ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

HUỲNH ĐỨC HUY

DỰ BÁO XU HƯỚNG CHỨNG KHOÁN DỰA VÀO TIN TỨC TÀI CHÍNH TẠI SÀN GIAO DỊCH TP.HỒ CHÍ MINH

LUẬN VĂN THẠC SỸ

NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2017

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

HUỲNH ĐỨC HUY

DỰ BÁO XU HƯỚNG CHỨNG KHOÁN DỰA VÀO TIN TỨC TÀI CHÍNH TẠI SÀN GIAO DỊCH TP.HỒ CHÍ MINH

LUẬN VĂN THẠC SỸ

NGÀNH KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. TIẾN SỸ DƯƠNG MINH ĐỨC

TP. HỒ CHÍ MINH, 2017

LỜI CẢM ƠN

Đầu tiên, tác giả xin gởi lời cảm ơn sâu sắc đến những người thân trong gia đình, những người đã không ngại vất vả để cho tác giả được theo đuổi con đường mà mình đã chọn. Đặc biệt, tác giả xin gửi lời cảm ơn và lòng biết ơn chân thành đến Tiến sĩ Dương Minh Đức, người hướng dẫn khoa học tận tâm và nghiêm túc. Thầy đã tạo điều kiện tốt nhất cho tác giả trong suốt quá trình thực hiện luận văn tốt nghiệp tại nhóm nghiên cứu bộ môn, truyền đạt cho tác giả những kinh nghiệm quý báu giúp tác giả có thể tự tin bước đi trên con đường nghiên cứu khoa học. Tác giả xin cảm ơn đến các thành viên nhóm nghiên cứu của bộ môn, các thành viên đã giúp đỡ và hỗ trợ rất nhiều để tác giả hoàn thành được luận văn này. Bên cạnh đó tác giả cũng xin cảm ơn giáo sư Takasu - viện nghiên cứu quốc gia Nhật Bản, tuy thời gian thực tập tại phòng thí nghiệm của viện có 5 tháng nhưng Giáo sư và các thành viên của phòng thí nghiệm đã tạo điều kiện cho tác giả tiếp xúc với môi trường nghiên cứu khoa học chuyên nghiệp, giúp tác giả định hướng trong quá trình thực hiện luận văn.

Trong thời gian hơn 6 tháng thực hiện đề tài, tác giả đã cố gắng vận dụng những kiến thức nền tảng đã tích lũy, đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới. Tuy nhiên, chắc chắn tác giả không tránh khỏi những thiếu sót, chính vì vậy tác giả rất mong nhận được những sự góp ý từ phía thầy cô nhằm hoàn thiện những kiến thức mà tác giả đã học tập để làm hành trang thực hiện tiếp các đề tài nghiên cứu khác trong tương lai. Những kiến thức đã tích lũy mà quý thầy cô truyền đạt sẽ mãi là những hành trang quý báu nhất để tác giả tự bước đi trên con đường mà mình đã chọn.

Xin chân thành tri ân!

Tp Hồ Chí Minh, tháng 01 năm 2017

Học viên

Huỳnh Đức Huy

LỜI CAM ĐOAN

Tác giả xin cam đoan đây là công trình nghiên cứu của bản thân dưới sự hướng dẫn của Tiến sĩ Dương Minh Đức. Các số liệu, kết quả trình bày trong luận văn là trung thực. Các tư liệu được sử dụng trong luận văn có nguồn gốc và trích dẫn một cách rõ ràng, đầy đủ.

Tp Hồ Chí Minh, tháng 01 năm 2017

Học viên

Huỳnh Đức Huy

MỤC LỤC

MỤC LỤC 3
Danh mục hình vẽ 6
Danh mục bảng 7
Danh mục các từ viết tắt 8
TÓM TẮT 10
MỞ ĐẦU 12
Chương 1. TỔNG QUAN 16
1.1. Đặt vấn đề 16
1.1.1. Phát biểu bài toán 16
1.1.2. Dữ liệu đầu vào 16
1.1.3. Dữ liệu đầu ra 17
1.2. Các nghiên cứu liên quan 17
1.2.1. Trong nước 17
1.2.2. Ngoài nước 17
1.2.3. Những vấn đề còn tồn tại 19
Chương 2. CƠ SỞ LÝ THUYẾT 20
2.1. Tổng quan về mạng nơ-ron (Neural Network) 20
2.1.1. Kiến trúc của mạng nơ-ron kết nối đầy đủ 20
2.1.2. Phương thức suy luận thông tin của mạng nơ-ron 22
2.1.3. Hàm kích hoạt 23
2.1.4. Mô phỏng hàm xác suất và hàm phân loại 23
2.1.5. Phương pháp ước lượng tham số của mạng nơ-ron 24
2.1.6. Hàm mất mát 25
2.1.7. Vấn đề Overfitting 26

2.2. Mạng Nơ-ron hồi quy 28
2.3. Vấn đề nắm bắt những thông tin dài hạn (Long-Term Memmory) 30
2.4. Mạng Gated Recurrent Unit (GRU) 31
Chương 3. MÔ HÌNH DỰ ĐOÁN XU HƯỚNG GIÁ CHỨNG KHOÁN BẰNG MẠNG NƠ-RON DỰA TRÊN TIN TỨC TÀI CHÍNH 33
3.1. Đề xuất mô hình mạng Gated Recurrent Unit hai chiều 33
3.2. Mô hình dự báo 35
3.2.1. Tiền xử lý văn bản 36
3.2.2. Word Embedding 38
3.2.3. Máy học với mô hình BGRU 40
3.2.4. Kỹ thuật Dropout 40
Chương 4. THỰC NGHIỆM 43
4.1. Cài đặt, côngcụ hỗ trợ 43
4.2. Phương pháp đánh giá 43
4.3. Bộ dữ liệu thực nghiệm 44
4.3.1. Sự tác động của tin tức lên giá chứng khoán theo thời gian 45
4.3.2. Dự báo sự chuyển động giá chứng khoán của mã S&P500 46
4.3.3. Dự báo mã chứng khoán riêng biệt 49
4.4. Dự báo chuyển động giá của VN-INDEX. 50
4.5. Đánh giá 53
Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 54
5.1. Kết quả đạt được 54
5.1.1. Về khoa học 54
5.1.2. Về thực tiễn 54
5.2. Hướng phát triển 54
5.3. Kết luận 55

TÀI LIỆU THAM KHẢO 56
PHỤ LỤC 60
A. Các khái niệm về thị trường chứng khoán 60
B. Mạng Long Short Term Memmory (LSTM) 62

Danh mục hình vẽ

Hình 2.1. Minh hoạ cho kết nối giữa các lớp trong một mạng nơ-ron. 21
Hình 2.2. Ví dụ minh họa cho việc tối ưu một hàm số.....25
Hình 2.3. Một ví dụ về overfitting.....27
Hình 2.4. Minh hoạ “learning curve” khi xuất hiện overfitting 28
Hình 2.5. Minh họa mô hình mạng nơ-ron hồi quy với hàm tanh.....30
Hình 2.6. Minh họa mô hình GRU 32
Hình 3.1. Minh họa mô hình BGRU 34
Hình 3.2. Minh họa mô hình dự báo chuyển động giá chứng khoán 35
Hình 3.3. Minh họa quá trình tiền xử lý văn bản 36
Hình 3.4. Giao diện tách nội dung tin tức từ file html 36
Hình 3.5. Tin tức sau khi được tách nội dung từ file HTML 37
Hình 3.6. Nội dung tin tức sau khi đã được tách từ.....37
Hình 3.7. Minh họa danh sách “từ dừng” của thư viện NLTK. 38
Hình 3.8. Minh họa vec-tơ của tên “quốc gia” và “thủ đô” [29].....40
Hình 3.9. Minh họa kỹ thuật dropout. [13].....41
Hình 3.10. So sánh mô hình BGRU khi áp dụng Dropout 42
Hình 4.1. Kết quả thực nghiệm đánh giá tác động của tin tức theo thời gian. 46
Hình 4.2. Biểu đồ kết quả các độ đo trên mô hình LSTM, GRU và BGRU 48
Hình 4.3. Biểu đồ đánh giá sự tác động tin tức lên từng mã cổ phiếu riêng biệt 49
Hình 4.4. Biểu đồ đánh giá kết quả thực nghiệm BGRU với SVM 52
Hình 4.5. Biểu đồ thể hiện các độ đo theo các mẫu thời gian 53

Danh mục bảng

Bảng 3.1. So sánh số lượng tham số cần ước lượng của các mô hình DL 34
Bảng 4.1. Ma trận kết hợp tính độ chính xác 44
Bảng 4.2. Kết quả thực nghiệm dự báo chuyển động giá mã S&P500 Index 47
Bảng 4.3. Kết quả các độ đo trên mô hình BGRU, GRU và LSTM 48
Bảng 4.4. Thống kê số lượng tin tức các mã cổ phiếu riêng biệt 49
Bảng 4.5. Chi tiết dữ liệu bài báo Tiếng Việt.....51

Danh mục các từ viết tắt

❖ **Tiếng Việt**

STT	Ký hiệu/ Chữ viết tắt	Ý nghĩa
1	CNTT	Công nghệ Thông tin
2	HoSE	Sàn giao dịch chứng khoán TP.HCM
3	KLCP	Khối lượng cổ phiếu
4	TTCK	Thị trường chứng khoán
5	VN-Index	Chỉ số giá cổ phiếu trong một thời gian nhất định (phiên giao dịch, ngày giao dịch) của các công ty niêm yết tại sàn giao dịch chứng khoán TP.HCM

❖ **Tiếng Anh**

STT	Ký hiệu/ Chữ viết tắt	Diễn giải	Ý nghĩa
1	ANN	Artifical Neural Network	Mạng nơ-ron nhân tạo
2	BGRU	Bidirectional Gated Recurrent Unit	Mạng GRU hai chiều
3	DNN	Deep Neural Network	Mạng nơ-ron sâu nhiều lớp
4	DL	Deep Learning	Deep Learning là một phương pháp dựa trên một số ý tưởng từ não bộ tới việc tiếp thu nhiều tầng biểu đạt, cả cụ thể lẫn trừu tượng, qua đó làm rõ nghĩa của các loại dữ liệu.
5	EMH	Efficient Market Hypothesis	Lý thuyết về thị trường
6	GRU	Gated Recurrent Unit	Một biến thể của mạng nơ-ron hồi quy (RNN)
7	LSTM	Long Short Term Memmory	Một biến thể của mạng nơ-ron hồi quy (RNN)
8	NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
9	NLTK	Natural Language Toolkit	Thư viện hỗ trợ xử lý ngôn ngữ tự nhiên trên Python.

10	S&P500	Standard & Poor 500	Chỉ số thị trường chứng khoán dựa trên thị trường vốn hóa của 500 công ty lớn có cổ phiếu phổ thông được niêm yết trên thị trường chứng khoán Hoa Kỳ.
11	RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
12	TF-IDF	term frequency – inverse document frequency	TF-IDF của một từ là một con số thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

TÓM TẮT

Thị trường chứng khoán (TTCK) ngày càng có vai trò quan trọng trong nền kinh tế của một quốc gia. Nhiều nghiên cứu hiện nay trong lĩnh vực TTCK cố gắng dự đoán chính xác giá trị của giá cổ phiếu hoặc dự đoán xu hướng giá cổ phiếu trong tương lai. Các dự đoán này thường dựa trên lịch sử giá, lịch sử giao dịch, khối lượng giao dịch và các phương pháp phân tích kỹ thuật. Tuy nhiên, các kết quả thu được còn nhiều hạn chế vì sự biến động phức tạp của chuỗi giá bởi lẽ TTCK chịu tác động từ rất nhiều yếu tố như tình hình chính trị, xã hội, kinh tế, hiệu suất của công ty,…

Gần đây, với sự thành công trên rất nhiều lĩnh vực của phương pháp máy học bằng Deep Neural Networks (DNN). Các nhà nghiên cứu đã bắt đầu áp dụng các mạng DNN kết hợp cùng với tin tức tài chính vào việc dự báo chuyển động giá chứng khoán. Trong phạm vi khóa luận, tác giả đã nghiên cứu và đề xuất mô hình dự báo Bidirectional Gated Recurrent Unit (BGRU) kết hợp cùng với các kỹ thuật huấn luyện mô hình máy học được sử dụng phổ biến gần đây nhất để dự đoán sự chuyển động giá của chứng khoán dựa vào tin tức tài chính. Khóa luận đã đề xuất các giải pháp để giải quyết các bài toán nhỏ cụ thể sau:

- Bài toán nguồn dữ liệu tin tức tài chính đầu vào là rất đa dạng với bộ từ điển lớn. Thứ tự xuất hiện các từ trong mỗi văn bản là khác nhau và độ dài mỗi văn bản là khác nhau. Đối với bài toán này, khoán luận đã đề xuất mô hình BGRU kết hợp với lớp word embedding có khả năng xử lý các sự đa dạng dữ liệu đầu vào và bộ dữ liệu lớn.
- Bài toán đòi hỏi mô hình máy học có khả năng học (lưu trữ) trên toàn bộ ngữ cảnh của văn bản để tăng độ chính xác. Khóa luận đã phân tích mô hình Gated Recurrent Unit (GRU) với những khả năng xử lý các vấn đề lưu trữ các ngữ cảnh dài hạn và ngắn hạn đối với mô hình văn bản. Đồng thời, mô hình đề xuất BGRU kế thừa những ưu điểm của GRU toàn diện trên cả ngữ cảnh văn bản.

- Bài toán tránh vấn đề quá vừa dữ liệu (overfitting) trong quá trính training dữ liệu với phương pháp máy học. Với vần đề này, khóa luận đã đề xuất áp dụng kỹ thuật dropout cho quá trình huấn luyện máy học để giảm việc quá vừa dữ liệu.
- Bài toán tiền xử lý văn bản với các ngôn ngữ khác nhau.
- Bài toán ứng dụng khả năng dự báo xu hướng chứng khoán trong rổ VN-Index thuộc sàn chứng khoán HoSE dựa trên tin tức tài chính và giá lịch sử của cổ phiếu theo ngày.
- Tìm cách tăng độ tin cậy, chính xác cho chương trình vì lý do hệ thống sử dụng nguồn tin tức có trên các trang báo nên sẽ có độ nhiễu lớn làm giảm độ tin cậy, chính xác.

Kết quả thực nghiệm được tác giả thực hiện trên 2 bộ dữ liệu. Bộ dữ liệu Tiếng Anh được dùng để so sánh với 2 nghiên cứu cùng hướng gần nhất hiện nay qua đó đánh giá phương pháp được đề xuất. Đồng thời, bộ dữ liệu Tiếng Việt được tác giả áp dụng vào sàn giao dịch thành phố Hồ Chí Minh, so sánh với phương pháp SVM để chứng tỏ tính khả thi của đề tài khi áp dụng cho thị trường chứng khoán Việt Nam.

MỞ ĐẦU

Ngày nay, TTCK ngày càng có vai trò quan trọng trong nền kinh tế, là thước đo hiệu quả các hoạt động và sự phát triển kinh tế của một quốc gia. TTCK tạo điều kiện thuận lợi cho việc thực hiện chính sách mở cửa, cải cách kinh tế thông qua việc phát hành chứng khoán ra nước ngoài. Giá trị cổ phiếu của các công ty tỷ lệ thuận với lợi nhuận mà công ty đạt được. Chỉ số chung của TTCK phản ánh mức tăng trưởng kinh tế của quốc gia đó trong thời gian ngắn, trung và dài hạn. Đồng thời, TTCK tạo điều kiện để sử dụng vốn có hiệu quả hơn đối với cả người có tiền đầu tư và người vay tiền để đầu tư. Thông thường lãi thu được qua đầu tư chứng khoán cao hơn lãi phiếu nhà nước hay lãi gửi tiết kiệm.

Tuy chứng khoán là kênh đầu tư có khả năng sinh lợi cao nhưng chứng khoán cũng tiềm ẩn nhiều rủi ro. Nhiều nghiên cứu hiện nay trong lĩnh vực TTCK cố gắng dự đoán chính xác giá trị của giá cổ phiếu hoặc dự đoán xu hướng giá cổ phiếu trong tương lai. Tuy nhiên, điều này là rất khó bởi sự biến động phức tạp của chuỗi giá, vì giá cổ phiếu chịu tác động bởi rất nhiều yếu tố như tình hình chính trị, xã hội, kinh tế, tin tức của công ty, hiệu suất, báo cáo hoạt động kinh doanh, [10]… Tuy nhiên, sự biến động của TTCK không ngẫu nhiên [22] mà có khả năng dự báo được. Một mô hình dự đoán có hiệu quả là mô hình dự đoán chính xác xu hướng của một mã cổ phiếu tăng hoặc giảm trong tương lai, giúp nhà đầu tư đưa ra quyết định đầu tư đúng đắn trong việc mua, bán cổ phần của cổ phiếu mà họ đang nắm giữ nhằm thu lợi nhuận cao nhất và giảm thiểu rủi ro đến mức thấp nhất. Do đó, việc dự báo xu hướng vận động của thị trường tài chính và giá cổ phiếu luôn được nhiều nhà đầu tư quan tâm. Đây là một vấn đề có tính thực tiễn và khả năng mở rộng rất cao, đã và đang được các viện và nhóm nghiên cứu quan tâm. Cũng chính vì thế, tác giả thực hiện đề tài luận văn này với mong muốn có thể đóng góp được phần sức vào sự phát triển chung và hy vọng có thể hữu dụng khi áp dụng vào TTCK Việt Nam. Những nghiên cứu có thể hỗ trợ các nhà đầu tư tham khảo những kênh dựa trên căn cứ có khoa học để thúc đẩy sự phát triển của TTCK Việt Nam, cũng như sự ứng dụng của CNTT vào sự phát triển của nền kinh tế nước nhà.

Phạm vi và đối tượng đề tài

Theo học thuyết thị trường (Efficient Market Hypothesis)[23] về thị trường tài chính "Trong thị trường chứng khoán, giá chứng khoán phản ánh đầy đủ mọi thông tin đã biết". Do đó những nhà đầu tư chứng khoán giỏi là những người nắm được nhiều thông tin nhất (thông tin đã biết như thông tin tổng quát của công ty, tin tức trong nội bộ của công ty hay những hình thái biến động của giá cả trong quá khứ của giá cổ phiếu, …). Ngày nay, với sự phát triển của công nghệ và truyền thông, tin tức được lan truyền rộng và nhanh hơn bao giờ hết, thông qua các kênh truyền hình, mạng xã hội hay cụ thể là những trang tin tức. Các thông tin, sự kiện của nền kinh tế trong và ngoài nước, các đánh giá của chuyên gia, thông tin các công ty đều được công khai rộng rãi. Các sự kiện tích cực lẫn tiêu cực của thị trường tài chính đều có thể trực tiếp gây tác động tốt hoặc xấu đến thị trường chứng khoán. Chẳng hạn như, sự kiện "Brexit" việc Vương quốc Liên hiệp Anh và Bắc Ireland rời khỏi Liên minh châu Âu ảnh hưởng đến thị trường chứng khoán thế giới, giá vàng hay ngoại tệ [19]. Giá xăng tăng hoặc giảm mạnh cũng sẽ tác động nền kinh tế và các nhà đầu tư, họ có thể tăng cường mua/bán các cổ phiếu có liên quan đến các công ty hay lĩnh vực đó và kết quả là giá chứng khoán cũng sẽ bị ảnh hưởng. Việc phân tích các thông tin này càng nhanh là rất quan trọng để giúp các nhà đầu tư ra quyết định đối với cổ phiếu mình nắm giữ nhằm mang lại lợi nhuận cao và giảm thiểu tối đa rủi ro. Đây là một công việc rất khó thực hiện thủ công vì khối lượng và tốc độ tin tức được xuất bản mỗi ngày. Vì vậy rất cần thiết có một hệ thống hỗ trợ đưa ra quyết định tự động dựa vào tin tức tài chính. Do đó, một giải pháp có thể bổ sung khá hiệu quả để giải quyết vấn đề dự báo chứng khoán đó là xem xét các tác động của tin tức đối với biến động của thị trường chứng khoán[7], [25], [2].

