BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

PHẠM MINH TUẤN – 19469421
TRƯƠNG NGUYỄN DUY TÂN – 19485441

KHÓA LUẬN TỐT NGHIỆP
MÔ HÌNH PHÂN TÍCH XÚC CẢM NHIỀU KHÍA CẠNH CHO TIN TỨC THỊ TRƯỜNG CHỨNG KHOÁN

Chuyên ngành: Khoa học dữ liệu

Giảng viên hướng dẫn: TS. Nguyễn Chí Kiên

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 12, NĂM 2023

INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY

FACULTY OF INFORMATION TECHNOLOGY

PHAM MINH TUAN – 19469421

TRUONG NGUYEN DUY TAN – 19485441

GRADUATION THESIS

ASPECT-BASED SENTIMENT ANALYSIS FOR STOCK MARKET NEWS

Major: Data Science

Instructor: PhD. Nguyen Chi Kien

HO CHI MINH CITY, DECEMBER 2023

TÓM TẮT

Tiêu đề: Mô Hình Phân Tích Xúc Cảm Nhiều Khía Cạnh Cho Tin Tức Thị Trường Chứng Khoán

Tóm tắt:

Hiện nay, ngày càng nhiều người, kể cả những người có ít kinh nghiệm, đang tham gia đầu tư vào thị trường chứng khoán. Tuy nhiên, việc thiếu thông tin cần thiết và khả năng đánh giá chính xác tác động của tin tức đến thị trường tạo ra thách thức không nhỏ. Để hỗ trợ người dùng trong các quyết định đầu tư, chúng tôi đề xuất cung cấp thêm thông tin bằng cách phân tích các khía cạnh khác nhau của tin tức tài chính. Ảnh hưởng mạnh mẽ của tin tức đối với nền kinh tế và thị trường tài chính là điều không thể phủ nhận. Tin tức tích cực thường khuyến khích đầu tư và tạo lòng tin trong thị trường, trong khi tin tức tiêu cực có thể gây biến động và giảm niềm tin của nhà đầu tư. Việc lan truyền thông tin nhanh chóng qua phương tiện truyền thông và mạng xã hội đã làm tăng sự nhạy cảm của thị trường, tạo ra một môi trường đầu tư phức tạp hơn. Đó là lý do tại sao chúng tôi đã xây dựng một công cụ để xác định chính xác tác động của các khía cạnh khác nhau của tin tức kinh tế.

Phân tích tình cảm là một lĩnh vực nghiên cứu phổ biến của Xử lý ngôn ngữ tự nhiên. Để lưu giữ thông tin theo ngữ cảnh và ý nghĩa, chúng tôi đã xây dựng các mô hình học sâu bao gồm RNN, GRU, LSTM để dự đoán tác động của tin tức trên thị trường tài chính.

Khi thử nghiệm trên các mô hình dự đoán tác động của tin tức, chúng tôi nhận thấy rằng mô hình GRU mang lại kết quả đầy hứa hẹn nhờ đặc tính lưu trữ và học hỏi thông tin, giúp cải thiện kết quả dự đoán. Những đặc điểm đáng chú ý này của GRU góp phần nâng cao tính chính xác của kết quả dự đoán trong hệ thống của chúng tôi, cung cấp công cụ hỗ trợ đáng tin cậy hơn cho người dùng khi đưa ra quyết định đầu tư.

Từ khóa: Phân tích xúc cảm nhiều khía cạnh, Học sâu, RNN, GRU, LSTM.

CONTENT SUMMARY

Title: Aspect-Based Sentiment Analysis For Stock Market News

Abstract:

Nowadays, an increasing number of individuals, including those with limited knowledge and experience, are venturing into stock market investments. However, a challenge arises from the lack of necessary information and accurate assessment of news impacts on the market. To assist users in investment decisions, it is proposed to provide additional information by analyzing different aspects of financial news. The strong influence of news on the economy and financial markets is obvious. Positive news often encourages investment and boosts market confidence, while negative news can quickly cause volatility and reduce investor confidence. The rapid and powerful dissemination of information through media and social networks has increased market sensitivity, creating a complex investment environment. Therefore, we build a tool to accurately determine the impact of different aspects of economic news.

Sentiment Analysis is a popular research area of Natural Language Processing. To preserve contextual information and meaning, we have built deep learning models including RNN, GRU, LSTM to predict the impact of news on financial markets.

When testing on news impact prediction models, we found that the GRU model produced promising results because of its information storage and learning characteristics, helping to improve predicted results. These notable characteristics of GRU contribute to improving the accuracy of prediction results in our system, providing a more reliable support tool for users when making investment decisions.

Keywords: Aspect-based sentiment analysis, Deep learning, RNN, GRU, LSTM.

LỜI CẢM ƠN

Lời đầu tiên cho phép chúng em gửi lời cảm ơn chân thành đến TS Nguyễn Chí Kiên. Thầy là người đã trực tiếp giảng dạy, dẫn dắt, góp ý em trong khoảng thời gian học tập tại trường, nhờ thầy mà em có thể có cơ hội tiếp cận và thử sức với một đề tài khó nhưng thú vị như thế này, và cũng nhờ thầy mà em có thể có cơ hội hoàn thành tốt hơn bài báo cáo này.

Em xin cảm ơn thầy Bùi Thanh Hùng và thầy Nguyễn Hữu Tình. Cảm ơn hai thầy vì đã đồng ý nhận phản biện đề tài của em. Em tin rằng những đánh giá phản biện của hai thầy sẽ góp phần quan trọng trong việc hoàn thiện luận văn này.

Em cảm ơn thầy Nguyễn Hữu Tình, thầy là giáo viên chủ nhiệm lớp DHKHDL15A của em, là người thầy đã dõi theo chúng em từ năm hai đến hiện tại, đã giúp đỡ em rất nhiều trong quá trình định hướng bản thân khi lựa chọn chuyên ngành Khoa Học Dữ Liệu này, thầy là một người đã cảm hứng cho chúng em để em hiểu rõ hơn và hứng thú hơn trong quá trình tiếp cận ngành học còn mới mẻ này và những giá trị mà nó mang lại, để biết được rằng, bản thân cần phải làm gì để có hướng phát triển đúng đắn trong chuyên ngành mà em đã lựa chọn.

Thêm nữa, em cũng xin gửi lời cảm ơn đến quý thầy, cô ở Khoa Công Nghệ Thông Tin – Trường Đại học Công Nghiệp Thành phố Hồ Chí Minh đã tận tình giảng dạy, giúp chúng em có được những kiến thức nền tảng cần thiết trong ngành lập trình trong suốt quãng thời gian em học tập tại trường và hơn hết là để chúng em có thể hoàn thiện được bài báo cáo lần này.

Em cũng xin bày tỏ lòng biết ơn đến ban lãnh đạo của Trường Đại học Công Nghiệp Thành phố Hồ Chí Minh và các Khoa, Phòng ban chức năng đã trực tiếp hoặc gián tiếp giúp đỡ em trong suốt quá trình em học tập và thực hiện báo cáo này.

Vì những kiến thức thiếu sót cũng như còn nhiều hạn chế về thời gian và công cụ nên kết quả đồ án tốt nghiệp của chúng em không thể tránh khỏi những thiếu sót. Chúng em xin nhận những ý kiến góp ý từ quý thầy, cô cũng như các bạn để chúng em có thể hoàn thiện đề tài tốt hơn.

Chúng em xin chân thành cảm ơn!

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIÁO VIÊN HƯỚNG DẪN

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIÁO VIÊN PHẢN BIỆN

MỤC LỤC

MỤC LỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC TỪ VIẾT TẮT

TỪ NGỮ	Ý NGHĨA
ABSA	Aspect-Based Sentiment Analysis - Là một nhiệm vụ trong lĩnh vực NLP nhằm mục đích xác định và trích xuất xúc cảm theo các khía cạnh cụ thể của văn bản.
Batch size	Số lượng dữ liệu mỗi lần đưa vào mô hình cho đến hết tập train
Epochs	Số lần mô hình được học trên toàn bộ dữ liệu tập train
EPS	Earnings per share - Lợi nhuận sau thuế của công ty phân bổ trên một cổ phiếu thông thường đang được lưu hành ở trên thị trường.
GRU	Gated recurrent units
Hyperparameter	Siêu tham số
IPO	Initial Public Offering - Phát hành lần đầu ra công chúng
Loss	Giá trị mất mát - chủ yếu được tính toán dựa trên giá trị thực tế với giá trị dự đoán
LSTM	Long-Short Term Memory
M&A	Mergers and Acquisitions - Mua bán và sáp nhập
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
NLP	Natural Language Processing - Xử lý ngôn ngữ tự nhiên
Overfitting	Dấu hiệu của loss trên tập train quá lệch với tập validation
P/B	Price to Book ratio - Tỷ lệ được sử dụng để so sánh giá của một cổ phiếu với giá trị sổ sách của cổ phiếu đó
P/E	Price to Earning ratio - Chỉ số đánh giá mối quan hệ giữa giá thị trường của cổ phiếu (Price) với thu nhập trên một cổ phiếu (EPS)
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SA	Sentiment Analysis - Phân tích xúc cảm

TỪ NGỮ	Ý NGHĨA
Supervised Learning	Học có giám sát
Test	Tập dữ liệu dùng cho việc đánh giá mô hình đã được huấn luyện
Tỉ lệ Dropout	Là tỉ lệ bỏ qua những đặc trưng lớp ẩn trước cho lớp ẩn tiếp theo
Train	Tập dữ liệu dùng cho việc huấn luyện mô hình
Unique	Giá trị duy nhất không bị trùng nhau
Validation	Tập dữ liệu dùng để đánh giá mô hình trong quá trình huấn luyện

CHƯƠNG 1. GIỚI THIỆU

1.1 Tổng quan

Trong chương này, chúng tôi sẽ trình bày sơ lược qua về bối cảnh chọn đề tài, lý do chọn đề tài, mục tiêu và phạm vi của nghiên cứu, ý nghĩa khoa học và thực tiễn mà đề tài.

1.1.1 Bối cảnh

Trong thời đại ngày nay, việc tham gia vào môi trường kinh doanh tài chính và đưa ra quyết định đầu tư dễ dàng hơn đối với người dùng, ngay cả khi họ chưa có quá nhiều kiến thức cũng như kinh nghiệm trong lĩnh vực, môi trường này. Một trong những vấn đề mà họ gặp phải trong trường hợp này là thiếu đi những thông tin cần thiết cũng như việc đánh giá chính xác các ảnh hưởng của các thông tin đến thị trường. Để hỗ trợ người dùng trong việc đưa ra quyết định đầu tư, chúng tôi đề xuất cung cấp thêm thông tin cho người dùng bằng cách phân tích xúc cảm nhiều khía cạnh trong tin tức tài chính. Để dự đoán mức độ tác động của một tin tức, bài báo đối với các khía cạnh tài chính được nhắc đến cần thực hiện một quá trình phân tích cẩn thận về nội dung bài viết. Bao gồm việc xác định thông tin chính, nguồn tin, ngữ cảnh thị trường và các tài sản tài chính khác liên quan hay được đề cập, nhắc đến trong nội dung bài báo, tin tức. Điều này giúp người đọc đánh giá được mức độ quan trọng của các thông tin trong bài báo đối với việc định hình quyết định đầu tư và các hành động có liên quan tác động đến thị trường tài chính.

Hiện nay, sự phổ biến của công nghệ và Internet đã thay đổi cách thức người dùng tiếp cận với các thông tin kinh tế. Họ có thể dễ dàng truy cập, tìm kiếm các thông tin từ nhiều nguồn khác nhau thông qua các công cụ, thiết bị di động, mạng xã hội và các trang web tin tức. Theo baochinhphu.vn, số liệu từ Trung tâm lưu ký Chứng khoán Việt Nam (VSD), lũy kế cả năm 2022, nhà đầu tư cá nhân trong nước đã mở mới gần 2,6 triệu tài khoản chứng khoán. Đây là con số kỷ lục trong 22 năm hoạt động [1]. Tuy nhiên cùng với sự tiện lợi và phổ biến đó, là việc có quá nhiều thông tin mà người dùng cần tiếp nhận đòi hỏi người dùng phải có sự hiểu biết và một lượng kiến thức nhất định về tài chính để có thể hiểu được chính xác những tác động của

những thông tin trên đối với nền kinh tế, đa số các nhà đầu tư mới chỉ tập trung vào phân tích kỹ thuật, điều này khiến nhà đầu tư không nắm rõ thông tin về cổ phiếu đang đầu tư và dẫn đến những phán đoán thiếu cơ sở. Để tránh rủi ro này, cần phải áp dụng kết hợp thông minh cả phân tích kỹ thuật và phân tích cơ bản [17], sử các nguồn thông tin truyền thống và trực tuyến để có cái nhìn tổng quan và đảm bảo tính chính xác của các thông tin kinh tế để đưa ra các quyết định đúng đắn.

