM, từ đó áp dụng để giải quyết bài toán dự báo giá chứng khoán.
   (b) Nội dung:

   - Phát biểu bài toán.
   - Xây dựng mô hình cho bài toán và thử nghiệm mô hình.
   - Phân tích kết quả thử nghiệm và đánh giá mô hình.
   - Lập tài liệu và báo cáo kết quả.

2. **Kết quả đạt được**:
   (a) Tìm hiểu về bài toán dự báo giá chứng khoán.
   (b) Nghiên cứu và tìm hiểu lý thuyết về những kiến thức có liên quan.
   (c) Tiến hành thu thập dữ liệu, xử lý dữ liệu, xây dựng, huấn luyện mô hình và thu được kết quả dự đoán giá chứng khoán.
   (d) So sánh giữa các mô hình lựa chọn.

3. **Ý thức làm việc của sinh viên**: Tốt.

Ngày đánh giá	Lần	Nội dung kế hoạch	Nội dung đã thực hiện	Điểm tích cực	Điểm nội dung	Ghi chú
03/11/2024	1	Thực hiện các nội dung 1) và 2)	Đảm bảo tiến độ	10	9	
09/12/2024	2	Các nội dung 3), 4) và 5)	Cơ bản hoàn thành đồ án	10	9.5	

Hà Nội, ngày 01 tháng 01 năm 2025
Giảng viên hướng dẫn

PGS.TS. Nguyễn Đình Hân

LỜI CẢM ƠN

"Chúng ta cần tìm thời điểm thích hợp để dừng lại và cảm ơn những người đã tạo nên sự khác biệt cho cuộc đời mình."

Lời đầu tiên của đồ án này, em muốn gửi lời cảm ơn và lòng biết ơn chân thành tới thầy **PGS.TS. Nguyễn Đình Hân**. Thầy là người trực tiếp hướng dẫn và hỗ trợ em trong suốt quá trình thực hiện đồ án. Không những là người thầy chỉ bảo em về kiến thức, về định hướng phát triển mà thầy còn là người truyền cảm hứng cho em rất nhiều. Sau một thời gian có cơ hội được làm việc cùng thầy, ngoài việc học hỏi được những kiến thức chuyên ngành, những kinh nghiệm về ngành học thì sự tận tâm, cách thầy quan tâm đến sinh viên thật sự khiến em cảm thấy vô cùng thân thuộc, gần gũi và đáng để học hỏi. Em chúc thầy thật nhiều sức khoẻ và những điều tuyệt vời nhất sẽ luôn đến với thầy.

Đồng thời, em cũng muốn gửi lời cảm ơn đến Khoa Toán Tin - Đại học Bách khoa Hà Nội, nơi đã tạo điều kiện cho em được học tập và phát triển bản thân trong một môi trường đầy năng động và thoả sức sáng tạo.

Cuối cùng, con cũng muốn bày tỏ sự biết ơn tới gia đình, bạn bè, đặc biệt là bố và mẹ đã không quản vất vả, cực nhọc để lo cho con được đầy đủ nhất, để con không cảm thấy thiệt thòi so với các bạn đồng trang lứa. Mọi người vừa là chỗ dựa của con, cũng là động lực to lớn để con cố gắng hơn mỗi ngày trên con đường học tập, tìm kiếm và chinh phục tri thức.

Do kiến thức và các kỹ năng của bản thân vẫn còn hạn hẹp nên đồ án của em không tránh khỏi những thiếu sót và sai sót. Bởi vậy, em rất mong nhận được những ý kiến đóng góp quý báu từ thầy cô để đồ án được hoàn thiện hơn.

Em xin chân thành cảm ơn!

TÓM TẮT ĐỒ ÁN

Dự báo giá chứng khoán là một bài toán quan trọng trong tài chính, nhằm dự đoán giá cổ phiếu tương lai dựa trên dữ liệu lịch sử. Đây là vấn đề phức tạp do thị trường chịu ảnh hưởng bởi nhiều yếu tố. Các phương pháp truyền thống thường gặp khó khăn với dữ liệu phi tuyến và biến động cao, trong khi mô hình học sâu như LSTM và RNN đã chứng minh hiệu quả vượt trội nhờ khả năng xử lý chuỗi thời gian.

Kết quả cuối cùng mà đồ án mong muốn đạt được là có thể đem lại giá trị dự đoán chính xác nhất, giúp hỗ trợ nhà đầu tư đưa ra quyết định đầu tư. Chính vì vậy, đồ án gồm 4 chương như sau:

- Chương 1: Bài toán dự báo giá chứng khoán. Chương này sẽ tập trung vào việc tìm hiểu về bài toán dự báo giá chứng khoán như các yếu tố chi phối giá, các phương pháp dự báo hiện nay, vai trò và tầm quan trọng của bài toán.
- Chương 2: Cơ sở lý thuyết. Chương này sẽ nghiên cứu và tìm hiểu về tất cả các kiến thức phục vụ cho đồ án từ các lý thuyết về chuỗi thời gian, mạng nơ-ron, các chỉ số đánh giá mô hình và đặc biệt là tìm hiểu về các mô hình dự báo là RNN và LSTM.
- Chương 3: Ứng dụng các mô hình dự báo giá chứng khoán. Sau khi đã tìm hiểu về lý thuyết, chương 3 sẽ tập trung vào việc áp dụng kiến thức lên bộ dữ liệu trực tiếp của Công ty Cổ phần Viễn Thông FPT để xây dựng và huấn luyện mô hình để đưa ra giá trị dự đoán chính xác nhất.
- Chương 4: Phân tích kết quả và đánh giá mô hình. Chương này sẽ tập trung vào việc phân tích các kết quả dự đoán và thực tế, tiến hành so sánh các chỉ số đánh giá giữa các mô hình dự báo. Cuối cùng, tổng kết và đưa ra hướng phát triển trong tương lai.

Hà Nội, ngày 01 tháng 01 năm 2025

Sinh viên thực hiện

Lê Ngọc Hà

Mục lục

Danh sách từ viết tắt

AI Artificial Intelligence

ANN Artificial Neural Network

BPTT Backpropagation Through Time

CNN Convolutional Neural Network

EPS Earnings Per Share

GDP Gross Domestic Product

HOSE Hồ Chí Minh Stock Exchange

LSTM Long Short-Term Memory

MA Moving Average

MACD Moving Average Convergence Diverge

MAPE Mean Absolute Percentage

ML Machine Learning

MSE Mean Squared Error

NN Neural Network

P/B Price-to-Book Ratio

P/E Price-to-Earnings Ratio

PCA Principal Component Analysis

R Coefficient of Determination

ReLU Rectified Linear Unit

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

RSI Relative Strength Index

SVM Support Vector Machines

TPE ree-structured Parzen Estimator

Danh sách hình vẽ

Danh sách bảng

Mở đầu

1. Lý do chọn đề tài

Trong bối cảnh phát triển mạnh mẽ của thị trường tài chính, dự báo giá chứng khoán trở thành nhu cầu cấp thiết không chỉ đối với các nhà đầu tư mà còn cho các doanh nghiệp và cơ quan quản lý. Biến động giá chứng khoán chịu ảnh hưởng bởi nhiều yếu tố phức tạp như các chỉ số kinh tế vĩ mô, tâm lý thị trường và yếu tố công ty cụ thể.

Việc sử dụng các mô hình học sâu sẽ giúp khai thác và phân tích tác động của nhiều yếu tố này một cách hệ thống và hiệu quả hơn, từ đó cung cấp các dự báo chính xác, hỗ trợ đưa ra các quyết định đầu tư có căn cứ. Chính vì vậy, đề tài "Dự báo giá chứng khoán dựa trên mô hình RNN và LSTM" nhằm nghiên cứu và ứng dụng các phương pháp phân tích dữ liệu để góp phần giải quyết bài toán phức tạp này.

2. Đối tượng và phạm vi nghiên cứu

Đề tài tập trung vào các công cụ và kỹ thuật dự báo giá chứng khoán của một số mã cổ phiếu được niêm yết trên thị trường chứng khoán Việt Nam.

- Đối tượng nghiên cứu: Các yếu tố tác động đến giá cổ phiếu, chẳng hạn như chỉ số lạm phát, tỷ giá, lãi suất cùng các yếu tố nội tại doanh nghiệp (lợi nhuận, doanh thu, giá trị tài sản).
- Phạm vi nghiên cứu: Giới hạn trong việc xây dựng và đánh giá mô hình hồi quy đa biến, sử dụng dữ liệu thực tế trong một khoảng thời gian nhất định để kiểm chứng tính chính xác và khả năng áp dụng của mô hình vào dự báo giá cổ phiếu.

3. Ý nghĩa khoa học và thực tiễn của đề tài

Đề tài có ý nghĩa khoa học khi nghiên cứu ứng dụng của mô hình học sâu, một trong những kỹ thuật phân tích dữ liệu phổ biến trong dự báo vào lĩnh vực tài chính, góp phần mở rộng ứng dụng của thống kê và phân tích dữ liệu vào thị trường chứng khoán.

Về mặt thực tiễn, kết quả của đề tài giúp các doanh nghiệp, nhà đầu tư hiểu rõ hơn mối liên hệ giữa các yếu tố ảnh hưởng đến giá cổ phiếu, từ đó hỗ trợ trong việc đưa ra các quyết định đầu tư có cơ sở. Kết quả nghiên cứu cũng có thể là tài liệu tham khảo hữu ích cho các doanh nghiệp tài chính, các tổ chức đầu tư và các nhà nghiên cứu quan tâm đến dự báo tài chính.

Chương 1

Bài toán dự báo giá chứng khoán

1.1 Tổng quan về tình hình giá chứng khoán

1.1.1 Tình hình giá chứng khoán trên thế giới

Tính đến tháng 10/2024, giá chứng khoán toàn cầu ghi nhận kết quả tích cực nhất trong 5 năm qua, với chỉ số MSCI toàn cầu tăng 7,7%, mức tăng cao nhất kể từ năm 2019. Sự tăng trưởng này chủ yếu nhờ vào nền kinh tế Mỹ và sự bùng nổ của công nghệ Trí tuệ nhân tạo (AI). Đặc biệt, sự tăng trưởng mạnh mẽ của Nvidia, với giá trị vốn hóa thị trường tăng hơn 1.000 tỷ USD trong ba tháng đầu năm đã thúc đẩy đà đi lên của giá chứng khoán.

Mặc dù lạm phát ở Mỹ tăng cao bất ngờ trong tháng 1/2024 và tháng 2/2024 nhưng giá chứng khoán vẫn tiếp tục duy trì đà tăng. Ngoài ra, các thị trường chứng khoán ở châu Âu và châu Á cũng có sự tăng trưởng vượt trội. Nhật Bản là nước dẫn đầu nhờ vào sự phục hồi mạnh mẽ của ngành công nghệ và niềm tin vào nền kinh tế trong nước.

Giá chứng khoán toàn cầu có sự tăng trưởng ấn tượng nhưng nếu suy thoái kinh tế diễn ra hoặc tỷ lệ thất nghiệp ở Mỹ tăng đột ngột thì những mức tăng trưởng này có thể bị ảnh hưởng. Mặc dù vậy, mức giá chứng khoán hiện tại có thể kéo dài đến năm 2029 hoặc thậm chí 2033 nếu không có yếu tố bất ngờ nào.

1.1.2 Tình hình giá chứng khoán tại Việt Nam

Xét đến tháng 10/2024, giá chứng khoán Việt Nam ghi nhận nhiều tín hiệu tích cực nhưng vẫn còn đối mặt với những thách thức đáng kể từ cả yếu tố trong nước và tác động của kinh tế toàn cầu. Trong tháng 9/2024, chỉ số VN-Index (chỉ số đại

diện cho thị trường chứng khoán Việt Nam) có nhiều biến động đáng chú ý. Sau khi chạm mốc 1.300 điểm, VN-Index không thể duy trì được mức này nhưng vẫn kết thúc tháng với một đợt tăng nhẹ.

Với sự tăng mạnh trở lại của tỷ giá USD/VND, áp lực thoái vốn từ các nhà đầu tư nước ngoài gia tăng, ảnh hưởng đến sự ổn định của giá chứng khoán. Tuy nhiên, nhờ các chính sách điều hành kinh tế vĩ mô ổn định giúp nguy cơ giảm sâu hơn của thị trường được đánh giá là không cao. Nhìn chung, thị trường chứng khoán Việt Nam nói chung và giá chứng khoán tại Việt Nam nói riêng vẫn còn rất nhiều tiềm năng để các nhà đầu tư, doanh nghiệp có thể khai thác và hy vọng trong tương lai.

1.2 Giới thiệu về bài toán dự báo giá chứng khoán

1.2.1 Giới thiệu bài toán dự báo

Trong những năm gần đây, thị trường chứng khoán đã nổi lên như một kênh đầu tư hấp dẫn, thu hút sự quan tâm của cả các nhà đầu tư cá nhân nhỏ lẻ lẫn các tổ chức lớn, chuyên nghiệp. Với sự đa dạng về phong cách đầu tư, từ những chiến lược ngắn hạn như giao dịch lướt sóng cho đến các chiến lược dài hạn, thị trường này mang lại cơ hội lớn để tìm kiếm lợi nhuận. Tuy nhiên, đi cùng với tiềm năng sinh lời cao là những rủi ro không hề nhỏ, khiến các nhà đầu tư phải đặc biệt thận trọng trong các quyết định của mình. Việc quản lý rủi ro và tối ưu hóa lợi nhuận trở thành mục tiêu trọng tâm, thúc đẩy nhu cầu phân tích và dự báo xu hướng thị trường ngày càng tăng cao.

Dự báo thị trường chứng khoán được coi là một công cụ quan trọng, không chỉ giúp các nhà đầu tư cá nhân giảm thiểu nguy cơ mất mát mà còn hỗ trợ các tổ chức tài chính trong việc lập kế hoạch chiến lược. Điều này đặc biệt cần thiết trong bối cảnh thị trường chứng khoán thường xuyên biến động mạnh và chịu ảnh hưởng từ nhiều yếu tố không thể đoán trước, bao gồm các sự kiện kinh tế, chính trị và tâm lý nhà đầu tư. Việc dự đoán chính xác xu hướng giá cổ phiếu không chỉ giúp nhà đầu tư đưa ra các quyết định sáng suốt mà còn tối ưu hóa hiệu quả danh mục đầu tư.

Hơn nữa, thị trường chứng khoán còn là một trong những lĩnh vực tài chính khó dự đoán nhất do tính phức tạp và nhạy cảm với các yếu tố ngoại cảnh. Để đối phó với sự khó lường này, các nhà đầu tư và chuyên gia tài chính không ngừng phát triển các phương pháp dự báo từ truyền thống như phân tích kỹ thuật đến các công cụ

hiện đại như mô hình máy học. Nhờ đó, khả năng nhận diện các xu hướng quan trọng trong thị trường được nâng cao, mang lại lợi thế cạnh tranh lớn cho những ai biết tận dụng. Chính vì vậy, việc dự báo giá chứng khoán không chỉ mang ý nghĩa chiến lược mà còn đóng vai trò thiết yếu trong việc đảm bảo sự thành công lâu dài trên thị trường.

1.2.2 Các loại giá chứng khoán phổ biến

Năm 1602, thị trường chứng khoán đầu tiên trên thế giới được thành lập tại Amsterdam, Hà Lan với sự ra đời của Công ty Đông Ấn Hà Lan (Dutch East India Company). Đây cũng là công ty đầu tiên phát hành cổ phiếu ra công chúng. Giá cổ phiếu lúc này được xác định bởi nhu cầu của các nhà đầu tư mua và bán cổ phần, đặt nền móng cho khái niệm giá chứng khoán sau này.

Giá chứng khoán là giá trị mà tại đó một loại chứng khoán (cổ phiếu, trái phiếu hoặc các công cụ tài chính khác) được giao dịch trên thị trường. Giá này thường được xác định bởi cung và cầu của thị trường, dựa trên các yếu tố như hiệu quả hoạt động của công ty, điều kiện kinh tế vĩ mô, tâm lý thị trường...

Giá chứng khoán có một số loại quan trọng mà bất kỳ ai nghiên cứu về chứng khoán cũng cần biết:

- Giá mở cửa: Đây là giá giao dịch đầu tiên của chứng khoán trong phiên giao dịch, nó được hình thành dựa trên cung và cầu tại thời điểm thị trường mở cửa. => Đóng vai trò quan trọng vì giá mở cửa phản ánh những biến động về tin tức, sự kiện ngoài giờ giao dịch.
- Giá đóng cửa: Đây là giá giao dịch cuối cùng của chứng khoán trong phiên giao dịch, thường được sử dụng để tham chiếu cho phiên giao dịch tiếp theo.
- Giá IPO (Initial Public Offering Price): Đây là giá cổ phiếu lần đầu ra mắt công chúng, thường được xác định bởi các tổ chức bảo lãnh phát hành dựa trên giá trị của doanh nghiệp và kỳ vọng từ thị trường.
- Giá tham chiếu: Đây là giá cơ sở để tính toán biên độ dao động giá trong ngày giao dịch đó. Ở thị trường chứng khoán Việt Nam, giá tham chiếu thường là giá đóng cửa của ngày giao dịch trước, dùng để xác định giá trần và giá sàn.
- Giá trần và giá sàn: Lần lượt là mức giá cao nhất và thấp nhất mà chứng khoán có thể được giao dịch trong một ngày.