DNN gần đây đang thu hút đông đảo sự chú ý của giới nghiên cứu về máy học, bởi vì những thành công của DNN trong nhiều lĩnh vực khác nhau đặc biệt là trong lĩnh vực xử lý ngôn ngữ tự nhiên [11]. Do đó, các nhà nghiên cứu đã áp dụng một số mô hình DNN để huấn luyện và học các đặc trưng từ các bản tin tài chính và lịch sử giá cổ phiếu như trong [7] và [25]. Nghiên cứu trước đây đã chứng minh hiệu quả của

các mạng DNN trong việc học các đặc trưng của các bản tin tức. Tuy nhiên, các đặc trưng này không nắm bắt được toàn diện mối quan hệ cấu trúc - thứ tự của các từ ngữ xuất hiện trong bài viết, đồng thời việc áp dụng lên các ngôn ngữ khác nhau là một thách thức lớn.

Tại Việt Nam, thị trường chứng khoán còn khá mới mẻ và sàn giao dịch lớn nhất của TP.HCM là sàn HoSE cũng mới được thành lập từ năm 2000, do đó việc dự đoán xu hướng chứng khoán sử dụng tin tức tài chính chưa được nhiều nhóm nghiên cứu so với thị trường ở các nước khác trên thế giới. Hơn nữa, vấn đề rào cản cho các nghiên cứu của thế giới áp dụng vào thị trường Việt Nam là ngôn ngữ, vì tiếng Việt có cấu trúc khác hoàn toàn với tiếng Anh [21], nên việc xử lý ngôn ngữ sẽ phức tạp hơn. Đó là những lý do và cũng chính là động lực để tác giả làm nghiên cứu này, mục tiêu nhằm đề xuất một mô hình dự đoán xu hướng chứng khoán cho thị trường Việt Nam, cụ thể là rổ chứng khoán VN-Index thông qua sử dụng tin tức tài chính và kết hợp thông tin dữ liệu lịch sử giá chứng khoán.

Để giải quyết các vấn đề còn tồn động, trong phạm vi đề tài luận văn, tác giả đã đặt ra những mục tiêu chính cụ thể như sau:

Mục tiêu đề tài

❖ *Về mặt khoa học:*
- Đề xuất mô hình mạng nơ-ron thích hợp cho mô hình dự báo dựa trên các nghiên cứu trước đó với dữ liệu đầu vào là các mô hình ngôn ngữ.
- Nghiên cứu áp dụng các kỹ thuật được áp dụng gần đây trong quá trình huấn luyện mạng nơ-ron đối với xử lý ngôn ngữ tự nhiên để tăng độ chính xác, tốc độ xử lý, khối lượng dữ liệu lớn, giảm số chiều văn bản và giảm thiểu các vấn đề trong quá trình huấn luyện như quá vừa dữ liệu (overfitting), …
- Một bài báo được công bố tại hội nghị quốc tế.

❖ *Về mặt thực tiễn:*
- Ứng dụng được mô hình đề xuất vào dự báo sự chuyển động của giá chứng khoán dựa trên các tin tức, sự kiện cho các mã cổ phiếu chung và riêng biệt.

- Áp dụng mô hình trên cở sở xử lý cả ngôn ngữ Tiếng Anh và Tiếng Việt, để ứng dụng cho TTCK trong và ngoài nước. Đồng thời so sánh và đánh giá mô hình với các đề tài tương tự nghiên cứu mới nhất hiện nay.

Bố cục luận văn

Nội dung của luận văn được chia thành 5 chương như sau:

Chương 1: TỔNG QUAN: Giới thiệu các hướng tiếp cận trong dự báo chứng khoán, mô tả bài báo dự đoán chuyển động giá chứng khoán dựa vào tin tức tài chính, khảo sát tình hình nghiên cứu liên quan sau đó đưa ra những vấn đề còn tồn tại cần giải quyết.

Chương 2: CƠ SỞ LÝ THUYẾT: Trình bày kiến thức tổng quan về mạng nơ-ron từ đó giới thiệu mô hình mạng nơ-ron hồi quy và biến thể GRU

Chương 3: MÔ HÌNH DỰ ĐOÁN XU HƯỚNG CHỨNG KHOÁN BẰNG MẠNG NƠ-RON DỰA TRÊN TIN TỨC TÀI CHÍNH: Trình bày mô hình đề xuất BGRU để giải quyết bài toán dự đoán xu hướng giá chứng khoán và quy trình thực hiện của mô hình.

Chương 4: THỰC NGHIỆM: Giới thiệu bộ dữ liệu thực nghiệm, phương pháp đánh giá, các cài đặt và kết quả thực nghiệm thu được, thông qua đó đưa ra các nhận xét và thảo luận về kết quả.

Chương 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN: Tổng kết những kết quả đạt được và trình bày hướng phát triển của đề tài trong tương lai.

Chương 1. TỔNG QUAN

Để hiểu rõ hơn về khóa luận, trong chương này tác giả sẽ mô tả chi tiết về bài toán dự báo chứng khoán dựa vào tin tức tài chính. Bên cạnh đó, tác giả sẽ đưa ra khảo sát các nghiên cứu trong và ngoài nước có liên quan, phân tích các vấn đề còn tồn tại của các nghiên cứu trước, từ đó định hướng những vấn đề cần giải quyết trong phạm vi luận văn.

1.1. Đặt vấn đề

1.1.1. Phát biểu bài toán

Dự báo xu hướng giá của chứng khoán dựa vào tin tức tài chính là việc xác định trong tương lai, ở một khoảng thời gian nhất định (ngắn, trung hoặc dài hạn), giá của chứng khoán sẽ chuyển động theo hướng *tăng* hay *giảm*. Xu hướng chuyển động giá chứng khoán được dự báo dựa trên phân tích ngữ nghĩa của các bản tin tài chính được đăng trong cùng thời gian. Đề tài nghiên cứu sử dụng phương pháp máy học giám sát đưa ra dự báo xu hướng giá nhằm hỗ trợ nhà đầu tư ra quyết định tối ưu để đạt được lợi nhuận cao và rủi ro thấp nhất.

Ở đây, tác giả không đề cập đến xu hướng *giữ nguyên* (tức là giá chứng khoán tại thời điểm mở cửa xấp xỉ bằng giá tại thời điểm đóng cửa) bởi vì 3 lí do. Thứ nhất, xu hướng *giữ nguyên* không mang lại giá trị nhận biết thời cơ hay rủi ro cho nhà đầu tư. Thứ hai, việc giá chứng khoán tại thời điểm mở cửa bằng giá lúc đóng cửa chiếm tỉ lệ rất nhỏ trên tập mẫu vì thế có thể làm giảm tỉ lệ chính xác khi tăng thêm một phân lớp khi dự báo. Cuối cùng, để hướng tiếp cận tương đồng với các nghiên cứu hiện tại, lấy cơ sở để so sánh và đánh giá.

1.1.2. Dữ liệu đầu vào

- Danh sách các bản tin tài chính
- Tập nhãn, trong đó nhãn 1 đại diện cho xu hướng giá *tăng*, nhãn 0 là xu hướng giá *giảm*.

1.1.3. Dữ liệu đầu ra

Các bản tin sẽ được gán nhãn 0 hoặc 1 tương ứng với kết quả dự đoán là xu hướng tăng hoặc giảm của giá chứng khoán trong cùng ngày với bản tin được phát hành của dữ liệu đầu vào.

1.2. Các nghiên cứu liên quan

1.2.1. Trong nước

Dự báo thị trường chứng khoán từ lâu đã thu hút nhiều nghiên cứu từ các nhà kinh tế học và các nhà khoa học máy tính. Gần đây, dự báo thị trường chứng khoán sử dụng khai phá văn bản là một lĩnh vực mới nổi và đã thu hút một số nghiên cứu trên thế giới nói chung và Việt Nam nói riêng. Đây là lĩnh vực liên ngành giữa ngôn ngữ học, học máy, và tài chính hành vi. Gần đây ở Việt Nam, nhóm tác giả Phạm Xuân Dũng và Hoàng Văn Kiếm [8] đã đề xuất mô hình kết hợp dữ liệu số và dữ liệu văn bản cùng với các luật về tài chính để nâng cao khả năng dự báo xu hướng của chỉ số giá chứng khoán VN-Index. Trong nghiên cứu, [8] đã đề xuất các bước tiền xử lý văn bản đối với Tiếng Việt đồng thời ứng dụng thuật toán SVM và mạng nơ-ron nhân tạo để rút trích các đặc trưng văn bản. Tuy nhiên, số lượng dữ liệu và kết quả thực nghiệm còn hạn chế. Cùng cách tiếp cận như trên, nhóm tác giả Đặng Liên Minh và Nguyễn Đức Toàn [9] đã cho thấy việc sử dụng tin tức tài chính có ảnh hưởng đến giá cổ phiếu tại Việt Nam rất khả quan. Thực nghiệm được triển khai bởi thuật toán máy học SVM kết hợp với phương pháp đánh trọng số từ TF-IDF trên sàn HoSE – nơi có chỉ số tài chính tốt và tính thanh khoản cao với độ chính xác là 73,66%. Đề tài đã đề xuất bộ dữ liệu thực nghiệm chuẩn được thu thập từ các website tin tức chứng khoán ở Việt Nam.

1.2.2. Ngoài nước

Từ nhiều năm nay, các nhà nghiên cứu trên thế giới có nhiều quan tâm trong việc ứng dụng các mô hình máy học vào dự báo chứng khoán như: thuật toán di truyền [18], Support Vector Machine [16], [17], Artificial Neural Network [18], [12] và

Random Forest [28] được sử dụng để dự đoán xu hướng chuyển động giá chứng khoán trên các dữ liệu giá theo chuỗi thời gian. Tuy nhiên hầu hết các giải pháp trên vẫn chưa đưa ra kết quả đầy đủ thỏa đáng với độ chính xác cao và hoạt động ổn định trên dự đoán cổ phiếu [1]. Sự hạn chế của việc áp dụng các kỹ thuật học máy của các nghiên cứu trước đây để dự đoán thị trường chứng khoán cho thấy rằng cần có thêm thông tin hữu ích hơn cho những dự đoán tốt hơn và cần các mô hình mạnh mẽ hơn để phù hợp với dữ liệu kết hợp phức tạp và với số chiều cao (high dimensional) [20].

Khoảng thập niên đầu của thế kỉ 21, các nhà nghiên cứu đã bắt đầu ứng dụng rộng rãi mạng nơ-ron vào việc dự báo chứng khoán. Ban đầu do sự thiếu hụt của dữ liệu huấn luyện, các mạng nơ-ron "nông" được triển khai với dữ liệu chuỗi thời gian (time series), dần sau mở rộng với nhiều loại dữ liệu như lịch sử giá, khối lượng thông tin giao dịch để dự đoán giá chứng khoán trong tương lai. Những năm gần đây, với sự thành công của các mô hình DNN và sự bùng nổ của các thông tin. Các nhà nghiên cứu đã bắt đầu ứng dụng các mạng DNN kết hợp một số thông tin bổ sung mà ảnh hưởng đến thị trường chứng khoán như các bản tin tài chính, tin tức[7], sentiment trên các mạng xã hội[27], micro blogs [4]…v.v. Trong số đó, [7], [25], [2] đã thu được một số kết quả đáng chú ý. Đại diện như nhóm nghiên cứu [7], họ đã đề xuất bộ dữ liệu Tiếng Anh với khối lượng lớn và xây dựng hệ thống để rút trích sự kiện về dạng E = (O1, P, O2) trong đó O1 thể hiện đối tượng thứ nhất, O2 thể hiện đối tượng thứ 2 (đối tượng ở đây có thể là mã cổ phiếu, tên công ty, tên nhân vật, …) và P thể hiện mối quan hệ giữa 2 đối tượng tạo thành sự kiện để biểu diễn cho một tin tức. [7] đã ứng dụng mạng mạng nơ-ron tiến (feedforward) để huấn luyện và thực nghiệm. Nhóm tác giả [25] đã dùng áp dụng word embedding để rút trích đặc trưng từ tin tức và triển khai mạng nơ-ron để dự báo chuyển động giá chứng khoán trong tương lai của chỉ số S&P500 [1] thị trường chứng khoán Hoa Kỳ. Gần đây nhất, [2] đã áp dụng mô hình Long Short Term Memmory (LSTM), một biến thể của Recurrent Neural Network (RNN) thu được các kết quả đáng khích lệ bước đầu chứng minh

[1] Standard & Poor 500 là chỉ số thị trường chứng khoán dựa trên thị trường vốn hóa của 500 công ty lớn có cổ phiếu phổ thông được niêm yết trên thị trường chứng khoán Hoa Kỳ.

được tiềm năng của việc ứng dụng các mô hình DNN vào trong việc dự báo chuyển động giá của chứng khoán.

1.2.3. Những vấn đề còn tồn tại

Thông qua những nghiên cứu trên, tác giả nhận thấy việc ứng dụng mạng nơ-ron ngày càng nhận được nhiều sự quan tâm từ phía các nhà nghiên cứu, đồng thời tin tức là một dữ liệu có căn cứ để tích hợp vào việc dự báo xu hướng của chứng khoán. Các nghiên cứu của các nhóm tác giả [7] và [25] đạt được những kết quả rất đáng mong đợi. Tuy nhiên, để có thể áp dụng những mô hình trên vào thị trường chứng khoán Việt Nam thì còn tồn tại những vấn đề cần được giải quyết như sau:

❖ *Về mặt khoa học:*

- Việc áp dụng các mạng nơ-ron chuẩn chưa thể khai thác được hết các đặc trưng của ngôn ngữ, thứ tự xuất hiện và ngữ nghĩa của từ. Ví dụ: giả sử ta xét sự kiện "Apple kiện Samsung". Mô hình nơ-ron chuẩn chỉ quan tâm đến đặc trưng của sự kiện, trong đó đánh đồng vai trò của Apple và Samsung. Tuy nhiên nếu xem xét thứ tự xuất hiện, và vị trí của hai chủ thể "Apple" và "Samsung" thì ý nghĩa hoàn toàn khác nhau. "Apple" xuất hiện với vai trò chủ động còn Samsung ở vai trò bị động sẽ có những tác động khác tới thị trường chứng khoán. Chính vì thế, trong khóa luận này, tác giả cố gắng để khắc phục hạn chế trên, đề xuất mô hình có khả năng học được đặc trưng trên toàn bộ ngữ cảnh của văn bản.
- Việc xử lý ngôn ngữ Tiếng Việt gặp nhiều khó khăn vì cấu trúc và cú pháp khác so với Tiếng Anh.

❖ *Về mặt thực tiễn:*

- Thị trường chứng khoán Việt Nam còn khá non trẻ. Việc nghiên cứu dự đoán giá chứng khoán vẫn chủ yếu tập trung vào phương pháp phân tích kỹ thuật, việc nghiên cứu theo hướng phân tích cơ bản vẫn chưa được khai thác rộng rãi.
- Việc tìm nguồn dữ liệu và tin tức từ các trang mạng ở Việt Nam gặp nhiều khó khăn. Nguồn tin tức chưa mang độ tin cậy cao.

Chương 2. CƠ SỞ LÝ THUYẾT

Để có thể hiểu rõ hơn về cơ sở khoa học của mô hình được đề xuất trong nghiên cứu này. Luận văn sẽ trình bày tổng quan về mạng nơ-ron, mô hình mạng nơ-ron hồi quy và biến thể Gated Recurrent Unit (GRU). Đây là những mô hình đang nhận được nhiều sự quan tâm của các nhà nghiên cứu trong việc áp dụng vào các mô hình máy học hiện nay. Đặc biệt, thế mạnh của các mô hình này trong việc huấn luyện và rút trích đặc trưng ngôn ngữ. Luận văn cũng phân tích để chỉ ra ưu thế và những vấn đề còn tồn tại trong các mô hình trên, từ đó làm cơ sở để đề xuất mô hình Bidirectional Gated Recurrent Unit sẽ được đề cập chi tiết trong chương 3.

2.1. Tổng quan về mạng nơ-ron (Neural Network)

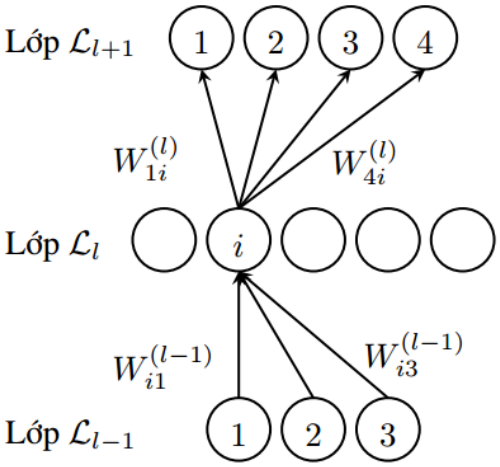
Mạng nơ-ron là một mô hình học máy có khả năng mô phỏng các hàm cực kỳ phức tạp, phi tuyến tính với một số lượng tham số vừa phải mà máy tính có khả năng tính toán ra được trong thời gian hợp lý. Dù đã ra đời từ khoảng 60 năm trước, thập niên 2006-2015 chứng kiến sự hồi sinh mạnh mẽ của mạng nơ-ron. Hiện nay, mô hình này được ứng dụng rộng rãi và đạt được nhiều kết quả tốt trong hầu như mọi lĩnh vực của trí tuệ nhân tạo, đặc biệt là trong xử lý ngôn ngữ tự [11].

Tuỳ vào ứng dụng cụ thể, mạng nơ-ron có thể mang các kiến trúc khác nhau, cho phép thông tin giữa các nơ-ron trong mạng được lan truyền theo nhiều phương pháp và định hướng thích hợp. Trong phần §2.1, tác giả giới thiệu tổng quan các kiến thức về mạng nơ-ron đầy đủ, sau đó sẽ trình bày tiếp mạng nơ-ron hồi quy, một mô hình mạng nơ-ron được đánh giá có nhiều ưu thế trong việc xử lý ngôn ngữ tự nhiên. Đây sẽ là những kiến thức nền tảng cho việc đề xuất mô hình dự báo xu hướng giá chứng khoán dựa trên tin tức tài chính.

2.1.1. Kiến trúc của mạng nơ-ron kết nối đầy đủ

Một mô hình mạng nơ-ron cơ bản thường bao gồm 3 lớp nơ-ron (layer) như **lớp dữ liệu vào** (input layer), **lớp ẩn** (hidden layer) và **lớp dữ liệu ra** (output layer). Một lớp thường bao gồm nhiều nơ-ron, tùy vào yêu cầu của mô hình mà số lớp ẩn có thể

là một hoặc nhiều lớp. Các nơ-ron giữa hai lớp liên tiếp được kết nối với nhau tạo thành một đồ thị hai phía đầy đủ với các cạnh có trọng số được biểu diễn bởi một ma trận trọng số. Có hai con đường lan truyền thông tin trong mạng nơ-ron kết nối đầy đủ. Trong bước lan truyền tới (feed-forwarding), thông tin được truyền từ lớp dữ liệu vào, qua các lớp ẩn rồi đến lớp dữ liệu ra. Lớp dữ liệu ra chính là kết quả của mạng, thể hiện giá trị của hàm mà mạng đang mô phỏng tại điểm dữ liệu nhận được ở lớp dữ liệu vào. Tất nhiên, mạng nơ-ron có thể cho kết quả không chính xác, tạo ra các lỗi sai lệch. Trong bước lan truyền ngược (back-propagation), các lỗi này sẽ được truyền qua các lớp của mạng theo trình tự ngược lại với bước lan truyền tới, cho phép mạng nơ-ron tính được đạo hàm theo các tham số của nó, từ đó điều chỉnh được các tham số này bằng một thuật toán tối ưu hàm số.



