HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

ĐẶNG THỊ NGỌC YẾN

PHÁT HIỆN LẬP TRƯỜNG SỬ DỤNG KỸ THUẬT HỌC SÂU

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2021

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

ĐẶNG THỊ NGỌC YẾN

PHÁT HIỆN LẬP TRƯỜNG SỬ DỤNG KỸ THUẬT HỌC SÂU

Chuyên ngành : **Khoa học máy tính**

Mã số : **8.48.01.01**

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:
TS. TRẦN THỊ OANH

HÀ NỘI – 2021

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của bản thân. Các số liệu, kết quả trình bày trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào trước đây.

Tác giả

Đặng Thị Ngọc Yến

LỜI CẢM ƠN

Em xin chân thành cảm ơn TS. Trần Thị Oanh, bộ môn Khoa học máy tính, Quốc tế, Đại học Quốc Gia Hà Nội đã tận tình chỉ dạy và hướng dẫn cho em trong việc lựa chọn đề tài, thực hiện đề tài và viết báo cáo luận văn, giúp cho em có thể hoàn thành tốt luận văn này.

Em xin chân thành cảm ơn các thầy cô giáo Khoa Công nghệ thông tin là những người giảng dạy em, đặc biệt PGS.TS. Ngô Xuân Bách và các thầy cô trong khoa Sau đại học đã tận tình dạy dỗ và chỉ bảo em trong suốt hai năm học.

Xin chân thành cảm ơn hai em Nguyễn Phương Ly và Đào Thanh Trang đã tham gia xây dựng kho ngữ liệu cho bài toán.

Cuối cùng em xin cảm ơn gia đình, bạn bè, những người đã luôn bên cạnh động viên em những lúc khó khăn và giúp đỡ em trong suốt thời gian học tập và nghiên cứu, tạo mọi điều kiện tốt nhất cho em để có thể hoàn thành tốt luận văn của mình.

Mặc dù đã cố gắng hoàn thành nghiên cứu trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự thông cảm của thầy cô và các bạn.

Em xin chân thành cảm ơn!

Hà Nội, 05/2021

Đặng Thị Ngọc Yến

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn mã hóa hai chiều từ Transformer
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
GRU	Gated Recurrent Units	Cổng tái Unit
LSTM	Long-Short Term Memory	Mạng bộ nhớ dài-ngắn
MLM	Masked language modeling	Mô hình ngôn ngữ bị che
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
SRM	Structural Risk Minimization	Cực tiểu hóa rủi ro có cấu trúc
SVM	Support Vector machine	Máy vector hỗ trợ

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

MỞ ĐẦU

Ngày nay công nghệ thông tin phát triển mạnh mẽ, hầu như đã xâm nhập toàn bộ các lĩnh vực đời sống xã hội. Xã hội ngày càng phát triển thì nhu cầu áp dụng các tiến bộ của công nghệ thông tin vào cuộc sống ngày càng cao để giải quyết những vấn đề phức tạp như y tế, giáo dục, pháp luật.Với nhu cầu trao đổi và tìm kiếm thông tin của con người ngày càng cao, thông tin tràn ngập trên mọi phương tiện truyền thông, đặc biệt là sự phát triển rộng rãi của mạng Internet, hằng ngày con người phải xử lý một lượng thông tin khổng lồ [1]. Như vậy,việc trích xuất và tổng hợp ý kiến dư luận có thể mang lại rất nhiều lợi ích cho những ai đặc biệt quan tâm. Để hỗ trợ việc trích xuất và tổng hợp ý kiến dư luận diễn ra hiệu quả và nhanh chóng, trí tuệ nhân tạo, đặc biệt là học máy và xử lý ngôn ngữ tự nhiên được hy vọng là tự động hóa đáng kể một số quy trình trong việc phân tích, nghiên cứu tình hình và xu hướng của dư luận xã hội.

Trích xuất thông tin tự động từ các văn bản là một chủ đề nghiên cứu quan trọng của xử lý ngôn ngữ tự nhiên (Natural language processing - NLP) trong nhiều thập kỷ. [2] Một số vấn đề nghiên cứu chính liên quan đến phân tích tự động các văn bản này bao gồm phân tình cảm (khai thác ý kiến), nhận dạng cảm xúc, khai thác lập luận (xác định lý do), phát hiện mỉa mai, phát hiện tin đồn và xác thực cũng như phát hiện tin tức giả. [2] Các giải pháp tự động và hiệu suất cao cho những vấn đề này sẽ tạo điều kiện thuận lợi cho các nhiệm vụ như phân tích xu hướng và thị trường, thu thập đánh giá của người dùng cho sản phẩm, khảo sát ý kiến, quảng cáo được nhắm mục tiêu, thăm dò ý kiến, dự đoán cho các cuộc bầu cử và trưng cầu dân ý, giám sát phương tiện truyền thông tự động và lọc ra nội dung chưa được xác nhận để có trải nghiệm người dùng tốt hơn, để giám sát sức khỏe cộng đồng trực tuyến.[2]

Trong luận văn này, chúng tôi sẽ tập trung nghiên cứu về vấn đề phát hiện lập trường cho tiếng Việt sử dụng phương pháp học máy giám sát, cụ thể là sử dụng một số mô hình truyền thống Decision Tree, Naïve Bayes, cũng như các mô hình học sâu hiện đại như LSTM , RNN.

Nội dung chính của luận văn được trình bày trong chương như sau:

- **Chương 1: Giới thiệu về bài toán phát hiện lập trường của người dùng tiếng Việt:** Nội dung của chương này giới thiệu tổng quan về bài toán phát hiện lập trường, trình bày một số nghiên cứu liên quan, mục tiêu và ý nghĩa của bài toán .

- **Chương 2: Các phương pháp học máy sử dụng trong bài toán phát hiện lập trường:** Chương 2 trình bày tổng quan một số phương pháp phân lớp truyền thống, phương pháp phân lớp dựa trên kỹ thuật học sâu được sử dụng trong bài toán phát hiện lập trường mà chúng tôi sẽ sử dụng.
- **Chương 3: Đề xuất phương pháp, giải pháp:** Chương 3 mô tả bài toán phát hiện lập trường, và đưa ra giải pháp đề xuất them hai hướng khảo sát các phương pháp học máy truyền thống và học sâu.
- **Chương 4: Thực nghiệm và đánh giá:** Trong chương 4, luận văn trình bày chi tiết các bước để xây dựng kho ngữ liệu về phát hiện lập trường cho tiếng Việt từ cách thu thập, tiền xử lý, xây dựng tập nhãn và thống kê kho dữ liệu. Sau đó, luận văn trình bày các thiết lập thử nghiệm, công cụ sử dụng và kết quả đạt được trên bộ dữ liệu được xây dựng đó. Chương này cũng so sánh và thảo luận kết quả thử nghiệm liên quan. Thực hiện huấn luyện hệ thống với bộ dữ liệu và tập nhãn đã xây dựng, thống kê và đánh giá kết quả thực nghiệm.

CHƯƠNG I: BÀI TOÁN PHÁT HIỆN LẬP TRƯỜNG

1.1 Giới thiệu bài toán phát hiện lập trường

Internet đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của mỗi người trên thế giới ngày nay và đóng một vai trò đặc biệt trong việc thúc đẩy sự phát triển mạnh mẽ của các kênh truyền thông mạng xã hội,diễn đàn, website tin tức. Tại Việt Nam, các kênh truyền thông mạng xã hội (MXH) ngày càng trở nên gần gũi và thân thuộc với rất nhiều người, kể cả đối với thanh, thiếu niên và người già. Theo báo cáo thường niên "Digital 2021" được công bố bởi WeAreSocial và Hootsuite, Việt Nam có 68.72 triệu người dùng Internet chiếm 70.3% dân số và 72 triệu người dùng mạng xã hội chiếm khoảng 73.6%.

Các kênh truyền thông MXH được sử dụng phổ biến bao gồm: Facebook, YouTube, Instagram, Zalo,Twitter… ALEXA cũng thông báo top 10 trang web có lượng truy cập nhiều tại Việt Nam, ngoài các website phổ biến nằm đầu danh sách như Facebook, Youtube, các loại website tin tức cũng chiếm phần lớn trong danh sách này bao gồm vnexpress.net, laodong.vn, thethao247.vn, vtv.vn, 24h.com.vn. Các diễn đàn như Webtretho, Beat.vn…cũng thu hút số lượng thành viên đông đảo và có tác động mạnh mẽ đến cộng đồng và dư luận xã hội ở Việt Nam. Thực tế cho thấy, internet và các kênh kể trên đem tới cho người dùng rất nhiều tiện ích nhờ tốc độ thông tin nhanh, gần như tức thời, nội dung phong phú, hình thức sinh động, hấp dẫn.

Điều này cũng tạo điều kiện cho mỗi cá nhân chia sẻ kiến thức, thảo luận, trao đổi những ý kiến về các chủ đề cùng quan tâm, đồng thời cũng là những kênh thông tin quan trọng góp phần đưa các chủ trương, đường lối của Đảng, chính sách, pháp luật của nhà nước đến với nhân dân nhanh chóng, kịp thời. [3] Tại đây cơ quan chức năng các cấp có thể nắm bắt thường xuyên thái độ của người sử dụng mạng xã hội đối với các vấn đề, sự kiện đang được dư luận xã hội quan tâm. [4] Như vậy dựa vào việc tổng hợp và phân loại ý kiến của dư luận trên nhiều kênh có thể là một nguồn tài nguyên quan trong việc phân tích, nghiên cứu tình hình và xu hướng của dư luận xã hội.

Nhìn vào bối cảnh trên, lập trường (stance) có thể hiểu là một ý kiến được thể hiện bởi một cá nhân hướng tới chủ đề hoặc sự kiện hoặc nhân vật nào đó. [5]Bài toán phát hiện lập trường thu hút rất nhiều sự chú ý của các nhà nghiên cứu bởi nó

mang lại rất nhiều ứng dụng thiết thực. Một trong những ứng dụng phổ biến nhất trong đó là khảo sát, thăm dò ý kiến được thực hiện trên các văn bản nội dung trực tuyến với các chủ đề khác nhau bao gồm các cuộc tranh luận chính trị, tư tưởng, xã hội, đánh giá sản phẩm và bầu cử, trưng cầu dân ý. [2] Ví dụ chúng ta có thể lấy ý kiến dư luận về một vấn đề chính trị - xã hội như hợp pháp hóa việc phá thai ở một số quốc gia. Dựa vào tự động hóa phát hiện lập trường, chúng ta có thể tổng hợp và phân loại ra có bao nhiêu người tán thành và không tán thành với vấn đề này.

Một ứng dụng quan trọng khác của phát hiện lập trường cũng thu hút nhiều sự chú ý là phát hiện tin giả. Đánh giá tính xác thực của một bài tin là một công việc phức tạp và cồng kềnh, ngay cả đối với các chuyên gia được đào tạo [6]. Mục tiêu áp dụng bài toán lập trường cho phát hiện tin tức giả là để hỗ trợ kiểm tra tính xác thực của thông tin giúp họ nhanh chóng thu thập thông tin cần thiết để đưa ra đánh giá của mình. Một bài toán phát hiện lập trường tốt sẽ cho phép người kiểm tra tính xác thực của thông tin nhập một tuyên bố hay tiêu đề và ngay lập tức họ nhận được các truy xuất về các bài báo: đồng ý, không đồng ý hoặc thảo luận về tuyên bố/tiêu đề được đề cập.

Sau đó, họ có thể xem xét các lập luận ủng hộ và chống lại tuyên bố, đồng thời sử dụng khả năng phán đoán và lập luận của mình để đánh giá tính hợp lệ của tuyên bố được đề cập. Một công cụ như vậy sẽ cho phép người kiểm tra thực tế nhanh chóng và hiệu quả. [7] Ngoài ra phát hiện lập trường còn áp dụng nhiều ứng dụng khác như: phân loại tin đồn, phân tích, dự báo xu hướng và thị trường, tạo hệ thống khuyến nghị, hỗ trợ giám sát sức khỏe cộng đồng, truy xuất thông tin, khảo sát ý kiến góp ý người tiêu dùng.

1.2 Một số nghiên cứu liên quan

Trong những năm gần đây, đã có rất nhiều nghiên cứu về lĩnh vực Xử lý Ngôn ngữ Tự nhiên (Natural language processing - NLP) liên quan đến lĩnh vực phát hiện lập trường. Phát hiện lập trường nhằm mục đích xác định lập trường của tác giả văn bản đối với mục tiêu (một thực thể, khái niệm, sự kiện, ý tưởng, ý kiến, tuyên bố, chủ đề, v.v.).

Các nghiên cứu ban đầu về phát hiện lập trường tính đến năm 2013 chủ yếu tập trung vào các cuộc tranh luận trong quốc hội [Thomas et al. 2006]; các cuộc thảo luận nội bộ công ty [Murakami và Raymond 2010]; các cuộc tranh luận xã hội ,chính

trị và các hệ tư tưởng trực tuyến [Anand et al. 2011; Somasundaran and Wiebe 2010; Walker et al. 2012a]; các cuộc tranh luận trực tuyến về sản phẩm [Somasundaran và Wiebe 2009]. Các công trình sau năm 2013 được thực hiện trên các bài luận của sinh viên [Faulkner 2014], và trên các tweet [Rajadesingan và Liu 2014]. Những chủ đề phổ biến được lấy trên các cuộc tranh luận trực tuyến ấy cho các mục tiêu lập trường (stance targets) bao gồm sự tiến hóa, quyền sử dụng súng, quyền của người đồng tính, phá thai, chăm sóc sức khỏe, án tử hình và sự tồn tại của Chúa[Dilek Küçük và Fazli Can. 2020]. Các cách phân loai lập trường trong các nghiên cứu đó cũng rất đa dạng, thay vì sử sụng bộ phân loại lập trường (Favor, Against) các nhà nghiên cứu còn sử dụng một số bộ khác như (Support, Oppose), (Pro, Con), và (Pro, Anti).

Các phương pháp được được sử dụng bao gồm thuật toán dựa trên quy tắc (rule-based algorithms) như JRip) [Anand et al. 2011; Murakami và Raymond 2010; Walker et al. 2012a, 2012b]; các thuật toán giám sát (supervised algorithms) SVM [Hasan và Ng 2013; Somasundaran và Wiebe 2010; Thomas et al. 2006; Walker et al. 2012b], naïve Bayes [Anand et al.2011; Hasan và Ng 2013; Rajadesingan và Liu 2014; Walker et al. 2012b], boosting [Levow et al.2014], cây quyết định (decision tree) và random forest [Misra và Walker 2013], Hidden Markov Models (HMM) và trường điều kiện ngẫu nhiên (Conditional Random Fields -CRF) [Hasan và Ng 2013]; thuật toán đồ thị (graph algorithms) MaxCut [Murakami và Raymond 2010; Walker et al. 2012a], và các cách tiếp cận khác như quy hoạch tuyến tính (Integer Linear Programming -ILP) [Somasundaran và Wiebe 2009] [2].

Bên cạnh đó, các cuộc thi về phát hiện lập trường cũng thu hút rất nhiều sự quan tâm. Theo như tìm hiểu thì cho đến hiện tại có 3 cuộc thi liên quan về chủ đề trên như (1) SemEval-2016 Task 6: phát hiện lập trường trong các tweet tiếng Anh, (2) phát hiện lập trường trong các blog nhỏ của Trung Quốc tại NLPCC-ICCPOL-2016 , (3) Phát hiện lập trường trên các Tweet tiếng Tây Ban Nha và Catalan tại IberEval-2017. Các cuộc thi đó cung cấp chi tiết các hướng dẫn chú thích, tập dữ liệu chú thích, chỉ số đánh giá và mô tả về các tác phẩm tham gia.

Cuộc thi thứ nhất gồm 2 nhiệm vụ: nhiệm vụ A và nhiệm vụ B. Nhiệm vụ A gồm tập dữ liệu đào tạo có chú thích gồm 2.814 tweet và tập dữ liệu thử nghiệm gồm 1.249 tweet được cung cấp cho tổng số năm mục tiêu. Nhiệm vụ B gồm tập dữ liệu đào tạo không được gắn nhãn (khoảng 78.000 tweet) và tập dữ thử nghiệm nhỏ hơn

(khoảng 707 tweet) cho một mục tiêu khác được cung cấp cho những người tham gia mà không có bất kỳ dữ liệu đào tạo nào được chú thích.