- Giá khớp lệnh: Mức giá mà tại đó một lệnh mua và một lệnh bán được khớp trên thị trường.
- Giá cao nhất và giá thấp nhất: Lần lượt là mức giá cao nhất và thấp nhất đạt được trong một phiên.
- Giá trung bình: Mức giá trung bình của tất cả các giao dịch được thực hiện trong một ngày giao dịch.
- Giá thị trường: Giá hiện tại của chứng khoán được giao dịch trên thị trường, phản ánh giá trị kỳ vọng của nhà đầu tư đối với chứng khoán tại thời điểm đó.
- Giá danh nghĩa: Mệnh giá của cổ phiếu được ghi trên giấy chứng nhận cổ phiếu. Ở Việt Nam, giá danh nghĩa phổ biến là 10.000 VNĐ/cổ phiếu.

Có thể thấy, mỗi loại giá chứng khoán cung cấp một góc nhìn cụ thể về diễn biến giao dịch, hỗ trợ nhà đầu tư trong việc phân tích và đưa ra quyết định phù hợp. Việc hiểu rõ từng loại giá chứng khoán sẽ giúp chúng ta tăng cường khả năng quản lý rủi ro và tối ưu hóa lợi nhuận.

1.2.3 Các yếu tố chi phối giá chứng khoán

Giá chứng khoán chịu sự chi phối bởi nhiều yếu tố bao gồm tình hình nội bộ của doanh nghiệp, yếu tố kinh tế vĩ mô, những tác động và sự kiện trên thế giới. Những yếu tố này không chỉ mang tính cục bộ mà còn thể hiện mối quan hệ mật thiết giữa chính trị, kinh tế và thị trường tài chính toàn cầu.

Đầu tiên, giá chứng khoán bị chi phối bởi tình hình nội bộ của doanh nghiệp. Mỗi doanh nghiệp sẽ có cách vận hành, tổ chức khác nhau, chính vì đặc điểm riêng biệt này sẽ là yếu tố quyết định tới việc xác định giá trị cổ phiếu:

- Các công ty sở hữu tài sản lớn thường có tính cạnh tranh cao hơn trên thị trường chứng khoán, điều này giúp góp phần nâng cao giá chứng khoán.
- Doanh thu và lợi nhuận càng cao thì giá cổ phiếu sẽ có xu hướng tăng nhờ lấy được niềm tin của các nhà đầu tư vào sự phát triển bền vững của doanh nghiệp.
- Các công ty chi trả cổ tức hấp dẫn giúp tăng sự thu hút trên thị trường, qua đó giá cổ phiếu cũng được đẩy lên cao hơn.
- Mức thu nhập trên mỗi cổ phần (EPS) càng cao sẽ là minh chứng cho hoạt

động tiềm năng của doanh nghiệp đó, tạo ra tác động tích cực đến giá cổ phiếu.

Tiếp theo, yếu tố thứ hai chi phối giá chứng khoán là yếu tố kinh tế vĩ mô. Những biến động của nền kinh tế có tác động không hề nhỏ đến toàn bộ thị trường chứng khoán nói chung và giá chứng khoán nói riêng. Tiêu biểu có thể kể đến như:

- Việc tăng trưởng GDP mạnh mẽ giúp doanh nghiệp đạt được lợi nhuận cao từ đó tạo điều kiện thuận lợi cho giá cổ phiếu tăng theo.
- Tình trạng lạm phát cao có thể làm giảm sức mua và giá trị thực của doanh nghiệp, dẫn đến tác động tiêu cực lên giá cổ phiếu.
- Lãi suất vay vốn cao khiến chi phí tài chính tăng, làm giảm lợi nhuận và kéo giá cổ phiếu đi xuống.
- Các chính sách tiền tệ ảnh hưởng trực tiếp đến thanh khoản và lãi suất, tạo ra những thay đổi lớn đối với thị trường chứng khoán.

Cuối cùng, những tác động trên thế giới và sự kiện mang tính toàn cầu cũng là yếu tố chi phối giá chứng khoán trong nước. Khi các giai đoạn tăng trưởng hoặc suy thoái kinh tế trên thế giới có thể gây ảnh hưởng trực tiếp đến các thị trường trong nước. Bên cạnh đó, biến động giá dầu cũng là một yếu tố quan trọng, bởi giá dầu thay đổi sẽ tác động đến chi phí sản xuất và lợi nhuận của nhiều ngành công nghiệp, đặc biệt là các doanh nghiệp phụ thuộc vào nguyên liệu đầu vào. Một yếu tố toàn cầu khác là lợi suất trái phiếu quốc tế có thể ảnh hưởng đến dòng vốn đầu tư và gián tiếp tác động đến giá cổ phiếu. Ngoài ra, giá cổ phiếu còn bị chi phối bởi các yếu tố khác như giá trị USD và vàng, bởi những biến động của chúng thường ảnh hưởng đến tâm lý nhà đầu tư và quyết định đầu tư. Chính sách của ngân hàng trung ương với các biện pháp như điều chỉnh lãi suất hoặc cung ứng tiền tệ có khả năng định hướng xu hướng thị trường chứng khoán.

Việc hiểu rõ và phân tích tỉ mỉ về các yếu tố ảnh hưởng này sẽ giúp nhà đầu tư, doanh nghiệp đưa ra quyết định hiệu quả hơn, nắm bắt được tình hình giá chứng khoán lên xuống ra sao, đồng thời tránh được các rủi ro không mong muốn.

1.2.4 Các phương pháp dự báo giá chứng khoán

a) Phân tích kỹ thuật

Phân tích kỹ thuật dựa trên giả thuyết rằng giá chứng khoán trong tương lai có thể được dự đoán từ các mẫu hình và xu hướng đã xuất hiện trong quá khứ. Phương pháp này tập trung vào dữ liệu giá và khối lượng giao dịch mà không quan tâm đến yếu tố cơ bản của doanh nghiệp.

Trong phương pháp này, giá phản ánh tất cả thông tin nghĩa là mọi yếu tố kinh tế, chính trị, tâm lý hay các tin tức liên quan đến công ty đều được phản ánh trực tiếp vào giá cổ phiếu. Giá chứng khoán không biến động ngẫu nhiên mà sẽ tuân theo các xu hướng ngắn hạn, trung hạn hoặc dài hạn.

Phương pháp phân tích kỹ thuật thường sử dụng các công cụ phân tích bao gồm:

- Biểu đồ giá giúp hiển thị các biến động của giá chứng khoán theo thời gian để giúp các nhà đầu tư, doanh nghiệp có cái nhìn tổng quan nhất.
- Đường trung bình để tính giá chứng khoán trung bình trong một khoảng thời gian cụ thể (theo giờ, theo ngày, theo tháng...) để từ đó xác định xu hướng giá trong tương lai.
- Các chỉ số kỹ thuật như chỉ số RSI (Relative Strength Index) để đo lường mức độ dao động của giá, chỉ số MACD (Moving Average Convergence Diverge) giúp cung cấp các biến động của thị trường giá và chỉ số MA (Moving Average) làm mượt biến động giá để xác định xu hướng chính..

Phân tích kỹ thuật đặc biệt hữu ích trong giao dịch ngắn hạn hoặc trung hạn. Các nhà giao dịch có thể sử dụng các tín hiệu từ chỉ số kỹ thuật để đưa ra quyết định mua hoặc bán một cách nhanh chóng, tối ưu hóa lợi nhuận từ biến động giá.

b) Phân tích cơ bản

Phương pháp phân tích cơ bản là một trong những kỹ thuật chủ yếu trong việc đánh giá giá trị thực sự của một cổ phiếu, giúp các nhà đầu tư đưa ra quyết định đầu tư thông minh và hiệu quả. Phân tích cơ bản tập trung vào việc đánh giá tình hình tài chính, hoạt động kinh doanh và các yếu tố vĩ mô có ảnh hưởng đến giá trị của cổ phiếu. Mục tiêu của phương pháp này là xác định giá trị thực của cổ phiếu và so sánh với giá thị trường để tìm ra các cơ hội đầu tư.

Yếu tố đầu tiên và quan trọng nhất trong phân tích cơ bản là khả năng sinh lời của công ty. Các nhà đầu tư cần phải đánh giá khả năng của công ty trong việc tạo ra lợi nhuận, từ đó xác định liệu cổ phiếu của công ty có đáng đầu tư hay không. Việc phân tích doanh thu trong quá khứ và dự báo doanh thu trong tương lai giúp nhà đầu tư đánh giá được khả năng tăng trưởng của công ty. Lợi nhuận được chia thành lợi nhuận gộp, lợi nhuận trước thuế và lợi nhuận ròng. Các nhà đầu tư cần chú ý đến lợi nhuận ròng, vì đây là khoản lợi nhuận cuối cùng mà công ty thu được sau khi trừ đi tất cả các chi phí hoạt động, thuế và các chi phí khác. Một công ty có khả năng tạo ra lợi nhuận ổn định và tăng trưởng thường có triển vọng đầu tư tốt.

Bên cạnh việc đánh giá doanh thu và lợi nhuận, các nhà đầu tư còn sử dụng các chỉ số tài chính để đo lường tình hình tài chính và hiệu quả hoạt động của công ty. Các chỉ số này giúp nhà đầu tư có cái nhìn tổng quan về tình hình tài chính của công ty và đánh giá khả năng sinh lời trong tương lai. Các chỉ số tiêu biểu như là EPS, P/E, P/B...

Một yếu tố quan trọng khác trong phân tích cơ bản là đánh giá vị thế cạnh tranh của công ty trong ngành. Sự cạnh tranh trong ngành sẽ ảnh hưởng trực tiếp đến khả năng duy trì và tăng trưởng lợi nhuận của công ty. Các công ty có vị thế cạnh tranh mạnh mẽ, khả năng chiếm lĩnh thị trường và sáng tạo sản phẩm thường có khả năng sinh lời bền vững hơn trong dài hạn. Các yếu tố cần phân tích bao gồm: Thị phần, sự khác biệt hóa sản phẩm, đối thủ cạnh tranh...

Các yếu tố vĩ mô cũng đóng vai trò quan trọng trong việc đánh giá giá trị của cổ phiếu. Các yếu tố này bao gồm những yếu tố kinh tế và chính trị ảnh hưởng đến toàn bộ thị trường và có thể tác động mạnh mẽ đến giá cổ phiếu của công ty:

- Lãi suất: Mức lãi suất do Ngân hàng Trung ương quyết định có tác động lớn đến hoạt động đầu tư. Khi lãi suất tăng, chi phí vay vốn của các công ty cũng tăng, làm giảm lợi nhuận và giá trị cổ phiếu. Ngược lại, khi lãi suất giảm, chi phí vay vốn giảm, thúc đẩy hoạt động kinh doanh và nâng cao giá trị cổ phiếu.
- Tăng trưởng GDP: Một nền kinh tế tăng trưởng mạnh mẽ tạo ra nhiều cơ hội đầu tư và tiêu dùng. Nếu nền kinh tế có tốc độ tăng trưởng GDP cao, các doanh nghiệp có thể dễ dàng tăng trưởng doanh thu và lợi nhuận. Vì vậy, khi GDP tăng trưởng, các nhà đầu tư sẽ có xu hướng đầu tư vào cổ phiếu của các

công ty có khả năng hưởng lợi từ sự phát triển này.

- Tỷ lệ thất nghiệp: Tỷ lệ thất nghiệp thấp cho thấy nền kinh tế đang phát triển mạnh mẽ, dân cư có thu nhập ổn định và chi tiêu tiêu dùng cao. Điều này tạo điều kiện thuận lợi cho các công ty tăng trưởng doanh thu và lợi nhuận.
- Tình hình chính trị và pháp lý: Các chính sách và quy định pháp lý có thể tác động đến môi trường kinh doanh của công ty. Những thay đổi trong chính sách thuế, chính sách thương mại hoặc các quy định về bảo vệ môi trường có thể làm thay đổi chiến lược và lợi nhuận của công ty.

Phương pháp phân tích cơ bản cung cấp cho các nhà đầu tư một công cụ mạnh mẽ để đánh giá giá trị thực sự của cổ phiếu. Bằng cách phân tích kỹ lưỡng các yếu tố tài chính của công ty, vị thế cạnh tranh và các yếu tố vĩ mô, nhà đầu tư có thể đưa ra các quyết định đầu tư hợp lý và dài hạn. Tuy nhiên, phương pháp này yêu cầu sự kiên nhẫn và khả năng phân tích sâu sắc, vì việc đánh giá chính xác giá trị của cổ phiếu cần thời gian và sự nghiên cứu tỉ mỉ.

c) Mô hình thống kê

Mô hình thống kê là một phương pháp quan trọng trong việc dự báo giá chứng khoán, dựa trên việc phân tích các dữ liệu lịch sử và các mối quan hệ thống kê giữa các biến. Các mô hình này giúp nhà đầu tư đưa ra dự đoán về sự biến động của giá cổ phiếu trong tương lai, qua đó hỗ trợ việc ra quyết định đầu tư. Phương pháp thống kê có thể áp dụng nhiều kỹ thuật khác nhau từ đơn giản đến phức tạp, tùy thuộc vào tính chất của dữ liệu và mục tiêu dự báo.

Một trong những mô hình thống kê phổ biến là hồi quy tuyến tính, được sử dụng để dự báo giá chứng khoán bằng cách tìm kiếm mối quan hệ tuyến tính giữa giá cổ phiếu và một hoặc nhiều yếu tố ảnh hưởng (biến độc lập). Mô hình này giúp xác định xem giá cổ phiếu có thể tăng hoặc giảm khi các yếu tố này thay đổi.

Mô hình hồi quy logistic được sử dụng khi mục tiêu dự báo không phải là giá trị liên tục mà là xác suất xảy ra một sự kiện cụ thể như sự tăng giá cổ phiếu hay giảm giá cổ phiếu. Mô hình này phổ biến trong việc phân loại và dự báo các xu hướng thị trường. Mô hình hồi quy logistic rất hữu ích trong việc phân tích các sự kiện nhị phân và xác định xu hướng thị trường, chẳng hạn như "tăng" hoặc "giảm" trong ngắn hạn, điều này có thể giúp các nhà đầu tư đưa ra các quyết định chiến lược.

ARIMA là một trong những mô hình thống kê mạnh mẽ và phổ biến nhất trong dự báo giá chứng khoán, đặc biệt là trong phân tích chuỗi thời gian. ARIMA kết hợp ba thành phần chính: Tự hồi quy (AR), tích hợp (I) và bình quân trượt (MA) để dự báo các giá trị trong tương lai dựa trên dữ liệu quá khứ.

Mô hình GARCH chủ yếu được sử dụng để dự báo sự biến động (volatility) của giá chứng khoán, thay vì giá trị cổ phiếu cụ thể. Mô hình này đặc biệt hữu ích trong việc phân tích rủi ro tài chính, vì nó cho phép dự báo biến động giá trong tương lai dựa trên sự thay đổi trong quá khứ. Mô hình GARCH có thể được mở rộng thành các phiên bản như EGARCH hay IGARCH để cải thiện độ chính xác.

Các mô hình thống kê trong dự báo giá chứng khoán đóng vai trò quan trọng trong việc giúp nhà đầu tư và chuyên gia tài chính đưa ra quyết định sáng suốt. Tuy nhiên, mỗi mô hình đều có những ưu nhược điểm và phù hợp với các loại dữ liệu và mục tiêu dự báo khác nhau. Khi lựa chọn mô hình, các nhà đầu tư cần xem xét đặc điểm của dữ liệu và yêu cầu của việc dự báo, từ đó kết hợp các phương pháp thống kê để đạt được kết quả tối ưu.

d) Phương pháp học máy

Học máy (Machine Learning - ML) là một nhánh của trí tuệ nhân tạo (AI), cho phép máy tính học hỏi từ dữ liệu mà không cần phải lập trình rõ ràng từng bước. Trong lĩnh vực dự báo giá chứng khoán, học máy được sử dụng để xây dựng các mô hình có khả năng phân tích, nhận dạng xu hướng, và đưa ra dự đoán về sự biến động của giá cổ phiếu. Các phương pháp học máy có thể áp dụng cho dữ liệu quá khứ của cổ phiếu, kết hợp với các yếu tố vĩ mô, giúp dự đoán xu hướng giá trong tương lai.

1. Học có giám sát:
Học có giám sát là một trong những phương pháp phổ biến trong học máy, trong đó mô hình được huấn luyện dựa trên một tập dữ liệu đã được gắn nhãn (labeled data). Trong dự báo giá chứng khoán, dữ liệu được sử dụng có thể bao gồm các thông tin như giá cổ phiếu, khối lượng giao dịch, chỉ số thị trường, các yếu tố kinh tế vĩ mô và các chỉ số tài chính khác. Mục tiêu của học có giám sát là xây dựng mô hình có thể dự đoán giá cổ phiếu hoặc xu hướng thị trường dựa trên những yếu tố này.

Các thuật toán học có giám sát phổ biến:

- Hồi quy tuyến tính (Linear Regression): Hồi quy tuyến tính là một phương pháp đơn giản nhưng hiệu quả trong dự báo giá chứng khoán. Mô hình này tìm mối quan hệ tuyến tính giữa các yếu tố đầu vào (như giá cổ phiếu trước đó, các chỉ số vĩ mô) và giá cổ phiếu dự đoán. Vì sự đơn giản nên mô hình hồi quy tuyến tính có thể không chính xác với dữ liệu có tính biến động cao như thị trường chứng khoán.

- Cây quyết định (Decision Trees): Cây quyết định là một mô hình học máy có thể được sử dụng để phân loại hoặc dự đoán các giá trị liên tục (như giá cổ phiếu). Mô hình này phân chia dữ liệu thành các nhánh, với mỗi nhánh dựa trên các câu hỏi về đặc điểm của dữ liệu. Cây quyết định có thể dễ dàng giải thích và cung cấp thông tin về các yếu tố quan trọng nhất ảnh hưởng đến giá cổ phiếu.