Các tác động, ảnh hưởng mạnh mẽ của các tin tức đối với nền kinh tế, thị trường tài chính. Điển hình như các tin tức tích cực thường thúc đẩy sự đầu tư và tạo tin tưởng trong thị trường, trong khi đó tin tức tiêu cực dễ dàng có thể gây nên các biến động và làm giảm lòng tin của nhà đầu tư đối với thị trường. Sự lan truyền của thông tin qua các phương tiện truyền thông và mạng xã hội diễn ra một cách nhanh chóng và mạnh mẽ đã làm cho thị trường trở nên nhạy cảm hơn với sự biến động và tạo ra môi trường đầu tư phức tạp. Vì thế, việc xây dựng một công cụ để có thể xác định chính xác các tác động, ảnh hưởng của từng khía cạnh trong tin tức kinh tế đóng vai trò rất quan trọng trong việc hỗ trợ hình thành quyết định đầu tư và quản lý rủi ro trong môi trường kinh doanh hiện nay.

1.1.2 Lý do chọn đề tài

Như đã đề cập ở bối cảnh trước đó, do sự dễ dàng tiếp cận thị trường đầu tư, chúng tôi mong muốn mang đến một công cụ hữu ích để giúp, hỗ trợ người dùng dễ dàng hơn trong việc tiếp cận và đưa ra những quyết định đầu tư hợp lý.

Số lượng thông tin các bài báo kinh tế hiện tại có rất nhiều tuy nhiên việc phân tích các khía cạnh để khai thác các ảnh hưởng của bài báo đấy đến thị trường lại ít được phổ biến và khai thác đối với các tin tức kinh tế ở Việt Nam.

1.2 Mục tiêu nghiên cứu

- Tìm hiểu về kiến trúc mô hình Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long-Short Term Memory (LSTM) trong bài toán xử lý ngôn ngữ tự nhiên.
- Tìm hiểu về các mô hình xử lý ngôn ngữ tự nhiên (NLP) áp dụng các mô hình học máy (RNN, GRU, LSTM) cho nhiệm vụ xử lý ngôn ngữ tiếng Việt ở lĩnh vực tài chính.

- Tìm hiểu về kỹ thuật Supervised Learning.
- Tìm hiểu về các công cụ hỗ trợ chat GPT.
- Áp dụng kết hợp công cụ chat GPT trong việc xử lý nhãn dữ liệu, sau đó đưa vào mô hình LSTM để thực hiện quá trình huấn luyện đối với các tin tức của các bài báo kinh tế ở Việt Nam.
- Đề xuất phương pháp xây dựng mô hình dự đoán mức độ tác động tin tức tài chính trên nhiều khía cạnh để giải quyết vấn đề cung cấp thêm thông tin từ tin tức tài chính cho quyết định đầu tư.

1.3 Phạm vi nghiên cứu

- Kiến thức và hiểu biết về các phương pháp phân tích thống kê để áp dụng trong việc xử lý dữ liệu.
- Kiến thức và hiểu biết về các mô hình Recurrent Neural Network, Gated Recurrent Unit, Long-Short Term Memory.
- Nguồn dữ liệu được sử dụng để nghiên cứu được thu thập từ các trang báo về tin tức kinh tế của các công ty hoạt động ở Việt Nam cũng như các công ty có ảnh hưởng đến thị trường Việt Nam.

1.4 Ý nghĩa khoa học và thực tiễn

- Ý nghĩa khoa học: đề xuất mô hình phân tích các khía cạnh của bài báo.
- Ý nghĩa thực tế: cung cấp giải pháp giúp nhà đầu tư có cái nhìn tổng quan và rõ ràng hơn về các khía cạnh và ảnh hưởng của các khía cạnh đó trong bài báo, từ đó hỗ trợ đưa ra quyết định cho nhà đầu tư.
- Mở rộng: Nghiên cứu này góp phần làm tiền đề cho nghiên cứu về bài toán ABSA trong tin tức & dự đoán tài chính.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Bài toán

Chúng tôi tiến hành trình bày tổng quát về bài toán xử lý ngôn ngữ tự nhiên về phân tích xúc cảm trong lĩnh vực tài chính đối với ngôn ngữ tiếng Việt.

2.1.1 Khái niệm

Phân tích xúc cảm (SA) là nhiệm vụ phân loại nhãn/dự đoán giá trị xúc cảm dựa theo một đoạn văn bản. Ví dụ, một đoạn văn bản bình luận trên mạng xã hội có thể được phân loại thành nhãn "tích cực", "tiêu cực", "bình thường" hay một giá trị thực cụ thể trong khoảng từ -1 đến 1 [16].

2.1.2 Các nghiên cứu trước đó

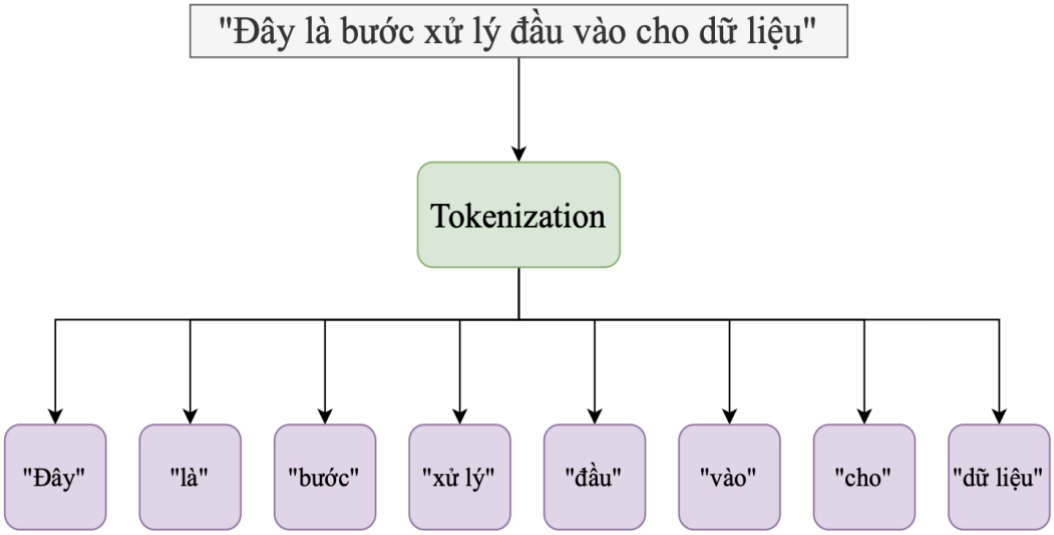
Nhiệm vụ nghiên cứu và phân tích xúc cảm (SA) trong văn bản hiện nay đóng vai trò hết sức quan trọng và rất cần thiết trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và phát triển trí tuệ nhân tạo. Mục tiêu là trích xuất thông tin có giá trị liên quan đến các khía cạnh được đề cập trong nhận xét của người dùng. Vấn đề này có thể được chia thành ba nhiệm vụ phụ: trích xuất thuật ngữ, phát hiện khía cạnh và phát hiện phân cực. Ở nghiên cứu của Minh-Hao Nguyen và cộng sự đã thực hiện đối với hai nhiệm vụ phụ là phát hiện khía cạnh và phát hiện phân cực trong nhận xét của người dùng [14]. Hay ở nghiên cứu của Quang-Linh Tran và cộng sự đã sử dụng các mô hình học sâu như Bi-GRU, Bi-LSTM để xây dựng mô hình phân loại khía cạnh của đánh giá và phân loại cảm tính của từng khía cạnh trong lĩnh vực thương mại điện tử đối với các đánh giá sản phẩm của người dùng [18]. Còn đối với lĩnh vực tài chính, ở nghiên cứu của Hitkul Jangid và cộng sự cũng đã sử dụng các mô hình học sâu như LSTM để phân tích khía cạnh trong văn bản, nhưng có sự chọn lọc trong các khía cạnh để tập trung vào một lĩnh vực cụ thể [8]. Tuy nhiên, đối với tiếng Việt, các mô hình dùng cho phân tích các khía cạnh tài chính vẫn chưa được áp dụng rộng rãi mặc dù lượng thông tin tài chính ở Việt Nam rất phổ biến và số lượng người dùng đầu tư vào thị trường tài chính ngày càng tăng. Vì thế, sẽ rất hứa hẹn khi áp dụng một mô hình học sâu để có thể phân tích được những khía cạnh, yếu tố ảnh hưởng của các tin

tức tài chính ở Việt Nam, giúp hỗ trợ người dùng trong việc đưa ra các quyết định trong đầu tư.

2.2 Kỹ thuật

2.2.1 Tokenization

Tokenization là quá trình chia nhỏ văn bản thành các đơn vị được gọi là "token", tương ứng với mỗi token có thể là một từ, một cụm từ hay đoạn văn tuỳ vào cách thực hiện của tokenization. Đây là một quá trình quan trọng và cần thiết để chuẩn bị xây dựng đầu vào cho mô hình máy học trong nhiệm vụ xử lý ngôn ngữ tự nhiên [9], cho phép hệ thống có thể hiểu được và xử lý một cách hiệu quả hơn.



Hình 1. Mô tả cách thức hoạt động của tokenization.

2.2.2 Stop Words Removal

Kỹ thuật Stop Words Removal là quá trình loại bỏ các từ không có quá nhiều ý nghĩa trong văn bản, các từ ngữ phổ biến không mang lại, đóng góp nhiều thông tin. Ví dụ: và", "hay", "hoặc", "nếu",... Mục tiêu của việc áp dụng kỹ thuật trên là để cải thiện hiệu quả của mô hình bằng cách đào tạo tập trung vào các từ khoá hiệu quả hơn và cũng để giảm kích thước của tập dữ liệu đào tạo.

Đây là một kỹ thuật phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên, tuy nhiên để tạo và sử dụng một bộ stop words một cách hiệu quả, cần phải xác định dựa trên ngữ cảnh bài toán cũng như mục tiêu xử lý của dữ liệu để góp phần tối ưu hoá kết quả đạt được [2].

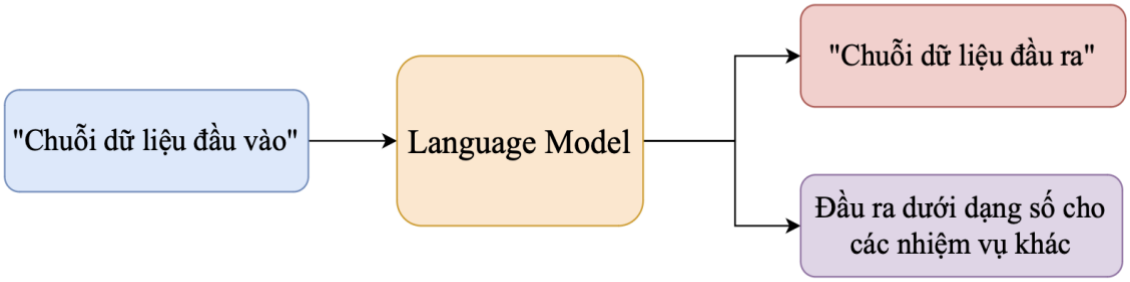
2.2.3 Word Embeddings

Word Embeddings là một kỹ thuật trong lĩnh vực xử lý ngôn ngữ tự nhiên dùng để biểu diễn các từ vựng dưới dạng vector trong không gian nhiều chiều, nó giúp máy tính hiểu được cách mà các từ ngữ tương tác với nhau trong văn bản. Đối với bước này, nhóm sử dụng Fasttext cho ngôn ngữ tiếng Việt để hỗ trợ xử lý [6].

Với điểm mạnh là có thể lưu giữ các đặc trưng của từ ngữ khi các từ có ngữ nghĩa tương tự sẽ được sắp xếp gần với nhau trong không gian véc-tơ, nên nó được ứng dụng rất nhiều trong việc tối ưu các mô hình học máy về xử lý ngôn ngữ tự nhiên [7].

2.2.4 Language Modeling

Language Modeling là quá trình mô hình hóa trong xử lý ngôn ngữ tự nhiên, có nhiệm vụ dự đoán các xác suất của từ hay cụm từ. Mục tiêu của kỹ thuật này chính là để máy có thể học được cấu trúc, quy luật và logic trong ngôn ngữ tự nhiên để có thể dự đoán kết quả đầu ra tương ứng với yêu cầu của người dùng.



Hình 2. Mô tả cách thức hoạt động của kỹ thuật Language Modeling.