Hình 2.1. Minh hoạ cho kết nối giữa các lớp trong một mạng nơ-ron.

Như đã nói ở phần trên, các nơ-ron trong một mạng nơ-ron kết nối đầy đủ được phân chia thành nhiều lớp. Mỗi nơ-ron trong một lớp nhận giá trị trả ra từ các nơ-ron ở lớp liền trước, kết hợp các giá trị này thành một giá trị trung gian, và sau cùng truyền giá trị trung gian qua một hàm kích hoạt để trả về kết quả cho nơ-ron ở lớp tiếp theo.

Cụ thể hơn, xét một mạng nơ-ron gồm $\mathcal{L} - 1$ lớp ẩn. Ta sẽ ký hiệu $\mathcal{L}^{(l)}$ là tập hợp các lớp nơ-ron nằm trong lớp thứ l, với $l = 0, 1,...,L$. Lớp $\mathcal{L}^{(0)}$ là lớp ***dữ liệu vào***. Lớp $\mathcal{L}^{(L)}$ là lớp ***dữ liệu ra***. Các lớp còn lại được gọi là các ***lớp ẩn***. Nơ-ron trong lớp thứ l

chỉ nhận thông tin từ các nơ-ron thuộc lớp thứ $l - 1$ và chỉ truyền thông tin cho các nơ-ron thuộc lớp thứ $l + 1$. Tất nhiên, các nơ-ron thuộc lớp $\mathcal{L}^{(0)}$ không nhận dữ liệu vào từ các nơ-ron khác và các nơ-ron thuộc lớp $\mathcal{L}^{(L)}$ không truyền dữ liệu ra cho các nơ-ron khác. Hình 2.1 minh hoạ liên kết xung quanh một nơ-ron mẫu trong một mạng nơ-ron. Tác giả quy ước về ký hiệu: trọng số giữa nơ-ron i thuộc lớp $\mathcal{L}_{l+1}$ và nơ-ron j thuộc lớp $\mathcal{L}_l$ được ký hiệu là $W_{ij}^{(l)}$.

Giữa hai lớp liên tiếp $\mathcal{L}^l$ và $\mathcal{L}^{l+1}$ trong mạng kết nối đầy đủ, ta thiết lập một ma trận trọng số $W^{(l)}$ với kích thước là $|\mathcal{L}^{l+1}|\times|\mathcal{L}^l|$. Phần tử $W_{ij}^{(i)}$ của ma trận này thể hiện độ ảnh hưởng của nơ-ron j trong lớp l lên nơ-ron i trong lớp $l + 1$. Tập hợp các ma trận trọng số $W = \{W^{(0)}, W^{(1)}, \cdots, W^{(L-1)}\}$ được gọi là tập hợp các tham số của mạng nơ-ron. Việc xác định giá trị của tập tham số được biết đến như việc học (learn) hay huấn luyện (train) mạng nơ-ron.

2.1.2. Phương thức suy luận thông tin của mạng nơ-ron

Giả sử rằng một khi các tham số của một mạng nơ-ron được xác định, làm thế nào để sử dụng mạng nơ-ron này như một hàm số thông thường? Thuật toán *lan truyền tới* cho phép mạng nơ-ron nhận một điểm dữ liệu vào và tính toán điểm dữ liệu ra tương ứng. Hàm $f : \mathbb{R} \to \mathbb{R}$ là một hàm kích hoạt mà ta sẽ tìm hiểu ở ngay phần sau. Mã giả thuật toán lan truyền tới được mô tả dưới đây:

```
FEED_FORWARD Algorithm
1. Function FEED_FORWARD(x^(0) ∈ R^|L_0|)
2.   for l = 1 to L do
3.       z^(l) ← W^(l−1).x^(l−1)
4.       x^(l) ← f(z^(l))
5.   end for
6.   return x^(L), Loss(z^(L))
7. end function
```

Ngoài giá trị của hàm số được mô phỏng, $x^{(L)}$, thuật toán lan truyền tới còn trả về giá trị của hàm mất mát (Loss), thể hiện độ tốt của tập tham số hiện tại.

2.1.3. Hàm kích hoạt

Hàm $f(z^{(l)})$ trong thuật toán 1 được gọi là hàm kích hoạt. Hàm kích hoạt có vai trò vô cùng quan trọng đối với mạng nơ-ron. Trên thực tế, những tiến bộ gần đây nhất trong các nghiên cứu về mạng nơ-ron chính là những công thức mới cho *f*, giúp tăng khả năng mô phỏng của mạng nơ-ron cũng như đơn giản hoá quá trình huấn luyện mạng. Hàm kích hoạt được sử dụng để loại bỏ khả năng tuyến tính hoá của mạng nơ-ron. Để biểu diễn được nhiều hàm số hơn, ta phải phi tuyến hoá mạng nơ-ron bằng cách đưa kết quả của mỗi phép nhân ma trận vec-tơ $W^{(l-1)} \cdot x^{(l-1)}$ qua một hàm không tuyến tính *f*. Một số hàm kích hoạt thường được sử dụng là:

- Hàm *sigmoid*: $f(x) = sigm(x) = \frac{1}{1+exp(-x)}$;
- Hàm *tanh*: $f(x) = \tanh(x)$;
- Hàm *đơn vị tuyến tính đứng (rectified linear unit – ReLU)*: $f(x) = \max(0, x)$;
- Hàm *đơn vị tuyến tính đứng có mất mát (leaky rectified linear unit – leaky ReLU)*: $f(x) = \begin{cases} x \text{ nếu } x > 0 \\ kx \text{ nếu } x \leq 0 \end{cases}$, với *k* là một hằng số chọn trước. Thông thường $k \approx 0.01$;
- Hàm *maxout*: $f(x_1, \dots, x_n) = \max\limits_{1 \leq i \leq n} x_i$;

2.1.4. Mô phỏng hàm xác suất và hàm phân loại

Mạng nơ-ron được ứng dụng rộng rãi để giải các bài toán phân loại, tức là xác định xem dữ liệu vào thuộc loại gì trong một tập các lựa chọn cho trước. Để giải bài toán này, ta dùng mạng nơ-ron để mô phỏng một phân bố xác suất trên tập các lựa chọn. Ví dụ ta muốn dùng mạng nơ-ron để giải bài toán xác nhận gương mặt (face verification). Tập các lựa chọn chỉ gồm hai phần tử: với một cặp ảnh chân dung bất kì, ta yêu cầu mạng nơ-ron trả lời "có" hoặc "không" cho câu hỏi rằng hai bức ảnh đó có phải cùng một người hay không. Mạng nơ-ron đưa ra câu trả lời dựa vào việc tính toán xác suất xảy ra của từng đáp án rồi chọn câu trả lời có xác suất cao hơn. Trong trường hợp này, giả sử rằng tổng xác suất của hai đáp án là 1, vậy thì ta chỉ cần tính xác suất cho một đáp án và suy ra xác suất của đáp án còn lại. Một mạng nơ-ron

sử dụng hàm sigmoid kích hoạt ở lớp cuối rất phù hợp để làm điều này, vì hàm sigmoid nhận vào một số thực trong khoảng $(-\infty, +\infty)$ và trả về một số thực trong khoảng (0,1). Tổng quát hơn, khi tập phương án lựa chọn có nhiều hơn hai phần tử, ta cần biến mạng nơ-ron thành một phân bố xác suất P(x) thỏa mãn hai điều kiện sau:

1. $P(x) \geq 0 \ \forall x \in \Omega$ (Ω là tập lựa chọn);

2. $\sum x \ P(x) = 1$.

Xét vec-tơ trước khi kích hoạt ở lớp cuối, $z^{(L)} = (z_0^{(L)}, z_1^{(L)}, .., z_{|L|-1}^{(L)})$. Thay vì sử dụng hàm *sigmoid*, ta dùng hàm phân lớp (softmax) để đưa vec-tơ này thành một phân bố xác suất. Hàm *softmax* có dạng như sau:

$$softmax(z^{(l)}) = (p_0, p_1, \dots, p_{|\mathcal{L}_L|-1}) \tag{2.1.1}$$

trong đó:

$$p_i = \frac{\exp(z_i^{(L)})}{\sum_{j=0}^{|\mathcal{L}_L|-1} exp(z_i^{(L)})} \tag{2.1.2}$$

với $exp(.)$ là hàm lũy thừa theo cơ số tự nhiên e và $0 \leq i \leq |\mathcal{L}_L| - 1$. Lưu ý là số lượng nơ-ron ở lớp cuối, $|\mathcal{L}_L|$, phải bằng với số các phương án lựa chọn. Dễ thấy là kết quả của hàm softmax thỏa mãn hai điều kiện của một phân bố xác suất và hàm *sigmoid* là một trường hợp đặc biệt của hàm *softmax*.

2.1.5. Phương pháp ước lượng tham số của mạng nơ-ron

Khi suy luận thông tin trên mạng nơ-ron, ta giả sử rằng các tham số (các ma trận $W^{(l)}$) đều được cho sẵn. Điều này dĩ nhiên là không thực tế; ta cần phải đi tìm các giá trị của tham số sao cho mạng nơ-ron suy luận càng chính xác càng tốt. Như đã nói ở trên, công việc này được gọi là *ước lượng tham số*, còn được biết đến như quá trình *huấn luyện* hay *học* của mạng nơ-ron.

Ta gọi *h(x; W)* và *g(x)* lần lượt là hàm biểu diễn bởi mạng nơ-ron (với tập tham số *W*) và hàm mục tiêu cần mô phỏng. Việc tìm ra công thức để tính ngay ra giá trị

của tập số tham số rất khó khăn. Ta chọn một cách tiếp cận khác, giảm thiểu dần khoảng cách giữa *h(x; W)* và *g(x)* bằng cách lặp lại hai bước sau:

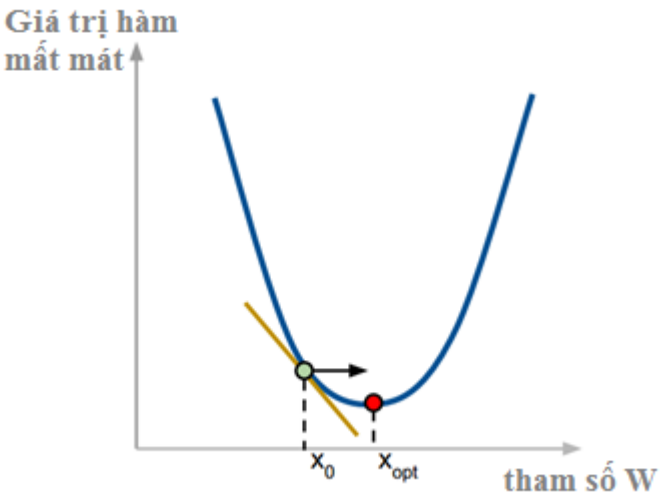
1. Đo độ sai lệch của suy luận của mạng nơ-ron trên một tập điểm dữ liệu mẫu *{(x_d, $g(x_d)$)}*, gọi là tập huấn luyện (training set).
2. Cập nhật tham số của mạng *W* để giảm thiểu độ sai lệch trên.

2.1.6. Hàm mất mát

Tổng của các độ sai lệch giữa dữ liệu ra của mạng nơ-ron, *h (x_d; W)*, và dữ liệu ra cần đạt được, *g(x_d)*, thể hiện độ tốt của tập tham số hiện tại. Nếu tập huấn luyện là cố định, tổng này về bản chất là một hàm số chỉ phụ thuộc vào tập tham số *W*, thường được biết đến với cái tên *hàm mất mát*:

$$Loss(W) = \sum_{d \in D} dist(h(x_d; W), g(x_d)) \qquad (2.1.3)$$

với *D* là tập huấn luyện, *dist* là một hàm tính độ chênh lệch giữa hai điểm dữ liệu ra.



Hình 2.2. Ví dụ minh họa cho việc tối ưu một hàm số.

Trong hình 2.2 đường thẳng màu vàng là đạo hàm tại điểm x_0. Mũi tên chỉ hướng x_0 cần được dịch chuyển để đến gần hơn với x_{opt}. Có nhiều cách định nghĩa độ chệnh lệch khác nhau. Người ta thường chọn những hàm liên tục, có đạo hàm ở (gần như) mọi nơi, và dễ tính để tính độ chênh lệch. Tất nhiên, với một mạng nơ-ron tối ưu, giá

trị của hàm mất mát sẽ bằng không. Trong thực tế, ta muốn tìm ra giá trị của tham số để giá trị hàm mất mát càng nhỏ càng tốt. Vì thế, bài toán ước lượng tham số của mạng nơ-ron về bản chất chính là bài toán tìm giá trị của biến *W* để cực tiểu hóa hàm số *Loss(W).* Tiếp theo, ta cần một thuật toán để có thể cực tiểu hóa hàm mất mát, thuật toán thường được sử dụng là *lan truyền ngược* (back propagation). Giải thuật lan truyền ngược có đoạn mã giả trình bày như sau:

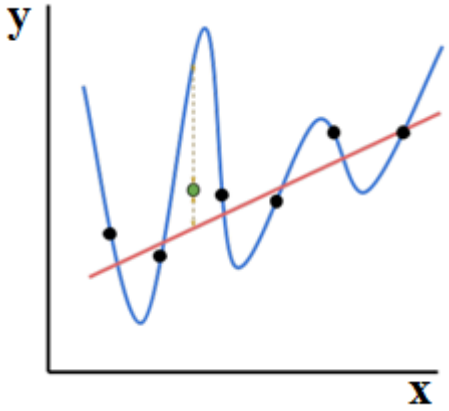
BACK PROPAGATION Algorithm

1. Function BACKPROP($x^{(0)} \in \mathbb{R}^{|\mathcal{L}_0|}, \{W^{(l)}\}$)
2. Áp dụng thuật toán lan truyền tới [2.1.2] với $x^{(0)}$ để tính các giá trị $z^{(1)}, \ldots, z^{(L)}$, $x^{(1)}, \ldots, x^{(L)}$ và hàm mất mát Loss($z^{(L)}$).
3. $\delta^{(L)} \leftarrow \frac{\partial}{\partial z^{(L)}} Loss$
4. for $l = L - 1 \; to \; 0 \; do$
5. $\frac{\partial}{\partial z^{(1)}} Loss \leftarrow f'(z^{(l)} \circ (W^{(l)\intercal}.\delta^{(l+1)})$
6. $\frac{\partial}{\partial W^{(1)}} Loss \leftarrow \delta^{(l+1)}.x^{(l)\intercal}$
7. end for
8. return $\{\frac{\partial}{\partial W^{(0)}} Loss, \ldots, \frac{\partial}{\partial W^{(L-1)}} Loss\}$
9. end function

2.1.7. Vấn đề Overfitting

Trong ứng dụng thực tế, ta thường sử dụng mạng nơ-ron để mô phỏng những hàm số mà cấu trúc của chúng vẫn chưa được xác định. Khi đó, ta chỉ có thể thu nhập được các bộ mẫu dữ liệu ra (vào) được sinh ra từ hàm số, nhưng lại không thể đặc tả quá trình sinh ra các bộ mẫu đó. Một ví dụ kinh điển đó là quá trình bộ não con người thu nhận thông tin từ hình ảnh của chữ viết tay, rồi suy luận ra chữ viết. Cơ chế bộ não biểu diễn hình ảnh và suy luận ra thông tin từ đó là một ẩn số đối với khoa học. Tuy nhiên, ta có thể dùng các bức ảnh cùng với nhãn đúng của chúng để huấn luyện mạng nơ-ron mô phỏng xấp xỉ được quá trình xử lý hình ảnh của bộ não. Cho dù cấu trúc

giữa bộ não và mạng nơ-ron khác nhau, với một thuật toán huấn luyện tốt, chúng sẽ đưa ra kết luận giống nhau với cùng một điểm dữ liệu vào.



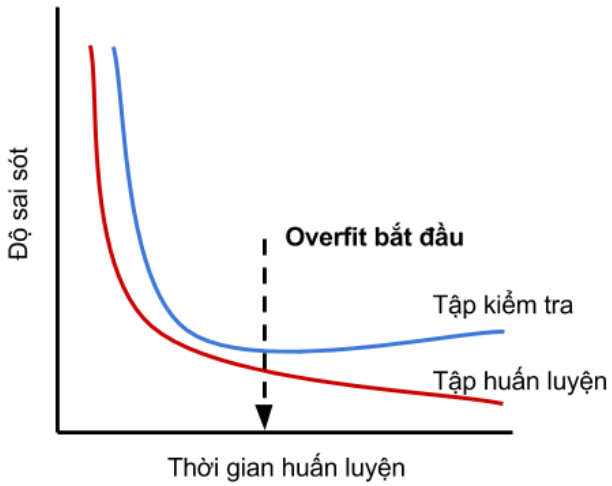
Hình 2.3. Một ví dụ về overfitting.

Đối với bài toán dự đoán, vì mục tiêu cuối cùng của ta là mô phỏng một hàm số ẩn, ta không nên cực tiểu hóa hàm mất mát trên tập huấn luyện. Nếu ta làm như vậy sẽ dẫn đến hiện tượng overfitting, tức là mạng nơ-ron sẽ học được một hàm phức tạp để mô phỏng hoàn hảo nhất tập huấn luyện. Tuy nhiên, cũng do cấu trúc phức tạp, hàm này không có tính tổng quát hóa cao, tức là nó rất dễ sai khi gặp một điểm dữ liệu không có trong tập huấn luyện. Theo ví dụ hình 2.3 thì đa thức có bậc cao hơn (xanh dương) vì quá chú trọng vào việc phải đi qua tất cả các điểm trong tập huấn luyện (đen) nên có hình dạng phức tạp, không "bình thường". Đa thức bậc thấp hơn (đỏ) cho giá trị hàm mất mát cao hơn trên tập huấn luyện nhưng lại phù hợp hơn với phân bố dữ liệu trong thực tế. Điều này thể hiện bằng việc đa thức bậc thấp ước lượng một điểm không có trong tập huấn luyện (xanh) chính xác hơn đa thức bậc cao. Overfitting là một vấn đề nghiêm trọng đối với mạng nơ-ron vì khả năng mô hình hóa của chúng quá cao, dễ dàng học được các hàm phức tạp. Khi ấy, mạng nơ-ron giống như một con người chỉ biết học tủ mà không biết cách vận dụng kiến thức để giải quyết những thứ chưa từng gặp phải.

Nếu ta áp dụng một phương pháp tối ưu hàm số hiệu quả, sai sót trên tập huấn luyện giảm theo thời gian. Tuy nhiên, sai sót trên tập kiểm tra không phải lúc nào cũng giảm. Nếu mô hình bị overfitting, đến một lúc nào đó, sai sót này sẽ bắt đầu

tăng trở lại. Thời điểm mà sai sót trên tập kiểm tra bắt đầu có xu hướng tăng được xem là thời điểm bắt đầu overfitting. Hình 2.4 thể hiện dấu hiệu nhận biết overfitting xảy ra khi mô hình dự đoán đạt được trên tập huấn luyện có độ lỗi nhỏ nhưng khi áp dụng lên tập dữ liệu test (dữ liệu mà mô hình chưa nhìn thấy) thì lại cho độ lỗi rất lớn, nên độ chính xác chung của mô hình bị giảm xuống.

Mục tiêu của các mô hình dự đoán đó là dự đoán chính xác những mẫu dữ liệu chưa nhìn thấy trong tương lai. Nếu ta có thể dự đoán chính xác trên dữ liệu thu thập được nhưng lại không thể dự đoán chính xác những dữ liệu trong tương lai thì nhìn chung mô hình của ta không đạt yêu cầu. Do vậy, trong phạm vi đề tài tác giả đã áp dụng một số kỹ thuật như là dropout trong quá trình huấn luyện mạng nơ-ron để tránh vấn đề overfitting.