- Random Forest: Random Forest là một thuật toán học máy dựa trên việc sử dụng nhiều cây quyết định để đưa ra dự đoán. Bằng cách kết hợp kết quả của nhiều cây quyết định, Random Forest giúp giảm thiểu nguy cơ overfitting (quá khớp dữ liệu) và cải thiện độ chính xác của dự báo. Đây là một thuật toán mạnh mẽ và phổ biến trong các bài toán dự báo giá chứng khoán.

- Support Vector Machines (SVM): SVM là một thuật toán học máy mạnh mẽ, thường được sử dụng cho các bài toán phân loại và hồi quy. SVM cố gắng tìm kiếm một siêu phẳng trong không gian đa chiều, giúp phân tách các lớp dữ liệu hoặc dự đoán các giá trị liên tục. Đối với dự báo giá chứng khoán, SVM có thể được sử dụng để phân loại các biến động giá (tăng hoặc giảm) hoặc dự đoán giá trị trong tương lai.

2. Học không giám sát:
Trong học không giám sát, mô hình học từ dữ liệu không có nhãn, tức là không có sự chỉ dẫn rõ ràng về kết quả đầu ra. Thay vào đó, mô hình tìm kiếm các mẫu, cấu trúc hoặc sự phân nhóm trong dữ liệu. Phương pháp học không giám sát có thể hữu ích trong việc phát hiện các nhóm cổ phiếu tương tự, hoặc phân tích các xu hướng thị trường mà không có sự chỉ dẫn trước.

Các thuật toán học không giám sát phổ biến:

- Clustering (Phân cụm): Thuật toán phân cụm như K-Means, DBSCAN hoặc Hierarchical Clustering có thể được sử dụng để nhóm các cổ phiếu có hành

vi giá tương tự vào các cụm. Phân cụm giúp nhà đầu tư nhận diện các nhóm cổ phiếu tiềm năng hoặc các xu hướng thị trường mà có thể chưa được nhận ra.

- Principal Component Analysis (PCA): PCA là một kỹ thuật giảm chiều dữ liệu, giúp giảm bớt số lượng đặc trưng đầu vào mà không làm mất đi quá nhiều thông tin. PCA có thể được sử dụng để phân tích các yếu tố chính ảnh hưởng đến giá cổ phiếu và giúp cải thiện độ chính xác của các mô hình.

3. Học sâu (Deep Learning):
Học sâu là một phương pháp mạnh mẽ trong học máy, với các mạng nơ-ron nhân tạo sâu (Deep Neural Networks - DNNs) có khả năng học các mẫu phức tạp trong dữ liệu. Mạng nơ-ron này có thể được huấn luyện để nhận diện các mối quan hệ phi tuyến tính và không gian tính toán phức tạp, điều này rất hữu ích trong việc dự báo giá chứng khoán, nơi có nhiều yếu tố tác động và biến động không thể mô hình hóa bằng các phương pháp truyền thống.

Các kiến trúc học sâu phổ biến:

- Mạng nơ-ron nhân tạo (ANN): Mạng nơ-ron nhân tạo là mô hình học sâu cơ bản, trong đó các "nơ-ron" được kết nối với nhau qua các lớp ẩn để học từ dữ liệu. ANN có thể học từ dữ liệu quá khứ về giá cổ phiếu và các yếu tố tác động để đưa ra các dự đoán về giá trong tương lai. Mô hình này có thể giải quyết các vấn đề phi tuyến tính mà các mô hình hồi quy tuyến tính không thể làm được.

- Mạng nơ-ron tích chập (CNN): Mặc dù CNN chủ yếu được sử dụng trong xử lý ảnh, nó cũng có thể được áp dụng trong phân tích chuỗi thời gian và dự báo giá chứng khoán. CNN có thể nhận diện các đặc điểm quan trọng trong dữ liệu chuỗi thời gian (như các xu hướng ngắn hạn của giá cổ phiếu), và từ đó dự đoán các biến động trong tương lai.

- Mạng nơ-ron tái hồi (RNN): Mạng nơ-ron tái hồi, đặc biệt là LSTM (Long Short-Term Memory) là một lựa chọn phổ biến trong dự báo giá chứng khoán khi dữ liệu có tính chuỗi thời gian. RNN và LSTM có khả năng học và dự đoán các xu hướng dài hạn và ngắn hạn trong dữ liệu, giúp dự báo giá cổ phiếu trong các tình huống biến động mạnh.

Có thể nói, các phương pháp học máy trong dự báo giá chứng khoán ngày càng trở nên phổ biến và mạnh mẽ nhờ khả năng xử lý và phân tích các dữ liệu phức tạp. Từ

các phương pháp học có giám sát, học không giám sát, học sâu, mỗi phương pháp đều có ưu điểm và ứng dụng riêng biệt, giúp các nhà đầu tư tối ưu hóa các quyết định đầu tư. Tuy nhiên, việc áp dụng học máy vào dự báo giá chứng khoán cần có sự kết hợp khéo léo giữa dữ liệu, thuật toán và các yếu tố thị trường để đạt được hiệu quả cao nhất.

Trong bài báo cáo này, em đã sử dụng mô hình RNN và LSTM để dự báo giá chứng khoán để giúp các công ty, doanh nghiệp có cái nhìn tổng quan và đưa ra quyết định dễ dàng hơn.

1.2.5 Vai trò của việc dự báo giá chứng khoán

Dự báo giá chứng khoán giúp nhà đầu tư có thể dự đoán sự biến động của thị trường và các cổ phiếu, từ đó đưa ra quyết định đầu tư hợp lý nhằm tối đa hóa lợi nhuận. Đồng thời, dự báo giúp phát hiện sớm các dấu hiệu của sự thay đổi trong xu hướng thị trường, giúp giảm thiểu rủi ro cho nhà đầu tư.

Việc dự báo giá chứng khoán cung cấp thông tin hữu ích để xây dựng và điều chỉnh danh mục đầu tư, giúp nhà đầu tư phân bổ vốn một cách hợp lý giữa các loại tài sản, giảm thiểu sự phụ thuộc vào một ngành hoặc cổ phiếu duy nhất.

Dự báo giá chứng khoán cũng giúp nhà đầu tư xác định thời điểm mua vào hoặc bán ra cổ phiếu, từ đó tối ưu hóa lợi nhuận. Các chiến lược giao dịch theo xu hướng hoặc theo phân tích kỹ thuật có thể được áp dụng dựa trên dự báo chính xác về giá cổ phiếu.

Chương 2

Cơ sở lý thuyết

2.1 Chuỗi thời gian

2.1.1 Khái niệm chuỗi thời gian

Định nghĩa 2.1. Mô hình chuỗi thời gian [1] là một dãy các giá trị được ghi nhận $\{x_t\}$, trong đó mỗi giá trị x_t được xem như một đại diện tại thời điểm t và là kết quả từ một chuỗi các biến ngẫu nhiên $\{X_t\}$. Những biến này tạo thành một quá trình ngẫu nhiên, được xác định bởi các tham số như phân phối xác suất, kỳ vọng hoặc tương quan giữa các biến trong chuỗi.

Chuỗi thời gian được hiểu đơn giản là tập hợp các quan sát $\{x_t\}$, mỗi quan sát ứng với một thời điểm cụ thể $t \in T$. Nếu miền thời gian T là tập rời rạc, chẳng hạn các số nguyên dương, thì chuỗi thời gian được gọi là rời rạc. Nếu T là tập liên tục (chẳng hạn như một khoảng trên tập số thực $\mathbb{R}$), thì chuỗi thời gian đó được gọi là liên tục.

Ví dụ 2.1. Chuỗi thời gian rời rạc: Giá cổ phiếu của một công ty được ghi lại vào cuối mỗi ngày giao dịch, số lượng khách hàng đến một cửa hàng được đo đếm vào cuối mỗi tháng...

Ví dụ 2.2. Chuỗi thời gian liên tục: Nhiệt độ được đo liên tục trong một khoảng thời gian hàng giây hoặc hàng phút trong suốt một ngày, tín hiệu điện tim được ghi lại liên tục trong một khoảng thời gian để theo dõi hoạt động của tim...

Việc phân biệt giữa chuỗi rời rạc và liên tục giúp xác định phương pháp xử lý phù hợp, từ việc phân tích thống kê đến dự báo giá trị trong tương lai.

2.1.2 Đặc điểm chuỗi thời gian

Dữ liệu chuỗi thời gian thường bao gồm bốn đặc điểm chính, mỗi đặc điểm đóng vai trò quan trọng trong việc phân tích và dự báo:

- Tính xu hướng (Trend): Thể hiện sự biến đổi dài hạn của dữ liệu theo thời gian, cho thấy xu hướng chung là tăng trưởng, suy giảm hoặc giữ ổn định. Đặc điểm này thường xuất hiện trong các dữ liệu liên quan đến sự phát triển kinh tế, dân số, sản lượng công nghiệp...
- Tính mùa vụ (Seasonality): Biểu thị chuỗi dữ liệu với sự biến động tăng hoặc giảm một cách đều đặn và lặp lại theo thời gian, phản ánh các đặc điểm chu kỳ. Những đặc điểm này có thể xuất hiện theo các chu kỳ quen thuộc, chẳng hạn như hàng ngày (thói quen mua sắm, tiêu dùng, hoạt động giải trí), hàng tháng (doanh thu, doanh số, du lịch, dịch vụ), hàng năm (GDP, kim ngạch xuất nhập khẩu).
- Tính chu kỳ (Cyclical): Phản ánh các dao động lớn hơn, không bị giới hạn bởi thời gian cố định, thường liên quan đến các chu kỳ kinh tế hoặc xã hội. Các giai đoạn tăng trưởng và suy thoái nối tiếp nhau trong chuỗi có thể kéo dài qua nhiều năm, chịu ảnh hưởng của các yếu tố vĩ mô và chính trị.
- Tính bất quy tắc (Irregular): Đây là các biến động không theo quy luật nào, xảy ra ngẫu nhiên do những yếu tố bất ngờ như thiên tai, khủng hoảng kinh tế, hoặc sai sót trong việc thu thập dữ liệu. Thành phần này khó dự đoán và thường được xem như nhiễu trong chuỗi thời gian.

2.2 Mạng nơ-ron

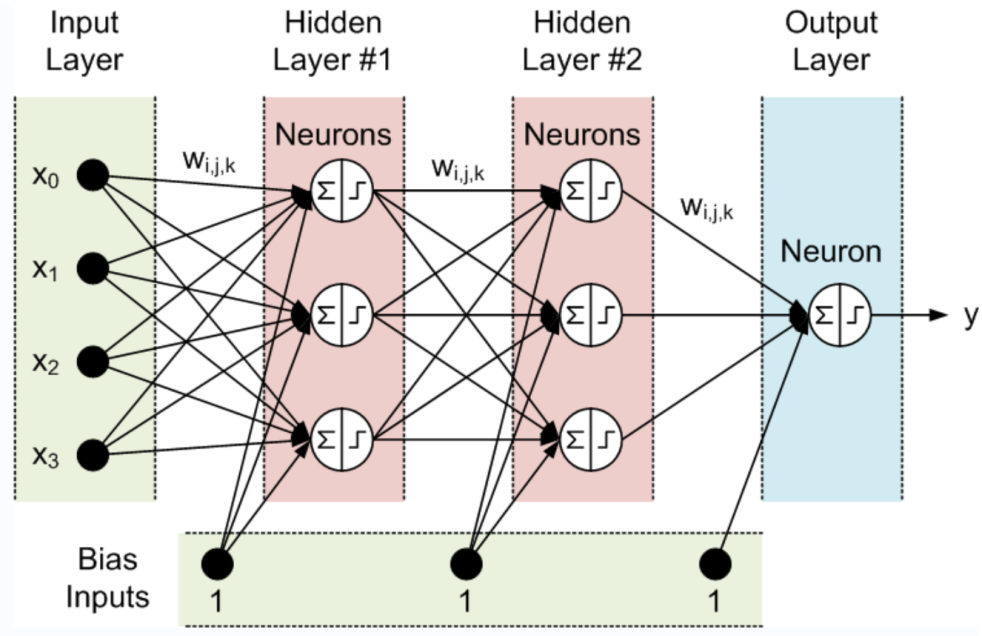
Neural Network (NN) về bản chất là tập hợp của nhiều nơ-ron được liên kết với nhau theo một trật tư, quy luật nhất định. Mạng nơ-ron là một mô hình tính toán được lấy cảm hứng từ cách hoạt động của não bộ con người, được tạo thành từ các nơ-ron liên kết với nhau thành từng lớp [2].

Mỗi nơ-ron nhận tín hiệu đầu vào từ các nơ ron khác, thực hiện tính toán rồi truyền kết quả đầu ra cho các nơ-ron tiếp theo. Nhờ vào việc điều chỉnh các nơ-ron liên kết trong quá trình huấn luyện mà mạng nơ-ron có thể học hỏi và giải quyết các vấn đề phức tạp trong thực tế như nhận diện hình ảnh, dự đoán ngôn ngữ, ký hiệu, dự đoán nhu cầu, xu hướng...

a) Cấu tạo của mô hình Neural Network

Cấu tạo của mô hình Neural Network được thể hiện trong hình (2.1) gồm các thành phần sau:

- Input Layer: Đây là layer đầu tiên chứa các giá trị đầu vào.
- Hidden Layer: Đây là layer ở giữa, layer này có thể gồm nhiều layer khác nhau hoặc không cần có. Mô hình càng có nhiều lớp ẩn thì càng phức tạp.
- Output Layer: Đây là layer cuối chứa các giá trị đầu ra.
  => Tổng số layer trong mô hình được quy ước là số layer - 1 (không bao gồm input layer).
- Node: Node là các hình tròn, mỗi node có hệ số bias b riêng. Mỗi node trong hidden layer và output layer sẽ liên kết với tất cả các node ở layer trước đó, mỗi liên kết có hệ số w riêng.



Hình 2.1: Cấu tạo của mô hình Neural Network

b) Thuật toán của mô hình Neural Network

Bước 1: Tại input layer sẽ chứa các giá trị x_i đầu vào tương ứng với một node.

Bước 2: Tính tổng Linear, đây là bước kết hợp thông tin từ các node ở layer trước

giúp tổng hợp thông tin theo cách tuyến tính để chuẩn bị cho bước xử lý phi tuyến tiếp theo.

$$z_j^{(l)} = \sum_{i=1}^{h^{(l-1)}} w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)} \tag{2.1}$$

Trong đó:

- $z_j^{(l)}$: Giá trị trước khi áp dụng activation function tại node thứ j, layer thứ l.
- $h^{(l-1)}$: Số lượng node trong layer trước đó $l-1$.
- $w_{ji}^{(l)}$: Trọng số weight dùng để nối từ node thứ i trong layer $l-1$ sang node thứ j trong layer l, mỗi kết nối giữa hai node có một w đặc trưng.
- $a_i^{(l-1)}$: Giá trị đầu ra từ node thứ i của layer $l-1$, tức là đầu ra sau khi áp dụng activation function ở layer trước.
- $b_j^{(l)}$: Hệ số điều chỉnh bias tại node thứ j trong layer l.

Bước 3: Áp dụng các hàm kích hoạt (Activation Function) để đưa tính phi tuyến lên tổng tuyến tính để tính đầu ra của node hiện tại, giúp mạng nơ-ron học các quan hệ phi tuyến phức tạp hơn đồng thời kiểm soát tín hiệu lan truyền và chuẩn hoá đầu ra vào khoảng giá trị cụ thể. Ở đây, đầu ra của hàm σ luôn nằm trong khoảng [0,1].

$$a_j^{(l)} = \sigma(z_j^{(l)}) \tag{2.2}$$

Trong đó:

- $a_j^{(l)}$: Giá trị đầu ra của node thứ j trong layer l sau khi áp dụng activation function.
- $\sigma(z_j^{(l)})$: Activation Function (Hàm kích hoạt) áp dụng lên giá trị $z_j^{(l)}$.

Bước 4: Output là giá trị dự đoán y của mô hình.