Kết quả của cuộc thi như sau: đối vơi nhiệm vụ A hệ thống tốt nhất dựa trên Mạng nơ-ron hồi quy (RNN - Recurrent Neural Network) đạt F-score là 67,82%, nhiệm vụ B hệ thống tốt nhất dựa trên Mạng nơ-ron tích chập (Convolutional Neural Networks -CNN) đạt F-score là 56.28%. Cuộc thi thứ hai cũng tương tự như cuộc thi SemEval-2016 với hai nhiệm vụ. Đối với nhiệm A: 4.000 blog nhỏ được gắn nhãn thủ công cho năm mục tiêu và 75% trong số đó được sử dụng làm tập dữ liệu đào tạo trong khi phần còn lại 25% trong số chúng được sử dụng làm tập dữ liệu thử nghiệm. Kết quả của cuộc thi như sau: hệ thống đạt F-score cao nhất (71.06%) với các phân loại được sử dụng dựa trên SVM and random forest trong khi hệ thống đạt F-score cao nhất của nhiệm vụ B chỉ đạt 46.87%.

Điều này là do người tham gia sử dụng nhiều cách phân loại và sử dụng hệ thống phân tích cảm tính hiệu suất cao có thể không đảm bảo hiệu suất phát hiện lập trường được cải thiện. Cuộc thi thứ ba cũng tương tự với 5,400 tweets tiếng Tây Ban Nha and 5,400 tweets tiếng Catalan. Hệ thống hoạt động tốt nhất việc phát hiện lập trường trên các tweet của Tây Ban Nha dựa trên cách tiếp cận dựa trên SVM với sự kết hợp của các tính năng khác nhau. Trong khi hệ thống hoạt động tốt nhất trên các tweet của Catalan dựa trên hồi quy logistic.

1.3 Tính thời sự của bài toán

Phát hiện lập trường là một chủ đề mới nổi trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (Natural language processing - NLP) thu hút rất nhiều sự quan tâm của các nhà nghiên cứu bởi các ứng dụng thực tế. Các nhà nghiên cứu hiện nay chủ yếu tiếp cận vấn đề phát hiện lập trường bằng tiếng Anh. Nhận thấy tầm quan trọng của chủ đề cùng với việc phát hiện lập trường cho tiếng Việt chưa được quan tâm nhiều, đã thôi thúc chúng tôi chọn nghiên cứu đề tài "Phát hiện lập trường cho tiếng Việt sử dụng kỹ thuật học sâu".

Ngoài ra lý do cần phải phát triển một hệ thống riêng cho Tiếng Việt là bởi Tiếng Anh và tiếng Việt có nhiều điểm khác biệt (do loại hình ngôn ngữ, do nền văn hoá,…). Khác về ngữ âm học, hình vị, ranh giới từ, sự từ vựng hoá (như: ox – bò đực, anh – elder brother ,…), từ loại, trật tự từ, kết cấu câu. [8] Trong tiếng Anh và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. [9] Do đó việc áp

dụng bài toán Tiếng Anh cho tiếng Việt có thể gây ra nhiều khó khăn. Trong nghiên cứu của chúng tôi, chúng tôi đã dành rất nhiều thời gian cho việc xây dựng một bộ ngữ liệu Tiếng Việt, dựa trên việc thu thập các đầu tin tức, các bình luận trên các kênh truyền thông mạng xã hội và các trang báo điện tử để thử nghiệm cho bài toán phát hiện lập trường của mình.

Với bộ ngữ liệu này chúng tôi hy vọng có thể đóng góp một phần nhỏ trong việc làm phong phú thêm tài nguyên ngôn ngữ trong lĩnh vực xử lý ngôn tự nhiên ứng dụng cho Tiếng Việt. Chúng tôi cũng hy vọng đề tài này có thể là tiền đề quan trọng cho các chuyên gia trong việc phân tích, nghiên cứu tình hình và xu hướng của dư luận xã hội.

1.4 Kết luận chương

Chương này đã giới thiệu tổng quan bài toán phát hiện lập trường, nêu bật được đặc điểm của dữ liệu Tiếng Việt, đưa ra được các nghiên cứu phát hiện lập trường liên quan và giới thiệu được một số phương pháp phát hiện lập trường.

CHƯƠNG II: CÁC PHƯƠNG PHÁP HỌC MÁY SỬ DỤNG TRONG BÀI TOÁN PHÁT HIỆN LẬP TRƯỜNG

Tiếp cận dựa trên học máy là cách tiếp cận được sử dụng phổ biến rộng rãi để giải quyết bài toán phát hiện lập trường. Cách tiếp cận này sẽ thay thế các kiến thức chuyên môn bằng một tập lớn các câu hỏi được gán nhãn (tập dữ liệu mẫu). Sử dụng tập này, một bộ phân lớp sẽ được huấn luyện có giám sát.

Cách tiếp cận dựa trên học máy chia làm hai nhóm là nhóm các phương pháp học máy truyền thống và nhóm các phương pháp sử dụng mạng nơ-ron (Neural NetWork). Nhóm các phương pháp học máy truyền thống thường được sử dụng như là tính xác suất Naïve Bayes, Maximum Entropy, cây quyết định (Decision Tree), lân cận (Nearest-Neighbors), Máy Vector hỗ trợ (Support Vector machine - SVM), K-nearest neighbors (KNN), Mô hình mạng nơ-ron hồi quy RNN, Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks) LSTM v.v. Cách tiếp cận bằng học máy đã giải quyết được các hạn chế trong cách tiếp cận dựa trên luật.

2.1 Phương pháp học máy truyền thống

2.1.1 Thuật toán phân lớp naïve Bayes

Naive Bayes Classification (NBC) – thuật toán phân loại Naive Bayes - là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê, được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán có độ chính xác cao, dựa trên một tập dữ liệu đã được thu thập.

Naive Bayes Classification thuộc vào nhóm học máy có giám sát. Kỹ thuật này dễ hiểu nhất khi được mô tả bằng các giá trị đầu vào nhị phân hoặc phân loại. Thuật toán Naive Bayes tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất. Tuy nhiên, ta cần lưu ý giả định của thuật toán Naive Bayes là các yếu tố đầu vào được cho là độc lập với nhau.

Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính A1, A2,...,An = {x1,x2,...,xn}. Giả sử có m lớp C1, C2,..,Cm. Cho một phần tử dữ liệu X, bộ phân lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân lớp Bayes sẽ dự đoán X thuộc vào lớp Ci nếu và chỉ nếu:

$$P(Ci|X) > P(Cj|X) \ (1<= i, j <=m, i != j)$$

Giá trị này sẽ tính dựa trên định lý Bayes.

Để tìm xác suất lớn nhất, ta nhận thấy các giá trị P(X) là giống nhau với mọi lớp nên không cần tính. Do đó ta chỉ cần tìm giá trị lớn nhất của P(X|Ci) * P(Ci). Chú ý rằng P(Ci) được ước lượng bằng |Di|/|D|, trong đó Di là tập các phần tử dữ liệu thuộc lớp Ci. Nếu xác suất tiền nghiệm P(Ci) cũng không xác định được thì ta coi chúng bằng nhau P(C1) = P(C2) = ... = P(Cm), khi đó ta chỉ cần tìm giá trị P(X|Ci) lớn nhất.

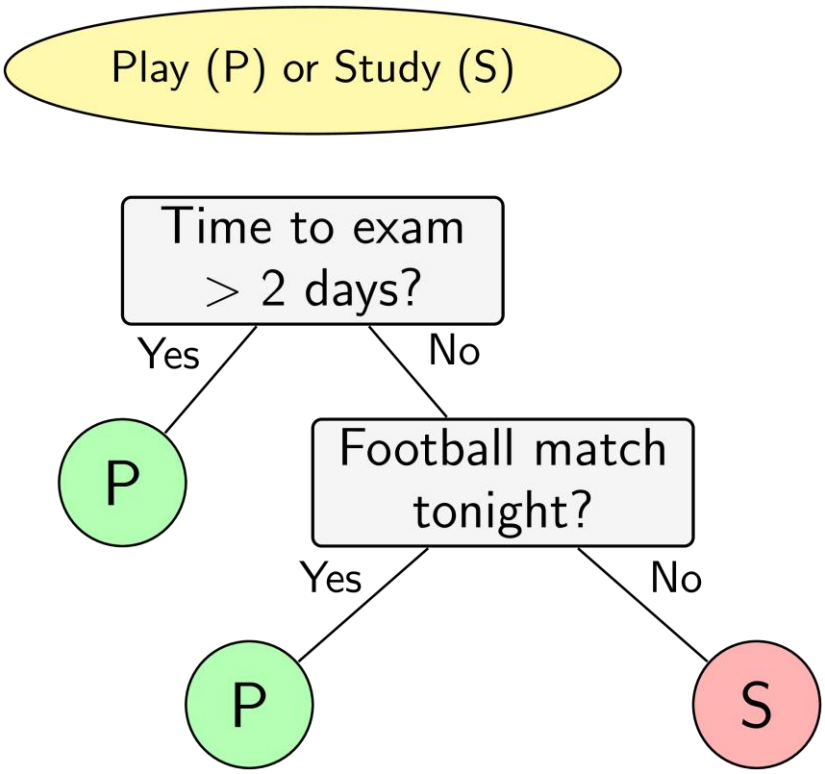
Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toàn P(X|Ci) là rất lớn, dó đó có thể giảm độ phức tạp của thuật toán Naive Bayes giả thiết các thuộc tính độc lập nhau. Khi đó ta có thể tính:

$$P(X|C_i) = P(x_1|C_i) \ldots P(x_n|C_i)$$

2.1.1 Cây quyết định (Decision tree)

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Khi cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết.

Cây quyết định gồm 3 phần chính: 1 node gốc, những node lá và các nhánh của nó. Node gốc là điểm bắt đầu của cây quyết định và cả hai node gốc và node chứa câu hỏi hoặc tiêu chí để được trả lời. Nhánh biểu diễn các kết quả của kiểm tra trên nút. Ví dụ câu hỏi ở node đầu tiên yêu cầu câu trả lời là “yes” hoặc là “no” thì sẽ có 1 node con chịu trách nhiệm cho phản hồi là “yes”, 1 node là “no”.



Hình 2.1: Minh họa cây quyết định[2]

Việc quan sát, suy nghĩ và ra các quyết định của con người thường được bắt đầu từ các câu hỏi. Machine learning cũng có một mô hình ra quyết định dựa trên các câu hỏi. Mô hình này có tên là cây quyết định.

Hiệu của phân lớp của cây quyết định phụ thuộc rất lớn vào training data. Chẳng hạn cây quyết định được tạo ra bởi chỉ giới hạn 10 dữ liệu training mẫu trong ví dụ trên thì hiệu quả ứng dụng cây quyết định để dự đoán các trường hợp khác là không cao (thường dữ liệu training phải đủ lớn và tin cậy). Cây quyết định là một phương pháp phân lớp rất hiệu quả và dễ hiểu.

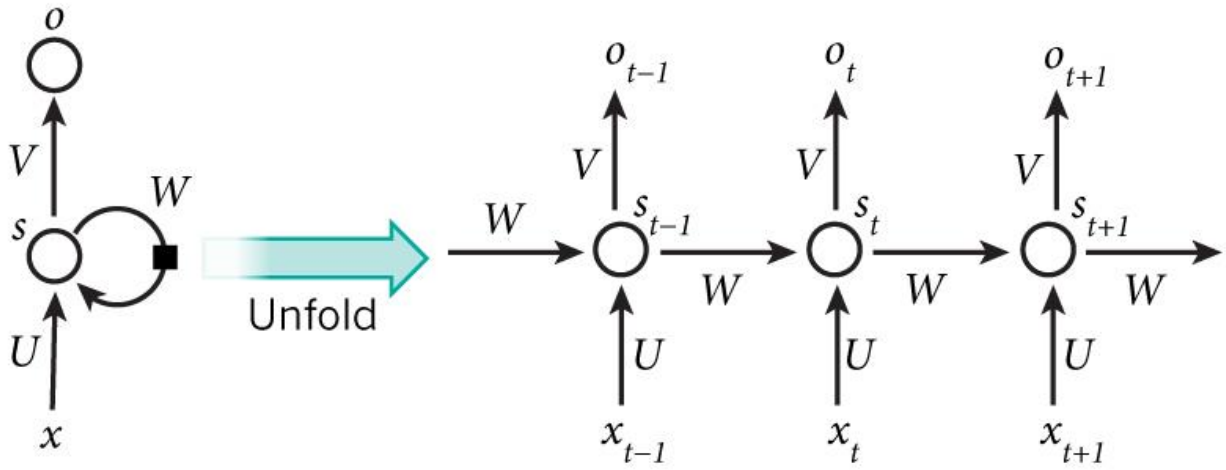
2.2 Phương pháp học sâu

2.2.1. Mô hình mạng nơ-ron hồi quy (RNN - Recurrent Neural Network)

RNN (Recurrent Neural Network) – Mạng nơ-ron hồi quy là một thuật toán được chú ý rất nhiều trong thời gian gần đây bởi các kết quả tốt thu được trong lĩnh vực xử lý ngôn ngữ tự nhiên, được thiết kế cho việc xử lý các loại dữ liệu có dạng chuỗi tuần tự.

Khác với các mạng nơ-ron truyền thống, mạng RNN tất cả các đầu vào và cả đầu ra có thể được thực hiện cùng một tác vụ cho các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó.

Ý tưởng chính của RNN (Recurrent Neural Network) là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau.



Hình 2.2: Mạng RNN và quá trình unfold liên quan đến tính toán chuyển tiếp[5].

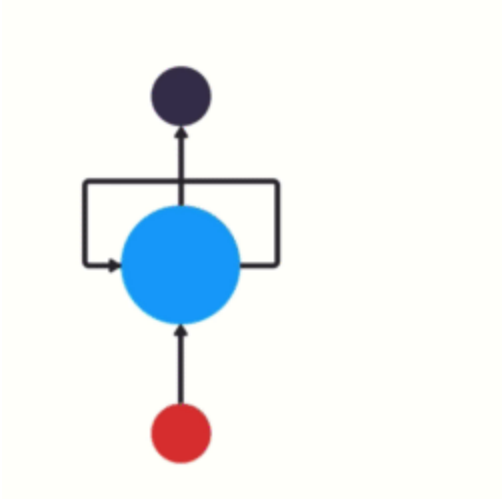
Ví dụ, nếu muốn đoán từ tiếp theo có thể xuất hiện trong một câu thì ta cũng cần biết các từ trước đó xuất hiện lần lượt thế nào.

RNN: Dựa vào những thông tin trước đó để dự đoán trạng thái tiếp theo.

Giả sử muốn dự đoán hướng mà quả bóng di chuyển. Vì vậy, chỉ với thông tin trên màn hình, ta có thể tiếp tục và đoán. Nếu không có kiến thức về các vị trí phía trước của quả bóng, sẽ không có đủ dữ liệu để dự đoán nó sẽ đi tiếp theo ở vị trí nào. Nếu ghi liên tiếp nhiều ảnh chụp nhanh vị trí của quả bóng, ta sẽ có đủ thông tin để đưa ra dự đoán tốt hơn. Vì vậy, đây là một trình tự, một trật tự cụ thể, trong đó trạng thái sau nối tiếp một trạng thái khác. Với thông tin này, ta có thể thấy rằng quả bóng đang di chuyển sang bên phải.