Hình 2.4. Minh hoạ "learning curve" khi xuất hiện overfitting

2.2. Mạng Nơ-ron hồi quy

Những nghiên cứu trước đó đã chứng minh hiệu quả của mạng nơ-ron trong lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và dự báo giá chứng khoán nói riêng [11], [12], [18]. Tuy nhiên, mạng nơ-ron thông thường vẫn còn tồn một số yếu điểm trong việc nắm bắt toàn bộ đặc trưng của một văn bản. Mạng nơ-ron đầy đủ có thể học ra

các đặc trưng của một văn bản, tuy nhiên thứ tự xuất hiện của các từ và mối quan hệ ngữ nghĩa chưa được học qua quá trình huấn luyện, vấn đề này được chỉ ra bởi [24].

Ví dụ khi xét ngữ cảnh của một sự kiện "Microsoft kiện Apple vì vi phạm bản quyền". Nếu việc huấn luyện mạng nơ-ron chỉ quan tâm đến các đặc trưng là "Microsoft", "kiện", "Apple" thì rất khó để dự đoán chính xác sự chuyển động giá của các công ty Microsoft và Apple bởi vì các đặc trưng không chỉ ra được công ty kiện và công ty bị kiện, bởi lẽ việc ngữ nghĩa trong ngữ cảnh được quyết định bởi thứ tự xuất hiện của từ có vai trò khác nhau. Theo tác giả nhận định, việc xác định vai trò ngữ nghĩa của từ, đối tượng cụ thể trong các bản tin tài chính sẽ có những tác động đến các nhà đầu tư chứng khoán. Chính vì vậy, để giải quyết hạn chế trên của mạng nơ-ron thông thường trong mô hình học dữ liệu ngôn ngữ tự nhiên, mô hình mạng nơ-ron hồi quy (Recurrent Neural Network - RNN) được cho là có khả năng giải quyết vấn đề này được khảo sát bởi [24]. Phần tiếp theo của chương này, luận văn sẽ trình bày chi tiết về mạng nơ-ron hồi quy để hiểu rõ hơn những ưu điểm của mô hình trong việc xử lý ngôn ngữ tự nhiên.

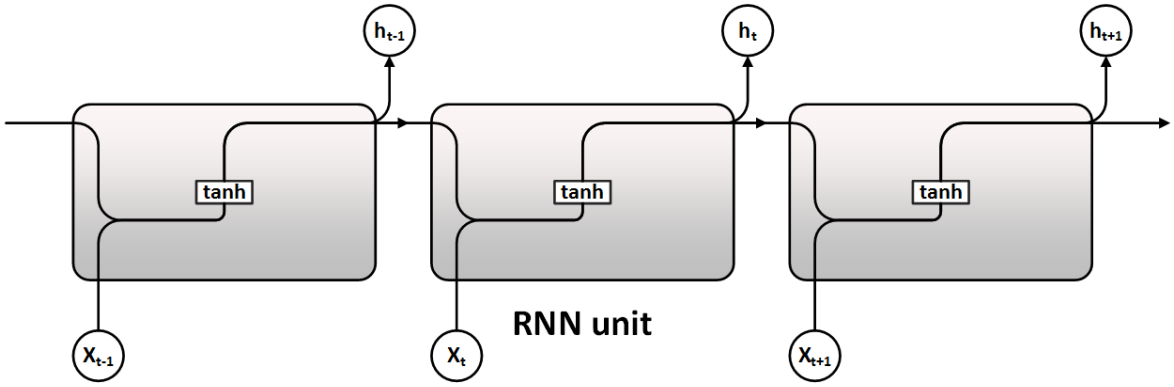
Mạng nơ-ron hồi quy (Recurent Neural Network) là một trong những mô hình DNN được đánh giá có nhiều ưu điểm trong các tác vụ xử lý ngôn ngữ tự nhiên [24]. Ý tưởng của RNN có thế mạnh xử lý thông tin dạng tuần tự (sequential information), ví dụ một câu là một chuỗi gồm nhiều từ. Recurrent có nghĩa là thực hiện lặp lại (hồi quy) cùng một tác vụ cho mỗi thành phần trong chuỗi. Trong đó, kết quả đầu ra tại thời điểm hiện tại bị ảnh hưởng bởi kết quả tính toán của các thành phần ở những thời điểm trước đó. Nói cách khác, RNN là một mô hình có bộ nhớ (memory), có khả năng lưu trữ các thông tin đã tính toán trước đó.

Không như các mô hình nơ-ron truyền thống đó là thông tin đầu vào (input) hoàn toàn độc lập với thông tin đầu ra (output). RNN nhận một chuỗi các từ đã được chuyển thành vec-tơ (x_1, x_2, ..., x_n) là đầu vào và trả ra một chuỗi vec-tơ (h_1, h_2, ..., h_n) đại diện cho thông tin tương ứng của mỗi thời điểm đầu vào.

Thông thường hàm kích hoạt của trạng thái ẩn ht sẽ biểu diễn bằng công thức:

$$ht = g(Wx_t + Uh_{t-1} + b) \quad (2.1.1)$$

Trong đó g thường là một hàm sigmoid hoặc hàm tanh. Tại mỗi thời điểm t, trạng thái của lớp ẩn h_t được tính bởi đầu vào x_t tại thời điểm đó và trạng thái của lớp ẩn trước h_{t-1}. Mô hình của RNN được minh họa qua hình 2.5.



Hình 2.5. Minh họa mô hình mạng nơ-ron hồi quy với hàm *tanh*

2.3. Vấn đề nắm bắt những thông tin dài hạn (Long-Term Memmory)

Như trình bày ở trên, RNN là mô hình có nhiều ưu điểm trong xử lý ngôn ngữ tự nhiên. Tuy nhiên, một vấn đề mà RNN được đưa ra thảo luận bởi [14], họ đã chỉ ra những khó khăn trong quá trình huấn luyện RNN và việc nắm bắt những thông tin dài hạn. Về lý thuyết, RNN có thể nhớ được thông tin của chuỗi có chiều dài bất kì, nhưng trong thực tế thực nghiệm mô hình này chỉ nhớ được thông tin ở vài bước trước đó bởi vấn đề "vanishing gradient"[14]. Ta thử cho một ví dụ về dự đoán từ tiếp theo trong câu:

"Tác giả sinh ra và lớn lên ở Việt Nam [...] vì thế tác giả có thể nói lưu loát tiếng [?]".

Trong tình huống này, RNN học các đặc trưng và dự đoán được [?] sẽ là một *loại ngôn ngữ*, tuy nhiên để dự đoán được chính xác ngôn ngữ nào thì cần phải xét đến ngữ cảnh "Việt Nam" để dự đoán ngôn ngữ cần dự đoán là tiếng Việt. Thông thường theo thực nghiệm RNN chỉ có thể nhớ những trạng thái của khoảng 5 bước tại các thời điểm trước đó, nếu như ở ví dụ trên bên trong […] là rất nhiều từ và ngữ cảnh

khác thì RNN sẽ bị chi phối bởi trọng số của những từ gần với [?] và khó có thể bắt được ngữ cảnh dài hạn "Việt Nam".

Để khắc phục vấn đề nắm bắt các thông tin dài hạn của ngữ cảnh. Trong phần tiếp theo của chương này, luận văn sẽ trình bày một biến thể của RNN là Gated Recurrent Unit (GRU). Đây là mô hình mạng nơ-ron rất mới dựa trên ý tưởng của RNN có bộ nhớ dài hạn.

2.4. Mạng Gated Recurrent Unit (GRU)

Mô hình GRU được đề xuất bởi Kyunghyun Cho năm 2014 [5]. Ở bước đầu tiên, GRU thực hiện tính *cổng* z_t dựa trên dữ liệu đầu vào tại thời điểm hiện tại x_t và đầu ra của trạng thái trước đó h_{t-1}. Về ý tưởng, tại bước này *cổng* z_t sẽ quyết định bao nhiêu bộ nhớ của các thời điểm trước đó được giữ lại.

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1} + b^{(z)}) \quad (2.3.1)$$

Ở bước tiếp theo, GRU sẽ tính *cổng* r_t, giống như *cổng* z_t nhưng khác về trọng số *W*. Tại bước này, cổng r_t sẽ xác định bao nhiêu giá trị mới sẽ kết hợp bộ nhớ của các thời điểm trước đó.

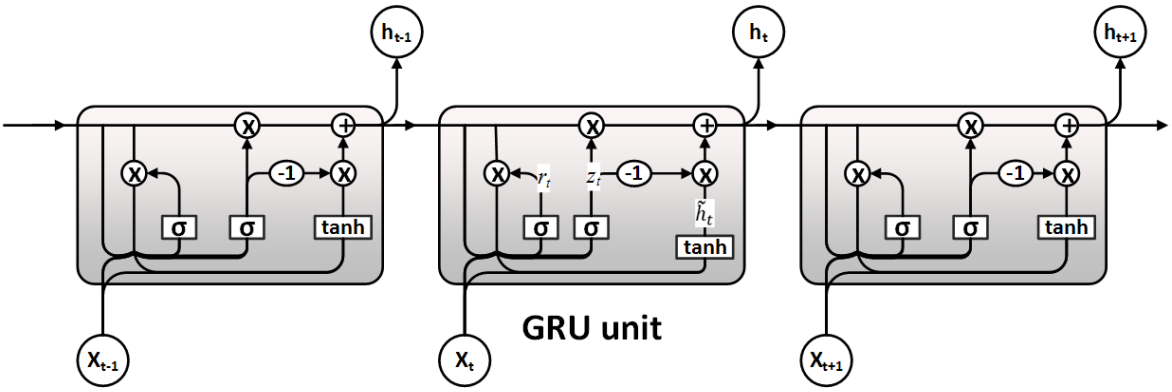
$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1} + b^{(r)}) \quad (2.3.2)$$

Ứng viên của hàm activation sẽ được tính:

$$\tilde{h}_t = tanh(Wx_t + r_t \odot Uh_{t-1} + b^{(h)}) \quad (2.3.2)$$

Tại bước cuối cùng, bộ nhớ tại thời điểm hiện tại sẽ được tính như sau:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (2.3.3)$$



Hình 2.6. Minh họa mô hình GRU

Dễ thấy, nếu cổng r_t của biểu thức 2.3.3 có giá trị gần bằng 0 thì cho phép mô hình loại bỏ các thông tin không liên quan. GRU có khả năng nắm bắt các thông tin dài hạn khi biểu thức 2.3.1 giá trị của cổng z_t gần bằng 1, đồng nghĩa với việc toàn bộ những thông tin được lưu trữ ở các bước trước đó được giữ lại ở bước hiện tại. Khi đó, GRU có thể lưu trữ thông tin qua nhiều bước và ngăn ngừa vấn đề vanishing gradient. Kiến trúc của mô hình GRU được minh họa qua hình 2.6.

Như vậy, có thể thấy rằng GRU có thể cải thiện khả năng lưu trữ các thông tin dài hạn so với RNN, tuy vậy trong mô hình ngôn ngữ, tác giả nhận thấy rằng mô hình GRU chỉ quan tâm đến những ngữ cảnh ở bên trái từ được xét. Tức là, khi đang xét tại từ thứ *t* trong câu thì bộ nhớ sẽ chỉ xét ngữ cảnh từ thứ 1 đến từ thứ *t-1* của mô hình. Vấn đề đặt ra là làm sao để học được những ngữ cảnh bên phải của của từ thứ *t.* Chính vì thế có thể nói rằng dữ liệu được học chưa thật sự đầy đủ bởi lẽ trong mô hình ngôn ngữ, một từ có thể bị ảnh hưởng bởi không những từ những từ bên trái mà còn bởi những từ bên phải của từ đang được xét. Nhằm cải thiện khả năng của học các đặc trưng từ mô hình ngôn ngữ của GRU, tác giả đã đề xuất một mô hình BGRU để khắc phục vấn đề này.

Chương 3. MÔ HÌNH DỰ ĐOÁN XU HƯỚNG GIÁ CHỨNG KHOÁN BẰNG MẠNG NƠ-RON DỰA TRÊN TIN TỨC TÀI CHÍNH

3.1. Đề xuất mô hình mạng Gated Recurrent Unit hai chiều

Như đã đề cập ở phần trên, nhằm cải thiện khả năng nắm bắt được toàn bộ đặc trưng của văn bản. Một từ khi được đưa vào mô hình sẽ biểu diễn đầy đủ tác động của những từ bên trái và những từ bên phải. Trong phần này, tác giả sẽ trình bày chi tiết mô hình mạng nơ-ron đề xuất mạng Bidirectional Gated Recurrent Unit (mạng GRU hai chiều).

Dựa trên ý tưởng từ bài báo [26] và [3] về mô hình RNN hai chiều, tác giả đề xuất mô hình BGRU với cải tiến đơn giản kết hợp thêm một chiều ngược lại từ phải sang trái đồng thời cùng chiều trái sang phải để học những đặc trưng bên phải của từ đang được xét. Có nghĩa là dữ liệu đầu vào của mạng GRU thông thường lần lượt là các từ từ trái sang phải của một câu, cùng lúc đó sẽ có thêm một chiều GRU từ phải sang trái được chạy song song và độc lập. Sau đó, kết quả giá trị của từ được xét là phép nối của kết quả đầu ra từ hai chiều GRU của từ đó. Mô hình BGRU được biểu diễn bởi các biểu thức bên dưới:

GRU tiến (từ trái sang phải):

$$\overrightarrow{z_t} = \sigma(\overrightarrow{W^{(z)}}x_t + \overrightarrow{U^{(z)}}h_{t-1} + \overrightarrow{b^{(z)}}) \quad (3.1.1)$$

$$\overrightarrow{r_t} = \sigma(\overrightarrow{W^{(r)}}x_t + \overrightarrow{U^{(r)}}h_{t-1} + \overrightarrow{b^{(r)}}) \quad (3.1.2)$$

$$\overrightarrow{\tilde{h}_t} = tanh(\overrightarrow{W}x_t + r_t \odot \overrightarrow{U}h_{t-1} + \overrightarrow{b^{(h)}}) \quad (3.1.3)$$

$$\overrightarrow{h_t} = \overrightarrow{z_t} \odot h_{t-1} + (1 - \overrightarrow{z_t}) \odot \overrightarrow{\tilde{h}_t} \quad (3.1.4)$$

GRU lùi (từ phải sang trái):

$$\overleftarrow{z_t} = \sigma(\overleftarrow{W^{(z)}}x_t + \overleftarrow{U^{(z)}}h_{t-1} + \overleftarrow{b^{(z)}}) \quad (3.1.5)$$

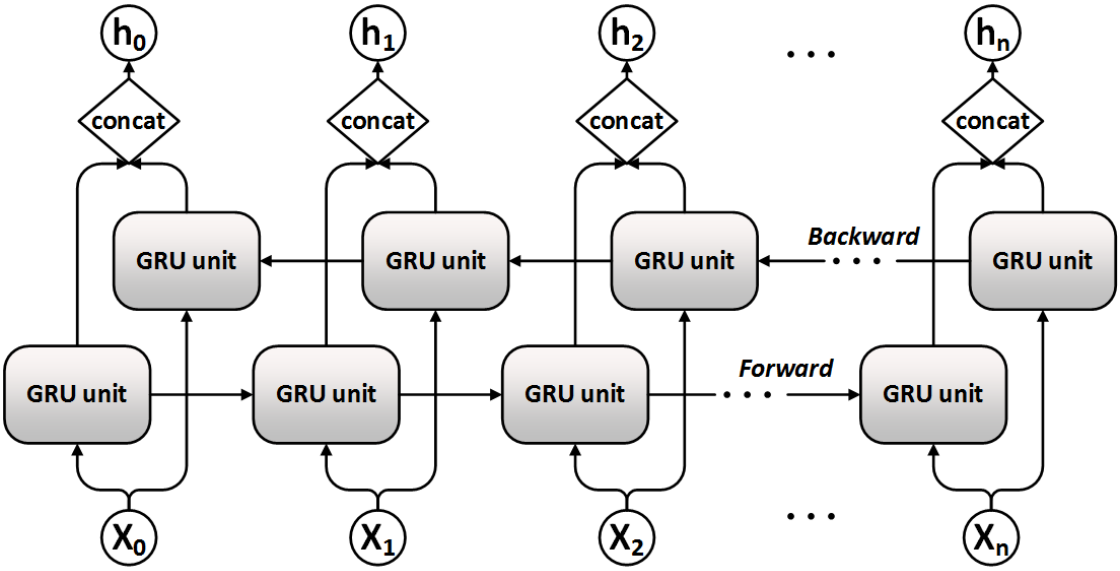
$$\overleftarrow{r_t} = \sigma(\overleftarrow{W^{(r)}}x_t + \overleftarrow{U^{(r)}}h_{t-1} + \overleftarrow{b^{(r)}}) \quad (3.1.6)$$

$$\overleftarrow{\tilde{h}_t} = tanh(\overleftarrow{W}x_t + r_t \odot \overleftarrow{U}h_{t-1} + \overleftarrow{b^{(h)}}) \quad (3.1.7)$$

$$\overleftarrow{h_t} = \overleftarrow{z_t} \odot h_{t-1} + (1 - \overleftarrow{z_t}) \odot \overleftarrow{\tilde{h}_t} \quad (3.1.8)$$

Kết quả $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$

Trong đó, một câu (x_1, x_2, …, x_n) có n từ, xét từ ở vị trí thứ t, *GRU tiến* sẽ biểu diễn sự ảnh hưởng ngữ cảnh bên trái $\vec{h}_t$ của từ thứ t và *GRU lùi* sẽ biểu diễn ngữ cảnh bên phải $\overleftarrow{h}_t$ của từ thứ t. BGRU sẽ biểu diễn từ thứ t bằng phép nối (concat) $h_t = [\vec{h}_t, \overleftarrow{h}_t]$ kết quả GRU tiến và GRU lùi. Hình 3.1 sẽ minh họa mô hình BGRU



Hình 3.1. Minh họa mô hình BGRU

Đánh giá độ phức tạp thuật toán BGRU:

Thông thường các mô hình DL được đánh giá độ phức tạp bằng O(W) trong đó W là số lượng tham số cần ước lượng của mỗi mô hình [31]. Để xác định được W của các mô hình, hai tham số được dụng để tính là số chiều của vector đầu vào **m-dimensional** và số chiều của lớp ẩn **n-dimensional**. Bảng 3.1 thể hiện số lượng tham số của các mô hình GRU, LSTM và BGRU.

Bảng 3.1. So sánh số lượng tham số cần ước lượng của các mô hình DL

Mô hình	**Số lượng tham số cần ước lượng**
GRU	$3\times(n^2 + nm + n)$.
LSTM	$4\times(n^2 + nm + n)$.
BGRU	$\mathbf{6\times(n^2 + nm + n)}$.

Trong đó độ phức tập của mô hình BGRU bằng 2 lần độ phức tập của GRU vì số lượng tham số cần ước lượng hơn 2 lần (2 chiều GRU). Chính vì vậy với số lượng dữ

liệu lớn, mô hình BGRU cần nhiều chi phi và thời gian tính toán hơn so với GRU và LSTM.