2.3 Các mô hình dự báo

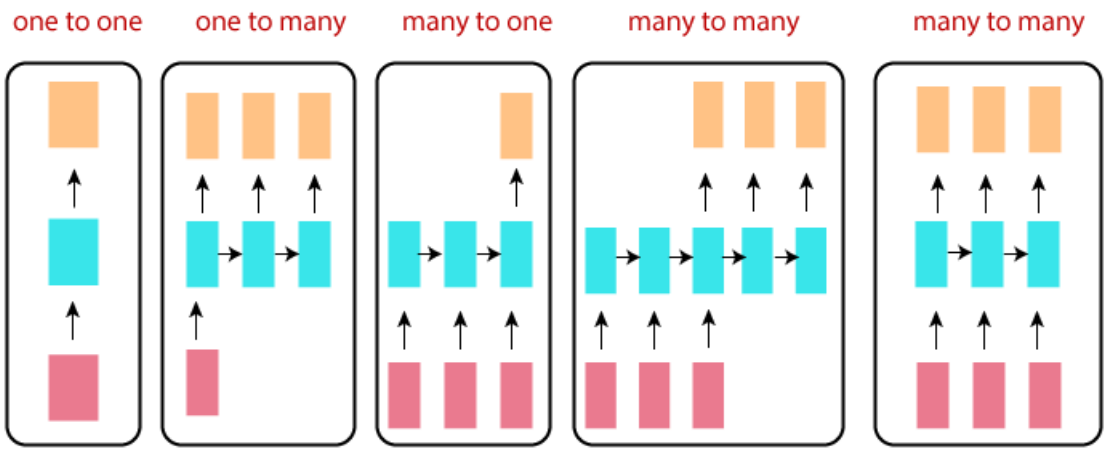
2.3.1 Mô hình Recurrent Neural Network

Recurrent Neural Network (RNN) được giới thiệu lần đầu tiên vào năm 1986 [3], đây là một loại mạng nơ-ron mà trong đó các kết nối giữa các node tạo thành một

chu trình, cho phép thông tin được duy trì và xử lý qua các bước thời gian liên tiếp. RNN có thể xử lý dữ liệu dạng chuỗi hoặc tuần tự, khác biệt so với các mạng nơ-ron truyền thống [4]:

- Mỗi node trong mạng RNN không chỉ nhận thông tin từ đầu vào mà còn từ trạng thái ẩn được cập nhật liên tục qua các bước thời gian.
- Mô hình RNN giúp mạng nơ-ron có khả năng ghi nhớ ngữ cảnh của dữ liệu trước và từ đó có thể đưa ra các dự đoán về hiện tại và tương lai. Chính vì thế, RNN được sử dụng phù hợp với các bài toán như dự đoán chuỗi, phân loại thời gian thực, dịch máy...

a) Phân loại mô hình Recurrent Neural Network



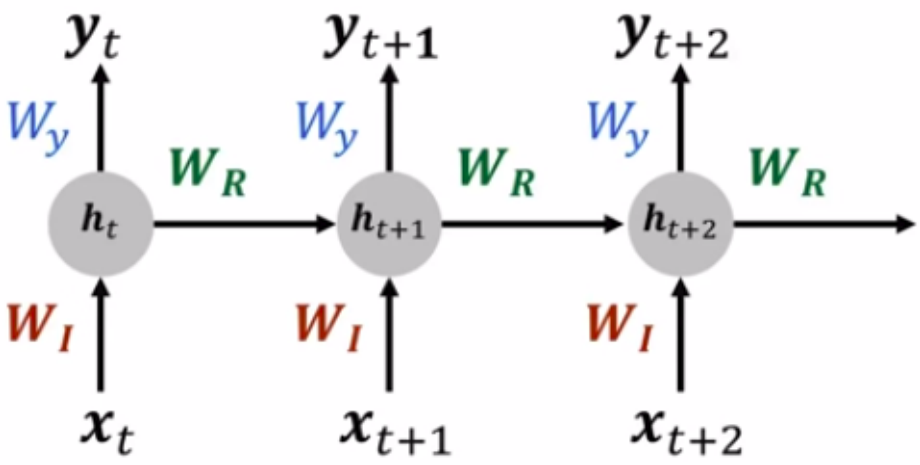
Hình 2.2: Phân loại mô hình Recurrent Neural Network

Phân loại mô hình Recurrent Neural Network được thể hiện trong hình (2.2) như sau:

- One - One: Đây là kiến trúc đơn giản nhất chỉ gồm một đầu vào và một đầu ra, thường dùng để phân loại hình ảnh.
- One - Many: Một đầu vào sẽ ánh xạ đến nhiều đầu ra như từ một bức ảnh có thể tạo ra một chuỗi văn bản.
- Many - One: Nhiều đầu vào ánh xạ đến một đầu ra duy nhất, thường dùng trong các bài toán phân loại cảm xúc, dự đoán chuỗi thời gian...
- Many - Many: Nhiều đầu vào ánh xạ đến nhiều đầu ra nghĩa là số lượng input

luôn bằng số lượng output, thường được dùng để dịch ngôn ngữ ký hiệu...

b) Thuật toán của mô hình Recurrent Neural Network



Hình 2.3: Mô hình Recurrent Neural Network [5]

Thuật toán của mô hình RNN được mô tả như trong hình (2.3) như sau:

Bước 1: Input là các giá trị x_t.

Mô hình RNN nhận một chuỗi dữ liệu đầu vào theo từng bước thời gian, các giá trị đầu vào x_t, x_{t+1}, x_{t+2} sẽ tương ứng với thời điểm t, $t+1$, $t+2$ và được gọi là các timestep. Mỗi bước thời gian này có thể là một từ trong câu, một giá trị trong chuỗi thời gian, một tín hiệu âm thanh tại một thời điểm...

Bước 2: Trạng thái ẩn h_t.

Tại mỗi thời điểm t thì RNN đều có một trạng thái ẩn h_t đóng vai trò như một "bộ nhớ" để lưu trữ thông tin từ các bước thời gian trước đó trong chuỗi. Trạng thái ẩn h_t sẽ được tính toán dựa trên hai yếu tố đó là thông tin từ bước thời gian trước h_{t-1} và thông tin đầu vào hiện tại x_t. Trạng thái ẩn sẽ được cập nhật qua một hàm phi tuyến tính để giúp mô hình học hỏi được các quan hệ phức tạp trong dữ liệu. Quá trình này được tính như sau:

$$h_t = f(W_r h_{t-1} + W_i x_t + b) \tag{2.3}$$

Trong đó:

- f: Hàm kích hoạt.
- W_r: Trọng số giữa h_{t-1} và h_t.
- W_i: Trọng số giữa h_t và x_t.
- b: Hệ số bias riêng.

Vì f là hàm kích hoạt chẳng hạn như σ, $tanh$ hoặc $ReLU$... tuỳ thuộc vào từng bài toán cụ thể. Từ đó, ta sẽ có công thức:

$$h_t = tanh(W_r h_{t-1} + W_i x_t) \tag{2.4}$$

Bước 3: Output là các giá trị y_t.

Đầu ra của mô hình tại mỗi thời điểm t sẽ được tính từ trạng thái ẩn h_t. Công thức tính đầu ra như sau:

$$y_t = W_y h_t \tag{2.5}$$

(Với W_y là trọng số giữa h_t và y_t).

Tóm lại, mỗi bước trong RNN không chỉ xử lý đầu vào hiện tại mà còn ghi nhớ thông tin từ quá khứ trước đó để đưa ra dự đoán cuối cùng sau khi chuỗi dữ liệu hoàn thành.

c) Vấn đề Loss Function

Khi làm việc với chuỗi thời gian dài, RNN thường sẽ gặp phải hai vấn đề đó là vanishing gradient và exploding gradient. Khi huấn luyện RNN, hàm mất mát cần phải được lan truyền ngược qua thời gian (BPTT) để tính toán gradient và cập nhật trọng số. Tuy nhiên, khi chuỗi dữ liệu dài, gradient có thể trở nên rất nhỏ, khiến việc cập nhật trọng số trở nên kém hiệu quả, đặc biệt đối với các bước thời gian sớm hơn trong chuỗi. Từ đó khiến cho mô hình sẽ không học được tốt các thông tin quan trọng từ các bước thời gian đầu của chuỗi.

Ngược lại với vanishing gradient, exploding gradient xảy ra khi các gradient trở nên quá lớn trong quá trình lan truyền ngược. Điều này thường xảy ra khi các trọng số của mô hình quá lớn, khiến cho các cập nhật trọng số bị phóng đại, dẫn đến sự mất ổn định trong quá trình huấn luyện.

2.3.2 Mô hình Long Short Term Memory

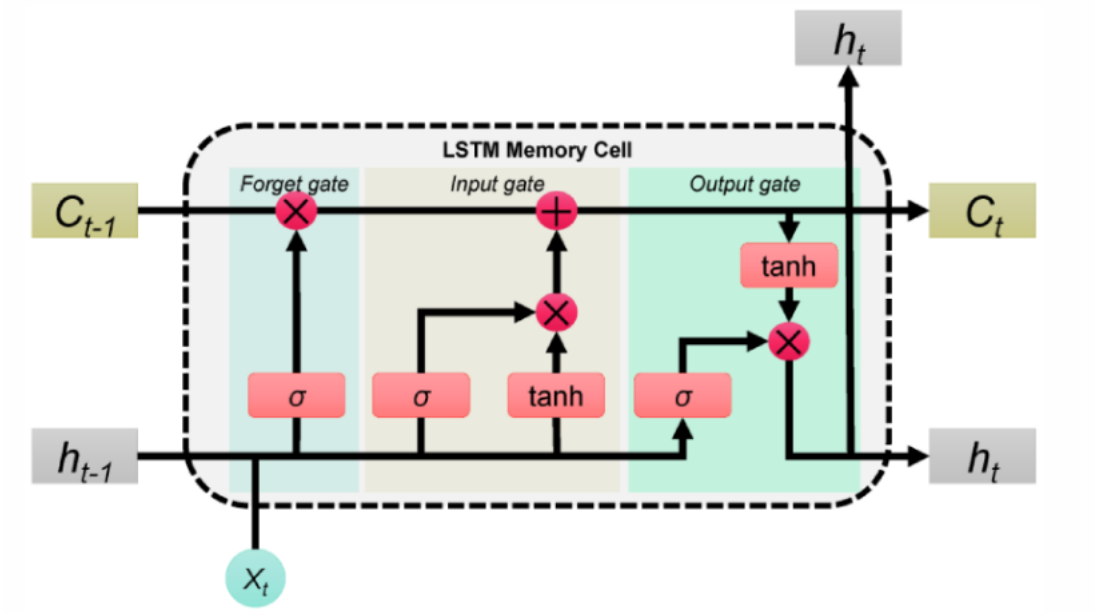
Mô hình Long Short - Term Memory (LSTM) được công bố lần đầu tiên bởi hai tác giả Sepp Hochreiter và Jürgen Schmidhuber vào năm 1997 [6]. LSTM được giới thiệu như một kiến trúc mạng RNN đặc biệt để giải quyết vấn đề vanishing gradient trong việc huấn luyện các mạng nơ-ron truyền thống qua các chuỗi dài.

Hiểu đơn giản, Long Short - Term Memory là một loại mạng nơ-ron hồi quy có cấu trúc đặc biệt với các cell states và các cổng (gates) điều khiển thông tin, cho phép mô hình học cách chọn lọc thông tin quan trọng để lưu trữ trong trạng thái "bộ nhớ" dài hạn và bỏ qua thông tin không cần thiết.

Ví dụ 2.3. Ta có chuỗi ký tự: "Tôi thích học nhưng hôm nay tôi bị ốm". Khi dùng LSTM, mô hình có thể bỏ qua thông tin "Tôi thích học" và chỉ tập trung vào về phía sau "tôi bị ốm" nhằm dự đoán hành động tiếp theo như là nghỉ ngơi, uống thuốc, khám bệnh...

a) Cấu tạo của mô hình Long Short - Term Memory

Cấu tạo của mô hình Long Short - Term Memory [7] được thể hiện như trong hình (2.4) như sau:



Hình 2.4: Mô hình Long Short - Term Memory

- Cell State: Kí hiệu c, thanh trạng thái bộ nhớ giúp lưu trữ thông tin quan trọng trong suốt thời gian dài kể cả thông tin của các timesteps ban đầu.

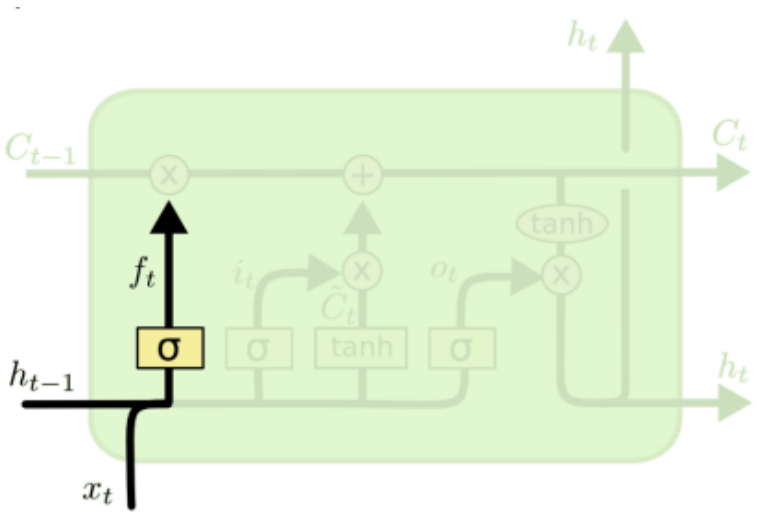
=> Khắc phục được nhược điểm Short - Term Memory.

- Hidden State: Kí hiệu h (có chức năng tương tự h trong mô hình RNN), đây là đầu ra chính tại mỗi bước thời gian được sử dụng trong các bước tính toán tiếp theo hoặc cho lớp đầu ra cuối cùng.
- Ba cổng chính: Forget Gate, Input Gate, Output Gate.

b) Thuật toán của mô hình Long Short - Term Memory

Bước 1: Input là các giá trị x_t, h_{t-1}.

Bước 2: Giá trị x_t, h_{t-1} sẽ được xử lý tại Forget Gate.



Hình 2.5: Forget Gate

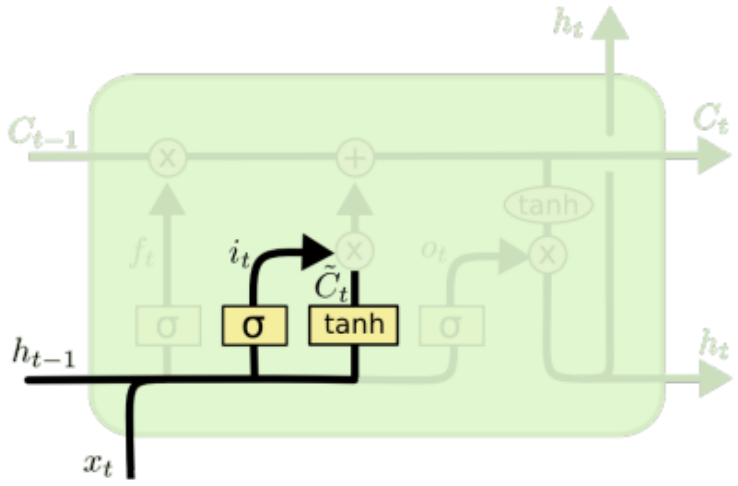
Trước tiên, LSTM sẽ xác định giữ lại hay loại bỏ phần thông tin nào trong trạng thái ô hiện tại c_t. Điều này giúp LSTM tập trung vào những thông tin cần thiết cho việc dự đoán tại thời điểm hiện tại và tương lai, đồng thời tránh lưu trữ thông tin không cần thiết, làm giảm hiệu quả mô hình. Tại đây sẽ sử dụng hàm σ (sigmoid) nhận đầu vào h_{t-1} và x_t để trả về các giá trị đầu ra nằm trong khoảng [0,1]. Khi giá trị gần tiến tới 0 nghĩa là thông tin sẽ bị loại bỏ còn giá trị gần bằng 1 tức là sẽ giữ lại hoàn toàn thông tin. Ta có công thức tính như sau:

$$f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f) \tag{2.6}$$

Trong đó:

- f_t: Đầu ra của forget gate tại thời điểm t.
- σ: Hàm kích hoạt sigmoid.
- U_f: Trọng số liên quan đến đầu vào x_t.
- x_t: Đầu vào ở bước hiện tại.
- W_f: Trọng số của forget gate giúp quyết định hành động ở bước hiện tại.
- h_{t-1}: Trạng thái ẩn từ thời điểm trước đó $t-1$.
- b_f: Hệ số bias của cổng quên.

Bước 3: Tại Input Gate sẽ quyết định thông tin nào cần được thêm vào trạng thái bộ nhớ c.



Hình 2.6: Input Gate

Cổng này quyết định mức độ quan trọng của dữ liệu mới, đồng thời tạo ra một trạng thái bộ nhớ tiềm năng để kết hợp với thông tin cũ. Điều này giúp mô hình liên tục cập nhật thông tin hữu ích.

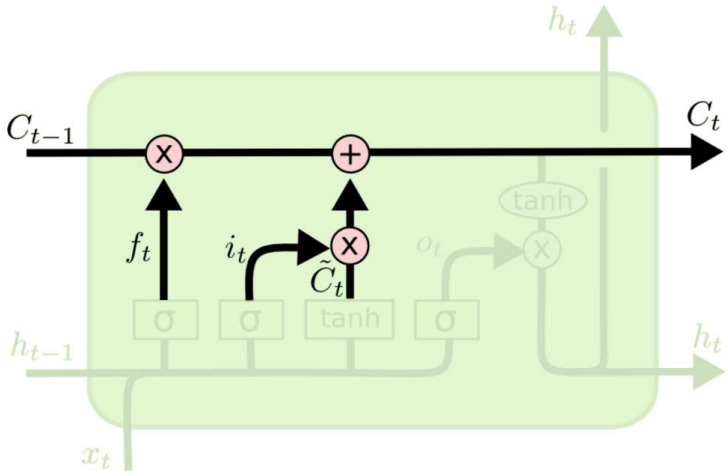
Đầu tiên, tại input gate sẽ sử dụng hàm σ (sigmoid) lần thứ hai để xác định mức độ thông tin mới sẽ được cập nhật. Một hàm *tanh* được dùng để tạo ra một vecto $\tilde{c}_t$, đại diện cho các giá trị trạng thái mới tiềm năng. Đầu ra của hàm sigmoid và hàm tanh được nhân với nhau để tạo thành thông tin cần được thêm vào cell state. Ta có công thức tính như sau:

$$i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i) \tag{2.7}$$

$$\tilde{c}_t = \tanh(U_c * x_t + W_c * h_{t-1} + b_c) \tag{2.8}$$

Trong đó:

- i_t: Đầu ra của input gate tại thời điểm t.
- U_i: Trọng số giữa đầu vào x_t và input gate.
- W_i: Trọng số giữa trạng thái ẩn h_{t-1} và input gate.
- $\tilde{c}_t$: Giá trị được điều chỉnh bởi i_t trước khi thêm vào c_t.
- $tanh$: Hàm kích hoạt nén giá trị vào khoảng [-1,1].
- U_c: Trọng số giữa đầu vào x_t và $\tilde{c}_t$.
- W_c: Trọng số giữa trạng thái ấn h_{t-1} và $\tilde{c}_t$.
- b_i, b_c: Hệ số điều chỉnh của i_t, $\tilde{c}_t$.