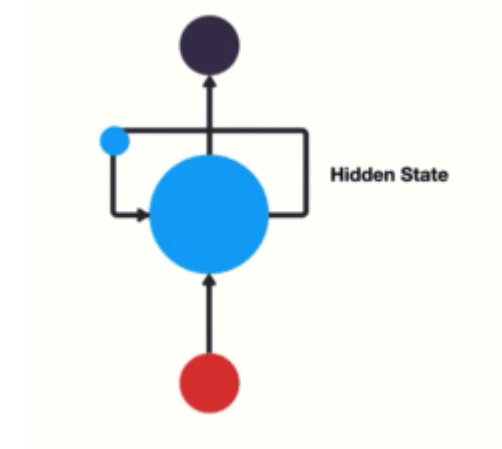
Hình 2.3: Mô tả ví dụ RNN với đường đi tiếp theo của quả bóng[5]

RNN được gọi là hồi quy (Recurrent) bởi vì chúng thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó. Nói cách khác, RNN có khả năng nhớ các thông tin được tính toán trước đó. Trên lý thuyết, RNN có thể sử dụng được thông tin của một văn bản



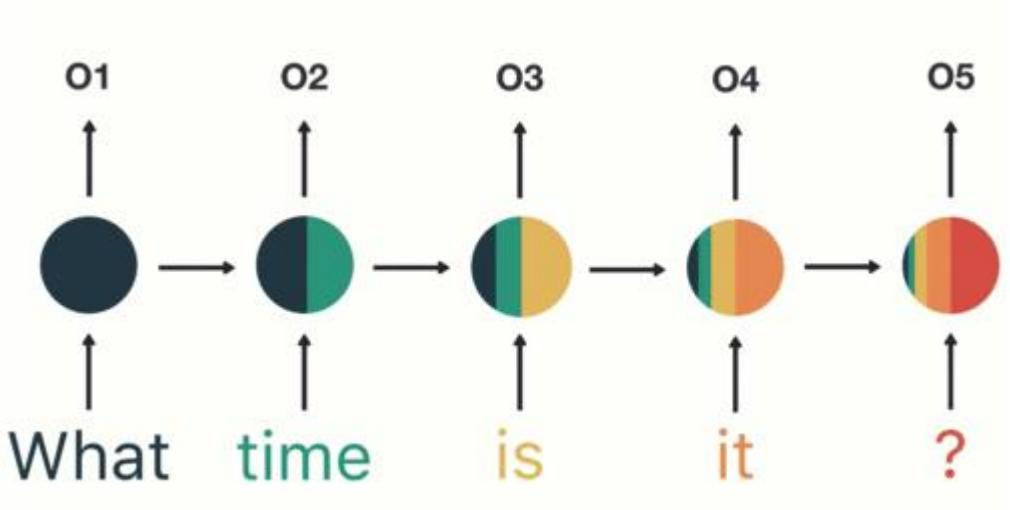
Hình 2.4: Mô hình Recurrent Neural Network[5]

RNN có cơ chế lặp hoạt động như một đường chính để cho phép thông tin truyền từ trạng thái này sang trạng thái tiếp theo.



Hình 2.5: Biểu diễn trạng thái ẩn đến bước tiếp theo[5]

Thông tin này là trạng thái ẩn, là đại diện của các đầu vào trước đó.



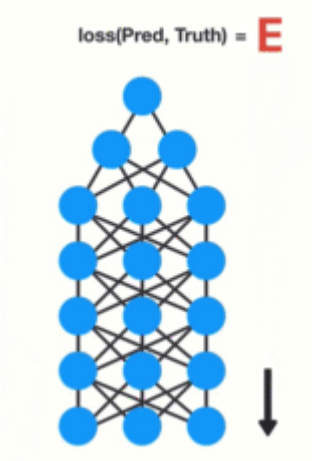
Hình 2.6: Ví dụ biểu diễn câu với RNN[5]

```
rnn = RNN()
ff = FeedForwardNN()
hidden_state =[0.0, 0.0, 0.0, 0.0]

for word in input:
    output, hidden_state = rnn(word, hidden_state)

prediction = ff(output)
```

Hình 2.7: Mã giả cho luồng điều khiển RNN[5]

Training một mạng nơ-ron có ba bước chính.

- Chuyển tiếp và đưa ra dự đoán.
- So sánh dự đoán với sự thật cơ bản bằng cách sử dụng một loss funtion. Loss funtion xuất ra một giá trị lỗi là giá trị ước tính mạng hoạt động kém như thế nào.
- Sử dụng giá trị lỗi đó để thực hiện lan truyền ngược, tính toán độ dốc cho mỗi nút trong mạng.



Hình 2.8: Traning mạng nơ-ron[5]

Gradient là giá trị được sử dụng để điều chỉnh trọng số nội bộ của mạng, cho phép mạng tìm hiểu. Gradient càng lớn thì điều chỉnh càng lớn và ngược lại. Khi thực hiện lan truyền ngược, mỗi nút trong một lớp sẽ tính toán gradient của nó đối với các hiệu ứng của gradient, trong lớp trước nó. Vì vậy, nếu các điều chỉnh cho các lớp trước nó là nhỏ, thì các điều chỉnh cho lớp hiện tại sẽ thậm chí còn nhỏ hơn.

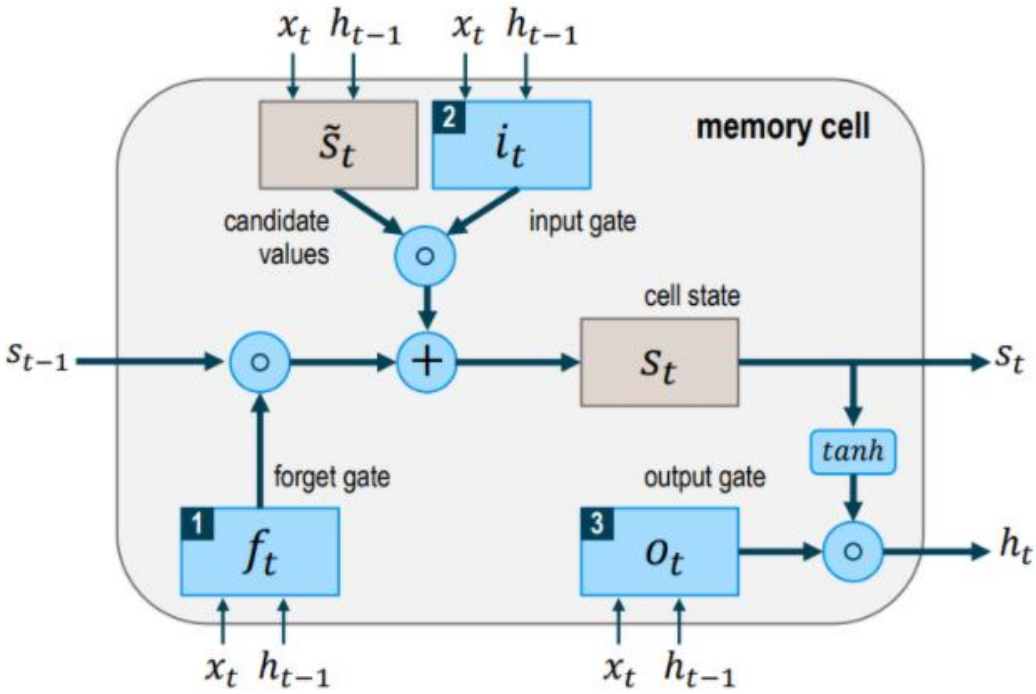
Hình 2.9: Minh họa trạng thái ẩn cuối của mạng RNN[5]

Điều đó làm cho gradient thu nhỏ lại theo cấp số nhân khi lan truyền trở lại. Các lớp trước đó không thực hiện được bất kỳ hoạt động nào vì giá trị bên trong hầu như không được điều chỉnh do độ dốc cực nhỏ. Và đó là vấn đề về gradient nhỏ lại và biến mất.

Do gradient biến mất, RNN qua các bước không phụ thuộc trong phạm vi văn bản dài. Như ví dụ bên trên, điều đó có nghĩa là có khả năng từ “What” và “time” không được xem xét khi dự đoán ý định của người dùng. Sau đó, mạng phải đưa ra dự đoán với “is it?”. Điều đó khá mơ hồ và sẽ khó ngay cả đối với con người. Vì vậy, RNN bị ảnh hưởng bởi trí nhớ ngắn hạn short-term memory.

2.2.2. Mạng bộ nhớ dài - ngắn (Long Short Term Memory)

LSTM là một kiến trúc mạng nơ ron lặp lai nhân tạo (RNN) được sử dụng trong lĩnh vực học sâu. Nó được thiết kế để giải quyết các bài toán về phụ thuộc xa trong mạng RNN do bị ảnh hưởng bởi vấn đề gradient biến mất. Không giống như các mạng nơ ron truyền thẳng tiêu chuẩn , LSTM có các kết nối phản hồi. Nó không chỉ có thể xử lý các điểm dữ liệu đơn lẻ (chẳng hạn như hình ảnh), mà còn toàn bộ chuỗi dữ liệu (chẳng hạn như lời nói hoặc video).



Hình 2.10: Mô hình kiến trúc tế bào LSTM memory cell[15]

Mạng LSTM có thể bao gồm nhiều tế bào LSTM memory cell liên kết với nhau, một LSTM có ba cổng sàng lọc các thông tin đầu vào và đầu ra cho tế bào bao gồm một cổng đầu vào (input gate i_t) có nhiệm vụ chọn lọc những thông tin cần thiết nào được thêm vào cell internal state, một cổng quên (forget gate f_t) xóa thông tin không cần thiết nhận được khỏi cell internal state và một cổng đầu ra (output gate o_t) quyết định những thông tin nào từ cell internal state được sử dụng như đầu ra.

Tại mỗi bước thời gian t, các cổng đều lần lượt nhận giá trị đầu vào x_t (đại diện cho một phần tử trong chuỗi đầu vào) và giá trị h_{t-1} có được từ đầu ra của memory cell từ bước thời gian trước đó t-1. Ba cổng này là cổng tương tự dựa trên chức năng sigmoid hoạt động trên phạm vi 0 đến 1.

Các cổng này quyết định thông tin nào là quan trọng để phân loại và thông tin nào có thể xóa được dựa trên giá trị đã bỏ. Trong quá trình lan truyền xuôi, cell internal state s_t và giá trị đầu ra h_t được tính như sau:

Bước một, Activation value f_t của forget gate tại bước thời gian t được tính dựa trên giá trị đầu vào hiện tại x_t, giá trị đầu ra h_{t-1} từ tế bào LSTM ở bước trước đó và bias b_f của forget gate.

$$f_t = \sigma\left(W_{f,x}x_t + W_{f,h}h_{t-1} + b_f\right)$$

Bước hai, quá trình tính toán của $\tilde{s_t}$ và f_t

$$\tilde{s_t} = \tanh\left(W_{\tilde{s},x}x_t + W_{\tilde{s},h}h_{t-1} + b_{\tilde{s}}\right)$$

$$i_t = \tanh(W_{i,x}x_t + W_{i,h}h_{t-1} + b_i)$$

Bước ba, giá trị mới của cell internal state s_t được tính dựa trên kết quả tính toán thu được từ bước trước với phép nhân Hadamard theo từng phân tử được ký hiệu bằng o:

$$s_t = f_t \circ s_{t-1} + i_t \circ \tilde{s_t}$$

Bước cuối, giá trị h_t của tế bào LSTM được tính toán dựa theo hai phương trình sau:

$$o_t = \sigma\left(W_{o,x}x_t + W_{o,h}h_{t-1} + b_o\right)$$

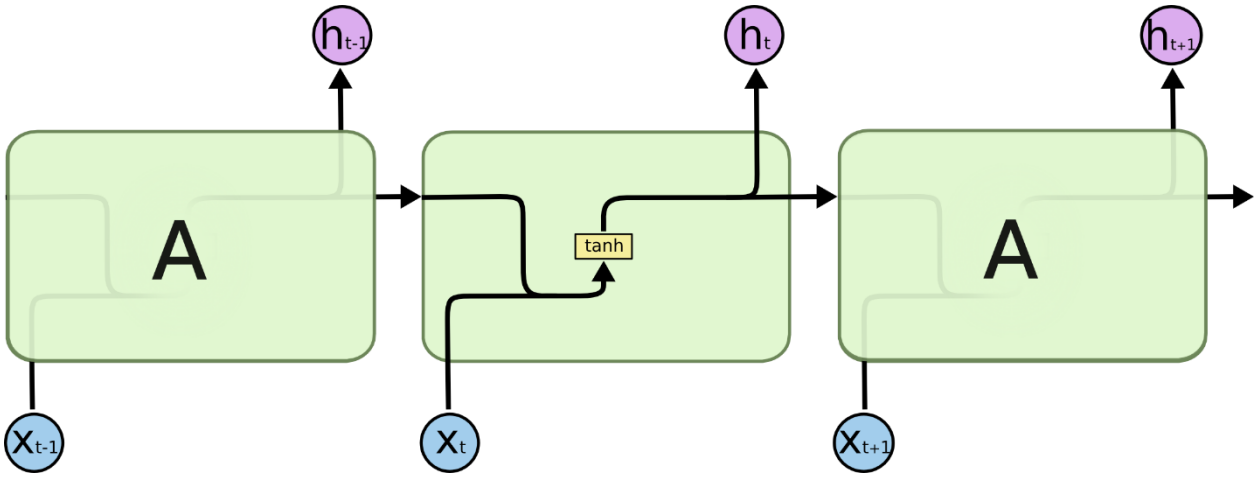
$$h_t = o_t \circ \tanh(s_t)$$

trong đó x_t là vector đầu vào tại mỗi bước thời gian t, $W_{f,x}$, $W_{f,h}$, $W_{s,x}$, $W_{s,h}$, $W_{i,x}$, $W_{i,h}$, $W_{o,x}$, $W_{o,h}$ là các ma trận trọng số trong mỗi tế bào LSTM. b_f, b_s, b_i, b_o là vector bias. f_t, i_f, o_t lần lượt chứa các giá trị kích hoạt lần lượt cho các cổng forget gate, input gate và output gate tương ứng. S_t, ~s là các vector đại diện cho cell internal state và candidate value. h_t là giá trị đầu ra của tế bào LSTM.

LSTM là một mạng cải tiến của RNN nhằm giải quyết vấn đề nhớ các bước dài của RNN. Về cơ bản mô hình của LSTM không khác mô hình truyền thống của RNN, nhưng chúng sử dụng hàm tính toán khác ở các trạng thái ẩn

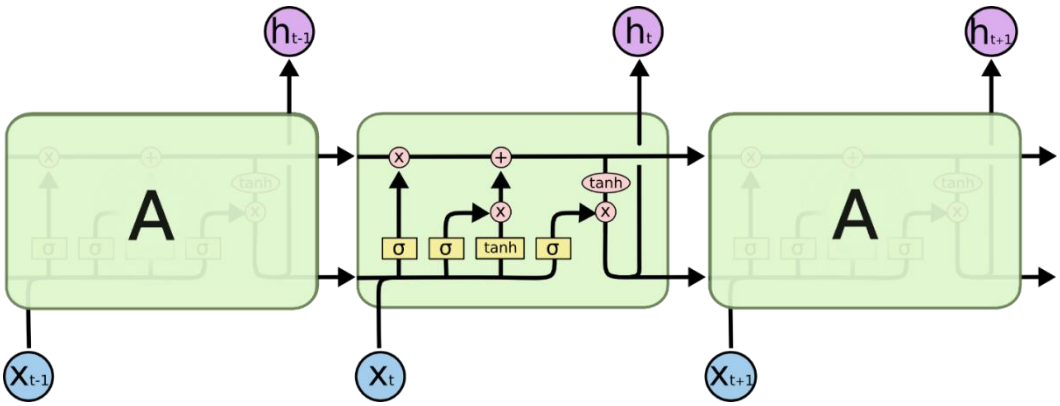
Bộ nhớ của LSTM được gọi là tế bào (Cell) và có thể coi chúng là các hộp đen nhận đầu vào là trạng thái phía trước h(t-1) và đầu vào hiện tại x(t). Bên trong hộp đen này sẽ tự quyết định cái gì cần phải nhớ và cái gì sẽ xoá đi. Sau đó, chúng sẽ kết hợp với trạng thái phía trước, nhớ hiện tại và đầu vào hiện tại. Vì vậy mà ta ta có thể truy xuất được quan hệ của các từ phụ thuộc xa nhau rất hiệu quả.