2.2.5 Min-Max Scaler

Là một phương pháp để chia tỷ lệ dữ liệu, trong đó giá trị tối thiểu được thực hiện bằng 0 và giá trị tối đa bằng một. Min-Max Scaler thu nhỏ dữ liệu trong phạm vi đã cho, thường từ 0 đến 1. Nó chuyển đổi dữ liệu bằng cách mở rộng các giá trị đến một phạm vi nhất định. Nó chia tỷ lệ các giá trị thành một phạm vi giá trị cụ thể mà không thay đổi hình dạng của phân phối ban đầu. Việc chia tỷ lệ Min-Max được thực hiện bằng cách sử dụng:

$$x_{std} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$x_{scaled} = x_{std} * (max - min) + min$$

Trong đó:

min, max : khoảng giá trị.

x_{min} : giá trị tối thiểu ban đầu

x_{max} : giá trị tối đa ban đầu

2.3 Phương pháp đánh giá

Các phương pháp đánh giá hiệu suất mô hình hồi quy phổ biến như MSE, RMSE, MAE [12].

Ở các phương pháp đánh giá y_i, $\hat{y}_i$ tương ứng là các giá trị thực tế và giá trị mà mô hình dự đoán được, n là số lượng quan sát của mô hình.

2.3.1 MSE

Sai số bình phương trung bình - MSE (Mean Square Error) của phép ước lượng là trung bình của bình phương các sai số, là sự khác giữa kết quả ước lượng được với những kết quả thực tế được đánh giá.

$$MSE = \frac{1}{n}\Sigma(y_i - \hat{y}_i)^2$$

2.3.2 RMSE

Sai số bình phương trung bình gốc - RMSE (Root Mean Square Error) là phương pháp đo lường và đánh giá mô hình hồi quy dựa trên độ lệch chuẩn của phần dư (lỗi dự đoán). Phần dư này là khoảng cách giữa các điểm dữ liệu đến đường hồi quy, RMSE là thước đo độ phân tán của các điểm dư này.

$$RMSE = \sqrt{\frac{1}{n}\Sigma(y_i - \hat{y}_i)^2}$$

2.3.3 MAE

MAE - Mean Absolute Error là phương pháp đo lường đánh giá mô hình hồi quy dựa trên trung bình tổng của các trị tuyệt đối giữa giá trị dự đoán và giá trị thực tế

$$MAE = \frac{1}{n}\Sigma|y_i - \hat{y}_i|$$

2.4 Phương pháp tối ưu

2.4.1 Grid Search

Grid Search là một thuật toán trong Machine Learning được áp dụng trong việc huấn luyện mô hình nhằm để tìm kiếm và tối ưu các tham số cho một mô hình học máy. Nó thường được sử dụng khi người dùng có một không gian các siêu tham số lớn, lúc này Grid Search sẽ xác định một tập hợp các giá trị của các siêu tham số, sau đó tạo ra các kết hợp có thể giữa các giá trị này. Mỗi kết hợp sẽ được áp dụng để huấn luyện mô hình và đánh giá bằng các phép đo hiệu suất như: accuracy, F1 score,... Tuỳ thuộc vào bài toán cụ thể để tìm ra giá trị tối ưu nhất dựa trên các phép đo. Vì vậy phương pháp này thường được áp dụng trong thực tế khi người dùng muốn tối ưu hoá các tham số trong mô hình của mình [3].

2.4.2 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) là một biến thể của thuật toán Gradient Descent [19] được sử dụng để tối ưu hóa các mô hình học máy. Nó khắc phục được những điểm kém hiệu quả trong tính toán của các phương pháp Gradient Descent truyền thống khi xử lý các bộ dữ liệu lớn trong các dự án học máy.

Trong SGD, thay vì sử dụng toàn bộ tập dữ liệu cho mỗi lần lặp, chỉ một ví dụ đào tạo ngẫu nhiên duy nhất được chọn để tính độ dốc và cập nhật các thông số mô hình. Lựa chọn ngẫu nhiên này giới thiệu tính ngẫu nhiên vào quá trình tối ưu hóa.

2.4.3 Adam

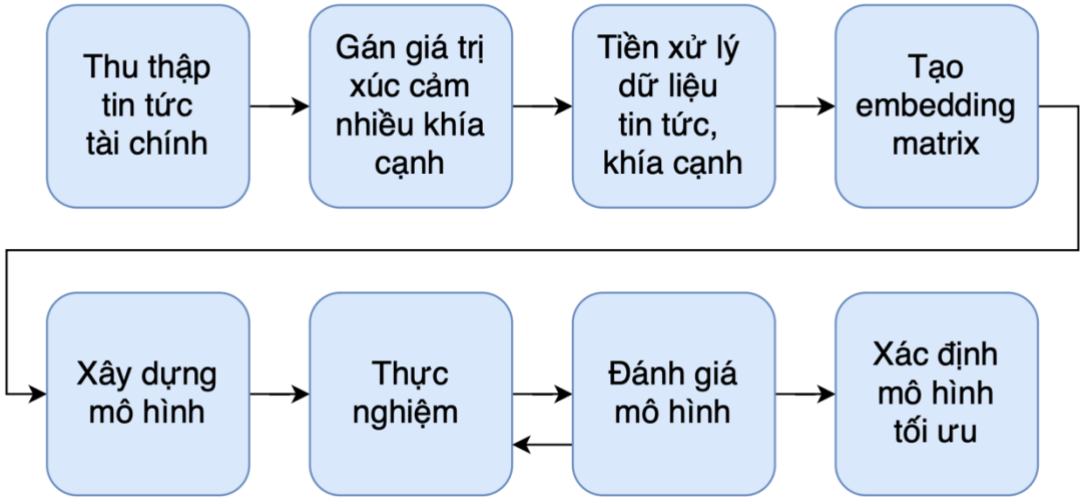
Thuật toán tối ưu hóa Adam là một phương pháp giảm gradient ngẫu nhiên dựa trên ước tính thích ứng của các khoảnh khắc bậc nhất và bậc hai.

Theo Kingma và cộng sự [4], phương pháp này "hiệu quả về mặt tính toán, có ít yêu cầu bộ nhớ, bất biến so với thay đổi kích thước chéo của gradient và rất phù hợp với các vấn đề lớn về dữ liệu/tham số". Vì vậy đây cũng là một thuật toán tối ưu phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và xử lý ảnh [11].

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN VÀ MÔ HÌNH ĐỀ XUẤT

3.1 Phương pháp thực hiện

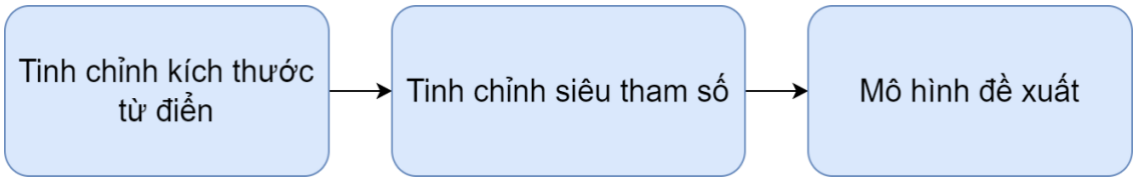
Khi nhóm đã xác định được bài toán cụ thể, sau đây nhóm sẽ đưa ra các bước cần thiết để giải quyết được bài toán này:



Hình 3. Các bước thực hiện để giải quyết bài toán.

3.2 Tổng quan thực nghiệm

Từ các bước thực hiện trên, nhóm sẽ đi sâu hơn vào bước thực nghiệm đã đề cập. Chúng tôi có dữ liệu nội dung bài báo "Content" như đầu vào và các dữ liệu về khía cạnh cần dự đoán là đầu ra được chia thành 3 tập train, test, validation theo tỉ lệ tương ứng 70% - 20% - 10% dùng trong các quá trình huấn luyện mô hình, đánh giá mô hình. Để thực nghiệm hiệu quả, nhóm nghiên cứu đã trao đổi với nhau và đưa ra kế hoạch thực nghiệm như sau:



Hình 4. Tổng quan kế hoạch thực nghiệm.

Khi đã có một kế hoạch thực nghiệm, chúng tôi tận dụng các công cụ có sẵn như:

- Python: Ngôn ngữ lập trình chính sử dụng xuyên suốt cả đề tài.

- Pytorch: Dùng cho việc xây dựng các cấu trúc mô hình RNN, GRU, LSTM và tạo các phương pháp đánh giá ở Mục 4.2 .

- Wandb: Một công cụ tiện ích cho việc quan sát kết quả thực nghiệm, hỗ trợ xác định bộ siêu tham số điều chỉnh tối ưu cho mô hình.

- Kaggle: Môi trường chính phục vụ cho việc chạy các mô hình Pytorch, dùng công cụ Wandb thông qua ngôn ngữ lập trình Python.

Hình 5. Các công cụ hỗ trợ thực nghiệm.

3.3 Mô hình đề xuất

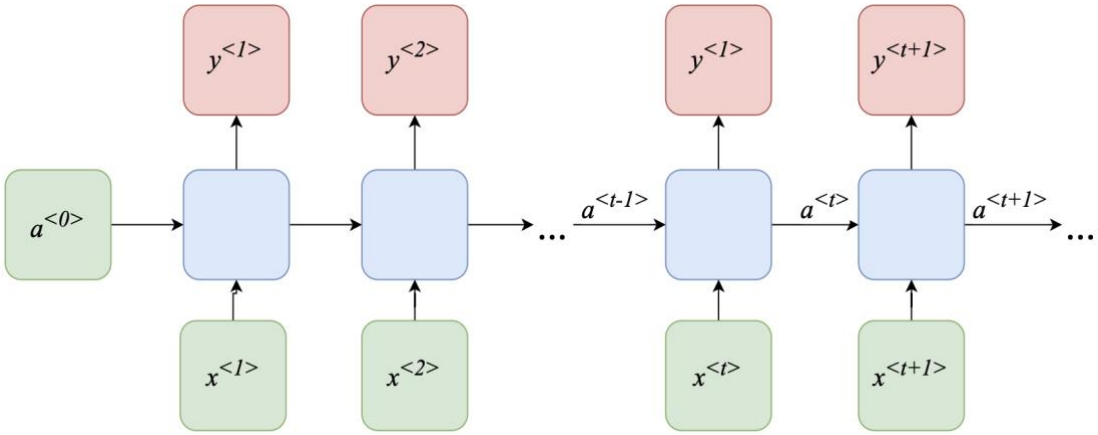
3.3.1 Mô hình RNN

Mô hình RNN (Recurrent Neural Network) được gọi là mô hình hồi quy (Recurrent) bởi vì chúng thực hiện tác vụ tuần tự cho từng phần tử của một chuỗi với đầu vào và đầu ra phụ thuộc vào các phép tính trước đó. Nói cách khác, RNN có khả năng nhớ được các thông tin tính toán trước để dự đoán cho bước hiện tại [5].

RNN được sử dụng tại các đơn vị mô hình hoá theo trình tự, việc có các kết nối tuần hoàn khiến nó mạnh mẽ hơn trong việc mô hình hoá các đầu vào của dữ liệu dạng chuỗi. Vì thế nó thường được sử dụng cho các nhiệm vụ dán nhãn và dự đoán trình tự trong các bài toán xử lý ngôn ngữ tự nhiên [10].

Các mạng neural hồi quy, còn được biết đến như là RNNs, là một lớp của mạng neural cho phép đầu ra được sử dụng như đầu vào trong khi có các trạng thái ẩn.

Thông thường cấu trúc mô hình có dạng tương tự như sau [21]:



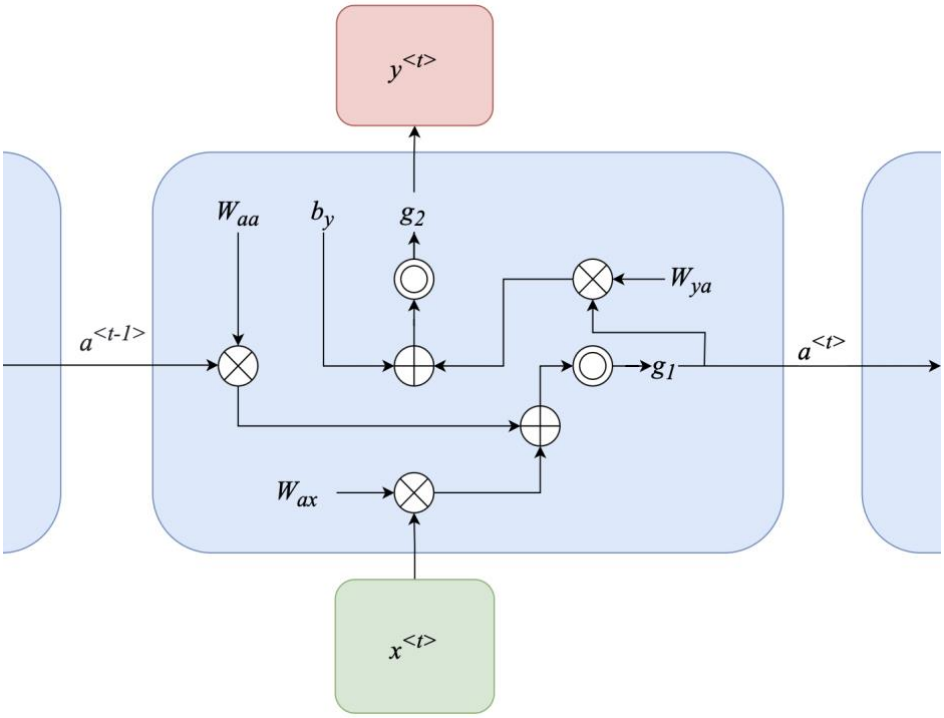
Hình 6. Cấu trúc của mô hình RNN [21].