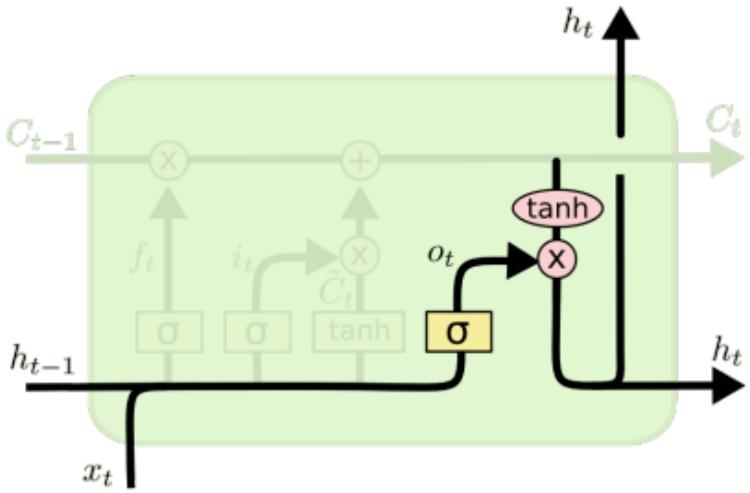


Hình 2.7: Input Gate

Sau đó, trạng thái thanh c_t được làm mới dựa trên thông tin từ forget input và input gate. Tại đây sẽ loại bỏ thông tin cũ bằng cách nhân trạng thái cũ c_{t-1} với đầu ra của forget input. Đồng thời, cập nhật thông tin mới với $\tilde{c}_t$ nhân với đầu ra của input gate. Cuối cùng, cộng tổng cả hai lại vào trạng thái đã được làm sạch. Công thức được viết như sau:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{2.9}$$

Bước 4: Tại Output Gate điều chỉnh lượng thông tin được xuất ra ngoài y_t và truyền đến trạng thái tiếp theo h_t.



Hình 2.8: Output Gate

Sử dụng hàm σ (sigmoid) để quyết định thông tin nào sẽ được đưa ra. Cell state c_t được đưa qua hàm *tanh* và nhân với đầu ra của sigmoid để tạo ra trạng thái ẩn h_t. Ta có công thức như sau:

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \tag{2.10}$$

$$h_t = o_t \tanh(c_t) \tag{2.11}$$

Trong đó:

- o_t: Đầu ra của output gate tại thời điểm t.
- U_o: Trọng số giữa đầu vào x_t và output gate.
- W_o: Trọng số giữa trạng thái ẩn h_t và output gate.

c) Mối quan hệ giữa Short-Term Memory và Long-Term Memory trong LSTM

Short-Term Memory (h_t) là trạng thái ẩn đại diện cho thông tin ngắn hạn và được tính dựa trên trạng thái Long-Term Memory (c_t). Điều này xảy ra vì $h_t = o_t \tanh(c_t)$ chứa thông tin tích lũy dài hạn đã được điều chỉnh bởi các cổng quên và đầu vào [8]. Cổng quên giúp loại bỏ thông tin không còn hữu ích trong khi cổng đầu vào thêm thông tin mới x_t, đảm bảo c_t lưu giữ được những yếu tố quan trọng nhất qua thời gian. Sự kết hợp này tạo ra sự cân bằng giữa c_t mang ngữ cảnh dài hạn và ít

bị ảnh hưởng bởi nhiễu, trong khi h_t phản ứng nhanh chóng với dữ liệu gần đây, giúp mô hình xử lý cả thông tin tổng quát và chi tiết. Điều này làm LSTM hiệu quả trong các chuỗi dữ liệu phức tạp, nơi cả quá khứ xa và gần đều quan trọng.

2.3.3 So sánh mô hình RNN và LSTM

Sự khác nhau giữa mô hình RNN và mô hình LSTM được thể hiện trong bảng (2.1) như sau:

Tiêu chí	**Mô hình RNN**	**Mô hình LSTM**
Cấu tạo	Đơn giản, mỗi bước thời gian chỉ có một trạng thái ẩn duy nhất	Phức tạp hơn với ba cổng (cổng quên, cổng đầu vào, cổng đầu ra) để điều khiển thông tin
Khả năng duy trì thông tin	Hạn chế trong việc duy trì thông tin qua nhiều bước thời gian	Có khả năng duy trì thông tin dài hạn, giúp xử lý chuỗi thời gian dài hiệu quả
Vấn đề gặp phải	Gặp vấn đề vanishing gradient và exploding gradient, khiến việc huấn luyện khó khăn trên chuỗi dài	Giải quyết được vấn đề vanishing gradient, giúp duy trì thông tin qua chuỗi dài
Thời gian huấn luyện	Huấn luyện nhanh hơn do cấu trúc đơn giản	Huấn luyện lâu hơn và tốn nhiều tài nguyên tính toán hơn
Hiệu suất	Hiệu suất giảm khi xử lý chuỗi dài hoặc có sự phụ thuộc dài hạn	Hiệu suất cao hơn trong việc xử lý các chuỗi dài và dữ liệu có mối quan hệ lâu dài
Ứng dụng	Hiệu suất cao hơn trong việc xử lý các chuỗi dài và dữ liệu có mối quan hệ lâu dài	Thích hợp cho chuỗi dài và bài toán cần giữ thông tin lâu dài

Bảng 2.1: So sánh mô hình RNN và LSTM

2.4 Các chỉ số đánh giá mô hình

2.4.1 Mean Squared Error

Mean Squared Error (MSE) đo lường sai số trung bình của các dự báo so với giá trị thực tế, với việc bình phương sai số của từng điểm dữ liệu trước khi tính. MSE luôn không âm và MSE càng nhỏ (gần 0) thì mô hình dự báo càng chính xác. MSE dễ bị ảnh hưởng bởi các sai số lớn (outliers), vì sai số được bình phương. Ta có công thức MSE cho bài toán dự báo giá chứng khoán như sau:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{2.12}$$

Trong đó:

- N: Số lượng mẫu.
- x_i: Giá trị thực tế của mẫu thứ i.
- $\hat{x}_i$: Giá trị dự đoán tương ứng với mẫu thứ i.

2.4.2 Root Mean Squared Error

Root Mean Squared Error (RMSE) là căn bậc hai của MSE và cũng đo lường sai số giữa giá trị thực tế và giá trị dự báo. RMSE có đơn vị giống như dữ liệu gốc, giúp việc diễn giải dễ dàng hơn so với MSE.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{2.13}$$

RMSE càng nhỏ thì mô hình càng chính xác. Tương tự như MSE, RMSE cũng rất nhạy cảm với các outliers, nhưng vì nó trả về một giá trị trong cùng đơn vị đo lường với dữ liệu, nên việc diễn giải dễ dàng hơn.

2.4.3 Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) đo lường sai số trung bình của dự báo dưới dạng tỷ lệ phần trăm so với giá trị thực tế. Nó là một chỉ số rất phổ biến vì giúp đánh giá độ chính xác của mô hình dự báo ở dạng dễ hiểu và có thể so sánh

giữa các mô hình, MAPE được xác định như sau:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \times 100\% \tag{2.14}$$

MAPE càng nhỏ thì mô hình càng chính xác. MAPE có ưu điểm là cho phép so sánh giữa các mô hình và các tập dữ liệu khác nhau, vì nó sử dụng tỷ lệ phần trăm.

2.4.4 Coefficient Of Determination

Coefficient of Determination (R^2) đo lường tỷ lệ phương sai của dữ liệu thực tế được giải thích bởi mô hình. Giá trị của R^2 dao động từ 0 đến 1, với 1 có nghĩa là mô hình giải thích hoàn hảo dữ liệu thực tế và 0 có nghĩa là mô hình không giải thích được gì. Công thức tính R^2 như sau:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.15}$$

Trong đó: $\bar{x}$ là giá trị trung bình của các giá trị thực tế.

2.5 Các hàm kích hoạt

Hàm kích hoạt (Activation Function) là một thành phần quan trọng trong các mạng nơ ron nhân tạo, giúp quyết định xem một nơ ron có nên kích hoạt (hoạt động) hay không. Hàm kích hoạt không chỉ có tác dụng thêm tính phi tuyến cho mô hình mà còn ảnh hưởng đến khả năng học và tính toán của mạng nơ ron.

2.5.1 Hàm sigmoid

Hàm sigmoid là một hàm phi tuyến được sử dụng để chuyển đổi giá trị đầu vào thành một giá trị trong khoảng $[0, 1]$.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.16}$$

Trong đó:

- σ: Hàm sigmoid.
- x: Giá trị đầu vào của nơ ron.
- e: Hằng số Euler (xấp xỉ 2.718).

Output nằm trong khoảng $[0,1]$, giúp dễ dàng giải thích kết quả dưới dạng xác suất, dễ dàng tính đạo hàm, giúp cho quá trình lan truyền ngược trong huấn luyện mạng nơ ron. Tuy nhiên, hàm sigmoid có vấn đề vanishing gradient (đạo hàm gần bằng 0) khi x có giá trị quá lớn hoặc quá nhỏ khiến mạng không thể học được hiệu quả. Chính vì thế, hàm σ thường được sử dụng trong các bài toán phân loại nhị phân (binary classification) hoặc trong các mạng nơ ron hồi tiếp (RNN) khi cần đầu ra xác suất.

2.5.2 Hàm tanh

Hàm tanh (Tanh Hyperbolic), tương tự như sigmoid nhưng với phạm vi đầu ra từ $[-1,1]$. Công thức của hàm tanh như sau:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.17}$$

Hàm tanh thường có xu hướng hoạt động tốt hơn sigmoid vì output có thể âm, giúp cải thiện sự thay đổi trong quá trình lan truyền ngược. Hàm tanh được sử dụng trong các bài toán phân loại đa lớp, mạng nơ ron hồi tiếp và các mô hình mạng nơ ron sâu.

2.5.3 Hàm ReLU

Hàm ReLU (Rectified Linear Unit) là hàm phi tuyến phổ biến trong các mạng nơ ron hiện đại. Hàm này trả về giá trị x nếu $x \geq 0$, và trả về 0 nếu $x < 0$.

$$\text{ReLU}(x) = \max(0,x) \tag{2.18}$$

ReLU có giá trị ≥ 0, giúp giải quyết vấn đề vanishing gradient, tính toán nhanh hơn các hàm sigmoid và tanh vì nó chỉ thực hiện một phép so sánh đơn giản. Một vấn đề lớn của ReLU là khi giá trị đầu vào quá nhỏ, mạng có thể "chết" vì hàm ReLU trả về 0, dẫn đến gradient bằng 0, làm cho nơ-ron đó không thể cập nhật được trong quá trình huấn luyện. ReLU được sử dụng phổ biến trong các mạng nơ ron sâu, đặc biệt là trong các mạng học sâu như CNN, RNN...

2.5.4 Hàm Softmax

Hàm Softmax chuyển đổi đầu ra của mô hình thành xác suất, thường được sử dụng ở lớp đầu ra của các bài toán phân loại đa lớp. Nó tính toán xác suất của mỗi lớp dựa trên các giá trị đầu vào.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{2.19}$$

(Trong đó, z_i: Đầu ra của nơ ron ở lớp cuối).
Output của Softmax luôn nằm trong khoảng $[0,1]$ và tổng của các giá trị đầu ra bằng 1, do đó chúng có thể được coi là xác suất. Hàm Softmax được sử dụng phổ biến trong các mạng nơ ron phân loại, đặc biệt là trong các mô hình phân loại đa lớp.

2.6 Các siêu tham số

Các siêu tham số [9] đóng vai trò quan trọng trong việc điều chỉnh quá trình huấn luyện mô hình học sâu. Chúng ảnh hưởng đến khả năng học của mô hình, tốc độ huấn luyện và chất lượng kết quả:

- Timesteps: Số ngày trước đó được sử dụng để dự đoán giá trị hiện tại, thường nằm trong khoảng [10, 60].
- Hidden Layers: Số lượng đơn vị ẩn trong mỗi lớp (10-200 đơn vị).
- Learning Rate: Điều chỉnh tốc độ thay đổi trọng số, quyết định tốc độ hội tụ và độ ổn định trong quá trình huấn luyện, thường nằm trong khoảng $[1e-5, 1e-2]$.
- Batch Size: Kiểm soát số lượng mẫu được xử lý trong mỗi lần cập nhật, ảnh hưởng đến tốc độ và độ ổn định của huấn luyện, thường có giá trị 16, 32, 64, 128.
- Epochs: Xác định số lần huấn luyện trên toàn bộ dữ liệu, ảnh hưởng đến việc tránh underfitting và overfitting (50-500 vòng).

2.7 Phương pháp Bayesian Optimization

Trong quá trình huấn luyện các mô hình, việc chọn lựa các tham số là một bước rất quan trọng. Việc chọn các giá trị tham số phù hợp thường dựa vào kinh nghiệm và mỗi bộ tham số như vậy cần phải được huấn luyện thử nghiệm, đánh giá kết quả, điều chỉnh tham số rồi lặp lại quá trình này. Để tự động hóa quy trình này, các thuật toán tìm kiếm như Grid Search và Random Search thường được sử dụng. Tuy nhiên, các thuật toán này có hiệu quả giới hạn khi số lượng tham số tăng lên.

Bayesian Optimization (BO) là một thuật toán tối ưu hóa hiệu quả cho những hàm mục tiêu tính toán cao, dựa trên lý thuyết xác suất Bayes. Tree-structured Parzen Estimator (TPE) là một phương pháp tối ưu hóa siêu tham số dựa trên Bayesian Optimization, được thiết kế để hoạt động hiệu quả trong các không gian tham số phức tạp và không có cấu trúc. TPE giải quyết các hạn chế của Gaussian Process trong không gian tham số hỗn hợp (liên tục, rời rạc, phân loại) [10].

Trong Bayesian Optimization, thay vì tối ưu hóa trực tiếp hàm mục tiêu $f(x)$, TPE xây dựng mô hình xác suất thay thế cho hàm mục tiêu thông qua công thức Bayes:

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} \tag{2.20}$$

TPE tách $P(y \mid x)$ thành hai nhóm:

- Nhóm tốt ($g(x)$): Gồm các tham số x có giá trị hàm mục tiêu y tốt hơn một ngưỡng l (ví dụ, $y < l$ khi tối thiểu hóa).
- Nhóm xấu ($l(x)$): Gồm các tham số x có giá trị hàm mục tiêu y kém hơn hoặc bằng l (ví dụ, $y \geq l$).

Công thức xác suất được viết lại thành:

$$P(x \mid y) = \begin{cases} g(x) & \text{nếu } y < l \\ l(x) & \text{nếu } y \geq l \end{cases} \tag{2.21}$$

Thay vì dự đoán $P(y \mid x)$ như Gaussian Process, TPE ước lượng:

$$g(x) = P(x \mid y < l) \quad \text{Xác suất tham số } x \text{ tạo ra kết quả tốt.}$$

$$l(x) = P(x \mid y \geq l) \quad \text{Xác suất tham số } x \text{ tạo ra kết quả kém.}$$

TPE chọn tham số tiếp theo x sao cho tỷ lệ giữa $g(x)$ và $l(x)$ đạt cực đại:

$$x^* = \arg\max \frac{g(x)}{l(x)} \tag{2.22}$$

Quy tắc này giúp TPE ưu tiên khám phá các vùng tham số có khả năng cải thiện giá trị hàm mục tiêu dựa trên dữ liệu đã thu thập.

Chương 3

Ứng dụng các mô hình để dự báo giá chứng khoán

3.1 Phát biểu bài toán

Trong đồ án này sẽ tập trung vào mã cổ phiếu FPT của Công ty Cổ phần Viễn thông FPT đã được giao dịch trên sàn HOSE. Mục tiêu là xây dựng một mô hình học sâu dựa trên kiến trúc RNN và LSTM để dự báo giá đóng cửa của cổ phiếu trong các phiên giao dịch tương lai, đồng thời dựa trên dữ liệu lịch sử như giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa và khối lượng giao dịch.

Dữ liệu lịch sử giá cổ phiếu sẽ được xử lý và chuyển đổi thành dạng chuỗi thời gian để làm đầu vào cho mô hình. Sau khi huấn luyện mô hình với các tham số tối ưu, kết quả dự báo sẽ được so sánh với giá trị thực tế để đánh giá độ chính xác. Bài toán này không chỉ giúp hiểu rõ hơn về xu hướng biến động giá cổ phiếu mà còn cung cấp công cụ hữu ích cho các nhà đầu tư trong việc lập kế hoạch và ra quyết định đầu tư.

3.2 Thu thập dữ liệu

Để thu thập dữ liệu giá chứng khoán của Công ty Cổ phần Viễn thông FPT, em đã sử dụng thư viện *vnstock*. Tại đây, em tiến hành lấy dữ liệu trực tiếp để thực hiện xử lý dữ liệu và xây dựng mô hình mà không cần phải tải dữ liệu về dưới dạng file excel gây tốn thời gian và phức tạp.

vnstock là một thư viện python mã nguồn mở, được phát triển để giúp người dùng dễ dàng tải về và phân tích dữ liệu từ thị trường chứng khoán Việt Nam một cách

nhanh chóng và miễn phí. Đây là nơi cung cấp một loạt các API cho phép truy cập và khai thác dữ liệu từ các sàn giao dịch chứng khoán tại Việt Nam bao gồm thông tin về giá cổ phiếu theo thời gian, dữ liệu tài chính của các công ty cũng như các chỉ số thị trường. Thư viện này lấy dữ liệu từ các nguồn đáng tin cậy như API của SSI (Công ty Chứng khoán Sài Gòn), TCBS (Chứng khoán Techcombank)... đảm bảo cung cấp thông tin chính xác và luôn được cập nhật về thị trường.

3.3 Mô tả dữ liệu

Bộ dữ liệu là thông tin về giá và khối lượng giao dịch hàng ngày của Công ty cổ phần Viễn thông FPT (Mã cổ phiếu: FPT) được lấy trong khoảng thời gian 5 năm từ 20/11/2020 đến 20/11/2024.