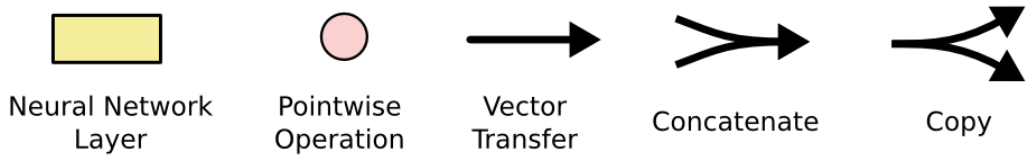
Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng *tanh.*



Hình 2.11*:* Mô-đun lặp lại trong RNN một lớp[15].

LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.

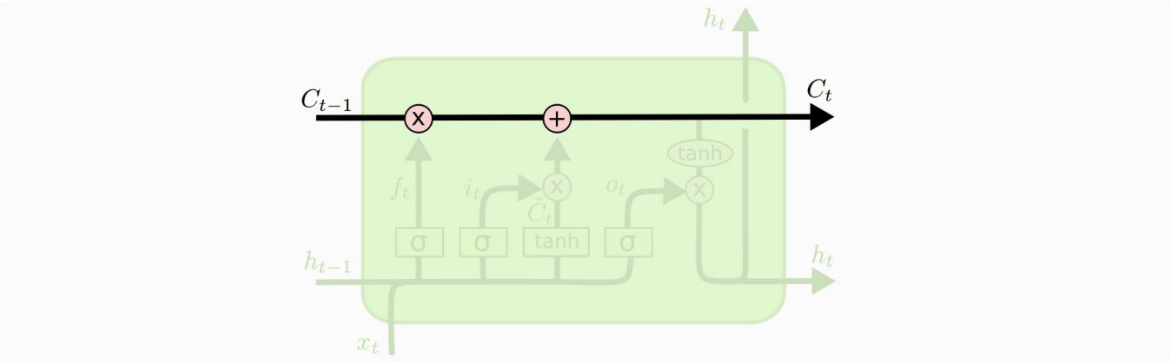




Hình 2.12: Mô-đun lặp lại trong một LSTM chứa bốn lớp tương tác[15]

Chìa khóa của LSTM là trạng thái tế bào (cell state) - chính đường chạy thông ngang phía trên của sơ đồ hình vẽ.

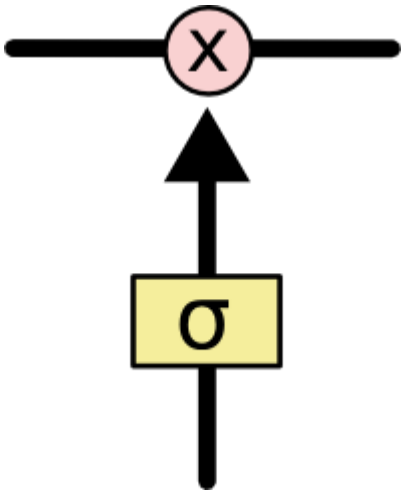
Trạng thái tế bào là một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) và chỉ tương tác tuyến tính đôi chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi.



Hình 2.13: Biểu diễn trạng thái tế bào[15]

LSTM có khả năng bỏ đi hoặc thêm vào các thông tin cần thiết cho trạng thái tế báo, chúng được điều chỉnh cẩn thận bởi các nhóm được gọi là cổng (gate).

Các cổng là nơi sàng lọc thông tin đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid và một phép nhân.

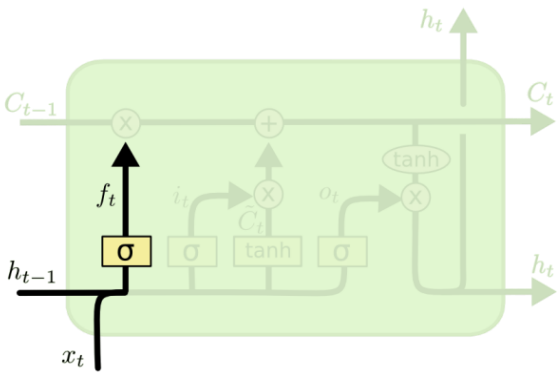


Hình 2.14: Biểu diễn cổng sàng lọc thông tin[15]

Tầng sigmoid sẽ cho đầu ra là một số trong khoản [0, 1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 0 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 1 thì có nghĩa là cho tất cả các thông tin đi qua nó.

Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.

Bước đầu tiên trong mô hình LSTM là việc quyết định thông tin nào sẽ được đưa đến trạng thái tế bào thông qua cổng. Quá trình này được thực hiện thông qua một lớp sigmoid gọi là "lớp cổng chặn" - cổng chặn với hai đầu vào là h_{t-1} và x_t, và cho đầu ra là một giá trị trong phạm vi [0, 1] cho mỗi đầu vào trạng thái ô C_{t-1}. 1 tương đương với "lưu giữ thông tin", 0 tương đương với "xóa thông tin".

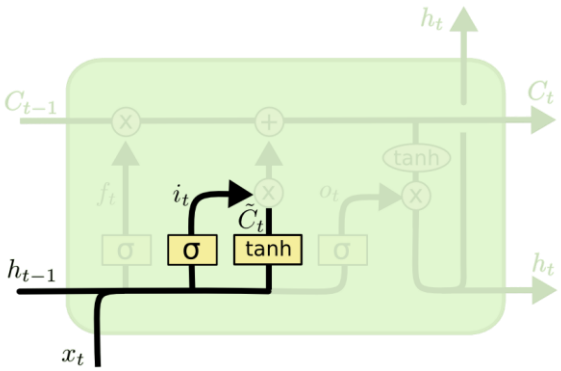


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Hình 2.15: LSTM focus f [22]

Bước tiếp theo là xác định loại thông tin mới nào cần được lưu lại trong cell state. Ta có hai phần. Một là single sigmoid layer được gọi là “input gate layer” quyết

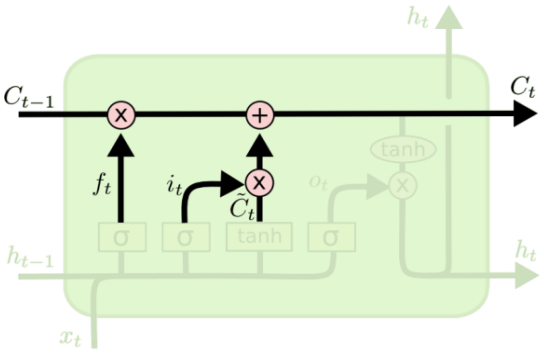
định các giá trị nào cần được cập nhật. Tiếp theo, một $tanh$ layer tạo ra một vector với giá trị mới có thể đưa vào cell state, C_t được thêm vào trong ô trạng thái.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Hình 2.16*:* LSTM focus I [22]

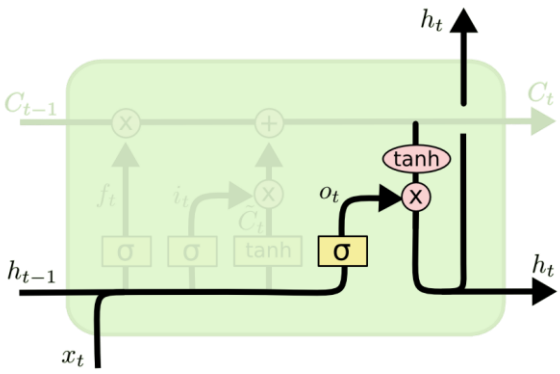
Ở bước tiếp theo, kết hợp hai thành phần này lại để cập nhật vào cell state. Lúc cập nhật vào cell state cũ C_{t-1} vào cell state mới C_t. Tại bước này thực hiện nhân trạng thái cũ với f_t, để cần nhớ hoặc quên đi những gì trước đó hay không. Sau đó, bổ sung $i_t * \tilde{C}_t$. Đây là giá trị ứng viên mới, co giãn (scale) số lượng giá trị mà ta muốn cập nhật cho mỗi state.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Hình 2.17*:* LSTM focus c [22]

Cuối cùng, cần quyết định xem thông tin output là gì. Output này cần dựa trên trạng thái của cell state, nhưng sẽ là giá trị được lọc bớt một số thông tin. Đầu tiên, chạy qua một single sigmoid layer để quyết định xem phần tử nào của cell state sẽ tác động đến output. Sau đó, ta sẽ đẩy cell state đi qua một function tanh giá trị khoảng [-1, 1] và nhân với một output sigmoid gate, để giữ lại những phần ta muốn output ra ngoài.



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Hình 2.1*:* LSTM focus o [22]

LSTM là một bước lớn trong việc sử dụng RNN. Ý tưởng của nó giúp cho tất cả các bước của RNN có thể truy vấn được thông tin từ một tập thông tin lớn hơn. Ví dụ, nếu sử dụng RNN để tạo mô tả cho một bức ảnh, nó có thể lấy một phần ảnh để dự đoán mô tả từ tất cả các từ đầu vào.

2.3. Kết luận chương

Nội dung chương đã giới thiệu được các phương pháp học máy sử dụng trong bài toán phát hiện lập trường, giới thiệu phương pháp học máy truyền thống, giới thiệu phương pháp học sâu và so sánh giữa hai phương pháp.

CHƯƠNG III: ĐỀ XUẤT PHƯƠNG PHÁP, GIẢI PHÁP

Trong chương này, luận văn giới thiệu bài toán phân loại phát hiện lập trường tiếng Việt, giới thiệu một số mô hình học sâu, giới thiệu phương pháp LSTM và trình bày mô hình phát hiện lập trường sử dụng mô hình LSTM.

3.1. Mô tả bài toán

Phát hiện lập trường là xây dựng một bộ phân loại để xác định lập trường của một nhận xét nhất định đối với một tuyên bố/ tiêu đề, với các bình luận có thể là đồng ý, không đồng ý, thảo luận và không liên quan.

Mô tả bài toán:

- **Input**: Một tuyên bố và một bình luận về tuyên bố đó.
- **Output**: Lập trường của nội dung bình luận liên quan đến tuyên bố được đưa ra thành một trong bốn loại:
  - *Agree*: Nội dung bình luận đồng ý với tuyên bố.
  - *Disagree*: Nội dung bình luận không đồng ý với tuyên bố.
  - *Discuss*: Nội dung bình luận mang tính chất thảo luận về chủ đề tương tự như tuyên bố, nhưng không đưa ra quan điểm.
  - *Unrelated*: Nội dung bình luậ thảo luận về một chủ đề khác với tuyên bố.

Ví dụ: Tuyên bố: "*Hà Nội khuyến cáo người dân đeo khẩu trang nơi công cộng.*" Bình luận: "*Quá đúng ,vẫn chưa hết dịch nên cần phòng tránh ko dc chủ quan.*" như vậy lập trường của người bình luận là *"đồng ý"* về tuyên bố, khuyến cáo nêu trên.

Trong phần này, luận văn đề xuất một mô hình để dự đoán lập trường của mỗi mội bình luận *d* sao cho: d ∈ D={d1,d2,d3, ..dn) hướng tới mục tiêu xác định lập trường của tuyên bố/ tiêu đề *h*. Ta có một tập cố định của các lớp s ∈ S= {agree, disagree, discuss, unrelated} là đầu ra lập trường của các bình luận tương ứng. Mục tiêu của bài toán là ánh xạ f : (d, h) → s {agree, disagree, discuss, unrelated}.

Nếu tuyên bố/ tiêu đề bao gồm các chủ đề khác nhau, lập trường là s = "unrelated". Ngược lại, *s* = "agree" nếu *d* đồng ý và *s* = "disagree" nếu *d* không đồng ý với *h*. Nếu *h* và *d* chỉ đơn thuần thảo luận về cùng một chủ đề, nhưng *d* không có quan điểm xác định thì *s* = "Discuss".

3.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một phần cực kỳ quan trọng trong việc xây dựng mô hình hoạt động tốt nhất cho các ứng dụng học máy. Trong nghiên cứu này, luận văn sử dụng phương pháp chuẩn hóa dữ liệu và mã hóa từ để tiền xử lý dữ liệu.

Chuẩn hóa dữ liệu

Chuẩn hóa văn bản giúp loại bỏ các ký tự đặc biệt như dấu chấm câu; đổi thành chữ thường.

Ví dụ: *"Hà Nội: Khuyến khích phân loại rác thải nhựa tại nguồn"*

Bằng cách sử dụng các phương thức Lower() để xử lý chuỗi viết thường và phương thức Replace() để loại bỏ các dấu câu.

Từ đó ta có đầu ra: *"hà nội khuyến khích phân loại rác thải nhựa tại nguồn"*

Thuật toán tách từ Tokenization

Tokenization là tách một cụm từ, câu, đoạn văn hoặc toàn bộ tài liệu văn bản thành các đơn vị nhỏ hơn thành các từ có ý nghĩa. Mỗi từ bao gồm 1 hoặc nhiều âm tiết. Khác với tiếng anh, mỗi từ được tách biệt với nhau bởi dấu cách, trong tiếng Việt các từ có thể bao gồm nhiều từ đơn.

Mã hóa là một phần cơ bản của quá trình xử lý NLP (dữ liệu văn bản) vì ý nghĩa của văn bản có thể dễ dàng được giải thích bằng cách phân tích các từ có trong văn bản. Và cũng một phần quan trọng của việc chuyển đổi dữ liệu từ văn bản (chuỗi) sang dữ liệu số vì các mô hình học máy cần dữ liệu số để được traning và đưa ra các dự đoán.

Ví dụ: *"hà nội khuyến khích phân loại rác thải nhựa tại nguồn"*

Sau khi thực hiện tách từ, ta được:
['hà_nội','khuyến_khích','phân_loại','rác_thải','nhựa','tại','nguồn']

3.3. Phát hiện lập trường sử dụng mô hình học máy truyền thống

3.3.1. Trích chọn đặc trưng

Trích chọn đặc trưng có ý nghĩa quan trọng, ảnh hưởng trực tiếp đến kết quả phân lớp. Các loại đặc trưng chính thường được sử dụng là tập từ (bag-of-word). Ngoài ra, trong phạm vi đồ án, chúng tôi còn sử dụng thêm các đặc trưng khác như đặc trưng âm tiết (Bag-of-syllables), âm tiết quan trọng, phân loại dựa trên Naïve bayes, biểu diễn từ bằng Vector (Vector glove), Log-count ratios của câu, từ phủ định.

Đặc trưng từ vựng

Với đặc trưng từ vựng, một câu sẽ được biểu diễn dưới dạng một tập các từ riêng biệt, không quan tâm tới ngữ pháp hay thứ tự của các từ trong câu, chỉ giữ lại số lần xuất hiện của từ trong câu.

Không giống như tiếng Anh, mỗi một âm tiết là một từ và được viết cách nhau bởi một khoảng trắng. Với tiếng Việt, một từ có thể được viết bởi một hoặc nhiều âm tiết, do đó không thể dùng khoảng trắng làm ranh giới phân cách các từ.

Ví dụ trong tiếng anh chúng ta có từ mobile, khi dịch ra tiếng Việt mobile có nghĩa là điện thoại, được tạo thành từ 2 âm tiết là điện và thoại.

Ví dụ với 2 câu:

Câu 1: "Thật sự là mới ở khách sạn lần đầu nhưng cảm giác rất thích thú."

Câu 2: "Khách sạn không bố trí chỗ gửi xe miễn phí cho du khách."

Khi tách từ ở bước tiền xử lý chúng ta có:

Câu 1: "Thật_sự là mới ở khách_sạn lần_đầu nhưng cảm_giác rất thích_thú."

Câu 2: "Khách_sạn không bố_trí chỗ gửi_xe miễn_phí cho du_khách."

Biểu diễn đặc trưng:

Xây dựng từ điển:

{

1 : Thật_sự , 2 : là , 3 : mới , 4 : ở , 5 : lần_đầu , 6 : nhưng , 7 : cảm_giác 8 : rất , 9 : thích_thú , 10 : khách_sạn , 11 : không , 12 : bố_trí , 13 : chỗ 14 : gửi_xe , 15 : miễn_phí , 16 : cho , 17: du_khách

}

Biểu diễn 2 câu trên dưới dạng vector đặc trưng, mỗi phần tử của vector có dạng: <vị trí của từ trong từ điển>: <1>, dựa vào chỉ số trong từ điển ta có 2 vector:

Câu 1: [1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1, 11:1]

Câu 2: [10:1, 11:1, 12:1, 13:1, 14:1, 15:1, 16:1,17:1]

Trong câu 1, từ "Thật_sự" xuất hiện trong câu, và nó có vị trí thứ 1 trong từ điển nên được biểu diễn là "1:1", từ "khách_sạn" xuất hiện trong câu, và có vị trí thứ 10 trong từ điển nên được biểu diễn là "10:1", các từ khác cũng được biểu diễn tương tự theo cách như vậy. Vector biểu diễn các từ không theo trật tự xuất hiện của từ trong câu, mà theo trật tự từ điển.