Tại mỗi bước t, giá trị kích hoạt $a^{<t>}$ và đầu ra $y^{<t>}$ được biểu diễn như sau:

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

Với W_{ax},W_{aa},W_{ya},b_a,b_y là các hệ số được chia sẻ tạm thời và $g1$,$g2$ là các hàm kích hoạt.



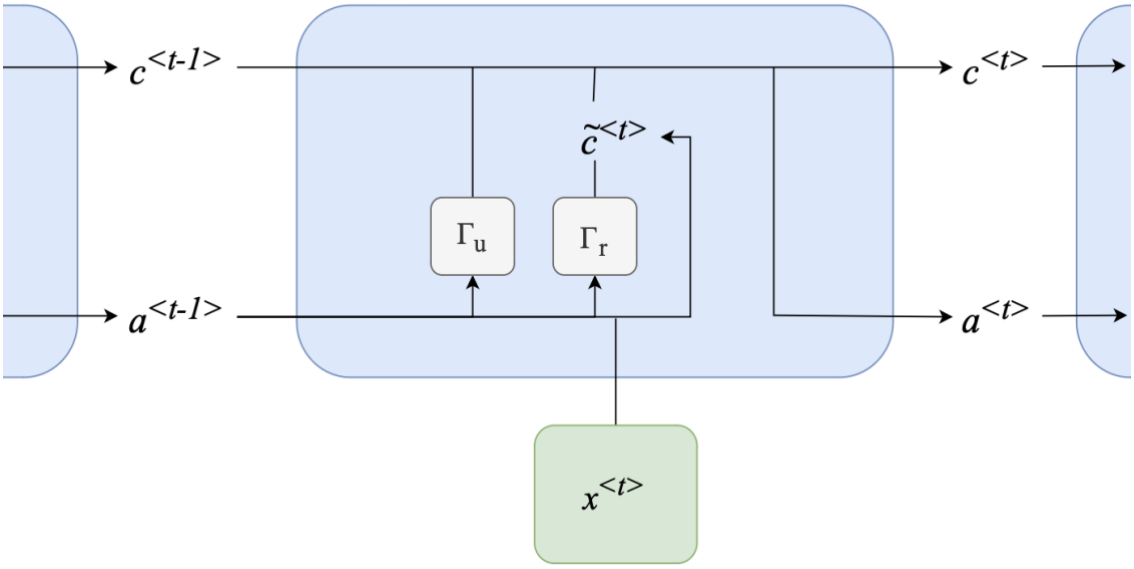
Hình 7. Cấu trúc của 1 đơn vị trong mô hình RNN [21].

3.3.2 Mô hình GRU

Mô hình GRU (Gated Recurrent Unit) là một trong những kiến trúc mạng nơ-ron hồi quy (RNN) phổ biến được sử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên và dữ liệu chuỗi thời gian. GRU được thiết kế để giải quyết các vấn đề về mất mát thông tin dài hạn trong quá trình huấn luyện mạng nơ-ron hồi quy truyền thống. Quy trình làm việc của GRU giống như RNN nhưng sự khác biệt nằm ở các hoạt động bên trong đơn vị GRU. Mô hình GRU cũng tương tự LSTM, cũng là mô hình RNN được thiết kế dùng để xử lý vấn đề mất mát thông tin dài hạn nhưng thường có cấu trúc đơn giản và sử dụng ít tham số hơn trong mỗi đơn vị, nên có thể khiến việc huấn luyện trở nên nhanh hơn.

Tương tự như RNN, mô hình GRU cũng hỗ trợ rất tốt trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), đáng chú ý hơn là nó còn xử lý tốt vấn đề học tập tuần tự và vanishing gradient descent trong mạng RNN tiêu chuẩn [15].

Mô hình GRU có cấu trúc bao gồm hai cổng chính: cổng cập nhật (update gate) và cổng đặt lại (reset gate). Các cổng này giúp điều chỉnh quá trình truyền thông tin thông qua các bước thời gian trong mạng nơ ron. Cổng cập nhật quyết định thông tin nào nên được cập nhật từ các bước thời gian trước đó và cổng đặc lại quyết định thông tin nào nên được bỏ qua. Mỗi cổng có trọng lượng (weight) và thành kiến (biases) riêng (nhưng trọng lượng (weight) và thành kiến (biases) cho tất cả các nút trong một lớp đều giống nhau) [13].



Hình 8. Cấu trúc của 1 đơn vị trong mô hình GRU [21].

Các phương trình đặc trưng của kiến trúc:

$$c^{\sim<t>} = tanh(W_c[\Gamma_r \star a^{<t-1>}, x^{<t>}] + b_c)$$

$$c^{<t>} = \Gamma_u \star c^{\sim<t>} + (1 - \Gamma_u) \star c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$

Trong đó:

$c^{<t>}$: trạng thái ẩn của đơn vị tại thời điểm t

$a^{<t>}$: giá trị kích hoạt của đơn vị tại thời điểm t

Γ_u : Cổng cập nhật

Γ_r : Cổng relevance

3.3.3 Mô hình LSTM

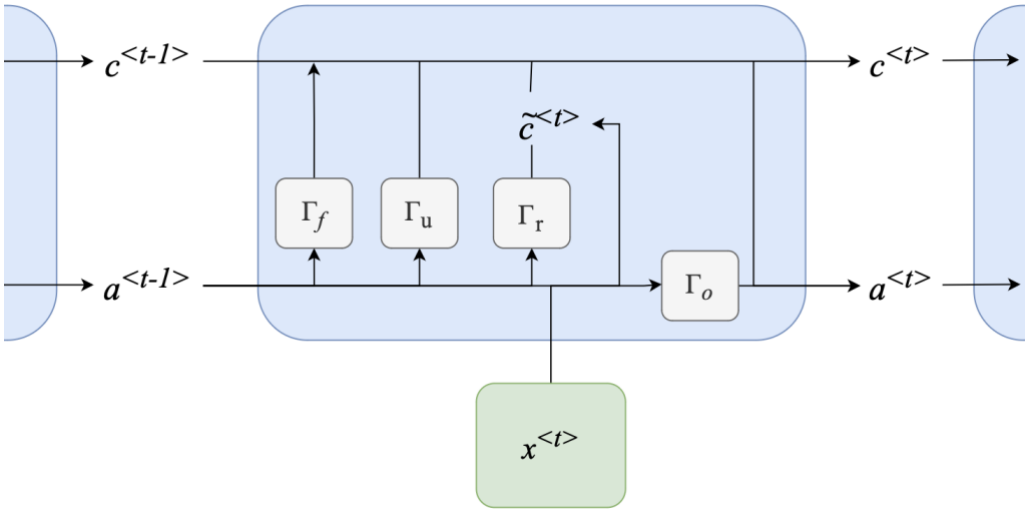
Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), hay thường được gọi là LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997) - là một mạng thần kinh hồi quy (RNN) được sử dụng trong lĩnh vực học sâu, LSTM là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa, giúp LSTM không chỉ xử lý các điểm dữ liệu đầu vào đơn lẻ mà còn xử lý được toàn bộ chuỗi dữ liệu. Trong nhiều tình huống ví dụ như các bài toán NLP cần xử lý những từ ngữ trong một đoạn

văn bản dài buộc ta phải sử dụng nhiều ngữ cảnh hơn để suy luận ra câu trả lời phù hợp với ngữ cảnh trước đó trong câu, đoạn văn. Và với yêu cầu về khoảng cách ngày càng lớn dần thì RNN đã bắt đầu không thể nhớ và học được nữa. Còn đối với LSTM việc phải nhớ những thông tin với một khoảng cách và thời gian dài là đặc tính mặc định của mô hình, không phải qua quá trình huấn luyện để mô hình có thể nhớ được và hoạt động tốt, mô hình LSTM hỗ trợ việc ghi nhớ mà không cần bất kỳ can thiệp nào [20].

Cũng giống như RNN và GRU, nhưng với đặc tính của mô hình, việc ghi nhớ và mô hình hoá những thông tin trong chuỗi đầu vào giúp LSTM thực hiện hiệu quả đặc tính của mình trong nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP) [8].

Một hệ thống mạng LSTM thường bao gồm các cell, một cổng vào (input gate), một cổng ra (output gate) và một cổng quên (forget gate). Cell có nhiệm vụ ghi nhớ các giá trị trong các khoảng thời gian tùy ý và ba cổng (input gate, output gate, forget gate) sẽ điều chỉnh luồng thông tin vào và ra khỏi cell.

Một mạng RNN tiêu chuẩn sẽ có kiến trúc bao gồm một tầng ẩn là hàm tanh, còn đối với LSTM cũng có một chuỗi tương tự như thế nhưng có phần khác biệt hơn ở cấu trúc của phần kiến trúc lặp lại. Thay vì chỉ có một tầng đơn như RNN, LSTM có đến 4 tầng ẩn (bao gồm 3 tầng sigmoid và 1 tầng tanh) tương tác với nhau theo một cấu trúc đặc biệt.



Hình 9. Cấu trúc của 1 đơn vị trong mô hình LSTM [21].

Các phương trình đặc trưng của kiến trúc:

$$c^{\sim<t>} = tanh(W_c[\Gamma_r \star a^{<t-1>}, x^{<t>}] + b_c)$$

$$c^{<t>} = \Gamma_u \star c^{\sim<t>} + \Gamma_f \star c^{<t-1>}$$

$$a^{<t>} = \Gamma_o \star c^{<t>}$$

Trong đó:

$c^{<t>}$: trạng thái ẩn của đơn vị tại thời điểm t

$a^{<t>}$: giá trị kích hoạt của đơn vị tại thời điểm t

Γ_u : Cổng cập nhật

Γ_r : Cổng relevance

Γ_f : Cổng quên

Γ_o : Cổng ra

CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT QUẢ

4.1 Dữ liệu

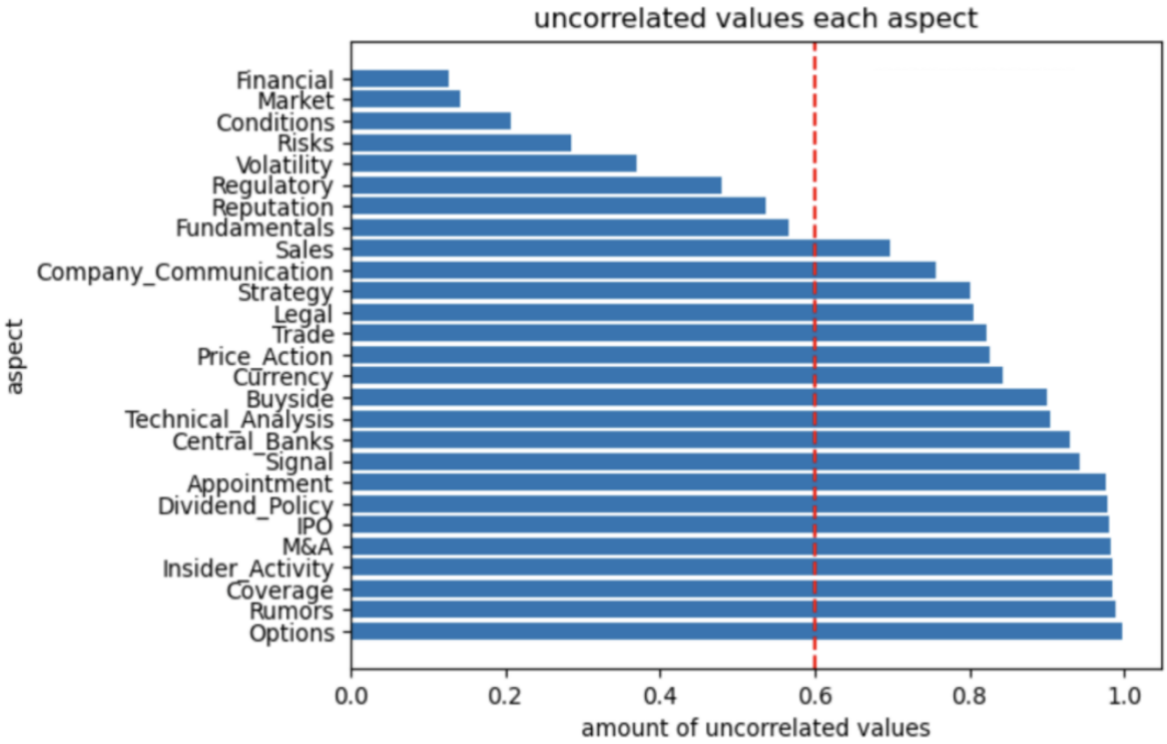
4.1.1 Giai đoạn thu thập dữ liệu

Giai đoạn thu thập dữ liệu, chúng tôi sử dụng công cụ selenium để tiến hành thu thập dữ liệu về tin tức tài chính trên thị trường của các công ty từ nhiều nguồn thông tin tài chính khác nhau như: cafef.vn, vtv.vn, tpo.vn, vnexpress.net, dantri.com.vn, nhandan.vn, baomoi.com, vietnamplus.vn. Lý do chúng tôi lựa chọn các nguồn này cho quá trình thu thập dữ liệu của mình do đây là các trang thông tin cập nhật liên tục, thường xuyên và đầy đủ các thông tin, tin tức tài chính của doanh nghiệp cần thiết để giải quyết cho vấn đề mà chúng tôi đã đặt ra. Về mặt lý thuyết, phương pháp quan sát là phương pháp thu thập dữ liệu bằng cách ghi lại có kiểm soát về các thông tin có hoặc không ảnh hưởng đến tổ chức, doanh nghiệp. Chúng tôi dựa theo các tin tức được đưa lên hằng ngày, lấy chủ đề những tin tức đó làm trọng tâm để tìm đến những bài báo cũng đề cập tin tức tương tự. Cụ thể trong trường hợp này là quan sát về các tin tức và các khía cạnh được đề cập đến trong từng tin tức để đưa ra quyết định phù hợp về tin tức đối với tổ chức, doanh nghiệp.

4.1.2 Mô tả khái quát bộ dữ liệu

Bộ dữ liệu ban đầu của chúng tôi bao gồm 27 khía cạnh tương ứng với nhiệm vụ phân tích các khía cạnh trong tài chính, tham khảo từ các nghiên cứu trước đó về ABSA [8]. Điểm khác biệt ở đây là bài toán nhóm nghiên cứu cho tiếng Việt. Điều này dẫn đến nhóm cần phải tạo được bộ dữ lisệu có 27 khía cạnh tương ứng với những bài báo tiếng Việt nhóm thu thập được. Dựa theo nguồn lực sẵn có và thời gian được phép làm khóa luận mà nhóm đưa ra quyết định tận dụng API của OpenAI để sử dụng ChatGPT cho việc đưa nhận định giá trị ảnh hưởng 27 khía cạnh (bao gồm tích cực,

tiêu cực) qua 2000 bài báo. Sau đây là kết quả thống kê dữ liệu ChatGPT sinh ra:



Hình 10. Tỉ lệ giá trị không ảnh hưởng (0.0) trong mỗi khía cạnh của dữ liệu.

Qua sơ đồ trên, thấy được 8 khía cạnh có tỉ lệ phần trăm giá trị không ảnh hưởng dưới 60% trên 2000 bài báo. Còn trong những khía cạnh còn lại, có những khía cạnh chứa giá trị không ảnh hưởng thậm chí lên đến gần 100%. Điều này dẫn đến việc nhóm cần loại bỏ những khía cạnh không có nhiều ý nghĩa ảnh hưởng. Sau khi đã loại bỏ các khía cạnh không cần thiết, bộ dữ liệu của chúng tôi còn lại 8 khía cạnh bao gồm: “Reputation” (Danh tiếng của công ty), “Financial” (Tài chính), “Regulatory” (Cơ quan quản lý, chính sách), “Risks” (Rủi ro), “Fundamentals” (Các chỉ số trong phân tích cơ bản như P/E, P/B, Liabilities to Asset ratio), “Conditions” (Điều kiện), “Market” (Thị trường), “Volatility” (Độ biến động, rủi ro).

	Content	Volatility	Market	Conditions	Fundamentals	Risks	Regulatory	Financial	Reputation
0	Đội ngũ phân tích dự phóng thu nhập của các do...	0.2	0.3	0.2	0.2	0.2	0.1	0.2	0.0
1	Sau 2 tuần điều chỉnh liên tiếp, thị trường ch...	-0.3	-0.3	-0.4	-0.2	-0.4	-0.2	-0.3	0.0
2	Sau hơn một năm, chứng khoán Việt Nam đã có đế...	-0.4	-0.5	-0.5	-0.3	-0.5	-0.4	-0.5	-0.2
3	Novaland phê duyệt việc bổ sung bên đảm bảo là...	0.0	-0.3	-0.5	-0.3	-0.5	0.0	-0.7	0.2
4	Trong 9 doanh nghiệp trả cổ tức tiền mặt, tỷ l...	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
...	...	...	...	...	...	...	...	...	...

Hình 11. Bộ dữ liệu sau khi chọn lọc các khía cạnh cần thiết.

Khi đã xử lý qua các biện pháp chọn lọc và trích xuất khía cạnh, chúng tôi tiếp tục thu thập dữ liệu nhiều hơn và chỉ dựa theo 8 khía cạnh đã chọn lọc. Sau cùng, nhóm có bộ dữ liệu với kích thước 10000 bài báo.

4.1.3 Giai đoạn xử lý dữ liệu

Chúng tôi xem xét về mức độ chính xác của API OpenAI cho vấn đề phân tích xúc cảm. Vì thế, chúng tôi tiến hành thực hiện gán nhãn nhiều lần trên một bài báo và đánh giá bằng cách tính toán phương sai giữa các lần để chọn lọc ra những dữ liệu phù hợp với phương sai thấp hơn 0.2 theo công thức:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Trong đó:

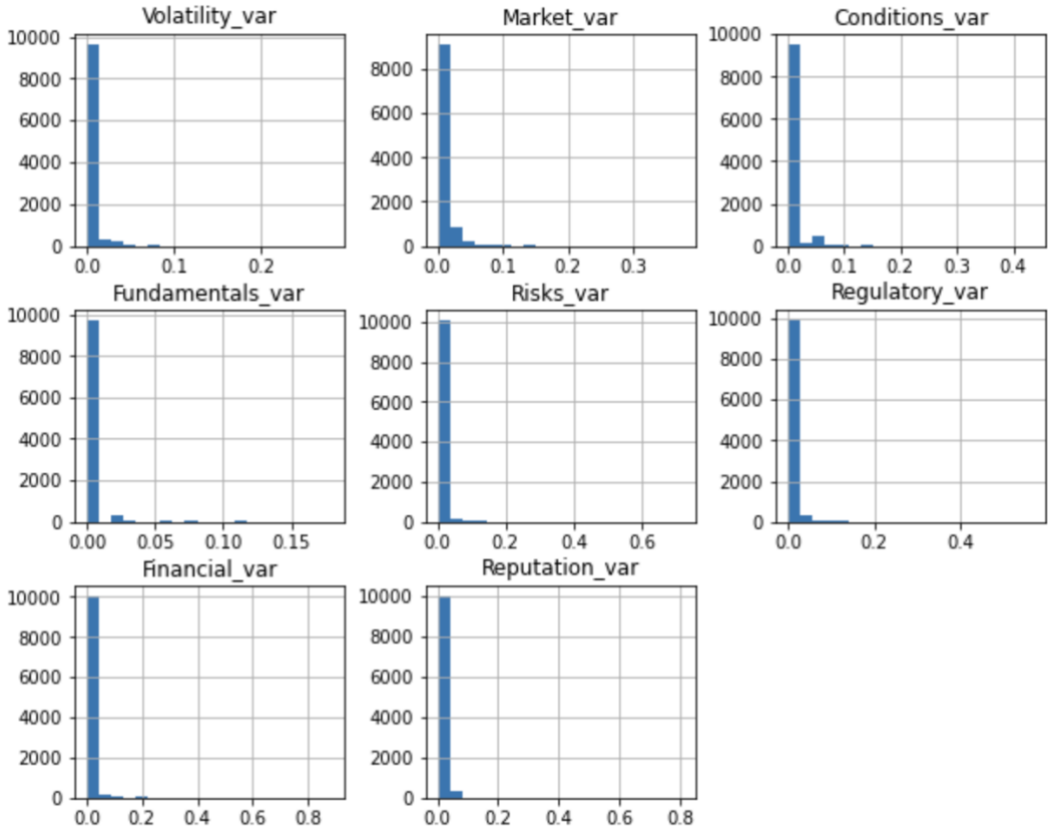
σ^2 : là phương sai của giá trị nhãn dữ liệu.

X : là giá trị nhãn dữ liệu.

μ : là giá trị trung bình của nhãn dữ liệu qua 3 lần gán nhãn.

N : là số lần gán nhãn dữ liệu.

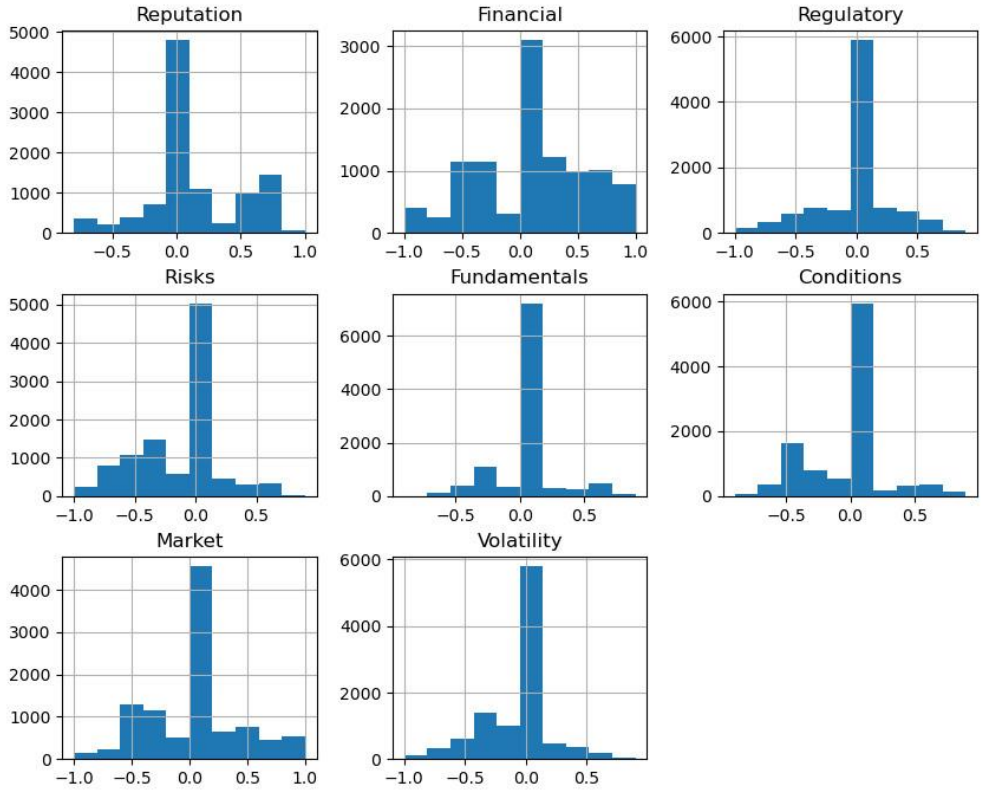
Qua đó chúng tôi thu được giá trị đánh giá khách quan nhất cho từng nhãn dữ liệu để đảm bảo qua nhiều lần gán nhãn, ChatGPT vẫn cho ra cùng một giá trị. Sau đây là kết quả tính toán phương sai qua nhiều lần gán nhãn lặp lại:



Hình 12. Phương sai của các nhãn qua ba lần gán nhãn bằng chat GPT.

Qua kết quả trên, nhìn chung qua 3 lần gán nhãn ChatGPT không có độ lệch phương sai vượt quá 0.2 . Cho nên, nhóm quyết định giữ lại giá trị gán nhãn trong đợt 1 làm đại diện cho bộ dữ liệu vì nhìn chung không có quá nhiều sự thay đổi.

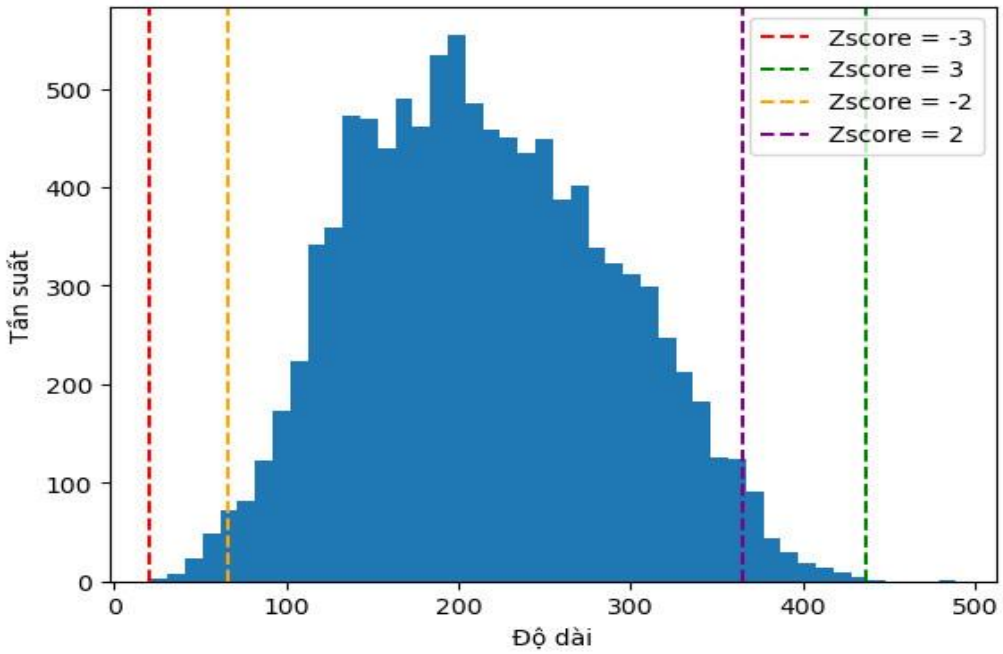
Sau khi hoàn thành gán nhãn dữ liệu, chúng tôi thống kê lại tần suất ảnh hưởng của các nhãn trong dữ liệu bằng cách trực quan hoá chúng với biểu đồ histogram.



Hình 13. Tần suất giá trị tác động của mỗi khía cạnh đến dữ liệu tin tức

Biểu đồ histogram trên cho nhóm nghiên cứu thêm một số thông tin về phân bố giá trị dữ liệu theo mỗi khía canh. Nhìn chung, đa phần dữ liệu ở mỗi khía cạnh đều tập trung xoay quanh giá trị không ảnh hưởng và trải dài 2 bên dữ liệu là -1, 1.

Sau đó chúng tôi tiến hành tiền xử lý dữ liệu ngôn ngữ tự nhiên bằng cách áp dụng các phương pháp như: Đánh dấu và tách từ (Tokenization), loại bỏ dấu câu và ký tự đặc biệt, chuyển đổi văn bản thành chữ thường (Lowercase), loại bỏ stopwords. Và vectơ hóa văn bản (Text Vectorization) để chuyển đổi văn bản thành các vectơ số học để mô hình học máy có thể hiểu và xử lý được. Tiếp theo đối với việc chọn lọc dữ liệu để đưa vào mô hình dựa theo độ dài phù hợp cho mỗi tin tức, để có thể lựa chọn khoảng dữ liệu phù hợp, ở đây chúng tôi sử dụng hệ số z-score từ đoạn (-2,2) (độ dài 66 từ đến 365 từ), với mức độ dài dữ liệu trung bình vào khoảng 216 từ và độ lệch chuẩn của bộ dữ liệu khoảng 75 để chọn ra khoảng dữ liệu thuộc khoảng 96.79% tin tức có độ dài phù hợp trong phân phối trung bình độ dài tổng thể của các tin tức trong bộ dữ liệu.



Hình 14. Khoảng dữ liệu phù hợp được lựa chọn

Cuối cùng chúng tôi thống kê và trực quan hoá lại các từ ngữ phổ biến xuất hiện trong bộ dữ liệu để đảm bảo các từ ngữ phổ biến vẫn đầy đủ ý nghĩa và tập trung vào đúng lĩnh vực, khía cạnh mà chúng tôi hướng đến trong việc phân tích và dự đoán trong mô hình. Sau đó tiến hành hình thành và xây dựng bộ từ điển để đưa vào mô hình học máy.

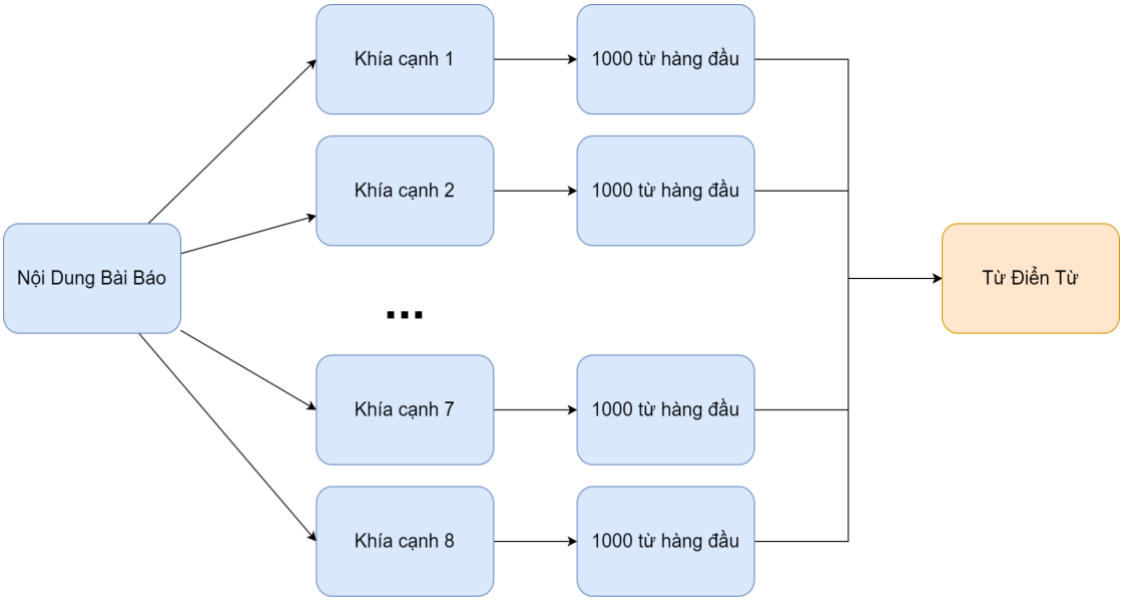
Hình 15. Các từ ngữ phổ biến trong bộ dữ liệu.

4.2 Kết quả thực nghiệm

Trong phần kết quả thực nghiệm, nhóm nghiên cứu sẽ trình bày chi tiết hơn các nội dung tinh chỉnh kích thước từ điển, tinh chỉnh siêu tham số, mô hình đề xuất đi kèm kết quả thực nghiệm cùng nhận định cho từng bước. Kết quả hiển thị loss xuyên suốt phần thực nghiệm sẽ được dùng từ phương pháp RMSE.

4.2.1 Tinh chỉnh kích thước từ điển

Trong lĩnh vực NLP, việc xây dựng bộ từ điển là một trong những bước đóng vai trò quan trọng quyết định những thông tin được đưa vào mô hình huấn luyện để học là gì? Nhận thấy tầm quan trọng này, nhóm đưa ra các bước xây dựng bộ từ điển phù hợp với bài toán cần giải quyết như sau:



Hình 16. Tổng quan các bước xây dựng từ điển từ.

Từ các bước nhóm đã liệt kê ở Hình trên, có thể thấy để xây dựng bộ từ điển cần chọn lọc những từ xuất hiện của mỗi khía cạnh. Điều này giúp cho việc khi mô hình học từ những bài báo sẽ không bị thiên vị nhận quá nhiều thông tin để dự đoán cho 1 khía cạnh hoặc mất đi nhiều thông tin để dự đoán khía cạnh khác.

Kích thước từ điển là 1300 từ trong tổng 29718 từ unique có thể có ở bộ dữ liệu. Ngoài ra còn có thêm con số, tần suất của từ/cụm từ ít xuất hiện nhất trong bộ từ điển là 142 lần qua các bài báo. Điều này cho biết, bộ từ điển có thể tránh đưa những từng có tấn suất xuất hiện thấp (1 hoặc 2 lần) vào mô hình sẽ làm mô hình học những thông

tin không cần thiết hoặc thậm chí là học không hiệu quả (có dấu hiệu overfitting nhanh).

Nhóm sử dụng bộ từ điển 1000 từ hàng đầu cho bước tinh chỉnh kích thước bộ từ điển để cân bằng cho việc tránh overfitting và dễ làm mất thông tin ngữ cảnh khi đưa vào mô hình. Đây cũng sẽ là bộ từ điển được dùng cho các bước thực nghiệm tiếp theo.

4.2.2 Tinh chỉnh siêu tham số

Ở bước tinh chỉnh siêu tham số, nhóm nghiên cứu nhận thấy để xác định được mô hình có thể đề xuất phải bao gồm: cấu trúc mô hình là gì?, tham số điều chỉnh bao nhiêu là tối ưu? Với câu hỏi đầu tiên để trả lời được, nhóm nghiên cứu cần đi so sánh kết quả đánh giá giữa các cấu trúc mô hình (RNN/GRU/LSTM) với nhau nhưng để so sánh được ta cần trả lời câu hỏi thứ 2 ở mức độ cụ thể hơn là tham số điều chỉnh ở mỗi cấu trúc bao nhiêu là tối ưu? Thì mới có thể đi đến so sánh kết quả giữa các cấu trúc khác nhau và đề xuất mô hình. Nhìn chung sẽ bao gồm 2 bước cụ thể trong tinh chỉnh siêu tham số:

1. Sử dụng Grid Search cho mỗi cấu trúc mô hình RNN/GRU/LSTM.

2. Lọc mô hình có bộ tham số tối ưu tương ứng từng cấu trúc trong tất cả mô hình mà Grid Search sinh ra.

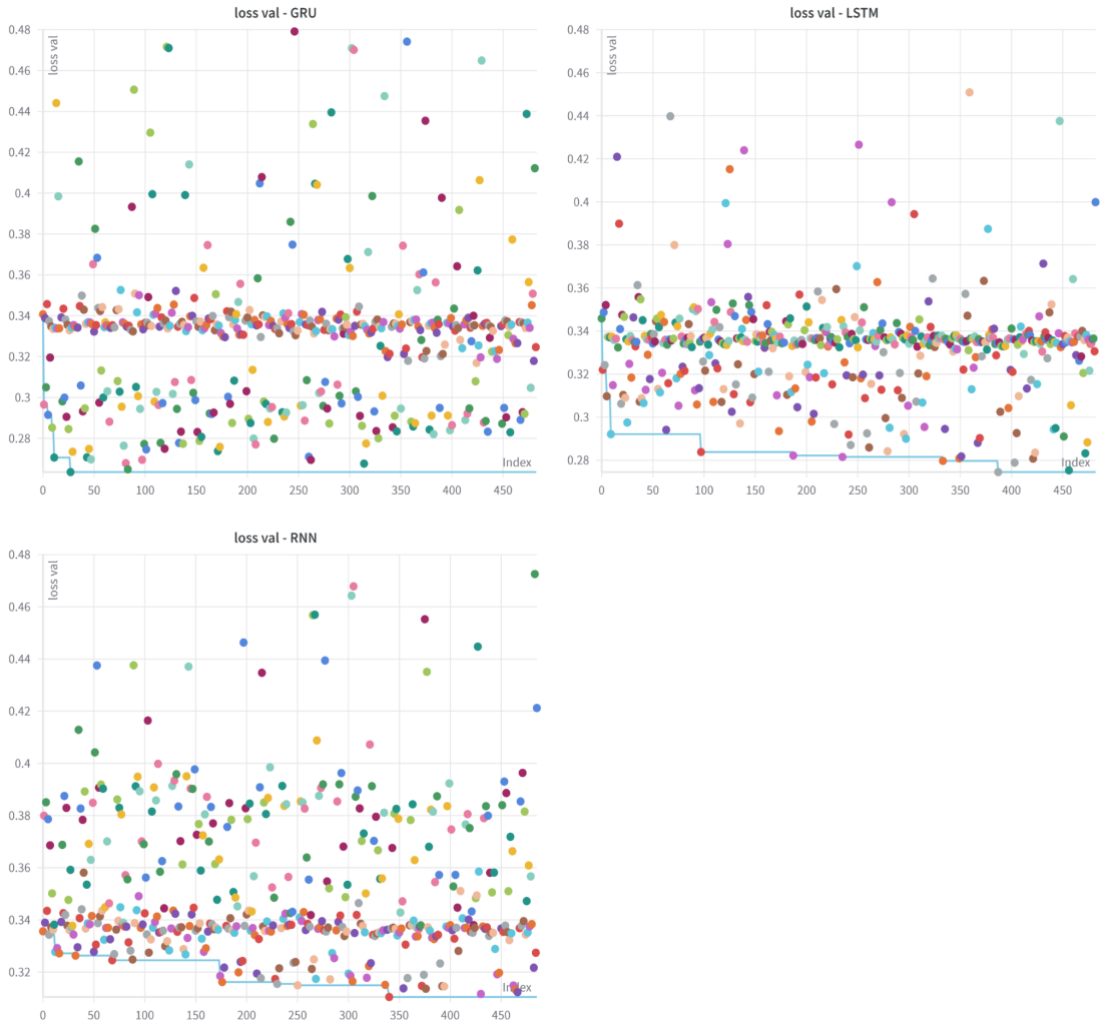
Bước 1, nhóm nghiên cứu sẽ tận dụng thuận toán tối ưu siêu tham số Grid Search ở mục 2.5 và đi kèm với các tham số điều chỉnh mà nhóm liệt kê trong bảng:

Bảng 1. Danh sách siêu tham số dùng cho Grid Search.

Tên tham số điều chỉnh	**Giá trị tối ưu bằng Grid Search**
Batch size	[64, 128, 256]
Số lượng lớp ẩn trong cấu trúc mô hình	[1, 2, 3]
Số lượng node trong mỗi lớp ẩn	[64, 128, 256]
Tốc độ học	[0.1, 0.01, 0.001]

Thuật toán tối ưu	[“Adam”, “Sgd”]
Tỉ lệ ở lớp Dropout	[0.2, 0.35, 0.5]

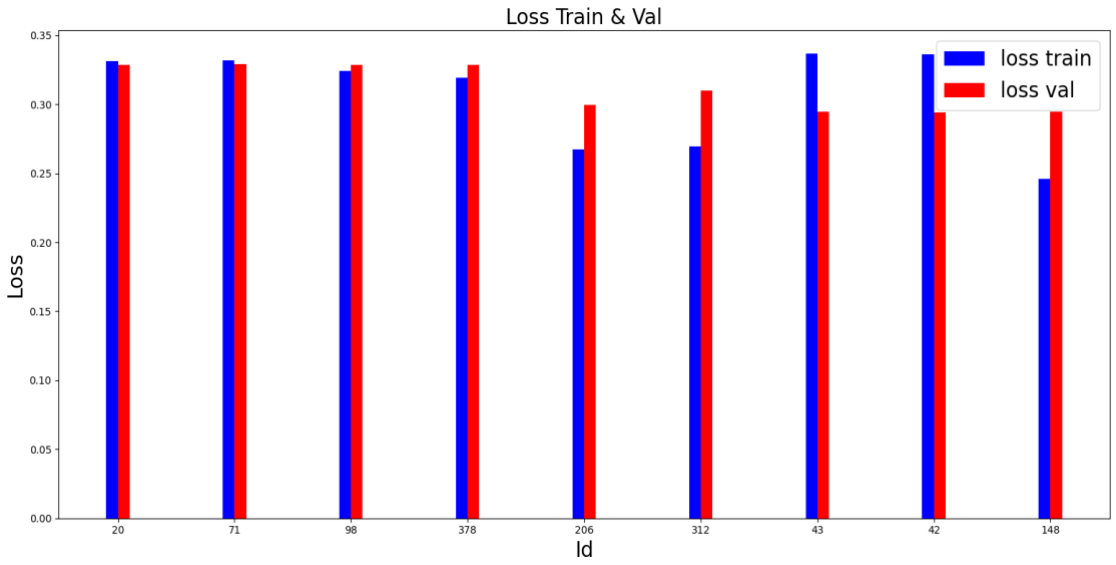
Ngoài những tham số cần tối ưu trên, nhóm nghiên cứu đã cân chỉnh giữa thời gian được phép thực nghiệm và tốc độ xử lý phần cứng có sẵn để xác định huấn luyện mỗi mô hình được sinh ra từ Grid Search với 20 epochs. Sau đây, nhóm sẽ trình bày kết quả thực nghiệm thu được:



Hình 17. Kết quả thực nghiệm loss validation Grid Search cho mỗi cấu trúc mô hình.

Từ kết quả trên, nhóm nghiên cứu có thể thấy được một số nhận định trên tổng 1458 mô hình được sinh ra từ Grid Search qua cả 3 cấu trúc mô hình. Dựa theo mức

độ phân bố kết quả loss validation, cấu trúc mô hình RNN đa phần tập trung giá trị ở gần 0.34 và tương tự với cấu trúc mô hình GRU, LSTM. Nhưng điểm khác ở những bộ siêu tham số tối ưu có thể đạt được dựa theo đường thẳng trên biểu đồ. Đối với RNN sẽ có bộ tham số tối ưu đạt loss validation trong khoảng 0.32 đến 0.3 đa phần cao hơn LSTM khi có thể đạt 0.28 và kết quả có thể đạt thấp nhất cho GRU so với 2 cấu trúc mô hình còn lại là vượt qua ngưỡng 0.28. Nhìn chung nhóm có một số cái nhìn sơ lược từ kết quả thực nghiệm trình bày trên nhưng để đánh giá cụ thể hơn thì nhóm sẽ cần đi đến bước 2. Lọc mô hình có bộ tham số tối ưu tương ứng từng cấu trúc trong tất cả mô hình mà Grid Search sinh ra.



Hình 18. Xếp hạng 9 mô hình cấu trúc LSTM tiềm năng hàng đầu sinh từ Grid Search.

Để có thể dễ dàng hình dung tiêu chí chọn lọc ở bước 2, nhóm sẽ lấy cấu trúc LSTM làm ví dụ cho việc chọn lọc. Hình trên là biểu đồ thể hiện những mô hình tiềm năng nhất từ trái sang phải mà nhóm chọn lọc lại từ các mô hình sinh ra từ Grid Search. Có bao gồm 2 tiêu chí đánh giá theo mức độ ưu tiên giảm dần:

1. Độ lệch giữa loss train với loss validation là nhỏ nhất. Đây là điều kiện tối thiểu để đảm bảo mô hình này khi đã được huấn luyện 20 epochs chưa có dấu hiệu overfitting. Vì thế ta có thể tiếp tục lấy mô hình này huấn luyện và kỳ vọng giá trị ở loss validation tiếp tục giảm.

2. Loss validation là thấp nhất. Không dừng lại chỉ đảm bảo mô hình chưa overfitting thì cần tối ưu hơn khi chọn lựa bộ siêu tham số tốt nhất với cấu trúc tương ứng.

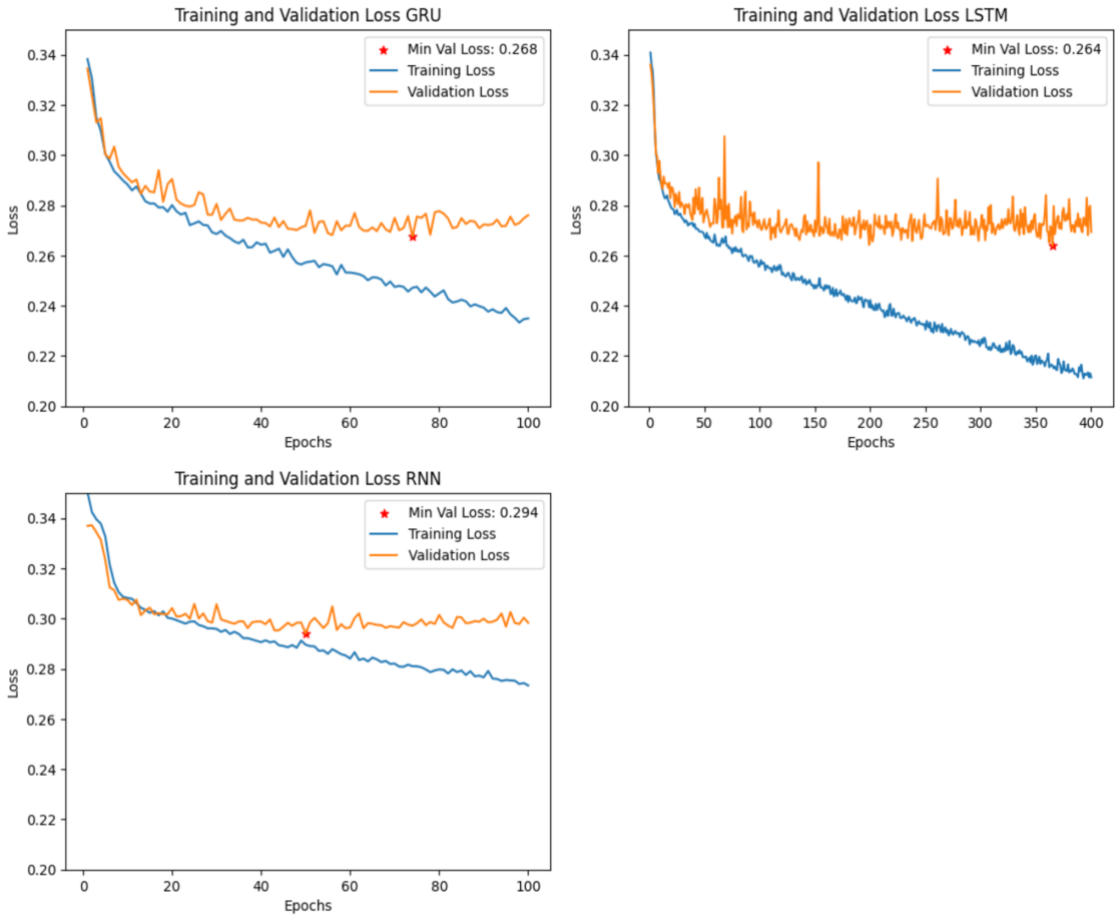
Theo 2 tiêu chí đã nêu trên và qua biểu đồ nhóm thể hiện, có thể thấy được mô hình ID 20 sẽ là mô hình tiềm năng vì có độ lệch giữa loss train với loss validation cũng như giá trị loss validation là thất nhất. Quá trình chọn này sẽ thực hiện tương tự với cấu trúc mô hình RNN, GRU. Sau cùng, nhóm sẽ trình bày những bộ siêu tham số tối ưu tương ứng với mỗi cấu trúc mô hình:

Bảng 2. Trình bày bộ siêu tham số tối ưu của mỗi cấu trúc mô hình.

	LSTM	GRU	RNN
Batch size	64	64	128
Số lượng lớp ẩn trong cấu trúc mô hình	1	2	3
Số lượng đặc trưng trong mỗi lớp ẩn	128	256	64
Tốc độ học	0.001	0.001	0.001
Thuật toán tối ưu	Adam	Adam	Adam
Tỉ lệ ở lớp Dropout	0.2	0.5	0.5

4.2.3 Mô hình đề xuất

Cuối cùng khi đã xác định được bộ siêu tham số cho mỗi cấu trúc, nhóm sẽ tiếp tục train với số lượng epochs nhiều hơn cho mỗi mô hình (RNN-100 epochs, GRU-100 epochs, LSTM-400 epochs), Sau đây là kết quả loss đạt được khi tiếp tục train mô hình qua các bộ siêu tham số tối ưu:



Hình 19. Loss train & validation các mô hình có bộ siêu tham số tối ưu.

Khi đã tiếp tục train, kết quả trên có thể cho thấy RNN đạt được kết quả loss validation tối ưu nhất của mình ở 0.294 . Nhìn chung đây là kết quả thấp hơn so với 2 cấu trúc còn lại. Đối với LSTM, tuy có kết quả không quá trên lệch với GRU nhưng có dấu hiệu overfitting (chênh lệch loss validation với loss train xấp xỉ 0.05) vì tốc độ giảm loss validation chậm dẫn đến cần train nhiều số lượng epoch hơn. Để có cái nhìn khách quan hơn hay so sánh có độ chênh lệch rõ ràng hơn, nhóm nghiên cứu tiếp tục dùng những mô hình trên và đánh giá trên tập test qua các chỉ số MSE, MAE. Sau đây là kết quả thực nghiệm nhóm thu được:

Bảng 3. Kết quả thực nghiệm đánh giá trên tập Test.

	MSE	MAE
LSTM	0.077	**0.195**
GRU	**0.075**	0.198
RNN	0.088	0.224

Vậy qua các bước thực nghiệm trên, nhóm có được kết quả thực nghiệm cấu trúc mô hình GRU có bộ tham số tối ưu đạt kết quả trên tập test qua 3 thông số đánh giá là thấp nhất. Điều này đồng nghĩa, đây cũng là mô hình thực nghiệm có kết quả tốt nhất trong những thực nghiệm của nhóm và có thể là mô hình đề xuất từ nhóm cho bài toán dự đoán mức độ tác động tin tức tài chính trên nhiều khía cạnh.

4.3 Kết quả dự đoán của mô hình

Qua quá trình thực nghiệm đánh giá và lựa chọn mô hình, chúng tôi tiến hành đưa vào dự đoán kết quả dựa trên các bài báo thực tế. Dưới đây là bảng kết quả của các dự đoán bằng mô hình đề xuất của chúng tôi.

Content	Reputation	Financial	Regulatory	Risks	Fundamentals	Conditions	Market	Volatility
Thị trường chứng khoán được dự báo sẽ rung lắc với biên độ hẹp trong các phiên tới, nhà đầu tư cân nhắc …	-0.046	-0.011	-0.056	-0.201	-0.018	-0.137	0.010	-0.110
Lượng tiền chờ giải ngân theo đó đạt mức 23,5 triệu USD (600 tỷ đồng) tại ngày 9/2. Hòa Phát lọt top 3 khoản đầu tư lớn nhất …	0.129	0.250	0.062	-0.039	0.094	0.025	0.210	0.024
Trong bối cảnh thị trường đầu năm 2023 đã xuất hiện một số dấu hiệu tích cực với áp lực bán giải chấp giảm mạnh so với 2022, một số công ty chứng khoán nội đã bắt đầu có động thái hạ lãi suất margin …	0.058	0.072	-0.069	-0.178	0.006	-0.104	0.041	-0.067
Sau hơn một năm, chứng khoán Việt Nam đã có đến 19 doanh nghiệp phải rời danh sách tỷ USD vốn hóa trong đó gần một nửa đến từ nhóm bất động sản…	-0.048	-0.056	-0.117	-0.247	-0.047	-0.184	-0.143	-0.139
…	…	…	…	…	…	…	…	…

Hình 20. Kết quả dự đoán bằng mô hình đề xuất.

Hình 21. Hình ảnh trực quan hoá kết quả dự đoán của chúng tôi.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Trong khóa luận này, nhóm đã thử nghiệm với 3 cấu trúc mô hình thuần RNN/GRU/LSTM để giải quyết bài toán dự đoán mức độ tác động tin tức tài chính trên nhiều khía cạnh. Từ những kết quả thực nghiệm sau cùng của GRU, nhìn chung có phần cải thiện hơn bài báo gốc qua các thông số MSE [8].

Nhóm mong những kết quả thực nghiệm cũng như những phương pháp đã được sử dụng có thể làm cơ sở để tiếp tục phát triển cách giải quyết bài toán tốt hơn cho các nhà nghiên cứu tiếp nối.

5.2 Hạn chế

Trong quá trình thực nghiệm, nhóm cũng thấy được một số điểm phát sinh dẫn đến hạn chế khi giải quyết bài toán. Cụ thể hơn, về mặt dữ liệu không thể lấy nhiều hơn do giới hạn về kinh phí nhóm nghiên cứu có thể chi trả cho API OpenAI gán nhãn (bao gồm cả bước gán nhãn nhiều lần trên 1 bài báo để tăng tính khách quan). Hay một hướng giải quyết tăng chất lượng dữ liệu hơn như thuê thêm các chuyên gia về tài chính để nhận định kết quả nhãn thu được. Mặc dù đây cũng là hướng sẽ tốn rất nhiều kinh phí. Một điểm hạn chế khác, mô hình đề xuất của nhóm nghiên cứu có

thể sẽ chỉ hoạt động tốt đối với các tin tức có độ dài trong khoảng 66 đến 365 từ. Vì đây là khoảng độ dài chiếm 96.79% tổng dữ liệu gốc. Ngoài ra, nếu có nhiều dữ liệu hơn, các nhà nghiên cứu tiếp nối có thể sử dụng cấu trúc mô hình phức tạp hơn như mô hình Transformer hoặc tận dụng thêm các mô hình đã được huấn luyện trên dữ liệu lớn để tinh chỉnh và giải quyết cho bài toán cụ thể này.

5.3 Kiến thức và kỹ năng

Việc thực hiện luận văn này chúng tôi đã phần nào hiểu rõ hơn được các công thức, kiến trúc, cách thức xây dựng, cài đặt và huấn luyện mô hình RNN, GRU, LSTM và áp dụng nó trong bài toán xử lý ngôn ngữ tự nhiên.

5.4 Hướng phát triển trong tương lai

Trong tương lai, chúng tôi sẽ tiếp tục tối ưu bài toán dự đoán mức độ ảnh hưởng của các khía cạnh đối với tin tức để từ đó dự đoán chính xác chiều hướng tác động của nó đối với thị trường.

Đồng thời, chúng tôi dự định kết hợp tên, giá trị cổ phiếu của các công ty tại thời điểm công ty được nhắc đến trong các tin tức để từ đó xây dựng bộ dữ liệu về tác động của tin tức đến giá trị của công ty nhằm dự đoán chiều hướng tác động của tin tức chính xác hơn đối với thị trường.

TÀI LIỆU THAM KHẢO

[1] baochinhphu.vn (Oct. 2023), https://baochinhphu.vn/gan-26-trieu-tai-khoan-chung-khoan-mo-moi-trong-nam-2022-10223011010052344.htm

[2] Daniel M. DiPietro (2022), "Quantitative Stopword Generation for Sentiment Analysis via Recursive and Iterative Deletion", *arXiv preprint arXiv:2209.01519.*

[3] Daniel Mesafint Belete, Manjaiah D H (2021), "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results", *International Journal of Computers and Applications* Vol.44 (1), pp.1-12.

[4] Diederik P. Kingma, Jimmy Ba (2017), "Adam: A Method for Stochastic Optimization".

[5] dominhhai.github.io (Oct. 2017), https://dominhhai.github.io/vi/2017/10/what-is-rnn/

[6] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov (2018), "Learning Word Vectors for 157 Languages", *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

[7] Felipe Almeida, Geraldo Xexéo (2023), "Word Embeddings: A Survey", *Computer and Systems Engineering Program (PESC-COPPE).*

[8] Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, Roger Zimmermann (2018), "Aspect-Based Financial Sentiment Analysis using Deep Learning", *WWW '18: Companion Proceedings of the The Web Conference 2018* pp.1961–1966.

[9] Jonathan J. Webster, Chunyu Kit (1992), "Tokenization as the initial phase in NLP", *COLING '92: Proceedings of the 14th conference on Computational linguistics* Vol.4, pp.1106-1110.

[10] Kanchan M. Tarwani, Swathi Edem (2017),"Survey on Recurrent Neural Network in Natural Language Processing", *International Journal of Engineering Trends and Technology* Vol.48 (6), pp.301-304.

[11] Lucas Weber, Jaap Jumelet, Paul Michel, Elia Bruni, Dieuwke Hupkes (2023), "Curriculum Learning with Adam: The Devil Is in the Wrong Details".

[12] medium.com (Dec 2020), https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

[13] medium.com (Nov 2018), https://blog.chappiebot.com/h%C6%B0%E1%BB%9Bng-d%E1%BA%ABn-chi-ti%E1%BA%BFt-v%E1%BB%81-c%C6%A1-ch%E1%BA%BF-c%E1%BB%A7a-lstm-v%C3%A0-gru-trong-nlp-a1bd9346b209

[14] Minh-Hao Nguyen, Tri Minh Nguyen, Dang Van Thin, Ngan Luu-Thuy Nguyen (2019), "A corpus for aspect-based sentiment analysis in Vietnamese", *Institute of Electrical and Electronics Engineers.*

[15] Muhammad Zulqarnain, Rozaida Ghazali, Yana Mazwin Mohmad Hassim, Muhammad Rehan (2020), "Text classification based on gated recurrent unit combines with support vector machine", *International Journal of Electrical and Computer Engineering* Vol.10 (4), pp.3734-3742.

[16] Nhan Cach Dang, María N. Moreno-García, Fernando De la Prieta, (2020), "Sentiment Analysis Based on Deep Learning: A Comparative Study", *Electronics 2020* Vol.9 (3), pp.483-511.

[17] prudential.com.vn, https://www.prudential.com.vn/vi/blog-nhip-song-khoe/nha-dau-tu-f0-la-gi-huong-dan-cho-nguoi-moi-bat-dau/

[18] Quang-Linh Tran, Phan Thanh Dat Le, Trong-Hop Do (2022), "Aspect-based Sentiment Analysis for Vietnamese Reviews about Beauty Product on E-

commerce Websites.", *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation* pp.767–776.

[19] Sebastian Ruder (2017), "An overview of gradient descent optimization algorithms".

[20] Sepp Hochreiter, Jürgen Schmidhuber (1997), "Long Short-term Memory", *Neural Computation* Vol.9 (8), pp.1735-1780.

[21] stanford.edu (May 2020), https://stanford.edu/~shervine/l/vi/teaching/cs-230/cheatsheet-recurrent-neural-networks

NHẬT KÝ LÀM VIỆC

Tuần	Từ ngày	Đến ngày	Nội dung
1	07/08/2023	13/08/2023	Tìm hiểu về từ khóa đề tài: Aspect-Based Financial Sentiment Analysis. Tìm - đọc các tài liệu liên quan đến đề tài, định hướng nghiên cứu.
2	14/08/2023	20/08/2023	Tiếp tục tìm hiểu về đề tài qua các nghiên cứu sẵn có trước đó. Đọc sâu hơn một số tài liệu nhất định của hướng nghiên cứu, hiểu rõ hơn hướng nghiên cứu đối với đề tài.
3	21/08/2023	27/08/2023	Tiếp tục nghiên cứu, đọc tài liệu liên quan đến ABSA.
4	28/08/2023	03/09/2023	Sau quá trình tìm hiểu các nghiên cứu trước đó, xác định các thuộc tính sử dụng cho tập dữ liệu. Tìm kiếm các nguồn thu thập dữ liệu phù hợp với yêu cầu của đề tài
5	04/09/2023	10/09/2023	Bắt đầu việc thu thập dữ liệu thử nghiệm từ các bài báo, trang báo ở Việt Nam.
6	11/09/2023	17/09/2023	Tiến hành tiền xử lý dữ liệu văn bản tiếng Việt. Thử nghiệm dán nhãn bậc 1 và bậc 2 cho bộ dữ liệu.
7	18/09/2023	24/09/2023	Thử nghiệm việc áp dụng công cụ chat GPT vào quá trình dán nhãn dữ liệu.
8	25/09/2023	01/10/2023	Tiến hành dán nhãn cho bộ dữ liệu sử dụng công cụ chat GPT và đánh giá độ chính xác của nhãn.

9	02/10/2023	08/10/2023	Thử nghiệm huấn luyện bộ dữ liệu trên mô hình LSTM với đầu ra gồm 8 đặc trưng.
10	09/10/2023	15/10/2023	Kết hợp huấn luyện bộ dữ liệu trên mô hình LSTM với đầu ra cho từng đặc trưng. Cải thiện mô hình huấn luyện và so sánh độ hiệu quả giữa cả 2 phương pháp huấn luyện.
11	16/10/2023	22/10/2023	Thử nghiệm huấn luyện bộ dữ liệu trên mô hình RNN với đầu ra với đầu ra cho từng đặc trưng và 8 đặc trưng tương ứng.
12	23/10/2023	29/10/2023	Điều chỉnh các tham số để cải thiện mô hình huấn luyện. Tiến hành làm giàu bộ dữ liệu bằng các thu thập và xử lý thêm dữ liệu từ nhiều nguồn, trang báo khác nhau ở Việt Nam.
13	30/10/2023	05/11/2023	Tiến hành dán nhãn cho bộ dữ liệu đã làm giàu thêm trước đó để đưa vào huấn luyện mô hình.
14	06/11/2023	12/11/2023	Thử nghiệm huấn luyện bộ dữ liệu trên mô hình GRU với đầu ra với đầu ra cho từng đặc trưng và 8 đặc trưng tương ứng.
15	13/11/2023	19/11/2023	Tiếp tục điều chỉnh các tham số để cải thiện kết quả của các bộ mô hình huấn luyện trước đó.
16	20/11/2023	26/11/2023	Tiến hành huấn luyện mô hình và tối ưu kết quả, viết báo cáo và xây dựng công cụ thực nghiệm.
17	27/11/2023	03/12/2023	Gửi báo cáo cho giáo viên hướng dẫn và nhận phản hồi, chỉnh sửa báo cáo.

18	04/12/2023	10/12/2023	Gửi báo cáo cho giáo viên hướng dẫn và nhận phản hồi, chỉnh sửa báo cáo. Hoàn thiện báo cáo và slide thuyết trình
19	11/12/2023	14/12/2023	Báo cáo đồ án