Bộ dữ liệu bao gồm các thông tin như sau và được mô tả chi tiết hơn trong hình (3.1) và hình (3.2):

- time: Ngày giao dịch.
- open: Giá mở cửa.
- high: Giá cao nhất.
- low: Giá thấp nhất.
- close: Giá đóng cửa.
- volume: Khối lượng giao dịch.

	time	open	high	low	close	volume
0	2019-11-20	24.64	24.64	23.96	24.05	2291760
1	2019-11-21	24.05	24.18	23.67	23.71	3303950
2	2019-11-22	23.84	24.18	23.50	23.71	1625820
3	2019-11-25	23.75	24.13	23.62	23.96	1087990
4	2019-11-26	24.13	24.22	24.05	24.18	968690
...	...	...	...	...	...	...
1247	2024-11-15	134.37	134.76	131.69	132.98	6574421
1248	2024-11-18	133.08	133.67	131.59	133.08	3910975
1249	2024-11-19	132.58	132.68	129.00	129.10	7652727
1250	2024-11-20	129.10	131.49	125.83	131.49	8474055
1251	2024-11-21	131.49	132.08	130.10	132.08	3183165

[1252 rows x 6 columns]

Hình 3.1: Dữ liệu giá và khối lượng giao dịch của cổ phiếu FPT

Bảng thống kê mô tả dữ liệu:

	open	high	low	close	volume
count	1251.000000	1251.000000	1251.000000	1251.000000	1.251000e+03
mean	60839.697042	61507.927258	60215.836930	60871.484412	2.241995e+06
std	31162.793875	31457.904366	30873.890747	31167.869224	1.673013e+06
min	17160.000000	17410.000000	17000.000000	17120.000000	1.239000e+05
25%	38520.000000	38770.000000	37885.000000	38470.000000	1.171050e+06
50%	57270.000000	57780.000000	56740.000000	57130.000000	1.821700e+06
75%	72690.000000	73415.000000	71585.000000	72860.000000	2.757500e+06
max	141020.000000	141810.000000	138630.000000	140720.000000	1.370680e+07

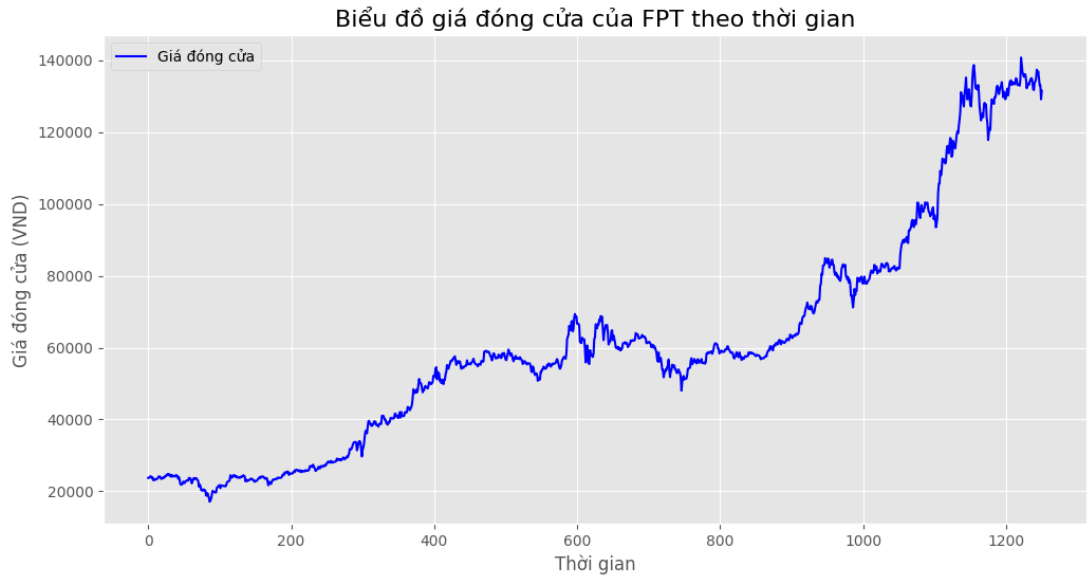
Hình 3.2: Bảng thống kê mô tả dữ liệu

Về cơ bản, bộ dữ liệu khá sạch với 0 giá trị null và 0 giá trị trùng lặp:

```
Số lượng giá trị null trong dữ liệu:
time      0
open      0
high      0
low       0
close     0
volume    0
ticker    0
dtype: int64
Số lượng bản ghi trùng lặp: 0
```

Hình 3.3: Thống kê giá trị null và trùng lặp của bộ dữ liệu

Dưới đây là biểu đồ thể hiện giá đóng cửa của FPT trong thời gian từ 20/11/2019 đến 20/11/2024:



Hình 3.4: Biểu đồ thể hiện giá đóng cửa của FPT trong thời gian 5 năm

Từ đồ thị hình (3.4), ta có một số nhận xét về giá chứng khoán của FPT trong vòng 5 năm qua như sau:

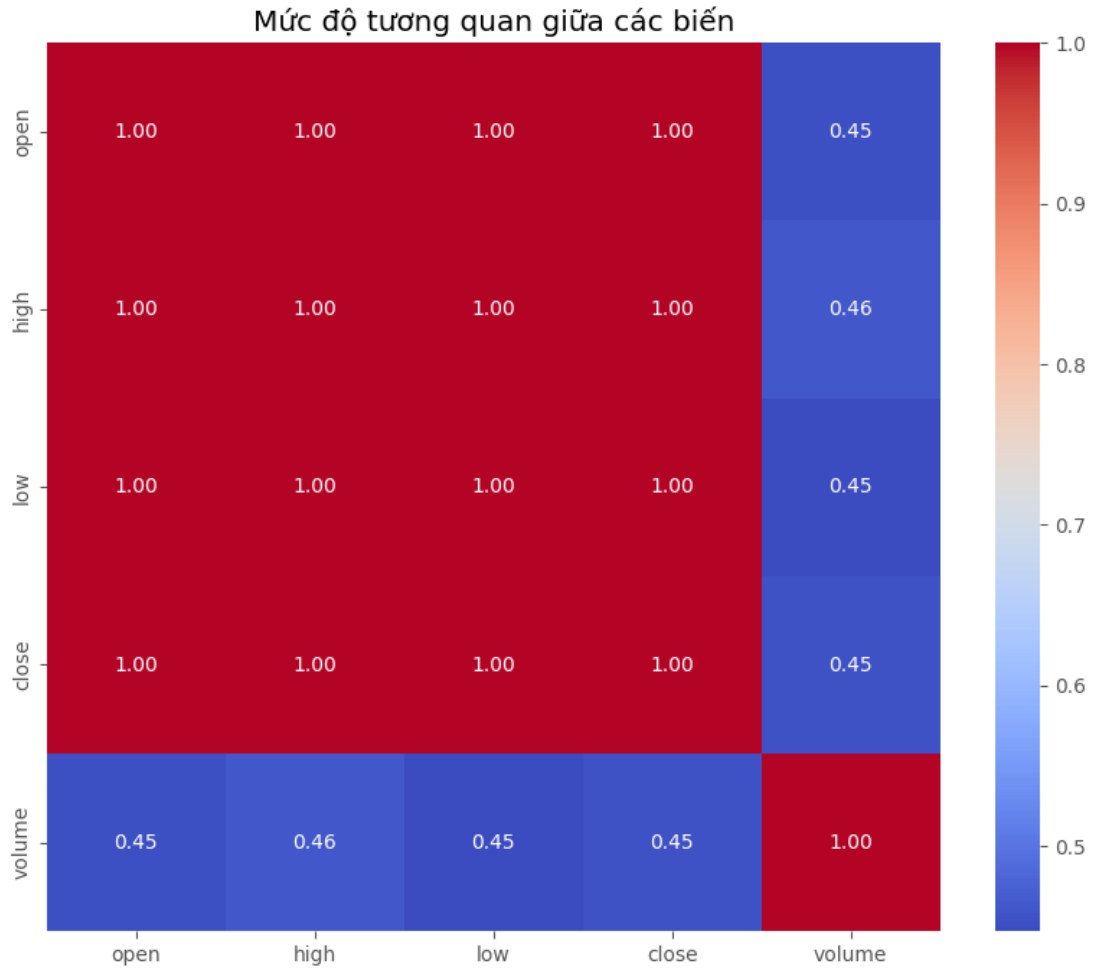
- Nhìn tổng thể, giá cổ phiếu FPT có xu hướng tăng mạnh qua thời gian. Đặc biệt là giai đoạn từ giữa biểu đồ trở đi, mức tăng trở nên rõ rệt hơn.
- Giá cổ phiếu có sự biến động trong ngắn hạn với các đỉnh và đáy rõ rệt, tuy nhiên, xu hướng chung vẫn là đi lên.
- Biểu đồ cho thấy một giai đoạn tăng giá đột biến trong phần cuối. Sau khi đạt đỉnh cao nhất, giá cổ phiếu có xu hướng giảm nhẹ, thể hiện sự điều chỉnh tự nhiên của thị trường.

Giá đóng cửa thường được chọn để huấn luyện mô hình dự đoán vì nó là giá cuối cùng trong ngày, phản ánh sự đồng thuận giữa người mua và bán, đồng thời mang tính ổn định và đại diện nhất. So với giá mở cửa, giá cao nhất và thấp nhất thì giá đóng cửa ít bị ảnh hưởng bởi biến động trong ngày.

Chính vì thế, trong bài toán dự báo này, em đã chọn biến "close" là biến chính để huấn luyện mô hình. Nhưng để có thể đánh giá về mô hình LSTM, RNN (đơn biến) và LSTM, RNN (đa biến) sẽ cho kết quả khác nhau như thế nào nên chúng ta cần chọn thêm các biến khác để kết hợp với biến "close" giúp so sánh dễ dàng hơn. Em đã tiến hành tính toán mức độ tương quan giữa các biến "open, high, low, volume" với biến "close" và thu được kết quả như sau:

open	0.999454	high	0.999762
low	0.999697	volume	0.454607

Bảng 3.1: Bảng kết quả tương quan của các biến với biến "close"



Hình 3.5: Biểu đồ tương quan giữa các biến

Từ kết quả trên, có thể thấy biến "high" và "volume" lần lượt có mức độ tương quan cao nhất và thấp nhất so với biến "close". Khi kết hợp biến "close" với các biến có tương quan cao nhất và thấp nhất này sẽ giúp mô hình huấn luyện hiệu quả hơn vì những biến này thường chứa thông tin quan trọng và bổ trợ cho nhau.

- Biến có tương quan cao với "close" thường phản ánh xu hướng hoặc mô hình chuyển động tương tự, giúp mô hình nhận diện các yếu tố ảnh hưởng trực tiếp đến giá.
- Ngược lại, các biến có tương quan thấp có thể cung cấp thông tin độc lập hoặc phản ánh mối quan hệ ngược chiều, giúp mô hình hiểu rõ hơn về các yếu tố đa chiều tác động đến giá cổ phiếu.

Sự kết hợp này giúp mô hình không chỉ tập trung vào yếu tố chính mà còn cân nhắc những tác động ngoại biên, từ đó tăng độ chính xác và khả năng khái quát hóa của mô hình.

3.4 Tiền xử lý dữ liệu

Trước khi xây dựng và huấn luyện mô hình, dữ liệu cần được xử lý qua một vài bước. Vì dữ liệu giá chứng khoán có tính liên tục theo thời gian, dữ liệu sau thường phụ thuộc vào dữ liệu trước đó. Do đó, việc giữ nguyên trình tự thời gian là rất quan trọng để mô hình học được xu hướng và sự phụ thuộc giữa các thời điểm. Nếu chia dữ liệu ngẫu nhiên theo % sẽ làm mất đi các mối quan hệ tuần tự trong dữ liệu, dẫn đến mô hình không hiểu đúng xu hướng. Vì vậy, ta tiến hành chia tập dữ liệu thành 3 phần như sau:

- Train (20/11/2019 - 20/11/2022): Đây là dữ liệu lịch sử mà mô hình sử dụng để "học" cách giá cổ phiếu thay đổi theo thời gian, mô hình sẽ tìm ra các quy luật hoặc xu hướng lặp lại. Đây là giai đoạn mô hình "được dạy" về hành vi giá trong quá khứ.
- Validation (21/11/2022 - 20/11/2023): Dữ liệu này nằm ngay sau tập train và được sử dụng để kiểm tra mô hình trong quá trình huấn luyện. Khi mô hình dự đoán giá trên tập validation, chúng ta điều chỉnh các tham số của mô hình sao cho dự đoán chính xác nhất. Validation giúp đảm bảo rằng mô hình không chỉ học thuộc lòng dữ liệu train mà còn có khả năng dự đoán tốt trên dữ liệu chưa từng thấy.
- Test (21/11/2023 - 20/11/2024): Đây là giai đoạn cuối cùng để đánh giá chất lượng của mô hình. Tập test không được sử dụng trong quá trình huấn luyện hoặc điều chỉnh mô hình nên nó đóng vai trò mô phỏng dữ liệu hoàn toàn mới (giống như những gì mô hình sẽ gặp trong thực tế). Kết quả trên tập test cho thấy mô hình thực sự dự đoán tốt đến đâu khi áp dụng vào dữ liệu thực tế trong tương lai.

Cùng với đó, em sẽ chuẩn hoá dữ liệu bằng phương pháp Min-Max. Mục tiêu của phương pháp này là đưa các giá trị dữ liệu về một phạm vi nhất định [0,1] để đảm bảo sự nhất quán và hiệu quả cho mô hình.

3.5 Xây dựng và huấn luyện mô hình RNN

3.5.1 Mô hình RNN đơn biến

a) Chọn biến huấn luyện

Đối với mô hình RNN đơn biến, em chỉ sử dụng một biến duy nhất là biến "close" để huấn luyện mô hình.

b) Chọn các siêu tham số

Trước khi quá trình huấn luyện mô hình diễn ra, chúng ta cần tìm kiếm bộ các siêu tham số để kết quả dự đoán đem lại tốt nhất. Em đã lựa chọn phương pháp Bayesian Optimization để thực hiện tối ưu hoá các siêu tham số và thu được kết quả như sau:

- timesteps = 20
- hidden layers = [50,40]
- learning rate = 0.0035246357544657 36
- batch size = 64
- epochs = 150

c) Xây dựng mô hình RNN

1. Tạo mô hình dưới dạng tuần tự nhiều lớp, với các lớp LSTM nối tiếp nhau để học các đặc trưng trong chuỗi thời gian.

2. Cấu trúc mô hình bao gồm:

- Lớp đầu tiên: Lớp này gồm 1 đơn vị RNN, nhận đầu vào với 20 bước thời gian và 1 đặc trưng tại mỗi thời điểm. Hàm kích hoạt ReLU được sử dụng để xử lý dữ liệu và lớp trả về toàn bộ chuỗi đầu ra, cho phép các lớp RNN tiếp theo học thêm từ dữ liệu này.
- Lớp thứ hai: Lớp này tiếp nhận chuỗi đầu ra từ lớp trước. Nó chứa 50 đơn vị ẩn, sử dụng hàm kích hoạt ReLU để xử lý dữ liệu và tiếp tục trả về toàn bộ chuỗi đầu ra để truyền đến lớp kế tiếp.
- Lớp cuối cùng: Lớp RNN cuối cùng bao gồm 40 đơn vị ẩn, trả về toàn bộ chuỗi đầu ra. Giá trị này sau đó được đưa qua một lớp Dense để thực hiện dự

đoán giá trị tiếp theo trong chuỗi.

d) Huấn luyện mô hình

- Sau khi xây dựng mô hình, mô hình được biên dịch với hàm mất mát MSE để đánh giá sai số dự đoán và bộ tối ưu hóa Adam với tốc độ học xác định bởi learning rate = 0.0035.
- Mô hình được huấn luyện trên dữ liệu train, đồng thời kiểm tra hiệu suất trên tập validation với số vòng lặp epoch = 150 và batch size = 64, được điều chỉnh thông qua tham số đầu vào.
- Trong quá trình huấn luyện, dữ liệu không bị xáo trộn để giữ nguyên trình tự thời gian.

3.5.2 Mô hình RNN đa biến

a) Chọn biến huấn luyện

Đối với mô hình RNN đa biến, em sử dụng các biến "close, high, volume" để huấn luyện mô hình.

b) Chọn các siêu tham số

- timesteps = 60
- hidden layers = [35,35]
- learning rate = 0.006202640842438905
- batch size = 64
- epochs = 300

c) Xây dựng mô hình RNN

1. Tạo mô hình dưới dạng tuần tự nhiều lớp, với các lớp LSTM nối tiếp nhau để học các đặc trưng trong chuỗi thời gian.

2. Cấu trúc mô hình bao gồm:

- Lớp đầu tiên: Lớp này có số đơn vị bằng số lượng đặc trưng của mỗi bước thời gian trong dữ liệu đầu vào, nhận đầu vào với 60 bước thời gian. Hàm

kích hoạt ReLU được sử dụng để xử lý dữ liệu và lớp trả về toàn bộ chuỗi đầu ra, cho phép các lớp RNN tiếp theo học thêm từ dữ liệu này.

- Lớp thứ hai: Lớp này tiếp nhận chuỗi đầu ra từ lớp trước. Nó chứa 35 đơn vị ẩn, sử dụng hàm kích hoạt ReLU để xử lý dữ liệu và tiếp tục trả về toàn bộ chuỗi đầu ra để truyền đến lớp kế tiếp.
- Lớp cuối cùng: Lớp RNN cuối cùng bao gồm 35 đơn vị ẩn, trả về toàn bộ chuỗi đầu ra cuối cùng. Giá trị này sau đó được đưa qua một lớp Dense để thực hiện dự đoán giá trị tiếp theo trong chuỗi.

d) Huấn luyện mô hình

- Dùng hàm mất mát MSE để đánh giá sai số dự đoán và bộ tối ưu hóa Adam với learning rate = 0.0062.
- Mô hình được huấn luyện trên dữ liệu train, đồng thời kiểm tra hiệu suất trên tập validation với số vòng lặp epoch = 300 và batch size = 64, được điều chỉnh thông qua tham số đầu vào.
- Trong quá trình huấn luyện, dữ liệu không bị xáo trộn để giữ nguyên trình tự thời gian.