Biểu diễn các từ bằng Vector Glove

Phương pháp biểu diễn Vector từ Glove là một phương pháp học không giám

sát, sử dụng để biểu diễn một từ thành một vector tương ứng. Glove là một thuật toán biểu diễn cho vector các từ huấn luyện được thực hiện trên số liệu thống kê từ các từ đồng xảy tổng hợp từ corpus, và kết quả biểu diễn là không gian vector từ N chiều.

Ví dụ:

Ta có câu: "*King is to queen as man is to woman*"

Câu này được mã hóa trong không gian vector bằng phương trình vector *king - queen = man - woman.*

Chương trình đánh giá này thuận lợi cho mô hình sản sinh chiều của ý nghĩa, qua đó nắm bắt được ý tưởng multi-clustering của biểu diễn phân phối. Số lần xuất hiện của từ-từ đồng xảy ra được biểu thị bởi ma trận X, trong đó X_{ij} biểu diễn số lần từ j xảy ra trong ngữ cảnh chứa từ i.

Khi đó: $X_i = \sum_K X_{ik}$: số lần bất kỳ từ nào xuất hiện trong ngữ cảnh chứa từ i.

Và $P_{ij} = P_{j|i} = X_{ij}|X_i$: là xác suất của từ j xuất hiện trong ngữ cảnh chứa từ i.

Ví dụ:

Bảng 3.1: Xác suất đồng xảy ra với các từ ice và steam với những từ được chọn trong corpus.

Probability and Ratio	k=solid	k=gas	k=water	k=fashion
P(k|ice)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P(k|steam)	2.2×10^{-4}	7.8×10^{-5}	2.2×10^{-3}	1.8×10^{-5}
P(k|ice)/ P(k|steam)	8.9	8.5×10^{-2}	1.36	0.96

Bắt đầu với một ví dụ đơn giản mà giới thiệu cách khía cạnh nhất định của ý nghĩa có thể được chiết xuất trực tiếp từ xác suất đồng xảy ra. Xem xét hai chữ i và j mà biểu lộ một khía cạnh cụ thể. Giả sử đang xét khái niệm về giai đoạn nhiệt động lực học, lấy i = ice và j = steam.

Các mối quan hệ của những từ này có thể được kiểm tra bằng cách nghiên cứu các tỷ lệ xác suất đồng xảy ra với các từ khác nhau, k. Đối với những từ k liên quan đến ice nhưng không liên quan đến steam, k = rắn, dự kiến tỷ lệ $P_{ik}|P_{jk}$ là rất lớn.

Tương tự như vậy, các từ k liên quan đến steam nhưng không liên quan đến ice, nói k = gas, tỷ lệ $P_{ik}|P_{jk}$ phải nhỏ. Đối với những từ k = water hay fastion, một trong

hai liên quan đến cả hai băng và hơi nước, nên tỷ lệ $P_{ik}|P_{jk}$ gần một. So với các xác suất ban đầu, tỷ lệ này có khả năng tốt hơn để phân biệt có liên quan từ (solid và gas) từ những từ không liên quan (water và fashion) và nó cũng có thể tốt hơn để phân biệt giữa hai từ có liên quan.

Trong bài toán phân loại quan điểm, ta sử dụng công cụ Glove Vector để biểu diễn các từ trong 1 câu. Khi đó để biểu diễn câu s chứa các từ w_1,w_2,..,$w_{|s|}$ với |s| là độ dài câu, ta cộng tương ứng các Vector của từng từ w_1,w_2,..,$w_{|s|}$.

Đặc trưng độ đo TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) là một kỹ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. Tf-idf cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.

Term Frequency (Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản(tổng số từ trong một văn bản).

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d): w \in d\}}$$

Trong đó:

$tf(t,d)$: tần suất xuất hiện của từ t trong văn bản d

$f(t,d)$: Số lần xuất hiện của từ t trong văn bản d

$\max\{f(w,d): w \in d\}$: Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

Inverse Document Frequency (Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ. Khi tính toán TF, tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như “is”, “of” và “that” thường xuất hiện rất

nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống.

$$idf(t,d) = log\frac{|D|}{|\{\mathrm{d} \in \mathrm{D}: \mathrm{t} \in \mathrm{d}\}|}$$

Trong đó:

$idf(t,d)$: giá trị idf của từ t trong tập văn bản

|D|: Tổng số văn bản trong tập D

$|\{\mathrm{d} \in \mathrm{D}: \mathrm{t} \in \mathrm{d}\}|$: thể hiện số văn bản trong tập D có chứa từ t.

Cơ số logarit trong công thức này không thay đổi giá trị idf của từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi một số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. Việc sử dụng logarit nhằm giúp giá trị tf- idfcủa một từ nhỏ hơn, do công thức tính tf- idf của một từ trong 1 văn bản là tích của tf và idf của từ.

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao.

3.3.2. Các bước thực hiện

- Hai phương pháp Naïve Bayes và Decision Tree yêu cầu dữ liệu được biểu diễn như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số.

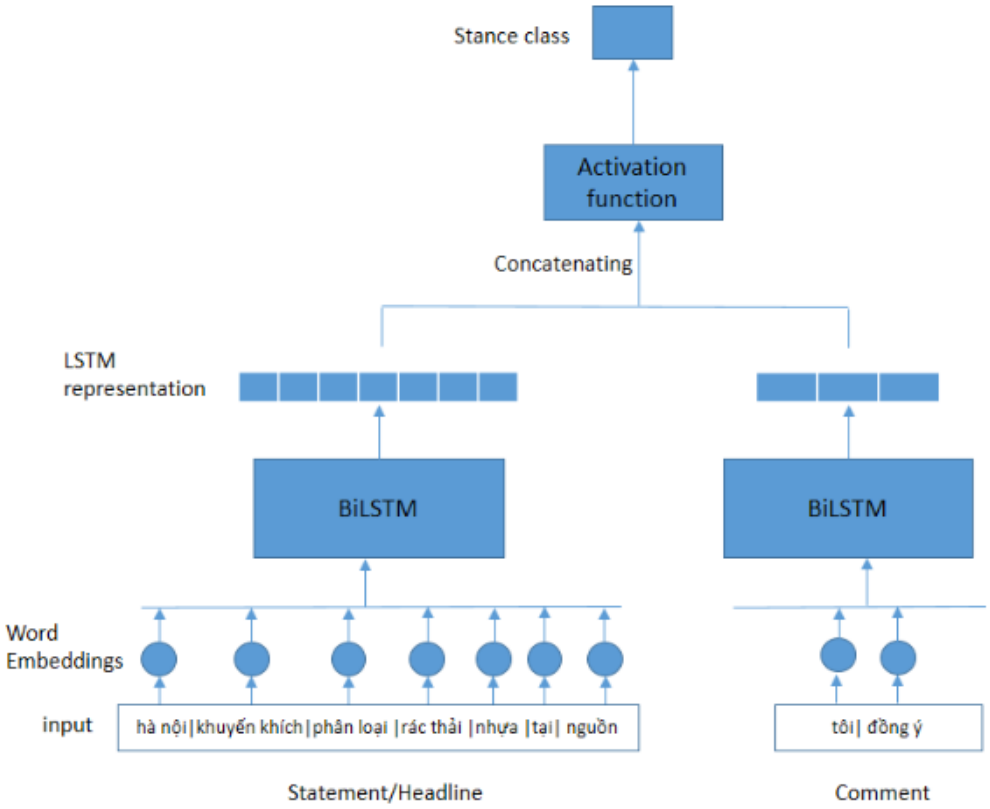
- Tiền xử lý dữ liệu: Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thường nên co giãn (scaling) dữ liệu để chuyển về đoạn [-1, 1] hoặc [0,1].

- Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.

- Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng dụng. Điều này cũng quyết định đến tính chính xác của quá trình phân lớp.

- Sử dụng các tham số cho việc huấn luyện với tập mẫu.

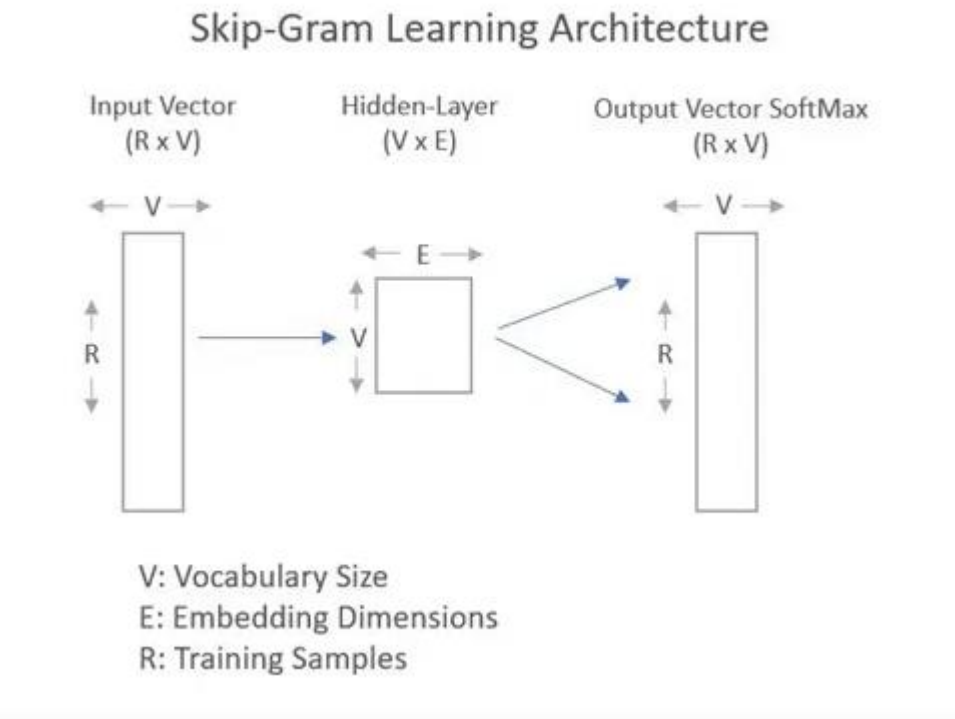
3.4. Phát hiện lập trường sử dụng mô hình học sâu



Hình 3.1: Mô hình phát hiện lập trường sử dụng kỹ thuật học sâu[17]

3.4.1. Word Embeddings là Word2vec

Dữ liệu văn bản được chuyển đổi thành biểu diễn vectơ trước khi đưa vào thuật toán học máy. Trong nghiên cứu này, các biểu diễn vector từ thử nghiệm là Word2Vec. Word2vec nhập một kho văn bản và xuất một tập hợp các vectơ. Là sự kết hợp của hai cách, sử dụng ngữ cảnh để dự đoán một từ mục tiêu (một phương pháp được gọi là bag of word liên tục, hoặc CBOW) và sử dụng một từ để dự đoán ngữ cảnh mục tiêu, được gọi là skip-gram. Trên thực tế, cả hai mô hình đều là mạng nơ-ron ba lớp với một lớp đầu vào, một lớp ẩn và một lớp đầu ra. Đối với hai mô hình này, word embedding là hidden layer, sau đó lớp đầu ra cuối cùng bị loại bỏ và chỉ giữ lại các lớp đầu vào và hidden layer. Ta có hidden layer trong Kiến trúc mô hình Skip-Gram được mô tả trong hình dưới



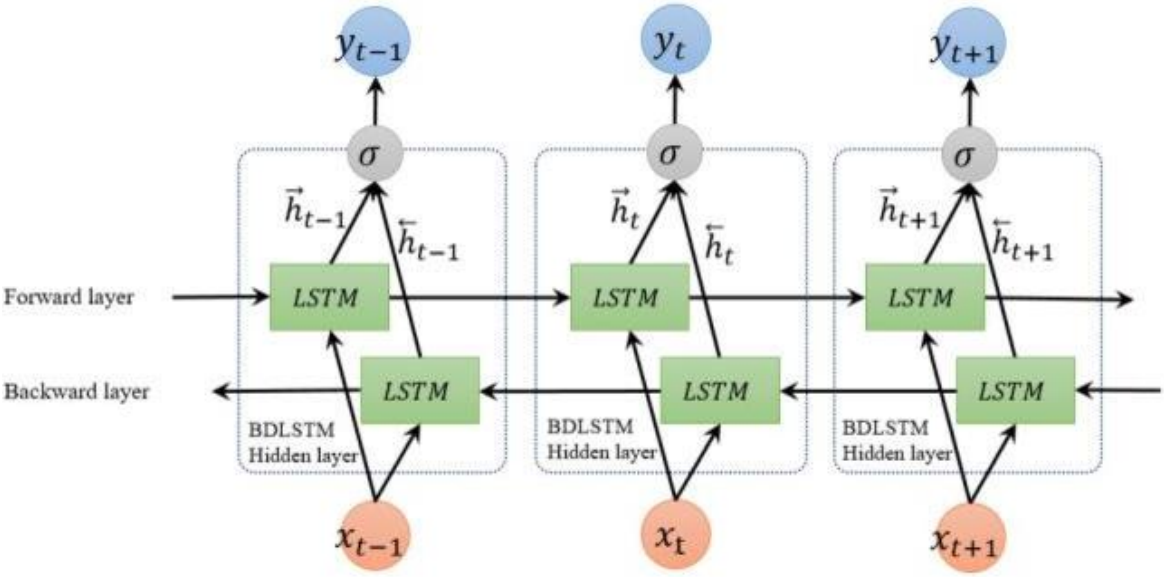
Hình 3.2: Skip-Gram Learning Architecture[18]

3.4.2. Mô hình BiLSTM

Việc phát hiện chính xác lập trường phụ thuộc không chỉ vào các thông tin phía trước của từ đang xét mà còn cả các thông tin phía sau. Tuy nhiên, một kiến trúc LSTM truyền thống với một lớp duy nhất chỉ có thể dự đoán nhãn của từ hiện tại dựa trên thông tin có được từ các từ nằm trước đó.

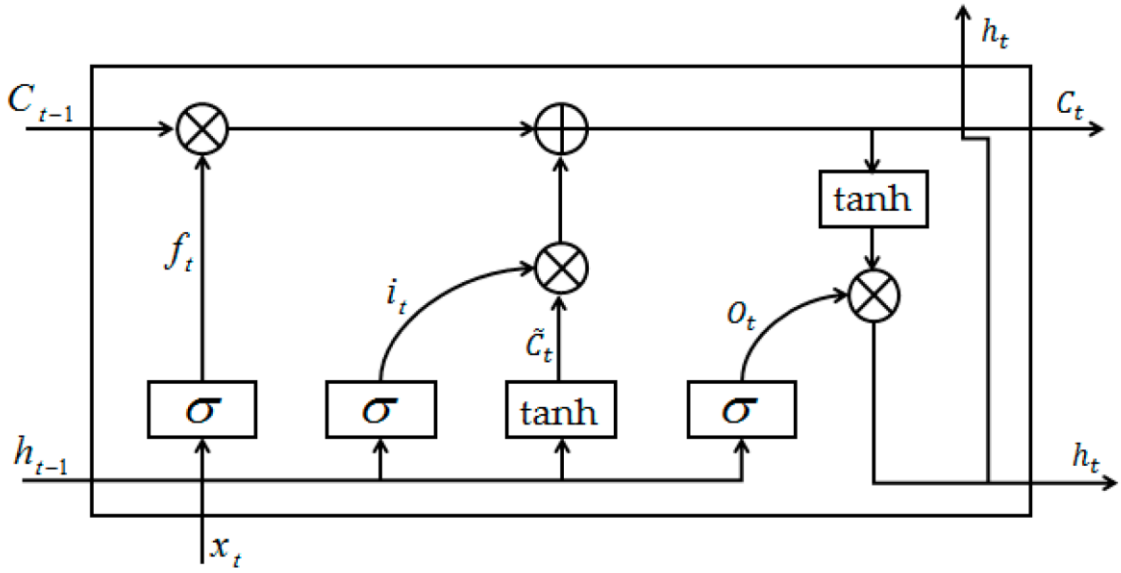
Trong nghiên cứu này, chúng tôi đề xuất sử dụng BiLSTM để xử lý câu và chuyển một số thông tin được trích xuất từ câu đó sang mô hình tiếp theo. LSTM hai chiều (BiLSTM) là một mô hình bao gồm hai LSTM độc lập, xử lý câu đầu vào bằng cách tổng hợp thông tin từ hai hướng của một câu và sau đó, hợp nhất thông tin đặc trưng.

BiLSTM làm tăng hiệu quả lượng thông tin có sẵn cho mạng, cải thiện ngữ cảnh có sẵn cho thuật toán (ví dụ: biết những từ nào ngay sau đó và đứng trước một từ trong câu). Cấu trúc mô hình được thể hiện trong Hình 6:



Hình 3.3: Mô Hình cấu trúc của BiLSTM[16]

Cấu trúc đơn vị bộ nhớ của mô hình BiLSTM được thể hiện trong hình 3.4:



Hình 3.4 : Cấu trúc đơn vị bộ nhớ của BiLSTM[16].

Phần quan trọng nhất của đơn vị bộ nhớ là trạng thái bộ nhớ C được truyền trực tiếp trên toàn bộ chuỗi cấu trúc và chỉ thực hiện một lượng nhỏ hoạt động tuyến tính để thông tin có thể dễ dàng được giữ nguyên trong quá trình truyền. Đồng thời, bộ nhớ có cấu trúc "gate" để thêm hoặc xóa thông tin chứa trong trạng thái bộ nhớ.

"Gate" là một phương pháp chọn thông tin, bao gồm phép toán nhân theo chiều của vectơ và hàm sigmoid. Một đơn vị bộ nhớ hoàn chỉnh chủ yếu bao gồm các phần sau: bộ nhớ C_{t-1} tại thời điểm $t-1$, đầu ra h_{t-1} tại thời điểm $t-1$, forget gate f_t, gate i_t, và output gate O_t, trong đó các giá trị của ba gate tất cả đều nằm trong khoảng từ 0 đến 1, trong khi trạng thái bộ nhớ C_{t-1} ghi lại thông tin lịch sử của tất cả các nút thời

gian trước đó, là bộ nhớ dài hạn của mô hình và h_{t-1} ghi lại thông tin của nút thời gian ngay trước thời gian hiện tại, là bộ nhớ ngắn hạn của mô hình.

Trạng thái bộ nhớ của đơn vị bộ nhớ thứ *j* tại thời điểm *t,* C^j_t là kết quả của input gate i^j_t, forget gate f^j_t và trạng thái bộ nhớ trước đó $C^j_t - 1$. [55] Công thức tính toán của đơn vị bộ nhớ như sau:

$$C^j_t = f^j_t \times C^j_{t-1} + i^j_t \times \tilde{C}^j_t$$

Khi

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_c . [h_{t-1}, x_t] + b_C)$$

Trong phương trình (2), W biểu thị ma trận trọng số tương ứng với mỗi cổng điều khiển, b biểu thị tham số bias, σ biểu thị hàm kích hoạt sigmoid, *tanh* biểu thị hàm tiếp tuyến hyperbol và x_t biểu thị đầu vào của mô hình tại thời điểm *t*. Input gate i^j_t, forget gate f^j_t lần lượt kiểm soát việc bổ sung thông tin mới và xóa thông tin cũ.

Khi đơn vị bộ nhớ được cập nhật, lớp ẩn sẽ tính toán lớp ẩn hiện tại h_t theo kết quả của output gate hiện tại O_t.

$$O_t = \sigma(W_O . [h_{t-1}, x_t] + b_O)$$
$$h_t = O_t \; x \tanh(C_t)$$

Từ quy trình xử lý dữ liệu của memory unit, có thể thấy rằng ý tưởng cốt lõi của cấu trúc memory unit là liên tục cập nhật thông tin dài hạn và ngắn hạn trong mô hình theo thông tin đầu vào của từ hiện tại để liên tục có được các đặc điểm ngữ cảnh trong văn bản.

Đặt đầu ra trạng thái ẩn của LSTM chuyển tiếp là $\overrightarrow{h_t}$ và đầu ra của LSTM lùi là $\overleftarrow{h_t}$ tại thời điểm *t,* đầu ra trạng thái ẩn h_t bởi BiLSTM sẽ là:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

3.4.3. Lớp phân loại ReLU

Hàm kích hoạt của một nút định nghĩa đầu ra của nút đó được cho bởi một đầu vào hay tập đầu vào. Với một đầu vào hay tập đầu vào, hàm kích hoạt sẽ cho ra đầu ra của một nút đó.

Đầu ra của trạng thái ẩn của ô cuối cùng trong mạng LSTM được sử dụng làm đầu vào cho một lớp được kết nối với nhau. Hàm kích hoạt (activation function) mô

phỏng tỷ lệ truyền xung qua axon của một neuron thần kinh. Trong một mạng nơ-ron nhân tạo, hàm kích hoạt đóng vai trò là thành phần phi tuyến tại output của các nơ-ron.

Hàm ReLU đang được sử dụng khá nhiều trong những năm gần đây khi huấn luyện các mạng neuron. ReLU đơn giản lọc các giá trị < 0. Ta sử dụng Activation functions với Neural Networks để xác định đầu ra của mạng thần kinh như 'Yes' hay 'No'. ReLU được sử dụng trong hầu hết các mạng nơ-ron phức hợp hoặc học sâu. ReLU sử dụng công thức đơn giản sau để biến đổi đầu vào:

$f(x)=max(0,x)$

Hàm ReLU là đạo hàm của nó cả hai đều là đơn điệu. Hàm trả về 0 nếu nó nhận bất kỳ đầu vào âm nào, nhưng với bất kỳ giá trị dương nào x, nó sẽ trả về giá trị đó. Do đó, cung cấp một đầu ra có phạm vi từ 0 đến vô cùng. Vì vậy, nếu đầu vào nhỏ hơn hoặc bằng 0, thì Relu sẽ xuất ra 0. Nếu đầu vào lớn hơn 0, thì relu sẽ chỉ xuất đầu vào đã cho. ReLu được đánh giá là hàm kích hoạt tốt hơn các hàm kích hoạt nổi tiếng trước đây như sigmoid và tanh.

3.5. Kết luận chương

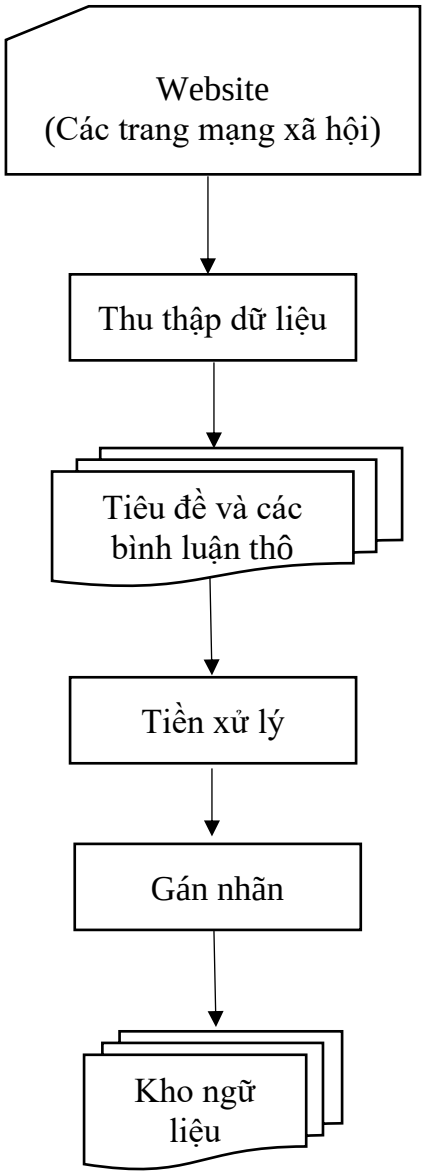
Nội dung chương đã mô tả bài toán phát hiện lập trường, và đưa ra giải pháp đề xuất them hai hướng khảo sát các phương pháp học máy truyền thống và học sâu.

CHƯƠNG IV: KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

Chương này trình bày cách xây dựng, thu thập kho ngữ liệu, mô tả cách thiết lập thực nghiệm, đưa ra các mô hình thực nghiệm, giới thiệu các công cụ được sử dụng trong bài toán và đánh giá kết quả thực nghiệm.

4.1. Xây dựng bộ ngữ liệu về phát hiện lập trường tiếng Việt

Việc thực hiện xây dựng kho ngữ liệu luận văn đã thực hiện theo từng giai đoạn trong mô hình dưới đây:



Hình 4.1*:* Mô hình xây dựng kho ngữ liệu.

4.1.1. Thu thập dữ liệu

Luận văn tập trung vào nghiên cứu bài toán phân loại ý kiến bình luận của người dùng trên tập dữ liệu Tiếng việt được thu thập từ các trang mạng xã hội như Facebook, Twitter và các trang báo mạng, để dựa vào đó phát hiện các tin tức không chính xác

hoặc hoặc phân tích tình cảm, đánh giá các chủ đề liên quan đến tin tức, chính trị, đời sống.

Thực hiện thu thập dữ liệu từ các nguồn: Vnexpress.vn, vtv24, dantri, beat.vn, baomoi.vn, trang Facebook: Beat.vn và Webtretho.

Dữ liệu gồm 500 tiêu đề là các chủ đề, tuyên bố. Tương ứng với đó là 11252 các bình luận tương ứng. Các bài viết liên quan đến chủ đề chính trị, đời sống, ý kiến, thời sự, giáo dục. Chọn lọc các dữ liệu có nhiều ý kiến bình luận, nêu ý kiến đồng tình, ủng hộ hoặc phản đối.

4.1.2. Tiền xử lý

Dữ liệu sau khi thu thập được từ các trang báo mạng sẽ được tiến hành tiền xử lý. Luận văn thực hiện tiền xử lý dữ liệu bằng cách loại bỏ một số nhiễu như: câu sai chính tả, lỗi font.

4.1.3. Gán nhãn

Thực hiện xem xét bình luận là liên quan hay không liên quan với *Tiêu đề*. Nếu không liên quan gắn nhãn *Unrelated*, còn các bình luận liên quan thì sẽ phân chia thành 3 loại nhãn là *Agrees, Disagrees, Discusses*. Và cụ thể trong từng trường hợp sau:

Quá trình gán nhãn:

Những bình luận gán nhãn Agrees:

- Thể hiện quan điểm đồng ý, đồng tình với 1 hoặc nhiều ý được nêu ra ở tiêu đề (sử dụng các từ và cụm từ cụ thể như '*đồng ý*', '*đúng rồi', 'chính xác', 'quá chuẩn', 'ủng hộ', "quá hay"*).

  Ví dụ 1: Lương dưới 10tr đừng vội cưới.

  Agrees: *Chuẩn bác ạ*.

  Agrees: *Đồng ý với ý kiến của tác giả.*

  Ví dụ 2: Sẽ có một lúc nào đó bạn nhận ra rằng" Làm vừa lòng thiên hạ là một việc vừa khó vừa vô nghĩa"

  Agrees: *Quá hay*

  Agrees: *Vô cùng ý nghĩa*

  Agrees: *Chính xác*

- Thể hiện quan điểm ủng hộ, đồng cảm, tuyên dương, tán thành với 1 hoặc nhiều ý được nêu ra ở tiêu đề.

Ví dụ 2: Giải biên tập viên dẫn chương trình ấn tượng nhất thuộc về Trần Việt Hoàng.

Agrees: *Chúc mừng BTV cà khịa mặn mà của VTV.*

Agrees: *Phát biểu nhận giải cũng hay, k dập khuôn.*

Ví dụ 3: Cô gái 'có khuôn mặt lạ' và mối tình online

Agrees: *Cặp đôi dễ thương và ấm áp. Chúc hai bạn mãi hạnh phúc bên nhau.*

Agrees: *Thật ngưỡng mộ tình yêu của 2 em.*

Ví dụ 4: Cử tri Mỹ gốc Việt 'phát khóc' khi nghe tin Trump nhiễm nCoV

Agrees: *Nếu được, tôi củng ủng hộ Ông 1 phiếu.*

Agrees: *Mong Tổng thống Trump sớm phục hồi.*

- Nêu ra dẫn chứng để gián tiếp ủng hộ 1 hoặc nhiều ý được nêu ra ở tiêu đề.

Ví dụ 5: Lương dưới 10tr đừng vội cưới.

Agrees: *Tôi 23 tuổi, lương 24 triệu/tháng ở Sài Gòn mà còn chẳng thấy dư dả gì nhiều nên cũng chưa nghĩ đến chuyện cưới xin.*

Agrees: *Lương cao hơn 10tr may ra còn dám cưới.*

Những bình luận gán nhãn Disagrees:

- Thể hiện quan điểm không đồng ý, phản đối với 1 hoặc nhiều ý được nêu ra ở tiêu đề (Sử dụng hoặc chứa các từ hoặc cụm từ sau: *"phét", "vớ vẩn" ,"không tin được", "không đồng ý", "không đúng"..*).

Ví dụ 1: Lương dưới 10tr đừng vội cưới:

Disagrees: *Theo cháu, ý kiến của bác không đúng. Vì vật chất rất quan trọng. Nhưng không có nghĩa là lương dưới 10 triệu thì chưa thể cưới.*

Disagrees: *Tôi không đồng tình với tác giả. Vật chất không phải là tất cả.*

- Thể hiện quan điểm không ủng hộ, không đồng quan điểm với 1 hoặc nhiều ý được nêu ra ở tiêu đề.

Ví dụ 3: Cô gái 'có khuôn mặt lạ' và mối tình online hạnh phúc.

Disagrees: *Tình online nhiều khi không bền đâu.*

Disagrees: *Cô gái ơi, có khi bị lừa đấy em ạ. Cẩn thận nhé.*

Ví dụ 4: Cử tri Mỹ gốc Việt 'phát khóc' khi nghe tin Trump nhiễm nCoV

Disagrees: *Vote 1 phiếu cho Biden chiến thắng.*

Disagrees: *Không cấm hành vi không đeo khẩu trang là một sai lầm khủng khiếp mà những người này vẫn cho là điểm tốt thì thua.*

- Nêu ra dẫn chứng để gián tiếp phản đối 1 hoặc nhiều ý được nêu ra ở tiêu đề.

Ví dụ 5: Lương dưới 10tr đừng vội cưới:

Disagrees: *Tôi 23 tuổi, lương 6 triệu/tháng ở Sài Gòn mà còn chẳng thấy mà vẫn cưới và cảm thấy hạnh phúc.*

Disagrees: *Chết em rồi. Em cưới vợ khi mà lương chưa được 5 triệu/tháng. Khi vợ có bầu, sinh con, tổng thu nhập của 2 vợ chồng chưa được 8 triệu/tháng. Cơ mà bọn em vẫn sống được, thiếu trước hụt sau nhưng tình cảm ngày càng sâu.*

Disagrees: *Tôi với vợ tôi bán cơm tấm ngày lời có 400k, hai vợ chồng. Nuôi 3 đứa con vẫn sống cuộc sống sung túc đấy.*

Ví dụ 6: Tin vui hôm nay: Giá thịt heo đồng loạt giảm trên cả nước, mẹ nội trợ mừng rỡ xách giỏ đi chợ

Disagrees: *Giảm ở trên mạng thôi.hôm nay đi chợ thịt vẫn đắt như thường nha*

Disagrees: *Thông tin chỉ mang t/c minh họa ra chợ họ bảo nên đó mà mua*

- Không đồng tình, lên án , phê phán, trách móc hành động , đối tượng được nêu trong tiêu đề.

Ví dụ 7: Chính thức khởi tố gã chồng bạo hành vợ suốt 11 năm vì không biết đẻ: 2 lần đò vẫn "lênh đênh".

Disagrees: *Thời nào rồi mà để nó đánh tận 11 năm nhưng vẫn cam chịu, ko hiểu nổi.*

Disagrees: *Chị thật đáng trách, thời buổi nào rồi còn để cho nta hành hạ, nta vì con mà chịu vài lần là quá rồi,đằng này k con cái gì mà cũng để nó hành hạ, thật tức chết.*

Ví dụ 8: Liều tháo rào chắn cảnh báo sạt lở, 2 thanh niên Quảng Nam bị vùi lấp: Vùng vẫy bới đất cứu thân.

Disagrees: *Ngu cho chết lần sau hết dám liều nhé.*

Disagrees: *Bó tay cảnh báo rồi mà còn tháo ra.*

Những bình luận gán nhãn Discusses:

- Thể hiện quan điểm trung lập với ý kiến được nêu ra ở tiêu đề.

Ví dụ 1: Lương dưới 10tr đừng vội cưới:

Discusses: *Tùy cách sống mỗi người thôi, sống đua đòi quá thì lương 20 triệu vẫn không đủ, mà nghèo quá cũng khó sống.*

Discusses: *Nếu như vậy thì tương lai chúng ta sẽ giống như Nhật Bản. Ở Nhật Bản rất nhiều người họ cũng nghĩ như vậy đến nỗi không dám lập gia đình.*

- Bổ sung, giải thích thêm thông tin, bàn luận với ý kiến được nêu ra ở tiêu đề.

Ví dụ 2: Nhiều người Việt hay nói chuyện quy trình nhưng lại làm việc theo cảm tính

Discusses: *Nằm ở phần đôn đốc, giám sát của cấp quản lý là chính. Đa phần chủ doanh nghiệp nhỏ muốn lượt bỏ, nhanh và tiện thì luôn kèm với rủi ro.(Không muốn bỏ thêm chi phí cho an toàn).*

Discusses: *Quy trình làm gì khi người vẫn hành không theo quy trình và dùng người vận hành vô trách nhiệm.*

- Nêu ra dẫn chứng thực tế để bàn luận về thông tin ở tiêu đề.

Ví dụ 5: 2,5 điểm mỗi môn vẫn đỗ vào lớp 10 công lập Hà Nội:

Discusses: *Đưa em mình nó học trung bình toán anh toàn 5, 6 đ điểm cả năm mà thi cũng đạt mỗi môn trên 5 điểm ko hiểu sao có học sinh thi 2, 3 điểm môn toán đươc?*

Discusses: *nhớ cô giáo dạy văn (người Hà nội cũ) dạy mình cấp 3 sau cô ra dạy trường cao đẳng sư phạm Thường tín cô nói đội ngũ giáo viên tự nhiên của Ứng hoà B cực kỳ giỏi và mình thấy đúng như vậy , có thầy dạy xong còn về đi đánh giậm nữa , nhớ thời gian khó .*

Những bình luận gán nhãn Unrelated:

- Bình luận về chủ đề không liên quan đến vấn đề đang nói ở tiêu đề.

Ví dụ 1: Lương dưới 10tr đừng vội cưới:

Unrelated: *Hôm nay trời đẹp thật.*

Unrelated: *Tôi hay cảm thấy khó thở là có nguy cơ bị làm sao hả các bác ?*

- Bình luận về các vấn đề lan man bên cạnh vấn đề đang nói ở tiêu đề.

Ví dụ 2: 'Nội chiến' tranh giành tài sản của gia tộc Trump

Unrelated: *Xưa ông ấy thật đẹp trai, như tài tử Hollywood vậy.*

Unrelated: *Con ông Trump đẹp trai, cao ráo thật.*

- Bài viết quảng cáo

Ví dụ 3: 'Nội chiến' tranh giành tài sản của gia tộc Trump

Unrelated: *Tiếp cận hàng triệu người dùng mới với TikTok For Business. Đạt mục tiêu kinh doanh. Bán hàng thông qua TikTok For Business đến hơn 20 thị trường toàn cầu.*

Unrelated: *Không cần vốn, không kinh nghiệm, nhiều thanh niên 9X giàu lên nhanh chóng. Làm giàu khôn ngoan. cách kiếm tiền an toàn. Truy cập: https://www.bimatcongtudo.club/*

Check chéo dữ liệu để kiểm tra tính thống nhất của gán nhãn

Giai đoạn gán nhãn thủ công luận văn thực hiện ba người gán nhãn. Vì vậy, luận văn cần biết được xem kết quả gán nhãn của ba người có tương đồng với nhau không. Để kiểm tra được điều đó, luận văn sử dụng độ đo Cohen's kappa [35] tính toán độ tương đồng gán nhãn giữa hai người.

Công thức:

$$K = \frac{p_o - p_e}{1 - p_e}$$

Trong đó: p_0 là xác suất tương đối giữa 2 người.

p_e là xác suất ngẫu nhiên giữa 2 người.

Ví dụ: Có 2 người A và B cùng duyệt một tập hồ sơ gồm 50 bộ, mỗi kết quả được đọc bởi 2 người, mỗi người nói "đủ" hoặc "thiếu" ám chỉ hồ sơ đủ giấy tờ hoặc thiếu giấy tờ. Ta có kết quả duyệt của 2 người như sau:

Bảng 4.1: Bảng ví dụ kết quả duyệt

		B	
		Đủ	Thiếu
A	Đủ	20	5
	Thiếu	10	15

Khi đó: p_0 = (20 + 15) / 50 = 0.70

Xác suất người A đọc "Đủ" là 50%

Xác suất người B đọc "Đủ" là 60%

Xác suất cả 2 người đọc "Đủ" là : 0.5*0.6=0.3

Xác suất cả 2 người đọc "Thiếu" là : 0.5*0.4=0.2

Áp dụng vào bộ dữ liệu bằng cách lấy ngẫu nhiên 2000 câu bình luận kiểm tra chéo, kết quả đo độ tương đồng phân loại trung bình giữa ba người là 99,2. Kết quả cho thấy ba người gán nhãn khá tương đồng với nhau.

4.1.4. Thống kê kho dữ liệu

Dữ liệu gồm 500 tiêu đề là các chủ đề, tuyên bố. Tương ứng với đó là 11250 các bình luận tương ứng. Các bài viết liên quan đến chủ đề chính trị, đời sống, ý kiến, thời sự, giáo dục.

- Tổng số câu tiêu đề và bình luận: 11750.
- Tổng số từ: 1314459.
- Tổng số từ trung bình / câu: 111.
- Số từ (không tính lặp) trên toàn bộ kho ngữ liệu: 8014.

Bảng 4.2: Liệt kê số lượng bình luận tương ứng với các nhãn phân loại

Nhãn phân loại	**Số lượng bình luận**
Agree	2938
Disagree	2574
Discuss	3334
Unrelated	2404

4.2. Thiết lập thực nghiệm

Với dữ liệu chuẩn bị cho thực nghiệm, luân văn lấy được 11750 câu bình luận về các tiêu đề, chủ đề tương ứng tiếng Việt. Từ dữ liệu này, luận văn chia thành 5 bộ dữ liệu, trong đó mỗi bộ dữ liệu xây dựng bằng cách ngẫu nhiên trong tập dữ liệu có. Kết quả thu được ở 5 lần thực nghiệm sẽ được tính trung bình để ra được kết quả của thực nghiệm.

Để đánh giá kết quả của việc xác định thực thể và thuộc tính ta đánh giá thông qua độ chính xác (precision), độ bao phủ (recall) và F1 được xác định như sau:

$$precision = \frac{số\ nhãn\ gán\ đúng}{tổng\ số\ nhãn\ được\ gán}$$

$$recall = \frac{số\ nhãn\ gán\ đúng}{tổng\ số\ nhãn\ thực\ tế}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

4.3. Công cụ thực nghiệm

Luận văn sử dụng Google Colab làm công cụ chính để xây dựng và đánh giá mô hình. Google Colab còn được gọi là Google Colaboratory, là sản phẩm do Google Research phát triển, dựa trên Jupyter Notebook, cho phép chạy mã python thông qua trình duyệt, đặc biệt thích hợp cho phân tích dữ liệu, học máy và giáo dục.

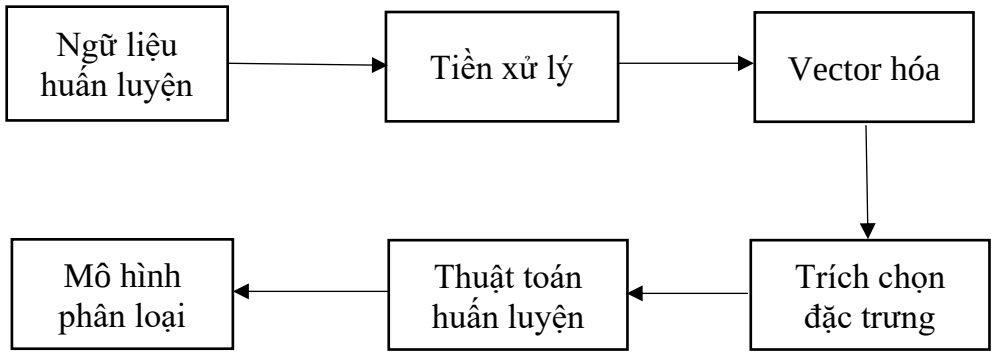
Google Colab cung cấp cho các thư viện phổ biến trong nghiên cứu Học sâu như PyTorch, TensorFlow, Keras và OpenCV. Colab không yêu cầu bất kỳ cài đặt hay cấu hình máy tính nào, mọi thứ đều có thể chạy thông qua trình duyệt, có thể sử dụng tài nguyên máy tính của mình từ CPU và GPU tốc độ cao, TPU có sẵn.

Yêu cầu duy nhất để sử dụng Google Colab là cần phải sử dụng tài khoản Google. Với Colab, chỉ cần sử dụng một các dòng mã, có thể nhập tập dữ liệu hình ảnh, traning cơ sở phân loại hình ảnh trên tập dữ liệu đó và đánh giá mô hình. Sổ tay Colab thực thi mã trên máy chủ đám mây của Google. Nhờ đó, có thể tận dụng sức mạnh của phần cứng Google, bao gồm GPU và TPU, bất kể cấu hình máy tính sử dụng nào.

4.4. Các mô hình thực nghiệm

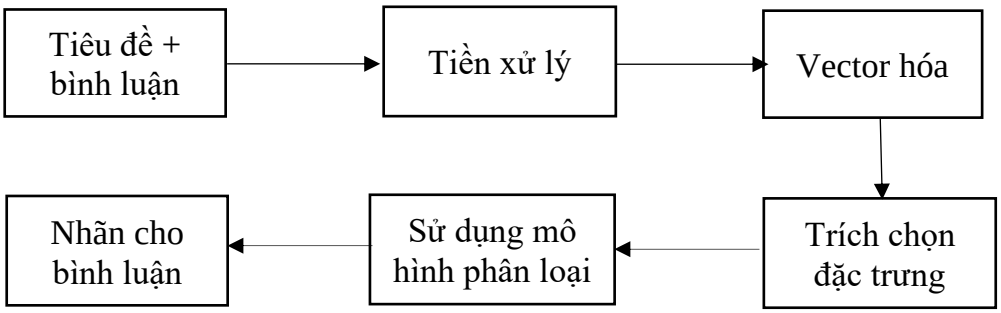
Với các phương pháp học máy truyền thống như Decision Tree, KNN, Naïve Bayes, v.v thì quá trình phân loại dữ liệu văn bản (văn bản, câu) thường gồm hai giai đoạn sau:

- Giai đoạn huấn luyện: Giai đoạn huấn luyện nhận đầu vào là tập ngữ liệu huấn luyện gồm các câu bình luận đã được gán nhãn, sau khi xử lý tập ngữ liệu và áp dụng các thuật toán huấn luyện sẽ cho đầu ra là một mô hình phân loại.



Hình 4.2: Mô hình giai đoạn huấn luyện

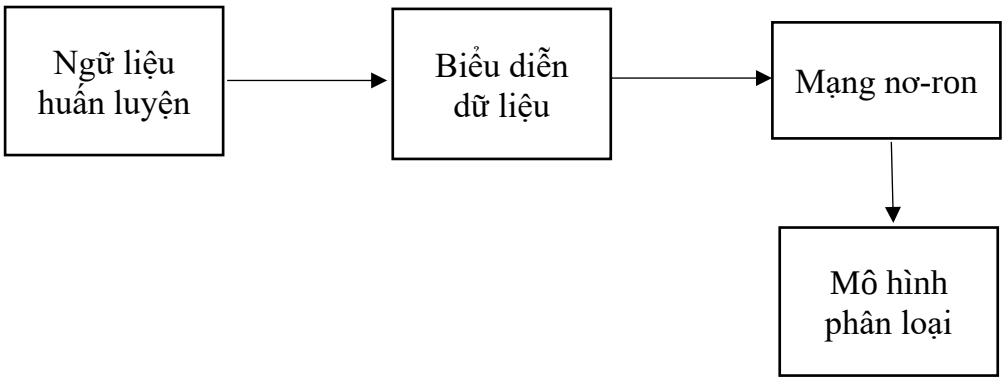
- Giai đoạn phân lớp: Giai đoạn phân lớp nhận đầu vào là câu tiêu đề và bình luận tương ứng của người dùng dưới dạng ngôn ngữ tự nhiên, sau quá trình tiền xử lý và áp dụng mô hình phân loại sẽ cho ra nhãn phân loại của câu bình luận đầu vào.



Hình 4.3 Mô hình giai đoạn phân lớp

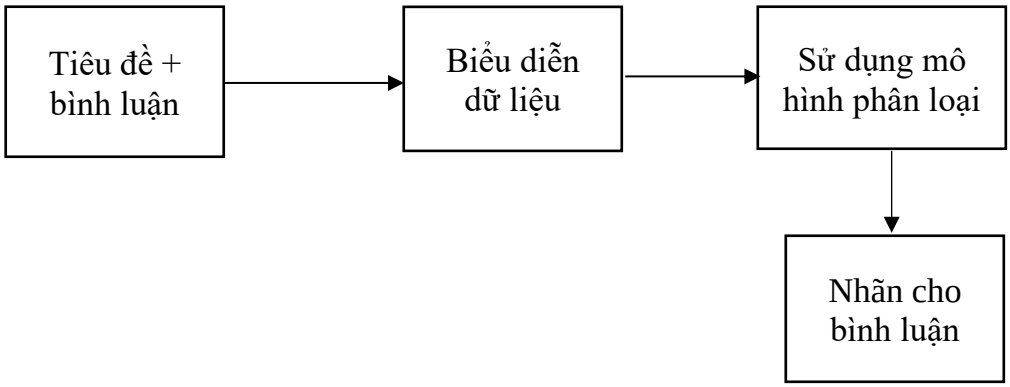
Với phương pháp sử dụng mạng nơ-ron như LSTM, CNN, RNN, v.v thì quá trình phát hiện lập trường gồm hai giai đoạn:

- Giai đoạn huấn luyện: Giai đoạn huấn luyện nhận đầu vào là tập ngữ liệu huấn luyện gồm các tiêu đề và bình luận tương ứng đã được gán nhãn, sau khi biểu diễn dữ liệu và đưa vào mạng nơ-ron sẽ cho ra đầu ra là một mô hình phân loại.



Hình 4.2: Mô hình giai đoạn huấn luyện sử dụng mạng nơ-ron.

- Giai đoạn phân lớp: Giai đoạn phân lớp nhận đầu vào là tiêu đề và bình luận của người dùng dưới dạng ngôn ngữ tự nhiên, sau quá trình biểu diễn dữ liệu và áp dụng mô hình phân loại sẽ cho ra nhãn phân loại của câu hỏi đầu vào.