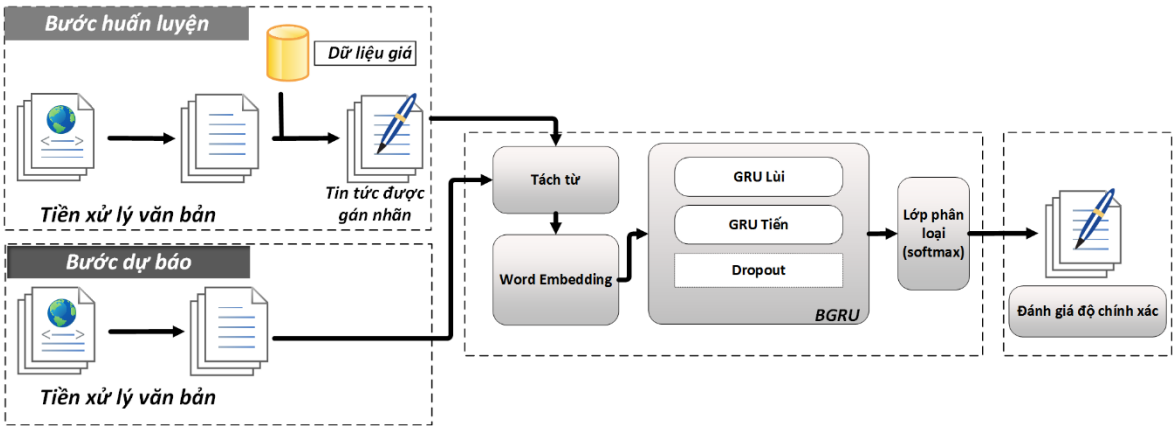
3.2. Mô hình dự báo

Luận văn đã áp dụng mạng nơ-ron BGRU vào mô hình dự báo chuyển động giá chứng khoán dựa vào tin tức tài chính thể hiện qua hình 3.2 bao gồm các bước sau:

Tiền xử lý văn bản: Tin tức sau khi được thu thập sẽ được trích xuất các nội dung chính và lưu vào tập tin dưới dạng text. Chi tiết xử lý được mô tả chi tiết ở mục 3.2.1 của chương này.

Gán nhãn văn bản: So sánh giá chứng khoán "mở cửa" và "đóng cửa" của mã được chọn trong từng khoảng thời gian tương ứng để phân làm 2 lớp tương ứng "tăng" (được gán nhãn **1** nếu giá đóng cửa lớn hơn giá mở cửa) và "giảm" (gán nhãn **0** nếu giá đóng cửa bé hơn hoặc bằng giá mở cửa) để gán nhãn cho từng tập tin text ở bước trên.

Huấn luyện: Sau đó, các file lần lượt được tách từ (Tokenize), chuyển từ thành vec-tơ (word embedding) và cuối cùng được đưa vào mô hình mạng nơ-ron BGRU để huấn luyện.

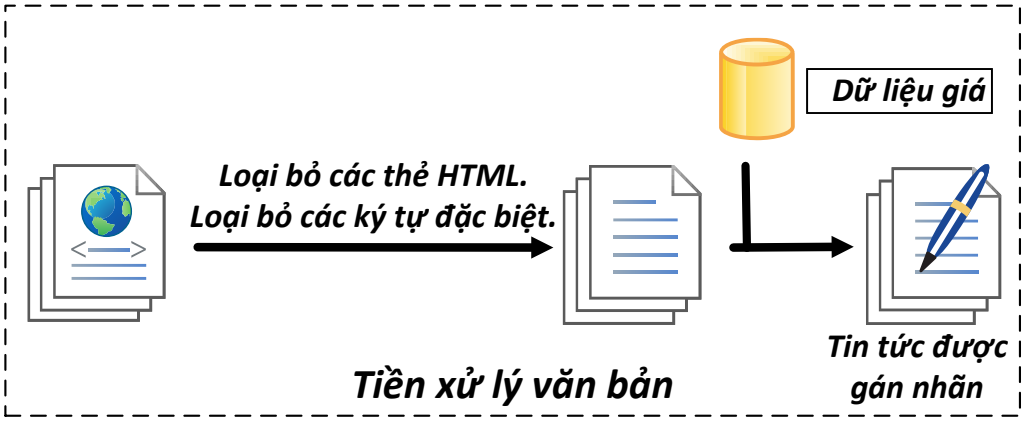


Hình 3.2. Minh họa mô hình dự báo chuyển động giá chứng khoán ứng dụng mạng nơ-ron BGRU.

Đánh giá độ chính xác: Để kiểm tra độ chính xác của mô hình dự báo, ở bước kiểm tra một hàm phân lớp softmax làm nhiệm vụ phân lớp để gán nhãn các tập tin kiểm tra và đánh giá độ chính xác của từng mẫu tập dữ liệu tương ứng.

Dưới đây luận văn sẽ trình bày chi tiết các bước trong mô hình trên.

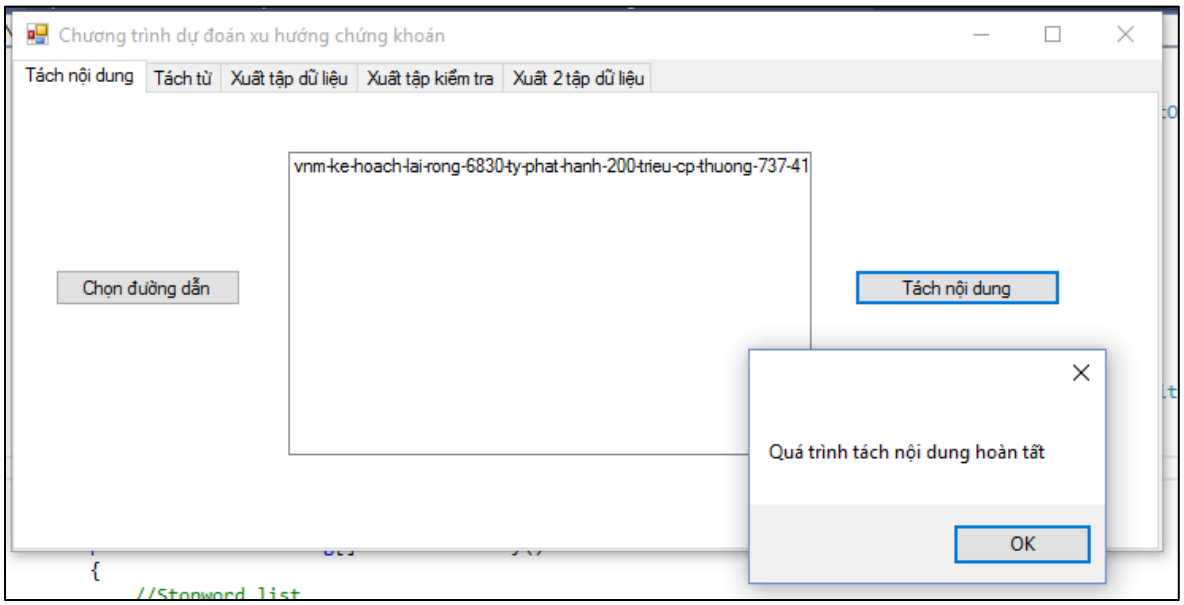
3.2.1. Tiền xử lý văn bản



Hình 3.3. Minh họa quá trình tiền xử lý văn bản

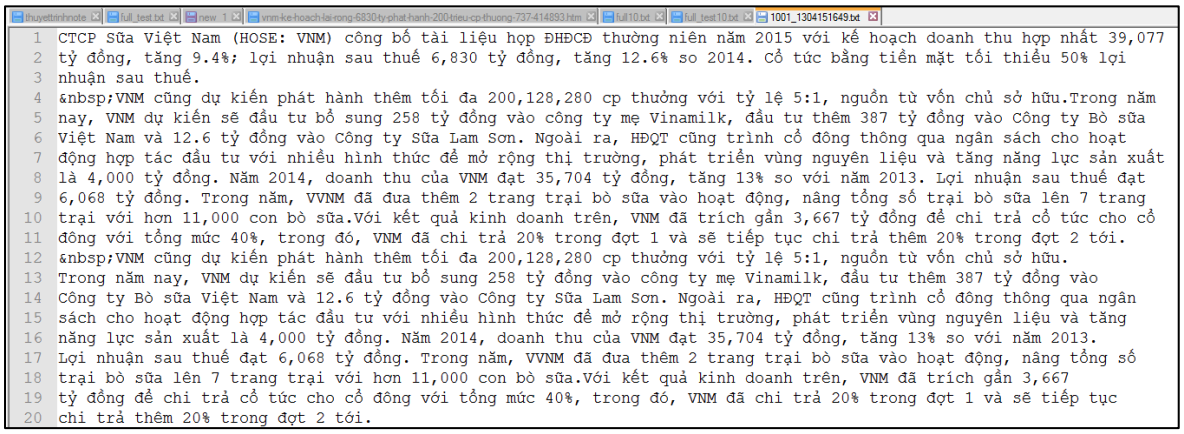
❖ *Loại bỏ các thẻ HTML*

Bước tiền xử lý văn bản loại bỏ các thẻ HTML được thực hiện cho dữ liệu Tiếng Việt, bởi vì bộ dữ liệu Tiếng Anh đã được chuẩn hóa. Tất cả tin tức thu thập được đều ở dưới định dạng HTML nên chứa rất nhiều thẻ không cần thiết của ngôn ngữ đánh dấu. Vì thế, đầu tiên cần loại bỏ tất cả nội dung không cần thiết để trích lấy nội dung chính và lưu dưới định dạng văn bản với tên tập tin là ngày, giờ bài báo được đăng. Bước tiền xử lý văn bản được minh họa qua hình 3.3.



Hình 3.4. Giao diện tách nội dung tin tức từ file html

Chương trình sẽ loại bỏ các thẻ định dạng của tập tin HTML, đồng thời tách lấy phần nội dung của tin tức và lưu lại với dạng tập tin có phần mở rộng ".txt". Hình 3.4 thể hiện giao diện chương trình tách nội dung văn bản từ file HTML. Hình 3.5 minh họa kết quả sau khi tách nội dung.

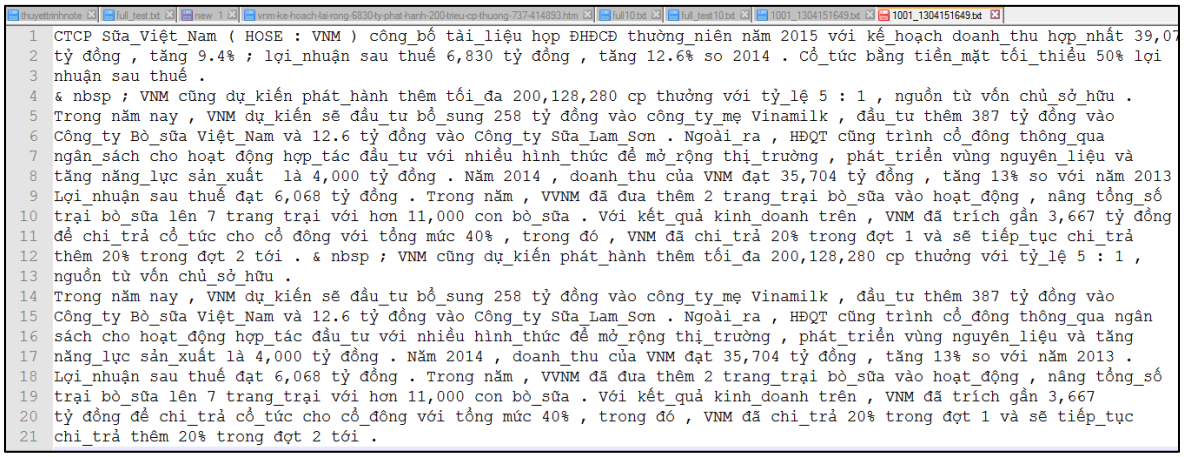


```
CTCP Sữa Việt Nam (HOSE: VNM) công bố tài liệu họp ĐHĐCĐ thường niên năm 2015 với kế hoạch doanh thu hợp nhất 39,077
tỷ đồng, tăng 9.4%; lợi nhuận sau thuế 6,830 tỷ đồng, tăng 12.6% so 2014. Cổ tức bằng tiền mặt tối thiểu 50% lợi
nhuận sau thuế.
 VNM cũng dự kiến phát hành thêm tối đa 200,128,280 cp thưởng với tỷ lệ 5:1, nguồn từ vốn chủ sở hữu.Trong năm
nay, VNM dự kiến sẽ đầu tư bổ sung 258 tỷ đồng vào công ty mẹ Vinamilk, đầu tư thêm 387 tỷ đồng vào Công ty Bò sữa
Việt Nam và 12.6 tỷ đồng vào Công ty Sữa Lam Sơn. Ngoài ra, HĐQT cũng trình cổ đông thông qua ngân sách cho hoạt
động hợp tác đầu tư với nhiều hình thức để mở rộng thị trường, phát triển vùng nguyên liệu và tăng năng lực sản xuất
là 4,000 tỷ đồng. Năm 2014, doanh thu của VNM đạt 35,704 tỷ đồng, tăng 13% so với năm 2013. Lợi nhuận sau thuế đạt
6,068 tỷ đồng. Trong năm, VVNM đã đưa thêm 2 trang trại bò sữa vào hoạt động, nâng tổng số trại bò sữa lên 7 trang
trại với hơn 11,000 con bò sữa.Với kết quả kinh doanh trên, VNM đã trích gần 3,667 tỷ đồng để chi trả cổ tức cho cổ
đông với tổng mức 40%, trong đó, VNM đã chi trả 20% trong đợt 1 và sẽ tiếp tục chi trả thêm 20% trong đợt 2 tới.
 VNM cũng dự kiến phát hành thêm tối đa 200,128,280 cp thưởng với tỷ lệ 5:1, nguồn từ vốn chủ sở hữu.
Trong năm nay, VNM dự kiến sẽ đầu tư bổ sung 258 tỷ đồng vào công ty mẹ Vinamilk, đầu tư thêm 387 tỷ đồng vào
Công ty Bò sữa Việt Nam và 12.6 tỷ đồng vào Công ty Sữa Lam Sơn. Ngoài ra, HĐQT cũng trình cổ đông thông qua ngân
sách cho hoạt động hợp tác đầu tư với nhiều hình thức để mở rộng thị trường, phát triển vùng nguyên liệu và tăng
năng lực sản xuất là 4,000 tỷ đồng. Năm 2014, doanh thu của VNM đạt 35,704 tỷ đồng, tăng 13% so với năm 2013.
Lợi nhuận sau thuế đạt 6,068 tỷ đồng. Trong năm, VVNM đã đưa thêm 2 trang trại bò sữa vào hoạt động, nâng tổng số
trại bò sữa lên 7 trang trại với hơn 11,000 con bò sữa.Với kết quả kinh doanh trên, VNM đã trích gần 3,667
tỷ đồng để chi trả cổ tức cho cổ đông với tổng mức 40%, trong đó, VNM đã chi trả 20% trong đợt 1 và sẽ tiếp tục
chi trả thêm 20% trong đợt 2 tới.
```

Hình 3.5. Tin tức sau khi được tách nội dung từ file HTML

❖ *Tách từ*

Do mỗi văn bản chứa nhiều câu, bước tiếp theo của hệ thống là tách từ từ các câu trong văn bản. Tác giả sử dụng thư viện tách từ Tiếng Việt Vntokenizer của nhóm tác giả Lê Hồng Phương [15]. Công cụ này được chứng minh đem lại độ chính xác hơn 90% trong việc tách từ Tiếng Việt. Chính vì vậy, tác giả đã tích hợp công cụ này vào trong chương trình tách từ.



```
CTCP Sữa_Việt_Nam ( HOSE : VNM ) công_bố tài_liệu họp ĐHĐCĐ thường_niên năm 2015 với kế_hoạch doanh_thu hợp_nhất 39,07
tỷ đồng , tăng 9.4% ; lợi_nhuận sau thuế 6,830 tỷ đồng , tăng 12.6% so 2014 . Cổ_tức bằng tiền_mặt tối_thiểu 50% lợi
nhuận sau thuế .
& nbsp ; VNM cũng dự_kiến phát_hành thêm tối_đa 200,128,280 cp thưởng với tỷ_lệ 5 : 1 , nguồn từ vốn chủ_sở_hữu .
Trong năm nay , VNM dự_kiến sẽ đầu_tư bổ_sung 258 tỷ đồng vào công_ty_mẹ Vinamilk , đầu_tư thêm 387 tỷ đồng vào
Công_ty Bò_sữa Việt_Nam và 12.6 tỷ đồng vào Công_ty Sữa_Lam_Sơn . Ngoài_ra , HĐQT cũng trình cổ_đông thông_qua
ngân_sách cho hoạt động hợp_tác đầu_tư với nhiều hình_thức để mở_rộng thị_trường , phát_triển vùng nguyên_liệu và
tăng năng_lực sản_xuất  là 4,000 tỷ đồng . Năm 2014 , doanh_thu của VNM đạt 35,704 tỷ đồng , tăng 13% so với năm 2013
Lợi_nhuận sau thuế đạt 6,068 tỷ đồng . Trong năm , VVNM đã đưa thêm 2 trang_trại bò_sữa vào hoạt_động , nâng tổng_số
trại bò_sữa lên 7 trang trại với hơn 11,000 con bò_sữa . Với kết_quả kinh_doanh trên , VNM đã trích gần 3,667 tỷ đồng
để chi_trả cổ_tức cho cổ đông với tổng mức 40% , trong đó , VNM đã chi_trả 20% trong đợt 1 và sẽ tiếp_tục chi_trả
thêm 20% trong đợt 2 tới . & nbsp ; VNM cũng dự_kiến phát_hành thêm tối_đa 200,128,280 cp thưởng với tỷ_lệ 5 : 1 ,
nguồn từ vốn chủ_sở_hữu .
Trong năm nay , VNM dự_kiến sẽ đầu_tư bổ_sung 258 tỷ đồng vào công_ty_mẹ Vinamilk , đầu_tư thêm 387 tỷ đồng vào
Công_ty Bò_sữa Việt_Nam và 12.6 tỷ đồng vào Công_ty Sữa_Lam_Sơn . Ngoài_ra , HĐQT cũng trình cổ_đông thông_qua ngân
sách cho hoạt_động hợp_tác đầu_tư với nhiều hình_thức để mở_rộng thị_trường , phát_triển vùng nguyên_liệu và tăng
năng_lực sản_xuất là 4,000 tỷ đồng . Năm 2014 , doanh_thu của VNM đạt 35,704 tỷ đồng , tăng 13% so với năm 2013 .
Lợi_nhuận sau thuế đạt 6,068 tỷ đồng . Trong năm , VVNM đã đưa thêm 2 trang_trại bò_sữa vào hoạt_động , nâng tổng_số
trại bò_sữa lên 7 trang_trại với hơn 11,000 con bò_sữa . Với kết_quả kinh_doanh trên , VNM đã trích gần 3,667
tỷ đồng để chi_trả cổ_tức cho cổ_đông với tổng mức 40% , trong đó , VNM đã chi_trả 20% trong đợt 1 và sẽ tiếp_tục
chi_trả thêm 20% trong đợt 2 tới .
```

Hình 3.6. Nội dung tin tức sau khi đã được tách từ

Văn bản sau khi được tách từ sẽ được lưu lại với dạng tập tin có phần mở rộng ".txt" với nội dung tương tự như nội dung đã được tách từ tập tin HTML. Điểm khác biệt ở đây là từ ngữ ở trong câu đã được phân biệt rõ ràng. Hình 3.6 minh họa kết quả sau khi tách từ.

❖ *Loại bỏ từ dừng*

Trong bước cuối cùng của giai đoạn này, tác giả cải thiện mức độ hiệu quả và tài nguyên hệ thống bằng cách loại bỏ các từ không cần thiết mà không đem lại thông tin có ích gì cho việc phân loại: các từ dừng (và, của, là,...), số, kí hiệu. Quy trình loại từ dừng sẽ được chương trình tự động thực hiện sau khi tiến hành tách từ trong văn bản xong. Tác giả sử dụng thư viện natural language toolkit (NLTK) [30], đây là 1 thư viện hỗ trợ mạnh mẽ trong việc xử lý ngôn ngữ tự nhiên. Thư viện NLTK cung cấp bộ dữ liệu từ điển của nhiều loại ngôn ngữ. Hình 3.7 dưới đây minh họa danh sách các từ dừng tiếng Anh và tiếng Việt trong thư viện NLTK

Hình 3.7. Minh họa danh sách "từ dừng" của thư viện NLTK.

3.2.2. Word Embedding

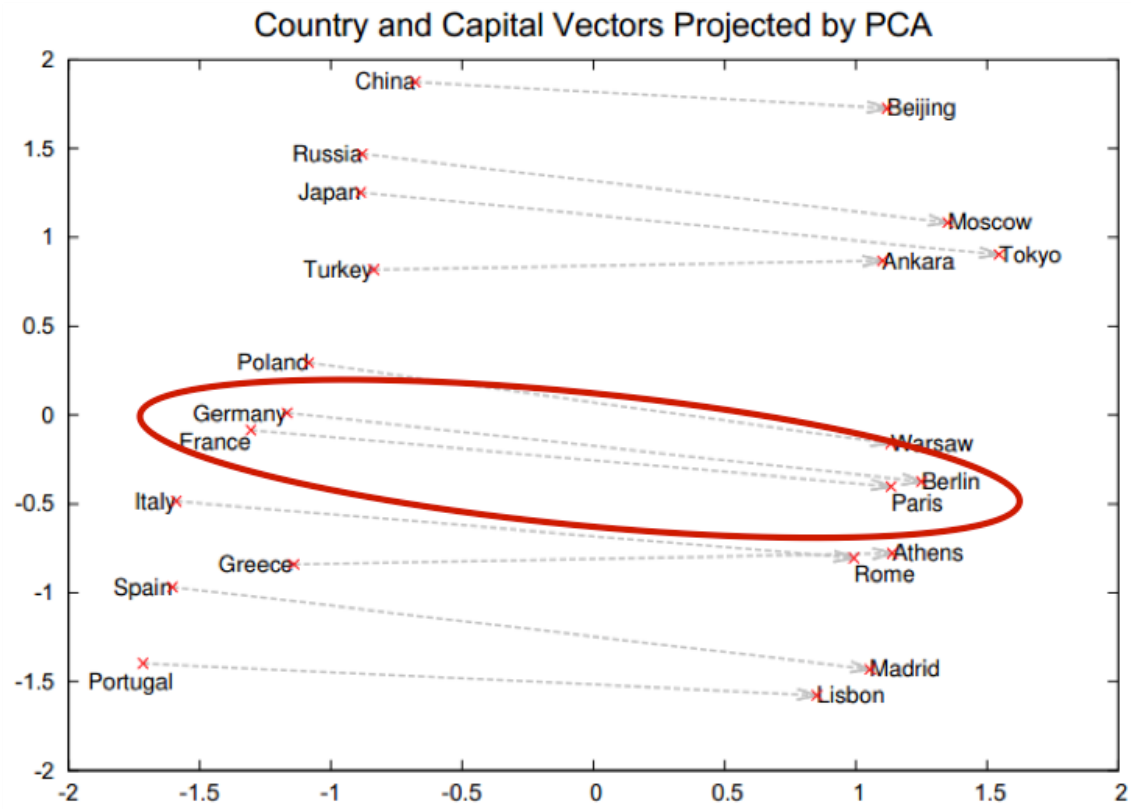
Bước word embedding là một bước để chuyển một từ thành vec-tơ, các vec-tơ này sẽ là dữ liệu đầu vào cho mô hình mạng nơ-ron. Trong nghiên cứu này tác giả ứng dụng dữ liệu đầu vào là một văn bản, xem như là tập hợp các từ (word). Đầu tiên, tương ứng với mỗi từ sẽ khởi tạo một vec-tơ ngẫu nhiên với số chiều được chỉ định. Sau khi đã có vec-tơ ngẫu nhiên, việc tiếp theo là thực hiện quá trình điều chỉnh vec-

tơ của các từ này để sao cho chúng có thể biểu diễn được liên hệ giữa các từ có quan hệ với nhau.

Giả sử chúng ta có câu văn sau: **Con mèo trèo cây cau.** Tương ứng với mỗi từ trong câu này, chúng ta sẽ khởi tạo một vec-tơ ngẫu nhiên với số chiều được quy định trước. Một mạng nơ-ron được dùng để điều chỉnh dần dần các vec-tơ thông qua quá trình huấn luyện dữ liệu. Nếu như thay từ "**trèo**" bằng từ "**ngủ**", rõ ràng chúng ta sẽ có 1 câu hoàn toàn vô nghĩa và hầu như không bao giờ xuất hiện trong văn bản bình thường: "**con mèo ngủ cây cau**". Bằng cách thay từ "**trèo**" bằng từ "**ngủ**" và nói cho mạng nơ-ron biết rằng câu mới sinh ra là không hợp lệ, mạng nơ-ron sẽ phải điều chỉnh các tham số trong mạng của nó một cách hợp lý để đưa ra được output đúng như chúng ta mong muốn.

Nhờ việc huấn luyện mạng nơ-ron trên một số lượng văn bản cực lớn, thì vec-tơ của mỗi từ sẽ được điều chỉnh càng chính xác và những từ có liên quan nhau cũng sẽ xuất hiện ở gần nhau hơn, khi đó giữa các từ có mối liên hệ với nhau rất thú vị.

Ví dụ, sơ đồ hình 3.8 mô tả mối liên hệ giữa thủ đô và quốc gia. Như ta thấy vec-tơ biểu diễn thủ đô Paris và Pháp thì tương tự như vec-tơ biểu diễn thủ đô Berlin và Đức. Sơ đồ đã cho thấy khả sức mạnh của mạng nơ-ron trong việc biểu ngữ cảnh giữa thủ đô và thành phố.



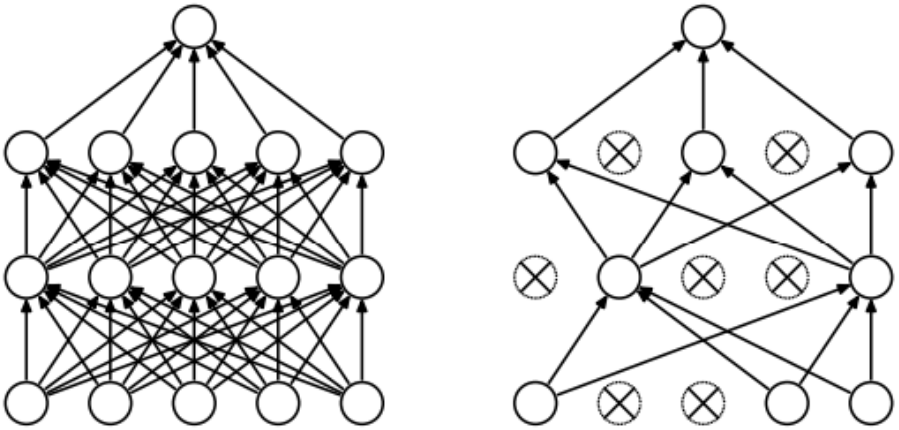
Hình 3.8. Minh họa vec-tơ của tên "quốc gia" và "thủ đô" [29].

3.2.3. Máy học với mô hình BGRU

Sau quá trình word embedding, thì vec-tơ đầu ra đại diện cho mỗi token (từ) của lớp word emdedding sẽ là đầu vào cho mạng BGRU. Trong đó, một bản tin tài chính ($x_1, x_2, \ldots, x_n$) có n token (từ) sẽ được nhập vào mạng BGRU. Đầu ra của mạng BGRU là 1 vevtor đặc trưng đại diện cho bản tin, sau đó vec-tơ này sẽ được đưa vào lớp softmax đã đề cập ở mục 2.1.3 để phân loại vào các lớp tương ứng.

3.2.4. Kỹ thuật Dropout

Mạng nơ-ron sâu với nhiều lớp ẩn và nhiều đơn vị tính toán với rất nhiều tham số được ước lượng là sức mạnh của hệ thống máy học. Tuy nhiên, một trong những vấn đề đau đầu đối với các nhà nghiên cứu trong lĩnh vực máy học là vấn đề quá vừa dữ liệu hay còn gọi là "overfitting". Một trong những kỹ thuật mà gần đây các nhà nghiên cứu thường ứng dụng để tránh overfitting được gọi là dropout [13].



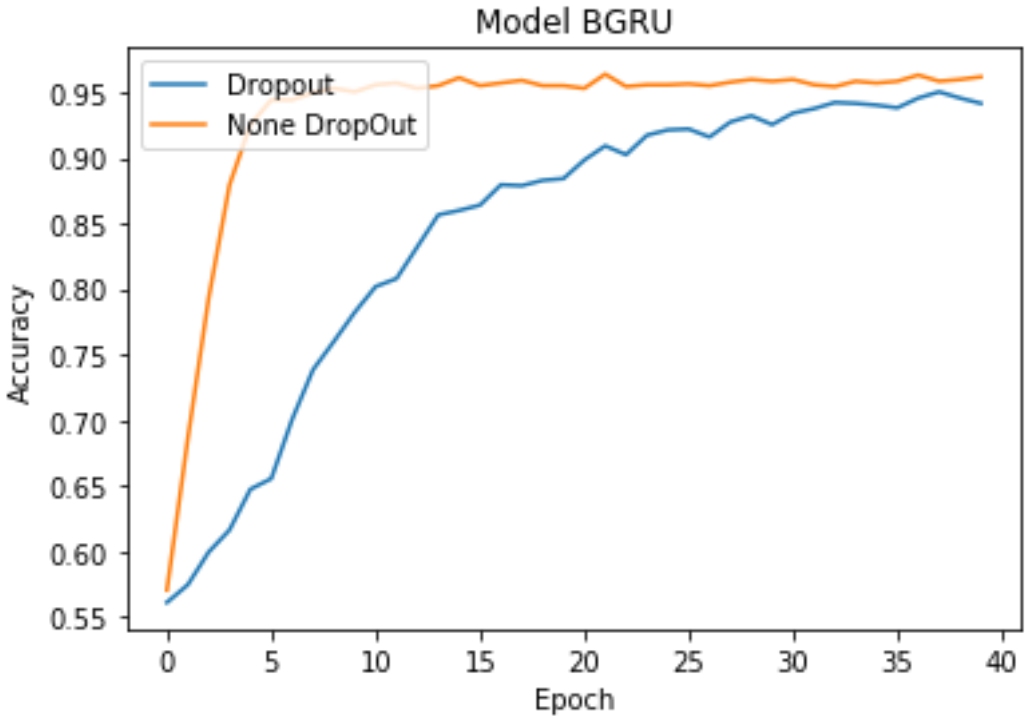
a) Mạng nơ-ron đầy đủ dropout

b) Mạng nơ-ron sau khi áp dụng

Hình 3.9. Minh họa kỹ thuật dropout. [13]

Ý tưởng chính của dropout là ngẫu nhiên ngắt các kết nối giữa các đơn vị tính toán trong suốt quá trình huấn luyện, việc này đảm bảo chống vấn đề quá vừa dữ liệu. Trong quá trình huấn luyện, dropout được xem như làm "mỏng" mạng nơ-ron. Quá trình thực nghiệm cho thấy rằng dropout cải thiện đáng kể kết quả đối với việc học có giám sát. Hình 3.9 minh họa kỹ thuật dropout, với mô hình bên trái mô tả một mạng nơ-ron với 2 lớp ẩn và bên phải mô tả mạng nơ-ron sau khi được áp dụng kỹ thuật dropout.

Đối với đề tài luận văn, kỹ thuật dropout được áp dụng trên việc "làm mỏng" mạng BGRU trên từng cổng của từng GRU tiến và lùi. Tác giả thực hiện in kết quả quá trình huấn luyện mô hình BGRU có áp dụng dropout - đường màu xanh dương và không sử dụng dropout - đường màu cam được minh họa qua hình 3.10 với bộ dữ liệu của thực nghiệm được mô tả ở mục 4.3. Ta có thể thấy rõ quá trình học của mô hình BGRU đường màu xanh dương có sự dịch chuyển độ chính xác đồng đều hơn đường màu cam. Đường màu cam có sự biến thiên nhanh gần chạm ngưỡng ở epoch thứ 5 và không thay đổi nhiều sau đó. Trong khi đó, đường màu xanh dương di chuyển từ từ với độ chính xác tăng dần. Chính vì thế, mô hình BGRU áp dụng dropout khi đánh giá với dữ liệu kiểm tra, chúng ta sẽ dễ dàng nhận diện được giai đoạn bị overfitting và đưa ra những tham số tốt nhất cho mô hình.



Hình 3.10. So sánh mô hình BGRU khi áp dụng Dropout

Chương 4. THỰC NGHIỆM

Để kiểm chứng tính khả thi và đánh giá mô hình BGRU, luận văn được thực nghiệm trên 2 bộ dữ liệu Tiếng Anh và Tiếng Việt. Đối với dữ liệu Tiếng Anh, luận văn so sánh kết quả thực nghiệm với 2 nghiên cứu liên quan [7], [25] với cùng bộ dữ liệu và tiêu chuẩn đánh giá mô hình BGRU. Đồng thời qua thực nghiệm để chỉ ra tin tức có tính tức thời lên sự chuyển động giá của chứng khoán và giảm dần theo thời gian. Đối với dữ liệu là Tiếng Việt, tác giả so sánh kết quả với nghiên cứu của nhóm tác giả [9] trên cùng bộ dữ liệu. Kết quả thực nghiệm rất đáng khả quan với độ chính xác cao hơn nghiên cứu trước đó hơn 5%.

4.1. Cài đặt, côngcụ hỗ trợ

Tác giả đã triển khai thực nghiệm bằng ngôn ngữ python phiên bản 2.7 và áp dụng các thư viện lập trình dưới đây:

- Sử dụng thư viện Theano 0.8.2[2];
- Thư viện deep learning Keras 1.2.2[3];
- Sử dụng công cụ Jupyter notebook[4] cho việc lập trình thực nghiệm;

4.2. Phương pháp đánh giá

Ma trận kết hợp, precision (độ chính xác giữa các mẫu), độ phủ (recall) và accuracy (độ chính xác) được sử dụng để đánh giá mô hình đề xuất. Accuracy là số mẫu được phân lớp chính xác so với tổng số mẫu. Nói chung, bộ phân lớp càng tốt thì accuracy càng cao. Tuy nhiên, nếu chỉ dựa vào accuracy để đánh giá hệ thống là không đủ vì nếu một lớp xuất hiện nhiều hơn đáng kể so với một lớp khác, hệ thống có thể có độ chính xác cao bằng cách gán nhãn tất cả mẫu vào phân lớp vượt trội hơn. Precision và recall là hai thước đo được sử dụng rộng rãi để đánh giá hiệu quả trong khai thác cũng như phân lớp văn bản. Chúng được xem là giá trị mở rộng của accuracy và bằng cách kết hợp các độ đo này sẽ cho ta một cách nhìn chi tiết hơn về hệ thống. Precision có thể được xem là thước đo độ chính xác chuẩn trong khi recall là thước

[2] https://github.com/Theano/Theano
[3] https://github.com/fchollet/keras
[4] http://jupyter.org

đo của độ hoàn chỉnh. Nói cách khác, precision cao, có nghĩa rằng hầu hết các mẫu được dán nhãn là tích cực được gán nhãn chính xác trong khi đó. Độ đo recall cao có nghĩa là hệ thống tìm thấy hầu hết mẫu tích. Trong ma trận kết hợp, TP và TN chỉ ra phân lớp đúng cho các lớp tương ứng, FP và FN chỉ ra phân lớp sai cho các lớp tương ứng.

Dựa vào ma trận trong bảng 4.1, accuracy là phần tăng dự đoán là tăng (TP) và giảm dự đoán là giảm (TN) chia cho tổng, precision được định nghĩa là tăng dự đoán là tăng (TP) chia cho tổng của tăng được dự đoán là tăng (TP) và giảm được dự đoán là tăng (FP). Recall được định nghĩa là tăng được dự đoán là tăng (TP) chia cho tổng của tăng dự đoán là tăng (TP) và tăng được dự đoán là giảm (FN). Các công thức tương ứng (CT4.1), (CT4.2), (CT4.3) như sau:

Bảng 4.1. Ma trận kết hợp tính độ chính xác

		Lớp dự đoán	
		Tăng	***Giảm***
Lớp thực tế	***Tăng***	TP	FN
	Giảm	FP	TN

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (4.1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4.3)$$

4.3. Bộ dữ liệu thực nghiệm

Bộ dữ liệu thực nghiệm được lấy từ các trang tin Reuters [5]và Bloomberg[6] từ tháng 10 năm 2006 đến tháng 11 năm 2013. Trong đó số lượng tin tức từ trang Reuters là 106,521 tin và 447,145 tin từ trang Bloomberg. Dữ liệu về giá cổ phiếu được công

[5] http://www.reuters.com
[6] https://www.bloomberg.com

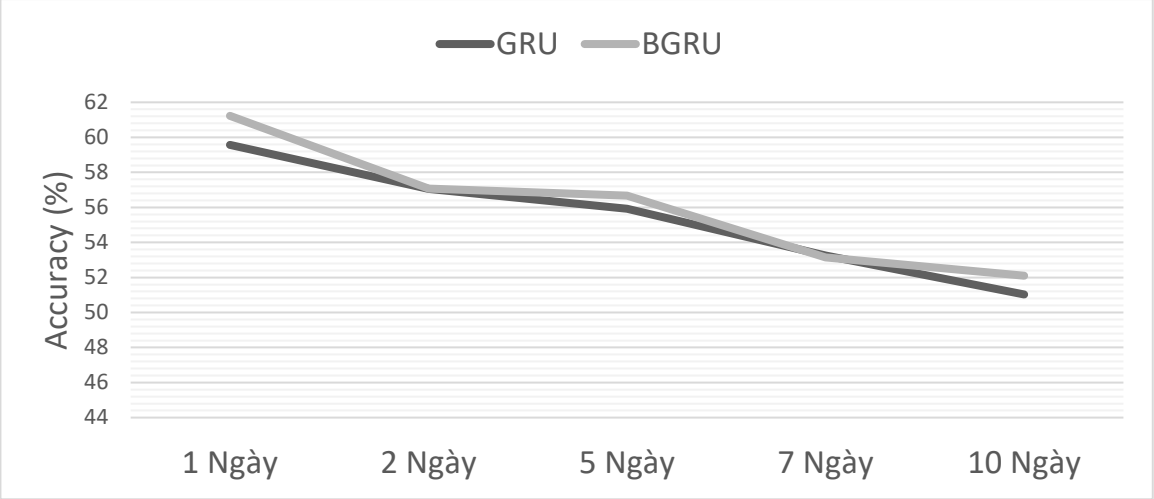
khai từ trang Yahoo Finance[7] được chọn trong khoảng thời gian tương ứng với thời gian tin tức được đăng để thực nghiệm.

4.3.1. Sự tác động của tin tức lên giá chứng khoán theo thời gian

Trước khi thực nghiệm so sánh với kết quả của 2 nghiên cứu [7], [25]. Tác giả tiến hành thực nghiệm đánh giá mức độ ảnh hưởng của tin tức lên giá của chứng khoán theo thời gian. Với thực nghiệm này, tác giả đã chọn bộ dữ liệu được thu thập từ trang Reuters và thực nghiệm với các khoảng thời gian khác nhau (1 ngày, 2 ngày, 5 ngày, 7 ngày và 10 ngày). Giá mã cổ phiếu S&P500 được chọn để thực nghiệm, trong đó với khoảng thời gian 1 ngày nghĩa là tin tức có tác động đến giá cổ phiếu trong 24 giờ kể từ khi bản tin được đăng, 2 ngày là khoảng thời gian 48 tiếng đồng hồ kể từ khi tin được đăng và tương tự vậy đối với các khoảng thời gian còn lại. Việc gán nhãn các bản tin vào lớp tăng hay giảm được thực hiện bằng cách so sánh giá mở cửa và đóng cửa của mã S&P500 trong ngày bản tin được đăng.

Kết quả thực nghiệm được thể hiện qua biểu đồ hình 4.1 với 2 mô hình được thực nghiệm là GRU và BGRU. Kết quả chứng minh trong khoảng 24 giờ đầu tiên (1 ngày) kể từ sau khi bản tin được đăng có độ chính xác cao nhất và giảm dần theo thời gian. Qua đó có thể thấy sự tác động gần như ngay lập tức của tin tức vào thị trường chứng khoán, tuy nhiên không thể phủ nhận tất cả các tin tức, sự kiện đều tác động tức thời. Bởi lẽ, ta lấy một trường hợp là sự kiện "Brexit" [19] năm 2016 với sự tác động đến thị trường chứng khoán trong khoản thời gian dài. Dù vậy, với bộ dữ liệu lớn được thực nghiệm thì có thể nhận định rằng hầu hết các tin tức có sự tác động tức thời đến chứng khoán.

[7] https://finance.yahoo.com



Hình 4.1. Kết quả thực nghiệm đánh giá tác động của tin tức theo thời gian.

Trong phần thực nghiệm đầu tiên có thể thấy rằng mô hình dự báo dùng mạng BGRU có độ chính xác cao hơn so với mô hình GRU. Đồng thời, sự ảnh hưởng của tin tức đến giá của chứng khoán trong cùng 1 ngày cao nhất (độ chính xác giảm dần qua các ngày), khoảng thời gian 1 ngày cũng là khoảng thời gian mà các nghiên cứu có liên quan chọn lựa để thực hiện các thực nghiệm. Chính vì thế, trong các thực nghiệm về sau tác giả chỉ chọn khoảng thời gian 1 ngày để cùng cơ sở so sánh và đánh giá kết quả.

4.3.2. Dự báo sự chuyển động giá chứng khoán của mã S&P500

Trong phần tiếp theo, luận văn sử dụng toàn bộ dữ liệu tiếng Anh để thực nghiệm và so sánh với 2 nghiên cứu có liên quan gần đây nhất. Trong đó, nhóm tác giả của nghiên cứu [7] đã xây dựng hệ thống để rút trích sự kiện về dạng E = (O1, P, O2) trong đó O1 thể hiện đối tượng thứ nhất, O2 thể hiện đối tượng thứ 2 (đối tượng ở đây có thể là mã cổ phiếu, tên công ty, tên nhân vật, …) và P thể hiện mối quan hệ giữa 2 đối tượng tạo thành sự kiện để biểu diễn cho một tin tức, sự kiện. Trong nghiên cứu này, [7] đã ứng dụng mạng nơ-ron feedforward tiêu chuẩn để thực nghiệm trong bộ dữ liệu trên với độ chính xác đạt 55,21%. Với cùng bộ dữ liệu này, nhóm tác giả [25] đã phát triển một mạng nơ-ron sâu và kết hợp với 1 lớp word embedding ban đầu để để dự báo chuyển động giá chứng khoán trong tương lai của mã S&P500 index. Độ chính xác của nghiên cứu này được cải thiện lên 56,87 %.

Nhằm đánh giá mô hình BGRU mà luận văn đã đề xuất ở trên, trong phần thực nghiệm này, tác giả thực hiện với các mô hình GRU và BGRU với cùng bộ dữ liệu và thời gian tương ứng. Bên cạnh đó, tác giả cũng triển khai một mô hình mạng nơ-ron LSTM[8], đây cũng được xem là một biến thể của RNN và rất ưa được sử dụng trong các mô hình dự báo với deep learning gần đây. Cụ thể, [2] đã áp dụng LSTM vào mô hình dự báo giá chứng khoán dựa vào tin tài chính. GRU và LSTM cùng là biến thể của RNN nên một câu hỏi đặt ra là mô hình nào sẽ hoạt động tốt hơn. Để trả lời câu hỏi trên [6] đã có một khảo sát và kết luận rằng cả hai mô hình đều cho kết quả gần như nhau nhưng GRU thường nhanh hơn vì ít tham số hơn LSTM. Bộ dữ liệu tin tức chứng khoán sau khi được phân lớp được chia thành 3 tập giống như cách chia của [25]. Tin tức trong khoảng 01-10-2006 đến 31-12-2012 được dùng để huấn luyện, từ 01-01-2013 đến 15-06-2013 được dùng để đánh giá mô hình tìm tham số và tin tức trong khoảng 16-06-2013 đến 31-12-2013 được dùng kiểm tra. Kết quả của thực nghiệm được thể hiện ở bảng 4.2. Qua đó, có thể thấy mô hình BGRU cho kết quả cao nhất với độ chính xác hơn 60%.

Bảng 4.2. Kết quả thực nghiệm dự báo chuyển động giá mã S&P500 Index

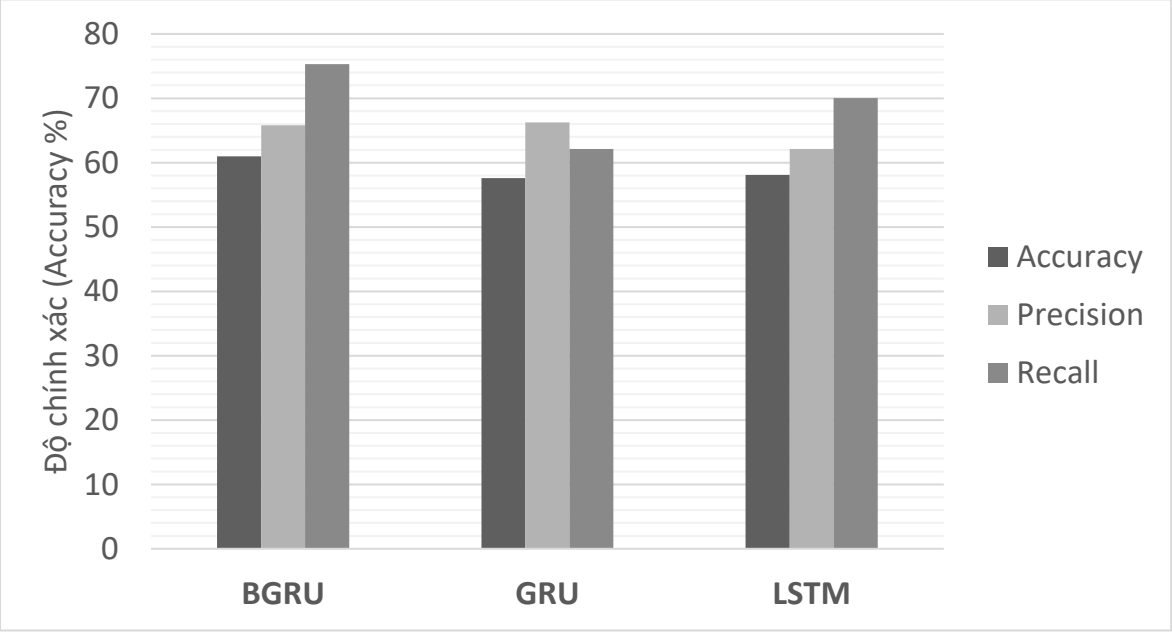
Mô hình	**Độ chính xác**
Ding et al. - 2014 [7]	55.21%
Pen and Hui Jiang - 2016 [25]	56.87%
LSTM	58.12%
GRU	57.59%
BGRU	**60.98%**

[8] Đặc tả chi tiết ở phụ lục B

Bảng 4.3. Kết quả các độ đo trên mô hình BGRU, GRU và LSTM

Độ đo	**Tỉ lệ**		
	BGRU	**GRU**	**LSTM**
Accuracy	**60.98%**	57.59%	58.12%
Precision	65.83%	**66.23%**	62.10%
Recall	**75.32%**	69.4%	70.01%

Bên cạnh đó, bảng 4.3 thể hiện kết quả các độ đo khác khi thực nghiệm trên mô hình LSTM, GRU và BGRU. Độ chính xác của dự báo của mô hình BGRU thể hiện sự tối ưu với kết quả độ chính xác cao nhất trong 3 mô hình. Tiếp theo, để thấy tỉ lệ dự báo đúng trên từng phân lớp của mô hình, độ đo precision thể hiện tỉ lệ chính xác của dự báo xu hướng giá tăng là đúng trên tổng số dự báo tăng của mô hình GRU đạt tỉ lệ cao nhất là 65%. Tuy nhiên, qua hình 4.2 cho thấy kết quả biết tỉ lệ độ phủ của mô hình BGRU là cao nhất đạt hơn 75%. Tỉ lệ này thể hiện tỉ lệ của dự báo xu hướng giá tăng là đúng trên tổng số các bài báo thực tế được gán nhãn tích cực.



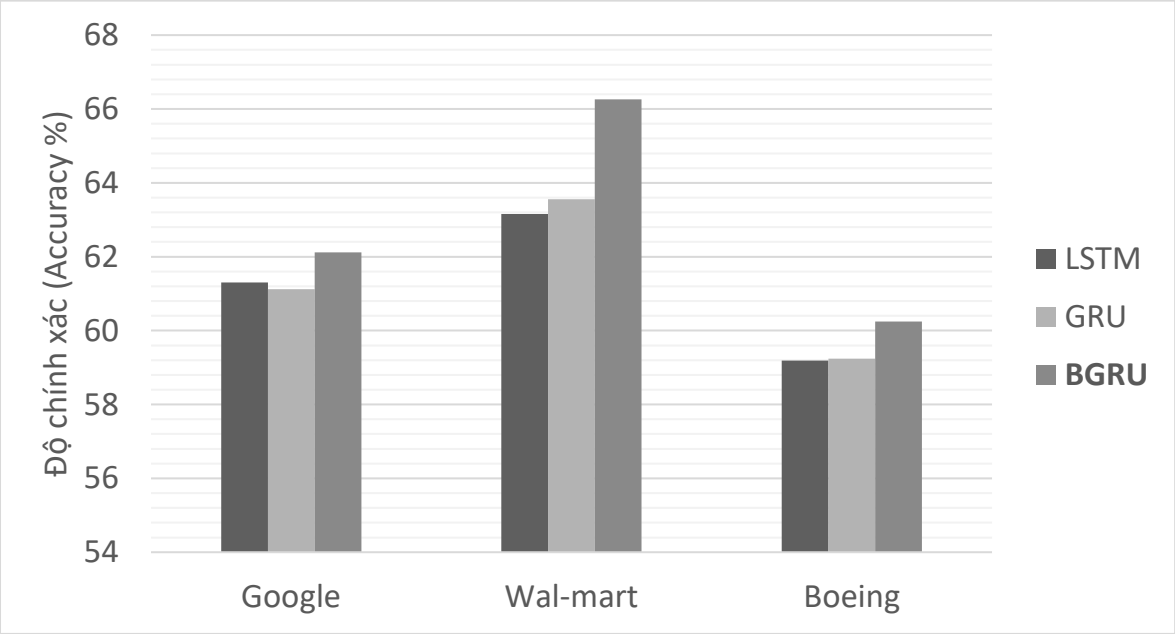
Hình 4.2. Biểu đồ kết quả các độ đo trên mô hình LSTM, GRU và BGRU

4.3.3. Dự báo mã chứng khoán riêng biệt

Để đánh giá các mô hình dự báo khi áp dụng cho từng mã chứng khoán riêng biệt, tác giả đã thực nghiệm với 3 công ty ở 3 lĩnh vực khác nhau. Tác giả chọn công ty Google[9] đại diện ngành công nghệ thông tin, công ty Wal-Mart[10] đại diện ngành thương mại và công ty Boeing[11] trong ngành công nghiệp chế tạo (được phân loại bởi The Global Industry Classification Standard[12]). Từ bộ dữ liệu tiếng Anh, tác giả trích lọc tất cả các tin tức có đề cập đến các công ty nêu trên trong bộ dữ liệu và tiến hành thử nghiệm. Chi tiết thống kê số lượng tin tức được thể hiện qua bảng 4.4

Bảng 4.4. Thống kê số lượng tin tức các mã cổ phiếu riêng biệt

Tên công ty	Số lượng tập tin huấn luyện	Số lượng tập tin kiểm tra
Google Inc	2,252	1,124
Wal-Mart	1,484	741
Boeing	2,080	1,039



Hình 4.3. Biểu đồ đánh giá sự tác động tin tức lên từng mã cổ phiếu riêng biệt

[9] https://www.google.com/ about/company
[10] https://www.walmart.com
[11] http://www.boeing.com
[12] Global Industry Classification Standard: Chuẩn phân loại các ngành công nghiệp toàn cầu (GICS) được đề xuất vào năm 1999 bởi MSCI và Standard & Poor (S & P) sử dụng trong ngành tài chính toàn cầu.

Qua kết quả thực nghiệm được thể hiện ở hình 4.3 có thể thấy rằng độ chính xác của mô hình BGRU cho kết quả tốt nhất, trong khi kết quả của 2 mô hình GRU và LSTM chênh nhau không quá đáng kể. Đặc biệt, trong đó độ chính xác dự báo cho mã cổ phiếu công ty Wal-Mart đạt hơn 66%.

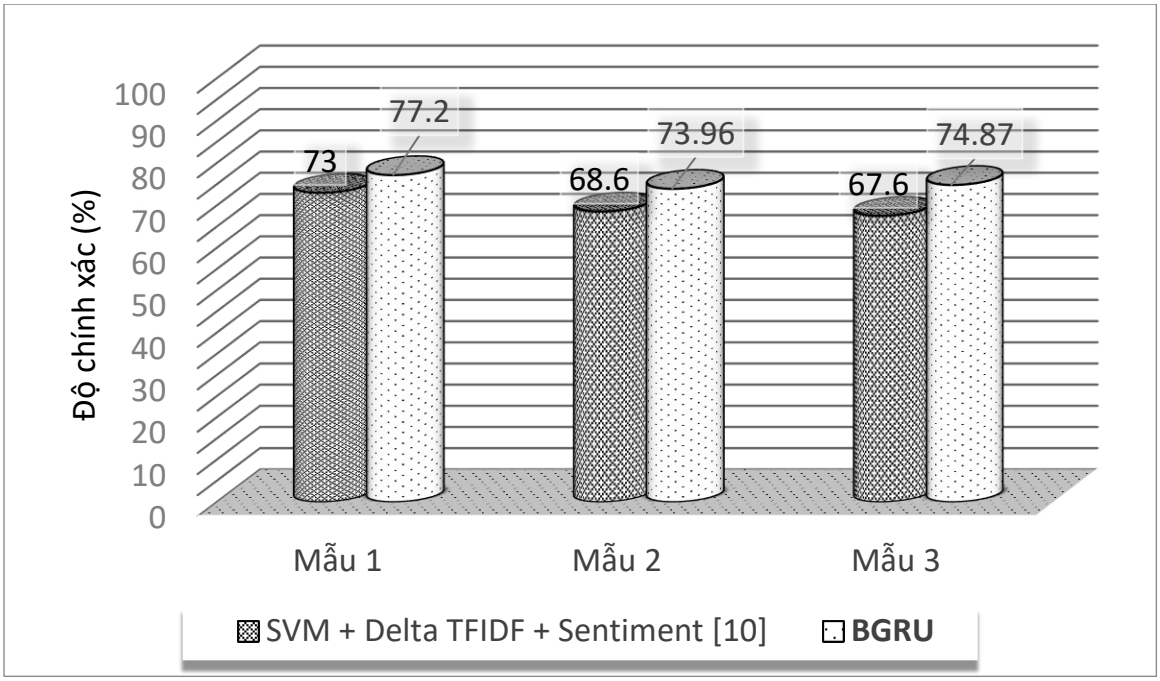
4.4. Dự báo chuyển động giá của VN-INDEX.

Tác giả thực nghiệm với bộ dữ liệu được thu thập bởi nhóm nghiên cứu [9]. Tin tức tài chính được thu thập từ nguồn trang vietstock.vn trong khoảng thời gian từ tháng 01/2014 đến tháng 05/2015 với số lượng 2,471 bài báo. Để khớp với so sánh của nhóm nghiên cứu [9] tác giả chia dữ liệu thực nghiệm làm 3 mẫu gồm: mẫu 1 chứa các tin tức từ tháng 01/2015 đến tháng 04/2015, mẫu 2 chứa các tin tức từ tháng 09/2014 đến tháng 04/2015, mẫu 3 chứa các tin tức từ tháng 05/2014 đến tháng 04/2015 như bảng 4.5

Bảng 4.5. Chi tiết dữ liệu bài báo Tiếng Việt

	Số lượng bài báo		
	Tập huấn luyện	***Tập kiểm tra***	***Tổng***
Mẫu 1 (01/2015-04/2015)	1092	463	**1555**
Mẫu 2 (09/2014-04/2015)	1499	640	**2139**
Mẫu 3 (05/2014-04/2015)	1715	756	**2471**

Qua biểu đồ hình 4.4 có thể thấy mô hình dự báo dựa vào mạng BGRU cho kết quả cao hơn so với phương pháp SVM kết hợp cùng phương pháp beta TF-IDF và bộ từ điển sentiment [9]. Với cùng bộ dữ liệu, ở mẫu số 1 kết quả cao hơn 4% chính xác, mẫu số 2 là hơn 5% và mẫu số 3 là hơn 7%. Qua đó, có thể nhận định rằng với bộ dữ liệu càng lớn thì mô hình dự báo với phương pháp BGRU có khả năng dự báo tốt hơn so với SVM. Nhìn chung độ chính xác (accuracy) của hệ thống luôn ở trên mức 70%, còn độ chính xác (accuracy) cao thấp giữa các mẫu là do có sự nhiễu trong các tin tức thu thập được.



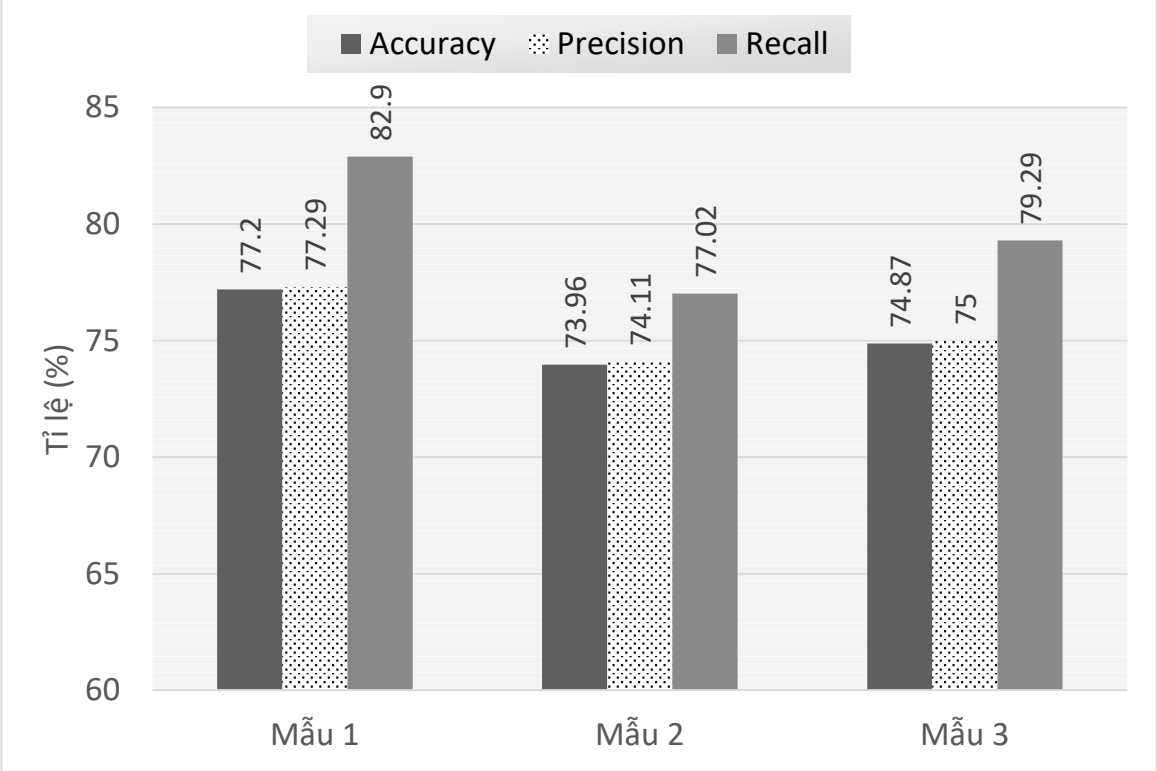
Hình 4.4. Biểu đồ đánh giá kết quả thực nghiệm BGRU với SVM

Tiếp theo, tác giả tiến hành so sánh độ đo của 3 mẫu thời gian ứng với mẫu dữ liệu 1, mẫu dữ liệu 2 và mẫu dữ liệu 3 để tìm hiểu độ đo precision và recall của hệ thống khi số lượng bài báo và khoảng thời gian được tăng lên.

Ở biểu đồ hình 4.5, độ đo precision thể hiện tỉ lệ chính xác của dự báo xu hướng giá tăng là đúng trên tổng số dự báo tăng của mô hình luôn cao hơn 75%. Riêng mẫu 3 có độ precision đạt tỉ lệ hơn 84%, qua đó cho thấy tỉ lệ dự báo đúng trong từng phân lớp của mô hình đạt tỉ lệ khả quan.

Bên cạnh đó kết quả độ đo recall cho biết tỉ lệ độ phủ của dự báo xu hướng giá tăng là đúng trên tổng số các bài báo thực tế làm xu hướng giá tăng luôn cao hơn 75%. Đặc biệt mẫu dữ liệu 1 đạt độ phủ hơn 82%.

Qua kết quả các độ đo trên các tập mẫu dự liệu theo thời gian tăng và số lượng tin tức tăng, chúng ta thấy có sự biến thiên nhỏ về độ chính xác, độ precision và recall. Tuy nhiên, mô hình vẫn cho thấy độ ổn định với các kết quả các độ đo đều lớn hơn 70% và độ phủ lớn hơn 75%.



Hình 4.5. Biểu đồ thể hiện các độ đo theo các mẫu thời gian

4.5. Đánh giá

Luận văn đã chứng minh được mạng nơ-ron và mô hình đề xuất BGRU là một công cụ mạnh trong việc dự báo chuyển động giá chứng khoán dựa vào tin tức nói riêng và mô hình ngôn ngữ nói chung.

Đồng thời, kết quả thực nghiệm áp dụng vào thị trường chứng khoán Việt Nam là rất khả thi. Để đạt được điều đó, tin tức tài chính cùng với giá chứng khoán được đánh giá. Và qua quá trình thực nghiệm đã chứng minh tin tức tài chính có sự tương quan với giá chứng khoán. Cụ thể, nghiên cứu của tác giả đã phản ánh đúng thực trạng sàn HoSE – nơi có chỉ số tài chính tốt và tính thanh khoản cao. Đặc biệt khi áp dụng vào rổ VN-INDEX đã mang lại độ chính xác khá cao và đáng ghi nhận lớn hơn 70%. Tuy nhiên, bộ dữ liệu Tiếng Việt vẫn còn khá ít, độ nhiễu và mức độ phản ánh thị trường của tin tức chưa cao đây cũng là một lý do chủ quan dẫn đến khi áp dụng vào thị trường Việt Nam chưa mang lại độ tin cậy cao.

Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả đạt được

5.1.1. Về khoa học

Trong khóa luận này, tác giả đã đạt được những kết quả về mặt khoa học như sau:

- Đề xuất một mô hình mạng nơ-ron nhân tạo được đặt tên là Bidirectional Gated Recurrent Unit (BGRU). Mô hình đặc biệt có thế mạnh trong việc xử lý dữ liệu dạng chuỗi và tuần tự, đồng thời BGRU có khả năng học được đặc trưng của cả 2 chiều ngữ cảnh đối với mô hình ngôn ngữ.
- Áp dụng các kỹ thuật như word embedding, dropout, các bước tiền xử lý dữ liệu để tăng độ chính xác và tin cậy cho mô hình.
- So sánh mô hình đề xuất BGRU với các mô hình LSTM, GRU và SVM.
- Xây dựng mô hình dự báo xu hướng chứng khoán dựa trên tin tức có độ chính xác cao, từ đó có thể kết hợp với tiếp cận phân tích kỹ thuật để giải quyết hiệu quả hơn trong vấn đề dự báo chứng khoán nói riêng và dự báo tài chính (vàng, ngoại tệ) nói chung.

5.1.2. Về thực tiễn

- Chứng minh việc sử dụng tin tức tài chính có ảnh hưởng đến giá cổ phiếu và cụ thể tại thị trường chứng khoán Việt Nam trong kết quả thực nghiệm là rổ VN-INDEX. Cụ thể, kết quả nghiên cứu đã phản ánh đúng thực trạng sàn HoSE – nơi có chỉ số tài chính tốt và tính thanh khoản cao. Đặc biệt khi áp dụng vào rổ VN-INDEX đã mang lại độ chính xác khá cao và đáng ghi nhận (77,2%).
- Ngoài ra việc nghiên cứu được áp dụng cho thị trường chứng khoán của cả 2 ngôn ngữ Tiếng Anh và tiếng Việt.

5.2. Hướng phát triển

Việc áp dụng tin tức tài chính vào dự báo giá chứng khoán còn khá mới mẻ ở Việt Nam, đồng thời các công nghệ deep learning khá mới mẻ với tác giả, chính vì thế

trong phạm vi thực hiện đề tài còn nhiều thiếu sót và dự định hướng phát triển trong thời gian tới.

- Xây dựng một hệ thống dự báo thời gian thực để ứng dụng vào thực tế.
- Tiếp cận hướng dự báo dựa vào sự kiện.
- Áp dụng thêm phương pháp phân tích kỹ thuật nhằm khai thác tối đa dữ liệu lịch sử giá đưa vào trong mô hình.
- Áp dụng các kỹ thuật tiền xử lý văn bản để giảm đặc trưng và các dữ liệu ngoại lai ảnh hưởng đến kết quả dự báo.
- Thu thập bộ dữ liệu tiếng Việt với số lượng lớn để tăng độ tin cậy.

5.3. Kết luận

Ngày nay, hướng dự báo chuyển động dựa vào tin tức tài chính ứng dụng các mô hình deep learning đang rất được các nhà nghiên cứu quan tâm, bằng chứng là rất nhiều nghiên cứu được công bố gần đây. Việc đề xuất và áp dụng mô hình BGRU trong phạm vi đề tài khóa luận đã mở ra một hướng tiếp cận khá mới mẻ và bước đầu có những kết quả đáng khích lệ khi áp dụng vào thị trường chứng khoán Việt Nam. Đề tài luận văn cơ bản đã giải quyết các mục tiêu đặt ra từ đầu. Song, vẫn còn nhiều điều cần phải cải tiến ở phía trước để có thể đưa vào ứng dụng thực tế.

TÀI LIỆU THAM KHẢO

[1] Agrawal, J. G., Chourasia, V. S., & Mittra, A. K. (2013), "State-of-the-art in stock prediction techniques", *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(4), 1360-1366.

[2] Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016, June), "Deep learning for stock prediction using numerical and textual information", *In Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference* on (pp. 1-6). IEEE.

[3] Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999), "Exploiting the past and the future in protein secondary structure prediction", *Bioinformatics*, 15(11), 937-946.

[4] Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011, July), "Identifying and following expert investors in stock microblogs", *In Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1310-1319). Association for Computational Linguistics.

[5] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014), "Learning phrase representations using RNN encoder-decoder for statistical machine translation". *arXiv preprint arXiv*:1406.1078.

[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014), "Empirical evaluation of gated recurrent neural networks on sequence modeling", *arXiv preprint arXiv*:1412.3555.

[7] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014, October), "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation", *In EMNLP* (pp. 1415-1425).

[8] Dũng, Phạm Xuân, and Hoàng Văn Kiếm. (2015). "Vietnamese Stock Market Prediction Using Text Mining", *Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin (FAIR)*

[9] Duong, D., Nguyen, T., & Dang, M. (2016, January), "Stock Market Prediction using Financial News Articles on Ho Chi Minh Stock Exchange". *In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication* (p. 71). ACM.

[10] Fama, E. F. (1965), "The behavior of stock-market prices", *The journal of Business*, 38(1), 34-105.

[11] Goldberg, Y. (2016), "A primer on neural network models for natural language processing", *Journal of Artificial Intelligence Research*, 57, 345-420.

[12] Guresen, E., Kayakutlu, G., & Daim, T. U. (2011), "Using artificial neural network models in stock market index prediction", *Expert Systems with Applications*, 38(8), 10389-10397.

[13] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014), "Dropout: a simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, 15(1), 1929-1958..

[14] Hochreiter, S. (1998), "The vanishing gradient problem during learning recurrent neural nets and problem solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02), 107-116.

[15] Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly, Tuong Vinh Ho. (2008, March), "A hybrid approach to word segmentation of Vietnamese texts*", In International Conference on Language and Automata Theory and Applications* (pp. 240-249). Springer Berlin Heidelberg.

[16] Huang, W., Nakamori, Y., & Wang, S. Y. (2005), "Forecasting stock market movement direction with support vector machine*", Computers & Operations Research*, 32(10), 2513-2522.

[17] Kaya, M. Y., & Karsligil, M. E. (2010, September), "Stock price prediction using financial news articles", *In Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on* (pp. 478-482). IEEE.

[18] Kim, K. J., & Han, I. (2000), "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", *Expert systems with Applications*, 19(2), 125-132.

[19] Krause, T., Noth, F., & Tonzer, L. (2016), "Brexit (probability) and effects on financial market stability"

[20] Längkvist, M., Karlsson, L., & Loutfi, A. (2014), "A review of unsupervised feature learning and deep learning for time-series modeling", *Pattern Recognition Letters,* 42, 11-24.

[21] Le-Hong, P., Roussanaly, A., Nguyen, T. M. H., & Rossignol, M. (2010, July), "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts*", In Traitement Automatique des Langues Naturelles-TALN 2010* (p. 12).

[22] Lo, A. W., & MacKinlay, A. C. (1988), "Stock market prices do not follow random walks: Evidence from a simple specification test", *Review of financial studies*, 1(1), 41-66.

[23] Malkiel, B. G. (1989), "Efficient market hypothesis", *The New Palgrave: Finance. Norton, New York*, 127-134.

[24] Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September), "Recurrent neural network based language model", *In Interspeech* (Vol. 2, p. 3).

[25] Peng, Y., & Jiang, H. (2015), "Leverage financial news to predict stock price movements using word embeddings and deep neural networks", *Proceedings of NAACL-HLT 2016*, pages 374–379,.

[26] Schuster, M., & Paliwal, K. K. (1997), "Bidirectional recurrent neural networks", *IEEE Transactions on Signal Processing*, 45(11), 2673-2681.

[27] Si, J., Mukherjee, A., Liu, B., Pan, S. J., Li, Q., & Li, H. (2014, October), "Exploiting Social Relations and Sentiment for Stock Prediction", *In EMNLP* (Vol. 14, pp. 1139-1145).

[28] Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011), "Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem", *Plant and soil*, 340(1-2), 7-24.

[29] W Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J., (2013), "Distributed representations of words and phrases and their compositionality". *In Advances in neural information processing systems* (pp. 3111-3119).

[30] W Bird, S. (2006, July), "NLTK: the natural language toolkit". *In Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69-72). Association for Computational Linguistics.

[31] Z Dey, R., & Salem, F. M. (2017). Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. arXiv preprint arXiv:1701.05923.

CÔNG TRÌNH CÔNG BỐ

[C1] 1 Paper được chấp nhận đăng tại Hội nghị quốc tế lần thứ VII về Information Science and Technology (ICIST 2017) sẽ được tổ chức tại Đà Nẵng, Việt Nam trong thời gian từ ngày 16- đến ngày 19 Tháng Tư, 2017.

PHỤ LỤC

A. Các khái niệm về thị trường chứng khoán

❖ ***Chứng khoán:***

Chứng khoán là bằng chứng xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với tài sản hoặc phần vốn của tổ chức phát hành. Chứng khoán được thể hiện bằng hình thức chứng chỉ, bút toán ghi sổ hoặc dữ liệu điện tử. Chứng khoán bao gồm các loại: cổ phiếu, trái phiếu, chứng chỉ quỹ đầu tư, chứng khoán phát sinh.

Chứng khoán là một phương tiện hàng hóa trừu tượng có thể thỏa thuận và có thể thay thế được, đại diện cho một giá trị tài chính. Chứng khoán gồm các loại: chứng khoán cổ phần (ví dụ cổ phiếu phổ thông của một công ty), chứng khoán nợ (như trái phiếu nhà nước, trái phiếu công ty...) và các chứng khoán phát sinh (như các quyền chọn, hợp đồng quy đổi, hợp đồng tương lai, hợp đồng kỳ hạn). Ở các nền kinh tế phát triển, loại chứng khoán nợ là thứ có tỷ trọng giao dịch áp đảo trên các thị trường chứng khoán. Còn ở những nền kinh tế nơi mà thị trường chứng khoán mới được thành lập, loại chứng khoán cổ phần lại chiếm tỷ trọng giao dịch lớn hơn.

Chứng khoán là một công cụ rất hữu hiệu trong nền kinh tế thị trường để tạo nên một lực lượng vốn tiền tệ khổng lồ tài trợ dài hạn cho các mục đích mở rộng sản xuất, kinh doanh của các doanh nghiệp hay các dự án đầu tư của Nhà nước và tư nhân.

Ví dụ:

- Một số cổ phiếu như: FPT, VNM, REE, …
- Các loại trái phiếu doanh nghiệp như: EVN, SHB, T&T Hà Nội, …
- Các chứng chỉ quỹ của các quỹ đầu tư như: VDF, VEH, ...

Cũng như các loại hàng hóa khác, chứng khoán là loại hàng hóa đặc biệt lưu thông trên thị trường riêng của nó: Thị trường chứng khoán.

❖ ***Sàn giao dịch chứng khoán***

Sàn giao dịch chứng khoán là một hình thức sàn giao dịch cung cấp các dịch vụ cho những người môi giới cổ phiếu và người mua bán cổ phiếu để trao đổi các cổ phiếu, trái phiếu và các loại chứng khoán khác. Sàn giao dịch chứng khoán cũng cung

cấp các dịch vụ cho việc phát hành và thu hồi chứng khoán cũng như các phương tiện tài chính và các sự kiện như việc chi trả lợi tức và cổ tức.

Chứng khoán được giao dịch trên sàn giao dịch chứng khoán gồm: các cổ phiếu do các công ty phát hành, các chứng chỉ quỹ và các sản phẩm hợp tác đầu tư và trái phiếu. Để có thể giao dịch trên một sàn giao dịch cổ phiếu, cổ phiếu cần phải được niêm yết ở đó.

Ví dụ: Sàn giao dịch chứng khoán thành phố Hồ Chí Minh (HoSE) được thành lập tháng 7 năm 2000, là một đơn vị trực thuộc Ủy ban Chứng khoán Nhà nước và quản lý hệ thống giao dịch chứng khoán niêm yết của Việt Nam. HoSE hoạt động như một công ty trách nhiệm hữu hạn một thành viên Nhà nước với số vốn điều lệ là một nghìn tỷ đồng. Hiện nay các sở giao dịch chứng khoán trên thế giới thường hoạt động dưới dạng công ty cổ phần.

Chức năng cơ bản của thị trường chứng khoán

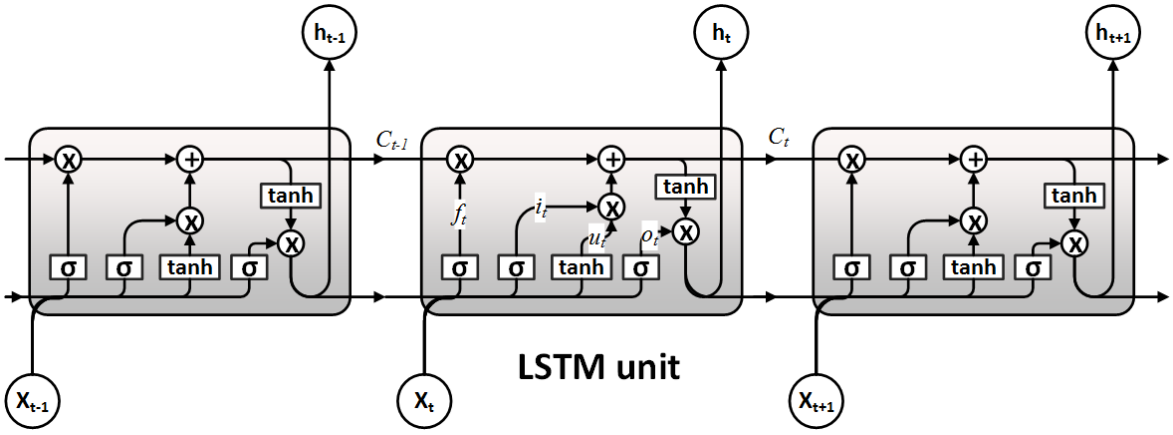
- Huy động vốn đầu tư cho nền kinh tế.
- Cung cấp môi trường đầu tư cho công chúng.
- Tạo môi trường giúp chính phủ thực hiện các chính sách kinh tế vĩ mô.
- Tạo tính thanh khoản cho các chứng khoán.
- Đánh giá hoạt động của các doanh nghiệp.

❖ ***Cổ phiếu***

Công ty cổ phần là một loại hình doanh nghiệp đặc biệt, vốn của nó được hình thành từ sự đóng góp vốn của rất nhiều người. Khi mới thành lập, công ty cổ phần chia vốn điều lệ thành những phần nhỏ bằng nhau, mỗi phần vốn là một cổ phần (Share), người góp vốn vào công ty qua việc mua cổ phần gọi là cổ đông (Shareholder), cổ đông nhận một giấy chứng nhận cổ phần gọi là cổ phiếu (Stock) và chỉ có công ty cổ phần mới phát hành cổ phiếu. Như vậy, Cổ phiếu là giấy chứng nhận cổ phần, nó xác nhận quyền sở hữu của cổ đông đối với công ty cổ phần.

B. Mạng Long Short Term Memmory (LSTM)

Mạng Long Short Term Memory thường được gọi là "LSTM" là một biến thể đặc biệt của RNN, cái mà có khả năng học và nắm bắt dữ liệu với những phụ thuộc dài hạn. LSTM được giới thiệu bởi Hochreiter và Schmidhuber từ năm 1997, tuy nhiên mãi đến những năm gần nay, các nhà nghiên cứu mới ứng dụng rộng rãi LSTM vào các lĩnh vực máy học và bất ngờ trước những kết quả đáng khích lệ. LSTM rõ ràng đã khắc phục được khuyết điểm của RNN trong việc nắm bắt các thông tin dài hạn. Trong phần tiếp theo, tác giả sẽ đi sâu để phân tích kiến trúc của LSTM để chỉ ra những đặc điểm giúp LSTM có thể nhớ dài hạn.



Minh họa mô hình mạng LSTM

Trong hình 2.4. Mỗi đơn vị (Unit) LSTM tại mỗi thời điểm t được biểu diễn bằng cách công thức dưới đây:

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)})$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)})$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)})$$

$$u_t = tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)})$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1}$$

$$h_t = o_t \odot \tanh(c_t)$$

Trong đó: σ và $tanh$ là hàm sigmoid và *tanh*, i_t được xem là *cổng nhập* (input gate) quyết định bao nhiêu thành phần mới được cập nhật vào bộ nhớ, f_t là *cổng quên* (forget gate) quyết định xem bao nhiêu thành phần bộ nhớ cũ sẽ được quên trước khi đưa vào bộ nhớ của trạng thái hiện tại. o_t là *cổng xuất* (output gate) kiểm soát lưu lượng bộ nhớ sẽ được cập nhật vào trạng thái bộ nhớ ở thời điểm *t*, u_t là ứng cử viên (candidate) cho cho lớp ẩn ở trạng thái *t*, c_t là trạng thái lớp ẩn hiện tại và h_t là giá trị xuất của lớp ẩn ở thời điểm *t*.