3.6 Xây dựng và huấn luyện mô hình LSTM

3.6.1 Mô hình LSTM đơn biến

a) Chọn biến huấn luyện

Đối với mô hình LSTM đơn biến, em chỉ sử dụng một biến duy nhất là biến "close" để huấn luyện mô hình.

b) Chọn các siêu tham số

- timesteps = 50
- hidden layers = [20,25]
- learning rate = 0.002537554785518352
- batch size = 64
- epochs = 300

c) Xây dựng mô hình LSTM

1. Tạo mô hình dưới dạng tuần tự nhiều lớp, với các lớp LSTM nối tiếp nhau để học các đặc trưng trong chuỗi thời gian.

2. Cấu trúc mô hình bao gồm:

- Lớp đầu tiên: Lớp này gồm 1 đơn vị LSTM, nhận đầu vào với 50 bước thời gian và 1 đặc trưng tại mỗi thời điểm. Hàm kích hoạt ReLU được sử dụng để xử lý dữ liệu và lớp trả về toàn bộ chuỗi đầu ra, cho phép các lớp LSTM tiếp theo học thêm từ dữ liệu này.
- Lớp thứ hai: Lớp này tiếp nhận chuỗi đầu ra từ lớp trước. Nó chứa 20 đơn vị ẩn, sử dụng hàm kích hoạt ReLU để xử lý dữ liệu, và tiếp tục trả về toàn bộ chuỗi đầu ra để truyền đến lớp kế tiếp.
- Lớp cuối cùng: Lớp LSTM cuối cùng bao gồm 25 đơn vị ẩn. Khác với các lớp trước, lớp này không trả về toàn bộ chuỗi mà chỉ trả về giá trị đầu ra cuối cùng. Giá trị này sau đó được đưa qua một lớp Dense để thực hiện dự đoán giá trị tiếp theo trong chuỗi.

d) Huấn luyện mô hình

- Dùng hàm mất mát MSE để đánh giá sai số dự đoán và bộ tối ưu hóa Adam với learning rate = 0.0025.
- Mô hình được huấn luyện trên dữ liệu train, đồng thời kiểm tra hiệu suất trên tập validation với số vòng lặp epoch = 300 và batch size = 64, được điều chỉnh thông qua tham số đầu vào.
- Trong quá trình huấn luyện, dữ liệu không bị xáo trộn để giữ nguyên trình tự thời gian.

3.6.2 Mô hình LSTM đa biến

a) Chọn biến huấn luyện

Đối với mô hình LSTM đa biến, em chọn các biến "close, high, volume" để huấn luyện mô hình.

b) Chọn các siêu tham số

- timesteps = 30
- hidden layers = [35,25]
- learning rate = 0.003275742243647281
- batch size = 64
- epochs = 250

c) Xây dựng mô hình LSTM

1. Tạo mô hình dưới dạng tuần tự nhiều lớp, với các lớp LSTM nối tiếp nhau để học các đặc trưng trong chuỗi thời gian.

2. Cấu trúc mô hình bao gồm:

- Lớp đầu tiên: Lớp này có số đơn vị bằng số lượng đặc trưng của mỗi bước thời gian trong dữ liệu đầu vào, nhận đầu vào với 30 bước thời gian. Hàm kích hoạt ReLU được sử dụng để xử lý dữ liệu và lớp trả về toàn bộ chuỗi đầu ra, cho phép các lớp LSTM tiếp theo học thêm từ dữ liệu này.
- Lớp thứ hai: Lớp này tiếp nhận chuỗi đầu ra từ lớp trước. Nó chứa 35 đơn vị ẩn, sử dụng hàm kích hoạt ReLU để xử lý dữ liệu, và tiếp tục trả về toàn bộ chuỗi đầu ra để truyền đến lớp kế tiếp.
- Lớp cuối cùng: Lớp LSTM cuối cùng bao gồm 25 đơn vị ẩn. Khác với các lớp trước, lớp này không trả về toàn bộ chuỗi mà chỉ trả về giá trị đầu ra cuối cùng. Giá trị này sau đó được đưa qua một lớp Dense để thực hiện dự đoán giá trị tiếp theo trong chuỗi.

d) Huấn luyện mô hình

- Dùng hàm mất mát MSE để đánh giá sai số dự đoán và bộ tối ưu hóa Adam với learning rate = 0.0025.
- Mô hình được huấn luyện trên dữ liệu train, đồng thời kiểm tra hiệu suất trên tập validation với số vòng lặp epoch = 250 và batch size = 64.
- Trong quá trình huấn luyện, dữ liệu không bị xáo trộn để giữ nguyên trình tự thời gian.

Chương 4

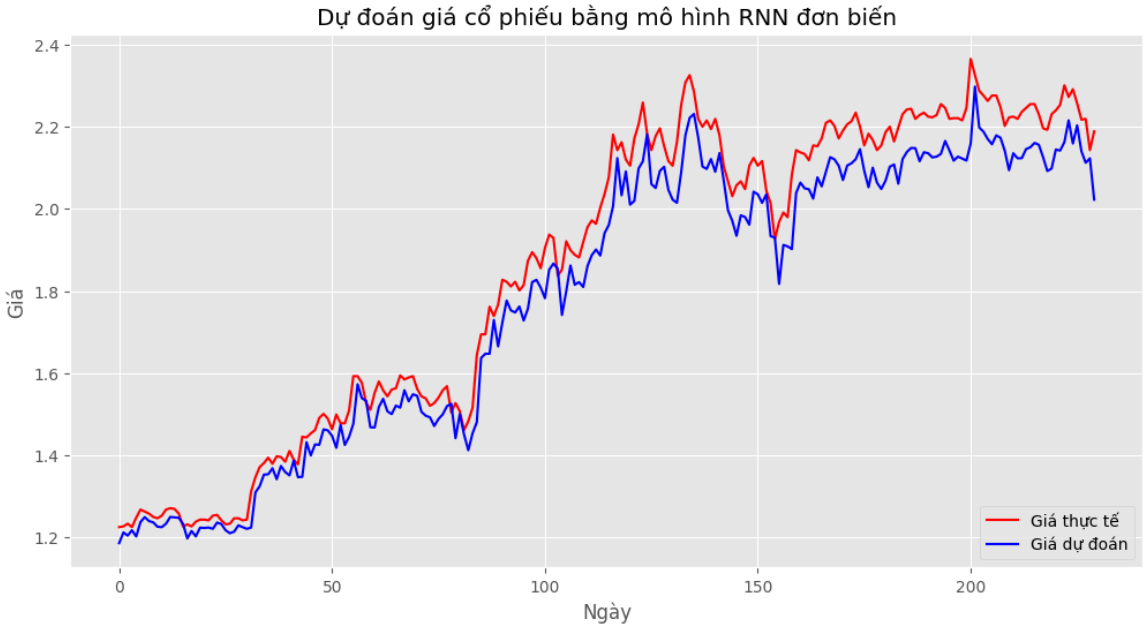
Phân tích kết quả và đánh giá mô hình

4.1 Kết quả của mô hình

4.1.1 Kết quả của mô hình RNN

a) Mô hình RNN đơn biến

Sau quá trình huấn luyện mô hình, ta thu được kết quả chạy mô hình RNN đơn biến được mô tả như hình (4.1):



Hình 4.1: Dự đoán giá cổ phiếu FPT bằng mô hình RNN đơn biến

Đánh giá mô hình qua các chỉ số, ta thu được kết quả như trong bảng (4.1):

Mô hình	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
RNN	close	0.0068	0.0823	0.0364	0.9518

Bảng 4.1: Kết quả chỉ số đánh giá của mô hình RNN đơn biến

Bảng so sánh giá trị thực tế và dự đoán (Đơn vị: nghìn đồng):

	Giá thực tế	Giá dự đoán	Chênh lệch
0	81.13	79.087257	2.042743
1	81.22	80.451859	0.768141
2	81.56	80.046516	1.513484
3	81.13	80.728592	0.401408
4	82.33	79.934563	2.395437
5	83.36	81.761452	1.598548
6	83.11	82.403206	0.706794
7	82.85	81.899895	0.950105
8	82.42	81.731552	0.688448
9	82.25	81.179398	1.070602
10	82.59	81.121460	1.468540
11	83.36	81.562233	1.797767
12	83.53	82.422722	1.107278
13	83.45	82.372620	1.077380
14	82.85	82.313721	0.536279
15	81.22	81.535271	-0.315271

Hình 4.2: Bảng so sánh giá trị thực tế và giá dự đoán bằng mô hình RNN đơn biến

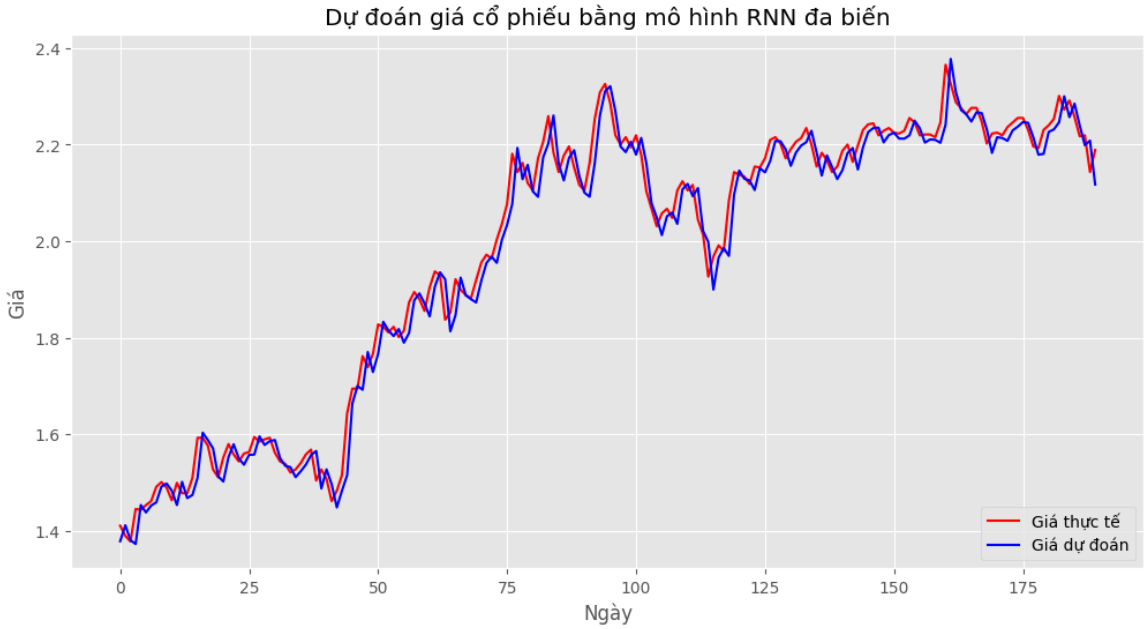
Dựa vào các kết quả thu được, ta có một số nhận xét về mô hình RNN đơn biến trong bài toán dự báo giá chứng khoán như sau:

- Mô hình dự đoán rất tốt với các chuỗi dữ liệu có xu hướng rõ ràng.
- Nhìn vào biểu đồ ở hình (4.1), có thể thấy giá trị dự đoán bám sát theo xu hướng của giá trị thực tế, đặc biệt ở các đoạn tăng hoặc giảm giá mạnh. Tuy nhiên, từ khoảng ngày 150 đến ngày 170 vẫn có một số điểm bất đồng nhỏ ở các giai đoạn biến động mạnh hoặc đột ngột.
- Với $R^2 > 95\%$, mô hình gần như khớp hoàn hảo với dữ liệu. Các chỉ số MSE, RMSE và MAPE đều ở mức thấp, khẳng định độ chính xác cao.
- Sự chênh lệch giữa giá thực tế và giá dự đoán chỉ dao động trong khoảng từ 0 đồng - 2.000 đồng, cho thấy mô hình RNN đơn biến đã hoạt động rất hiệu

quả trong việc dự đoán giá cổ phiếu. Với các tài sản tài chính như cổ phiếu, khoảng chênh lệch này được xem là rất nhỏ so với giá trị tuyệt đối của cổ phiếu.

b) Mô hình RNN đa biến

Sau quá trình huấn luyện mô hình, ta thu được kết quả chạy mô hình RNN đa biến được mô tả như hình (4.3):



Hình 4.3: Dự đoán giá cổ phiếu FPT bằng mô hình RNN đa biến

Đánh giá mô hình qua các chỉ số, ta thu được kết quả như trong bảng (4.2):

Mô hình	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
RNN	close	0.0014	0.0380	0.0151	0.9827

Bảng 4.2: Kết quả chỉ số đánh giá của mô hình RNN đa biến

```
Bảng so sánh giá trị thực tế và dự đoán (Đơn vị: nghìn đồng):
    Giá thực tế  Giá dự đoán  Chênh lệch
0         90.82    89.138504    1.681496
1         89.71    90.865616   -1.155616
2         89.11    89.250328   -0.140328
3         92.63    88.838554    3.791446
4         92.54    93.058655   -0.518655
5         93.05    92.258705    0.791295
6         93.48    92.997551    0.482449
7         95.03    93.346741    1.683259
8         95.54    95.061996    0.478004
9         94.94    95.382782   -0.442782
10        93.57    94.623383   -1.053383
11        95.46    93.057205    2.402795
12        94.34    95.569115   -1.229115
13        94.34    93.808304    0.531696
14        95.97    94.137894    1.832106
15       100.34    96.007263    4.332737
```

Hình 4.4: Bảng so sánh giá trị thực tế và giá dự đoán bằng mô hình RNN đa biến

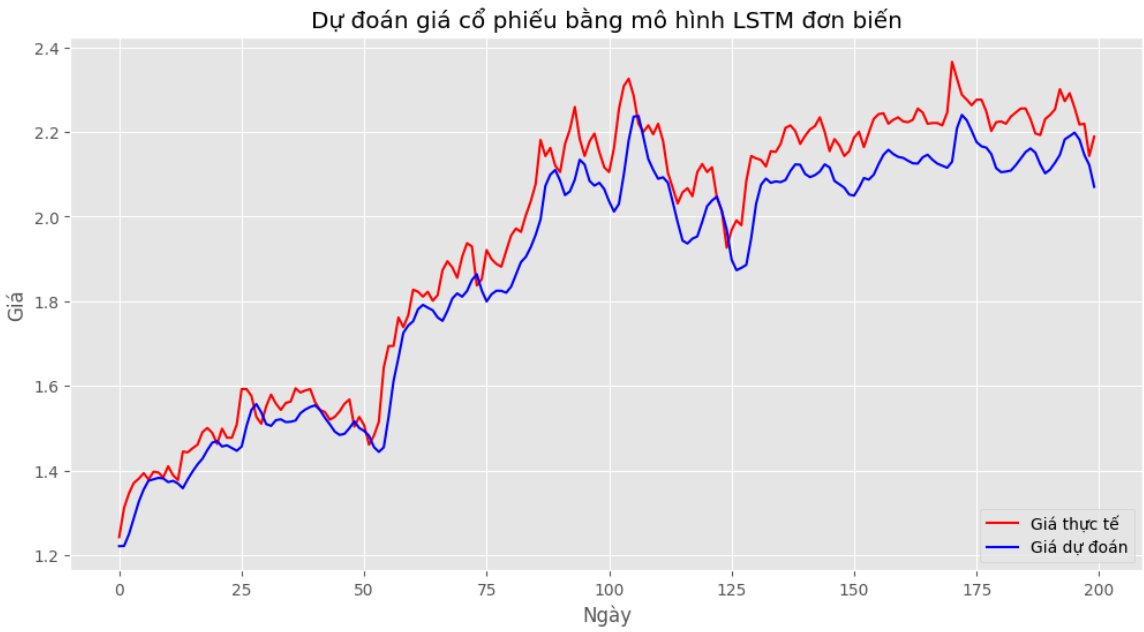
Dựa vào các kết quả thu được, ta có một số nhận xét về mô hình RNN đơn biến trong bài toán dự báo giá chứng khoán như sau:

- Nhìn vào biểu đồ ở hình (4.3), đồ thị của giá dự đoán và giá thực tế gần như là trùng nhau tuyệt đối, cho thấy mô hình đã dự đoán kết quả giá chứng khoán vô cùng chính xác. Các biến động trong dữ liệu thực tế được mô hình nắm bắt rõ ràng bao gồm các xu hướng tăng, giảm và kể cả các dao động nhỏ.
- Với giá trị R^2 lên tới 98%, mô hình gần như khớp hoàn hảo với dữ liệu. Các chỉ số MSE, RMSE đều ở mức rất thấp, chỉ số MAPE đạt kết quả 1.5% khẳng định độ lệch giữa dự đoán và giá trị thực tế khá nhỏ.
- Với các chỉ số đều đạt kết quả tối ưu nên sự chênh lệch giữa giá thực tế và giá dự đoán chủ yếu dao động trong khoảng từ 0 đồng - 1.500 đồng, cho thấy mô hình RNN đa biến đã hoạt động rất hiệu quả trong việc dự đoán giá cổ phiếu của Công ty Cổ phần Viễn thông FPT.