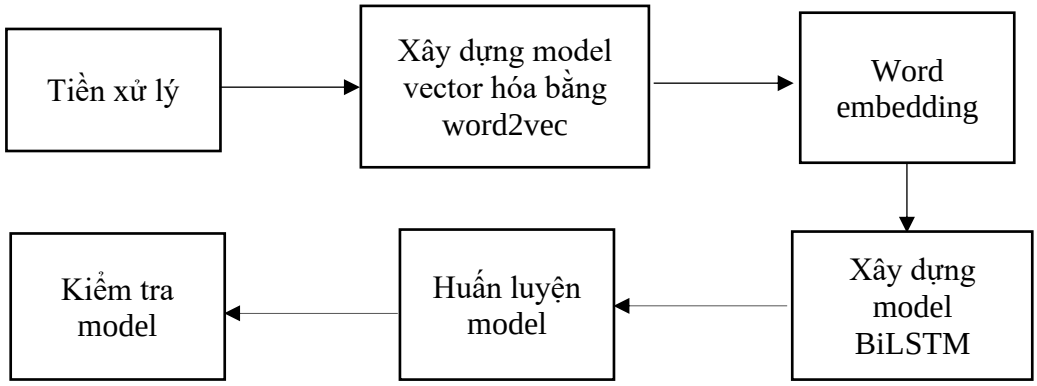


Hình 4.3: Mô hình giai đoạn phân lớp sử dụng mạng nơ-ron.

Ứng dụng trong bài toán phát hiện lập trường LSTM

Việc giải bài toán phát hiện lập trường sẽ bao gồm việc giải quyết một chuỗi các bài toán nhỏ hơn. Chuỗi các bài toán nhỏ hơn này được gọi là pipline của mô hình học máy.

Phát hiện lập trường sử dụng mô hình mạng RNN, LSTM gồm các bước sau:



Hình 4.6: Các bước của bài toán phát hiện lập trường sử dụng mạng nơ-ron.

- Tiền xử lý kho ngữ liệu.
- Xây dựng model vector hóa bằng Word2vec cho tập ngữ liệu văn bản đã được tiền xử lý. Mô hình Word2vec bản chất là việc huấn luyện một mạng nơ-ron nhân tạo - Artificial Neural Network (ANN) với một lớp ẩn. Các cặp từ được tách theo skip-gram và dựa trên xác suất để tính độ tương quan giữa các từ.
- Word embedding sử dụng mô hình kết quả của Word2vec để vector từng câu trong tập ngữ liệu.
- Áp dụng mạng nơ-ron để giải quyết bài toán bao gồm các bước nhờ: Xây dựng model, huấn luyện model, kiểm tra model.

4.5. Kết quả thực nghiệm

Phương pháp phân loại dựa trên học máy đươc chia làm 2 nhóm chính là phương pháp học máy truyền thống và phương pháp học máy sử dụng mạng nơ-ron. Do vậy, luận văn đã lựa chọn thực nghiệm hai mô hình chính đại diện cho hai nhóm phương pháp đó là mô hình Decision Tree, Naïve Bayes đại diện cho nhóm phương pháp học máy truyền thống, mô hình RNN, LSTM đại diện cho nhóm phương pháp học máy sử dụng mạng nơ-ron.

4.5.1. Mô hình LSTM (Long-Short Term Memory)

Bảng dưới đây cho thấy độ chính xác khi chạy kho dữ liệu trên 5 fold. Kết quả cho thấy Fold 3 đạt độ chính xác cao nhất với 68% và kết quả trung bình là 66,38%.

Bảng 4.3: Độ chính xác của từng fold sử dụng mô hình LSTM (%)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Accuracy	67.20	64.90	68.00	66.00	65.80	**66.38**

Đây là kết quả của precision, recall và điểm F1 của mỗi nhãn (bảng 4), ta có thể thấy, nhãn "Unrelated" có độ chính xác cao nhất, Recall và điểm F1 tương ứng với 85,1%, 82,38% và 83,62%. Nhãn "Agree" cho kết quả khoảng 70%. Tuy nhiên, 2 nhãn khác lại cho kết quả thấp hơn.

Bảng 4.4: Precision, recall và F1-score tương ứng với các nhãn (%)

Labels	Precision	Recall	F1-score
Agree	70.68	69.46	70.02
Disagree	56.48	55.52	55.96
Discuss	57.82	60.56	59.10
Unrelated	85.10	82.38	83.62

4.5.2. Mô hình RNN(Recurrent Neural Network)

Chạy dữ liệu trên 5 fold trong mô hình RNN. Bảng dưới cho thấy kết quả trung bình là 62.3, tương đối thấp so với hai phương pháp LSTM.

Bảng 4.5: Độ chính xác của từng fold sử dụng mô hình RNN (%)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Accuracy	61.00	64.40	62.00	64.50	61.20	**62.30**

4.5.3. Học máy Decision Tree và Naïve Bayes

Để xác minh tính hiệu quả của các mô hình được đề xuất, chúng tôi so sánh với một số phương pháp học máy phổ biến làm đường cơ sở. Word Embedding là một trong những kỹ thuật mà chúng ta có thể biểu diễn văn bản bằng cách sử dụng vectơ. Trong luận văn sử dụng Bag of Word, Glove Vector, Term Frequency-Inverse Document Frequency do hình thức biểu diễn văn bản dưới dạng số đơn giản và hiệu quả.

Bảng 4.6: Độ chính xác của từng fold sử dụng mô hình Decision Tree (%)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Glove Vector	52.15	54.33	52.40	53.02	53.51	**53.08**
Bag of Word	57.13	56.37	55.51	58.8	55.73	**56.71**
TF - IDF	55.48	55.53	54.22	52.76	55.07	**54.61**

Bảng 4.7: Độ chính xác của từng fold sử dụng mô hình Naïve Bayes (%)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Glove Vector	42.56	43.76	40.27	43.11	44.27	**42.79**
Bag of Word	45.45	47.00	46.04	46.89	45.29	**46.13**
TF - IDF	43.58	43.36	42.76	44.09	43.42	**43.44**

4.6. Thảo luận và phân tích lỗi

Dựa trên kết quả thực nghiệm thu được, nhận thấy rằng phương pháp học sâu LSTM cho kết quả nhất (66.38%) so với ba phương pháp còn lại là RNN (62.30%), Decision Tree (56.71%) và Naïve Bayes (46.13%). Tuy nhiên, trong một số nghiên cứu gần đây về tìm lập trường trong tiếng Anh, mô hình học sâu thường cho kết quả khá cao, có thể do sự khác biệt về ngôn ngữ và độ phức tạp trong tiếng Việt. Kết quả của hai phương pháp học sâu trên chỉ ở mức trung bình. Vì vậy, thời gian tới, chúng tôi sẽ dành nhiều thời gian hơn nữa để hoàn thiện và phát triển bộ môn này đạt kết quả cao hơn.

Một số lỗi điển hình được hiển thị trong bảng sau:

Bảng 4.8: Ví dụ một số lỗi điển hình khi gán dữ liệu

Tiêu đề	Bình luận	Original label	Predict label
1. Trump có thể tái tranh cử vào năm 2024.	Như vậy là bây giờ ông đã chịu thua rồi khi có kế hoạch tranh cử 2024. Thua thì nói sớm cho có hình tượng tốt trong công chúng chứ cãi riết làm nản lòng người ủng hộ	Discuss	Disagree

2. Đừng đánh giá sách giáo khoa lớp 1 bằng tư duy của người lớn.	Sao bạn có thể nói chúng tôi định kiến với tư duy người lớn và tư duy đám đông. Chúng tôi định kiến với những gì mới có à, tại sao không ai lên tiếng với môn toán mà là tiếng việt tư duy người lớn ư. Khi mà con trẻ đi học cả tháng trời chỉ học mỗi tiếng việt mà vẫn cô vẫn thấy chưa đủ còn tư duy đám đông cũng không đúng ngay từ hôm nhận sách mới về. Đừng trách phụ huynh thế nào các vị hãy đến các trường lấy ý kiến thực tế của các cháu lớp 1 và của các cô giáo chủ nhiệm lớp.	Disagree	Discuss
3. Chín điều cần cải cách giáo dục phổ thông.	Từ quan điểm của 1 ông bố có 2 con đang học cấp 1 tôi xin phản bác lại vài vấn đề vấn đề 1 6 7 bạn nên nhìn nhận cả về góc độ nhân lực giáo dục cơ sở vật chất tâm sinh lý học sinh. Nếu muốn đủ cả 3 vấn đề hãy cho con bạn học theo kiểu phong trào homeschooling ghép nhi đồng với thiếu niên thanh niên cùng 1 trường học là cực kỳ dở con tôi học lớp 2, nhưng cháu hóng hớt bố mẹ dạy anh cũng nhớ cách làm toán lớp 5. Nếu dạy bố mẹ dạy được nhưng nhà trường nào dạy được kiểu đó vấn đề 2 các cô giáo ở trường con tôi chỉ dám dạy chui và dạy cho vài học sinh mà muốn học thêm phụ huynh cũng phải gãy lưỡi mà nhờ cô mới	Disagree	Discuss

	nhận mà chỉ dám nhận học sinh gần nhà chứ không thì cô bị kỷ luật như chơi. Con học không theo kịp cũng có thể vì chương trình nặng và cha mẹ không kèm con chứ đừng đổ lỗi cho giáo viên ép học thêm vấn đề 3. Bạn nên xem nó như là dạy kỹ năng lãnh đạo từ nhỏ vậy có phải tốt hơn không chứ lớp trưởng lớp phó của tụi nhỏ đâu có lợi ích gì vấn đề 4 giáo dục lớp lớn thì tôi không rõ, nhưng với lớp nhỏ có mục tiêu rõ ràng và sgk hiện tại mỗi bài học đều có 1 mục tiêu nhỏ có gì phải bàn thêm đâu. Vấn đề 5 trường con tôi ngoài giờ môn học ngoại khóa theo tuần năm rồi cũng 4 lượt liên kết ngoại khóa chưa kể hoạt động văn hóa văn nghệ thể thao của nhà trường bạn thấy ít tôi thì không.		

Có thể thấy, đã có sự hiểu nhầm giữa hai nhãn (Disagree và Discuss), các bình luận bị gắn nhãn sai thường là các bình luận dài không có dấu hiệu nhận biết với các từ cụ thể như "phản đối", "không đồng ý", "không tán thành" .... Trong quá trình thu thập dữ liệu, chúng tôi cũng gặp phải những xung đột về nhãn, và phải thảo luận và phân tích phù hợp để chọn nhãn phù hợp cho mọi nhận xét, do đó máy móc hiểu nhầm là không thể tránh khỏi. Tuy nhiên, lỗi này chỉ chiếm một phần nhỏ trong tổng số tập dữ liệu.

4.7. Kết luận chương

Chương này đã trình bày được cách thiết lập thực nghiệm, mô tả được các mô hình thực nghiệm, giới thiệu được các công cụ thực nghiệm, đưa ra kết quả và phân tích đánh giá được kết quả thực nghiệm.

KẾT LUẬN

Sự phát triển của internet thực sự là một bước tiến lớn của nhân loại, những lợi ích to lớn của internet mang lại thật sự không thể phủ nhận. Tuy nhiên, nó giống như một con dao hai lưỡi khi nhiều người sử dụng nó với mục đích tiêu cực như bình luận xúc phạm, chửi bới, thậm chí là tung tin giả.

Mục đích của nghiên cứu này là giúp xác định và phân loại các bình luận và giúp ngăn chặn tin tức giả mạo. Trong luận văn này, chúng tôi sử dụng nhiều phương pháp khác nhau như Decision Tree, Naïve Bayes, RNN và LSTM để so sánh độ chính xác giữa các mô hình và lựa chọn mô hình có kết quả tốt nhất.

Luận văn sẽ khảo sát bài toán phát hiện lập trường về một chủ đề, đề xuất một phương pháp phù hợp. Đồng thời phương pháp đề xuất sẽ được phân tích và đánh giá bằng một số phương pháp đánh giá thông dụng trên tập dữ liệu đã thu thập.

Nhìn chung, luận văn đã đạt được:

- Nghiên cứu cho bài toán phát hiện lập trường Tiếng Việt là bài toán còn ít được nghiên cứu.

- Xây dựng được bộ dữ liệu cho bài toán.

- Nghiên cứu này chỉ là nghiên cứu ban đầu có thể đóng góp bộ dữ liệu cho các nghiên cứu tiếp theo.

- Nghiên cứu một số phương pháp phân loại dưa trên học máy sử dụng mô hình LSTM là một mô hình huấn luyện sẵn mà hiện tại đang đạt kết quả phương pháp hiện đại trong xử lý ngôn ngữ tự nhiên.

- Thực nghiệm, phân tích, đánh giá kết quả và tìm ra được trường hợp cho kết quả tốt nhất.

Về hướng phát triển tương lai, luận văn có thể tiếp tục được nghiên cứu trên bộ dữ liệu lớn hơn và nhiều mô hình khác góp phần cải thiện tốt nhất khả năng phát hiện lập trường tiếng Việt và nghiên cứu sử dụng thêm nhiều phương pháp, góp phần cải thiện tốt hơn khả năng phân loại. Ngoài ra luận văn sẽ nghiên cứu và thử nghiệm với một số mô hình khác để tìm ra mô hình phù hợp nhất với bài toán phân loại phát hiện lập trường tiếng Việt.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Nguyễn Đức Vinh, Phân tích câu hỏi trong hệ thống hỏi đáp tiếng Việt, Khóa luận tốt nghiệp đại học, Đại học quốc gia Hà Nội, 2009.

[2] Nguyễn Minh Thành, Phân loại văn bản, Đồ án môn học Xử lý ngôn ngữ tự nhiên, Đại học quốc gia Thành phố Hồ Chí Minh, 01/2011.

[3] Vu Thi Tuyen, Một số mô hình học máy trong phân loại câu hỏi, Đại học Công nghê, 2016

[4] Nguyễn Thị Hương Thảo. Phân lớp phân cấp Taxonomy văn bản Web và ứng dụng. Khóa luận tốt nghiệp đại học, Đại học Công nghệ, 2006.

[5] Phạm Văn Sơn, Tìm hiểu về Support Vector Machine cho bài toán phân lớp quan điểm

Tiếng Anh

[6] Shalmoli Ghosh⋆1, Prajwal Singhania⋆1, Siddharth Singh⋆1,Koustav Rudra⋆⋆2, and Saptarshi Ghosh1. Stance Detection in Web and Social Media: A Comparative Study.

[7] Dhruv Ghulati, Co-Founder, Factmata. “Introducing Factmata—Artificial intelligence for automated fact-checking”.

[8] Che-Wen Chen, OrcID,Shih-Pang Tseng, Ta-Wen Kuan and Jhing-Fa Wang. Outpatient Text Classification Using Attention-Based Bidirectional LSTM for Robot - Assisted Servicing in Hospital.

[9] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, Iryna Gurevych. A Retrospective Analysis of the Fake News Challenge Stance Detection Task.

[10] Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 40–46 Brussels, Belgium, November 1, 2018. c 2018 Association for Computational Linguistics.

[11] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, Iryna Gurevych, A Retrospective Analysis of the Fake News Challenge Stance Detection Task, Research Training Group AIPHES Computer Science Department, Technische Universität Darmstadt Smart Data Analytics, University of Bonn.

[12] Peter Krejzl, Barbora Hourová, Josef Steinberger. 2017. Stance detection in online discussions.

[13] Bilal Ghanem, Paolo Rosso, Francisco Rangel. 2018. Stance Detection in Fake News A Combined Feature Representation. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER).

[14] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding.

[15] Qingying Sun, Zhongqing Wang, Qiaoming Zhu, Guodong Zhou. 2018. Stance Detection with Hierarchical Attention Network. Proceedings of the 27th International Conference on Computational Linguistics.

[16] Mirko Laia, Alessandra Teresa Cignarellaab, Delia Irazú Hernández Farías, Cristina Boscoa Viviana Patti, Paolo Rossob. 2020. Multilingual stance detection in social media political debates. Computer Speech & Language Volume 63, September 2020, 101075.

[17] Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, Saptarshi Ghosh. 2019. Stance Detection in Web and Social Media: A Comparative Study. International Conference of the Cross-Language Evaluation Forum for European Languages CLEF.

[18] Qingying Sun, Zhongqing Wang, Shoushan Li, Qiaoming Zhu & Guodong Zhou. 2018. Stance detection via sentiment information and neural network model. Frontiers of Computer Science.