4.1.2 Kết quả của mô hình LSTM

a) Mô hình LSTM đơn biến

Sau quá trình huấn luyện mô hình, ta thu được kết quả chạy mô hình LSTM đơn biến được mô tả như hình (4.5):



Hình 4.5: Dự đoán giá cổ phiếu FPT bằng mô hình LSTM đơn biến

Đánh giá mô hình qua các chỉ số, ta thu được kết quả như trong bảng (4.3):

Mô hình	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
LSTM	close	0.0082	0.0907	0.0388	0.9157

Bảng 4.3: Kết quả chỉ số đánh giá của mô hình LSTM đơn biến

Dựa vào các kết quả thu được, ta có một số nhận xét về mô hình LSTM đơn biến trong bài toán dự báo giá chứng khoán như sau:

- Dựa vào biểu đồ ở hình (4.5), giá thực tế và giá dự đoán không khớp nhau hoàn toàn, đặc biệt trong các giai đoạn biến động lớn. Mô hình dự đoán các xu hướng chung (tăng hoặc giảm) khá tốt nhưng có độ lệch đáng kể ở một số đoạn, đặc biệt khi giá thực tế biến động mạnh từ ngày 100 đến 150.
- Đường dự đoán có xu hướng làm giảm mức độ dao động so với giá thực tế, điều này cho thấy mô hình chưa phản ánh đầy đủ các biến động ngắn hạn.
- Với $R^2 = 91\%$, có thể coi đây cũng là một kết quả khá cao (trên 90%), chứng tỏ mô hình có khả năng giải thích phần lớn biến động của dữ liệu. Các chỉ số MSE, RMSE và MAPE cũng đều ở mức thấp, khẳng định độ chính xác và tối ưu của mô hình.

- Sự chênh lệch giữa giá thực tế và giá dự đoán chỉ dao động phổ biến từ 0 đồng - 3.000 đồng. Một số ít trường hợp chênh tới 5.000 đồng có thể do dữ liệu chứa các biến động bất thường, tuy nhiên số lượng trường hợp này là không đáng kể.

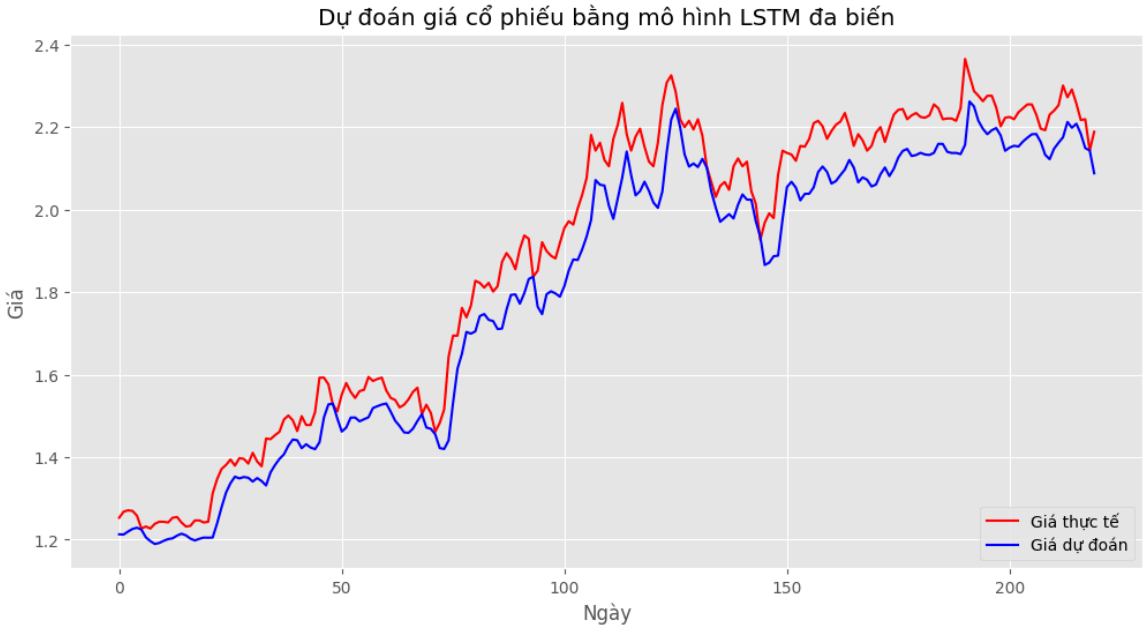
Bảng so sánh giá trị thực tế và dự đoán (Đơn vị: nghìn đồng):

	Giá thực tế	Giá dự đoán	Chênh lệch
0	82.08	80.960075	1.119925
1	85.68	80.983864	4.696136
2	87.48	82.438835	5.041165
3	88.77	84.430367	4.339633
4	89.28	86.431778	2.848222
5	89.97	87.946892	2.023108
6	89.19	89.049385	0.140615
7	90.14	89.201241	0.938759
8	90.05	89.385170	0.664830
9	89.45	89.320496	0.129504
10	90.82	88.865303	1.954697
11	89.71	89.021202	0.688798
12	89.11	88.686989	0.423011
13	92.63	88.102737	4.527263
14	92.54	89.205246	3.334754
15	93.05	90.186722	2.863278

Hình 4.6: Bảng so sánh giá trị thực tế và giá dự đoán bằng mô hình LSTM đơn biến

b) Mô hình LSTM đa biến

Sau quá trình huấn luyện mô hình, ta thu được kết quả chạy mô hình LSTM đa biến được mô tả như hình (4.7):



Hình 4.7: Dự đoán giá cổ phiếu FPT bằng mô hình LSTM đa biến

Đánh giá mô hình qua các chỉ số, ta thu được kết quả như trong bảng (4.4):

Mô hình	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
LSTM	close, high, volume	0.0086	0.0929	0.0433	0.9332

Bảng 4.4: Kết quả chỉ số đánh giá của mô hình LSTM đa biến

Bảng so sánh giá trị thực tế và dự đoán (Đơn vị: nghìn đồng):

	Giá thực tế	Giá dự đoán	Chênh lệch
0	82.59	80.488579	2.101421
1	83.36	80.465775	2.894225
2	83.53	80.842026	2.687974
3	83.45	81.171661	2.278339
4	82.85	81.324219	1.525781
5	81.22	81.088089	0.131911
6	81.48	80.106255	1.373745
7	81.22	79.632454	1.587546
8	81.82	79.261604	2.558396
9	82.08	79.394814	2.685186
10	82.08	79.671021	2.408979
11	81.99	79.884628	2.105372
12	82.59	79.982033	2.607967
13	82.68	80.326660	2.353340
14	81.99	80.572182	1.417818
15	81.48	80.359230	1.120770

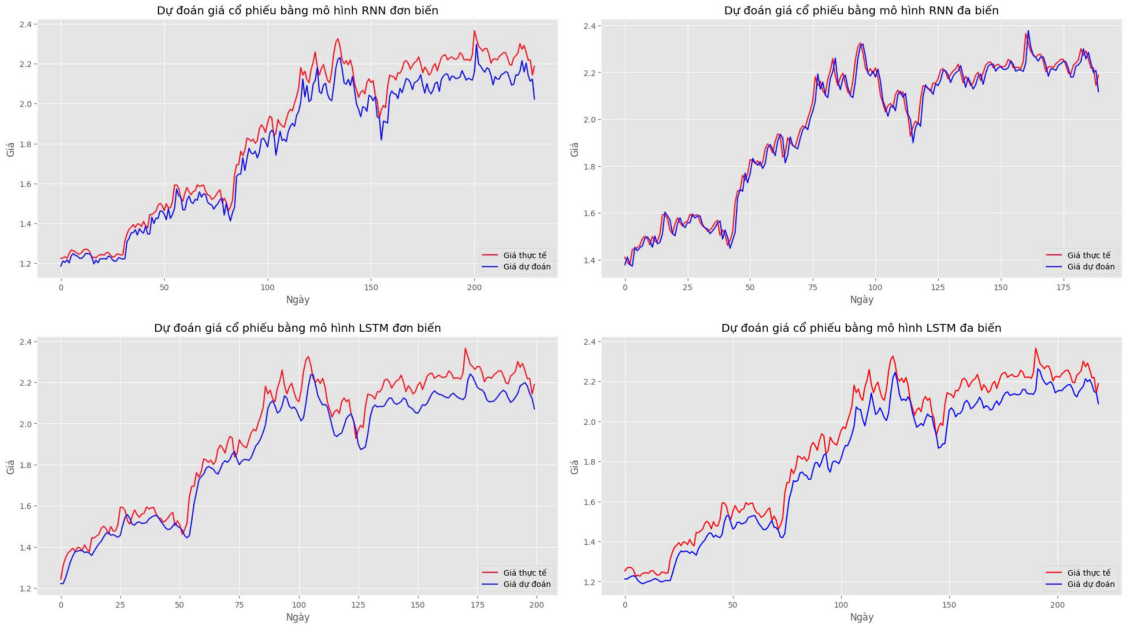
Hình 4.8: Bảng so sánh giá trị thực tế và giá dự đoán bằng mô hình LSTM đa biến

Dựa vào các kết quả thu được, ta có một số nhận xét về mô hình LSTM đa biến trong bài toán dự báo giá chứng khoán như sau:

- Nhìn vào biểu đồ ở hình (4.7), có thể thấy mô hình LSTM đa biến đem lại kết quả tốt hơn so với mô hình LSTM chỉ dùng một biến để huấn luyện. Tại các vùng có sự biến động nhẹ, đường dự đoán gần như bám sát đường giá thực tế.
- Một số biên độ dao động của giá dự đoán nhỏ hơn so với giá thực tế ở một số giai đoạn, cho thấy mô hình có thể chưa dự đoán tốt các biến động ngắn hạn. Giai đoạn từ ngày 150 đến 200 có một số điểm mà giá dự đoán lệch rõ rệt so với giá thực tế. Điều này có thể phản ánh rằng mô hình gặp khó khăn khi xử lý dữ liệu tại các giai đoạn biến động mạnh.
- Sự chênh lệch giữa giá thực tế và giá dự đoán chỉ dao động phổ biến từ 0 đồng - 2.500 đồng, đây là số tiền không đáng kể so với giá trị của một cổ phiếu.

4.2 Đánh giá mô hình

Sau khi đã xây dựng mô hình RNN và LSTM thành công, cả hai mô hình đều đem lại những kết quả rất tích cực được tổng hợp lại trong hình (4.9) và bảng (4.5) như sau:



Hình 4.9: Đồ thị biểu diễn dự đoán giá cổ phiếu bằng RNN và LSTM

Mô hình	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
RNN đơn biến	close	0.0068	0.0823	0.0364	0.9518
RNN đa biến	high, volume, close	0.0014	0.0380	0.0151	0.9827
LSTM đơn biến	close	0.0082	0.0907	0.0388	0.9157
LSTM đa biến	high, volume, close	0.0086	0.0929	0.0433	0.9332

Bảng 4.5: Chỉ số đánh giá kết quả dự báo giá chứng khoán cổ phiếu FPT

So sánh mô hình RNN đơn biến và RNN đa biến:

- Mô hình RNN đơn biến chỉ sử dụng giá đóng cửa (close) làm đầu vào đã bị giới hạn trong việc phân tích các yếu tố tác động đến giá cổ phiếu. Trong khi đó, RNN đa biến mở rộng tập dữ liệu đầu vào bằng cách bổ sung thêm các yếu tố bao gồm giá cao nhất (high) và khối lượng giao dịch (volume).
- Nhờ tích hợp các yếu tố này, RNN đa biến có thể khai thác được nhiều thông tin hơn từ dữ liệu, giúp mô hình nhận diện tốt hơn các tín hiệu phức tạp ảnh hưởng đến giá cổ phiếu. Các yếu tố bổ sung không chỉ cải thiện độ chính xác của dự đoán mà còn tăng khả năng của mô hình trong việc dự đoán các xu hướng biến động ngắn hạn hoặc các sự kiện bất thường.
- Hiệu suất vượt trội của RNN đa biến còn được minh chứng qua các chỉ số đánh giá. Các chỉ số MSE ($0.0014 < 0.0068$), RMSE ($0.0380 < 0.0823$), MAPE ($0.0151 < 0.0364$) đều thấp hơn rõ rệt so với mô hình RNN đơn biến. Chỉ số R^2 ($98\% > 95\%$) chứng tỏ mô hình đa biến giải thích tốt hơn phần lớn các biến động của giá cổ phiếu.

So sánh mô hình LSTM đơn biến và LSTM đa biến:

- Tương tự như mô hình RNN, mô hình LSTM cũng cho kết quả về yếu tố đa biến đem đến hiệu quả tối ưu hơn so với mô hình đơn biến.

Từ bảng kết quả các chỉ số, ta dễ dàng nhận thấy trong bài toán này mô hình RNN đa biến đem lại kết quả chính xác cao nhất, sau đó lần lượt là mô hình LSTM đa biến, RNN đơn biến và cuối cùng là mô hình LSTM đơn biến. RNN có thể là lựa chọn tốt hơn trong bài toán này vì RNN là mô hình học sâu nổi bật cho các chuỗi dữ liệu, nơi mà các yếu tố thời gian và mối quan hệ quá khứ có ảnh hưởng quan trọng đến dự đoán. Trong bối cảnh bài toán hiện tại, RNN có thể là lựa chọn tối ưu vì:

- Dữ liệu không yêu cầu khả năng ghi nhớ dài hạn của các chuỗi, tức là các thông tin trong quá khứ gần ở một khoảng thời gian ngắn sẽ có ảnh hưởng mạnh mẽ hơn đến dự đoán so với thông tin từ quá khứ xa. Trong trường hợp này, RNN có thể phát huy tối đa ưu điểm của mình vì nó có khả năng xử lý các chuỗi dữ liệu có độ dài vừa phải mà không cần đến khả năng ghi nhớ lâu dài của LSTM.
- Mặc dù LSTM có thể xử lý các chuỗi dữ liệu dài và phức tạp hơn nhưng trong

trường hợp này, RNN có thể hiệu quả hơn về mặt tính toán và tài nguyên, vì nó không cần phải duy trì các trạng thái bộ nhớ phức tạp như LSTM.

- Mô hình LSTM là phiên bản mở rộng của RNN, được thiết kế để khắc phục vấn đề "vanishing gradient" trong các chuỗi dữ liệu dài, cho phép mô hình ghi nhớ thông tin lâu dài. Trong bài toán này, LSTM có thể không thực sự cần thiết khi dữ liệu không có sự phụ thuộc lâu dài. Tuy nhiên, nếu sau này chuỗi dữ liệu trở nên dài hơn hoặc phức tạp hơn như thêm yếu tố như mùa vụ, chu kỳ dài hạn, sự kiện ngẫu nhiên ảnh hưởng tới chuỗi thời gian thì LSTM có thể sẽ phát huy được hiệu quả, đặc biệt là trong việc dự đoán các mẫu dữ liệu lâu dài.

Tóm lại, mặc dù mô hình RNN có thể là lựa chọn tốt hơn với dữ liệu hiện tại nhưng việc chuyển sang mô hình LSTM hoặc áp dụng dữ liệu đa biến có thể mở ra cơ hội cải thiện dự đoán giá chứng khoán trong các trường hợp phức tạp hơn.

Kết luận

Những điều đã đạt được sau khi hoàn thành đồ án:

- Trang bị thêm các kiến thức về kinh tế, đặc biệt là về thị trường chứng khoán. Đồng thời, hiểu được tầm quan trọng của các bài toán dự báo nói chung và bài toán dự báo giá chứng khoán nói riêng.
- Tìm hiểu về những kiến thức toán học như chuỗi thời gian, mạng nơ-ron, các siêu tham số, ôn lại các chỉ số đánh giá mô hình. Nghiên cứu chi tiết về hai mô hình là RNN và LSTM cùng các ưu, nhược điểm của từng mô hình.
- Tiến hành xây dựng mô hình bằng ngôn ngữ Python kết hợp với Google Colab để chạy chương trình. Xử lý các vấn đề vanishing gradient, overfitting trong quá trình huấn luyện mô hình. Kết hợp với các thư viện để tìm các siêu tham số giúp mô hình tối ưu hơn.
- So sánh và đánh giá kết quả mà hai mô hình RNN và LSTM đem lại.

Hướng phát triển của đồ án trong tương lai:

- Tối ưu thời gian để tìm các siêu tham số nhanh hơn bằng các phương pháp khác nhau. Thử nghiệm thuật toán trên nhiều bộ dữ liệu chứng khoán của các công ty khác nhau.
- Cải tiến mô hình hơn. Làm giàu dữ liệu, đồng thời xây dựng một cơ sở dữ liệu để lưu trữ và phân tích giá chứng khoán...

Một lần nữa, em muốn gửi lời cảm ơn đến giảng viên hướng dẫn em trong đồ án lần này là thầy **PGS.TS. Nguyễn Đình Hân**, cùng tất cả thầy cô Khoa Toán Tin đã dạy em rất nhiều kiến thức bổ ích từ các học phần khác nhau, góp phần giúp em hoàn thành đồ án. Đồ án không tránh khỏi những sai sót, vụng về. Em rất mong nhận được sự đánh giá và góp ý từ thầy cô để bản thân có thêm nhiều trải nghiệm hơn. Em xin chân thành cảm ơn ạ!

Tài liệu tham khảo

[1] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: Forecasting and control*, 5th. Wiley, 2015.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. DOI: 10.1038/323533a0.

[4] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989. DOI: 10.1162/neco.1989.1.2.270.

[5] T. Mikolov, M. Karafiát, L. Burget, and J. Cernocký, "Recurrent neural network based language model," *Interspeech*, pp. 1045–1048, 2010. DOI: 10.21437/Interspeech.2010-285.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994. DOI: 10.1109/72.279181.

[8] F. A. Gers and J. Schmidhuber, "Learning to forget: Continual prediction with lstm," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, Springer, 2000, pp. 850–855.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980.

[10] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 2546–2554. [Online]. Available: https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf.