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

TÓM TẮT LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

Hà Nội - 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS Trần Thị Oanh

Phản biện 1: PGS.TS Hoàng Hữu Hạnh

Phản biện 2: PGS.TS Nguyễn Linh Giang

Luận văn này được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 14 giờ ngày 28 tháng 8 năm 2021

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. Do 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 (sentiment analysis) (khai thác ý kiến), nhận dạng cảm xúc (emotion recognition), khai thác lập luận (xác định lý do), phát hiện mỉa mai / 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.

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.

Chương 2: Các phương pháp học máy sử dụng trong bài toán phát hiện lập trường.

Chương 3: Đề xuất phương pháp, giải pháp: Chương này trình bày chi tiết về giải pháp đề xuất.

Chương 4: Thực nghiệm và đánh giá.

Trong phần Kết luận, luận văn tóm tắt các kết quả nghiên cứu chính của luận văn cùng với những bàn luận xung quanh đóng góp mới cả về ưu điểm và hạn chế từ đó đưa ra những gợi mở cần tiếp tục nghiên cứu.

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%.

Nhìn vào bối cảnh trên, lập trường 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. 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. 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.).

Điều này là do người tham gia sử dụng nhiều Classifers 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".

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.

2.1 Phương pháp học máy truyền thống

2.1.1 Thuật toán phân lớp naïve Bayes

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.

2.1.2 Cây quyết định (Decision tree)

Cây quyết định 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”.

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.

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. 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.

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.

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.

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 (long-term dependencies) trong mạng RNN do bị ảnh hưởng bởi vấn đề gradient biến mất.

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

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.
- 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.

- Cuối cùng, cần quyếtđịnh 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

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

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.

- **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.

3.2 Giải pháp đề xuất

3.2.1 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.

Thuật toán tách từ

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.

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.

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

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.

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.

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.

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 đụng. 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.

3.2.3 Phát hiện lập trường sử dụng mô hình học sâu

Word Embeddings

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.

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.

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

Lớp phân loại ReLU

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:

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

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 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.

3.3 Kết luận chương

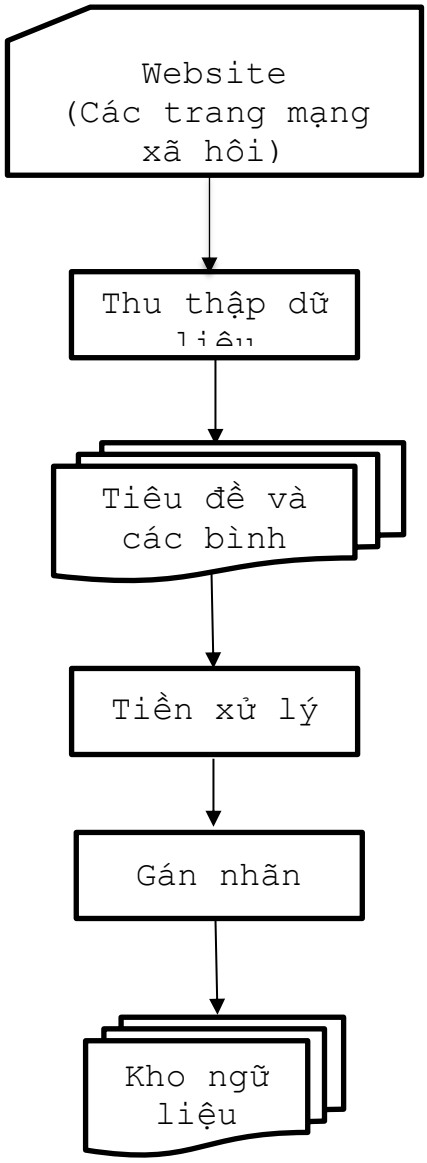
Nội dung chương đã mô tả bài toán phát hiện lập trường, và đưa ra giải pháp đề xuất them hai hướng khảo sát các phương pháp học máy truyền thống và học sâu.

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.

Thực hiện thu thập dữ liệu từ các nguồn: Vnexpress.vn, vtv24, dantri, beat.vn, baomoi.vn, trang Facebook: Beat.vn và Webtretho.com. Dữ liệu gồm 500 header là các chủ đề, tuyên bố. Tương ứng với đó là 11252 các bình luận tương ứng.

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 Header. 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.

Quá trình gán nhãn:

Những comment gán nhãn Agrees:

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

Những comment gán nhãn Disagrees:

- Thể hiện quan điểm không đồng ý, phản đối với 1 hoặc nhiều ý được nêu ra ở header.
- Thể hiện quan điểm không ủng hộ, không đồng quan điểm với 1 hoặc nhiều ý được nêu ra ở header
- Nêu ra dẫn chứng để gián tiếp phản đối 1 hoặc nhiều ý được nêu ra ở header.
- Không đồng tình, lên án , phê phán, trách móc hành động , đối tượng được nêu trong Header.

Những comment gán nhãn Discusses:

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

- Bổ sung, giải thích thêm thông tin, bàn luận với ý kiến được nêu ra ở header.
- Nêu ra dẫn chứng thực tế để bàn luận về thông tin ở header

Những comment gán nhãn Unrelated:

- Bình luận về chủ đề không liên quan đến vấn đề đang nói ở header.
- Bình luận về các vấn đề lan man bên cạnh vấn đề đang nói ở header
- Bài viết quảng cáo

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 tài liệu bài báo về độ đo này tính toán độ tương đồng gán nhãn giữa hai người.

4.1.3 Thống kê kho dữ liệu

Dữ liệu gồm 500 header 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.

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.

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.

4.4 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.

4.4.1 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	67.2	64.9	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

4.4.2 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 4-6 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**

4.4.3 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 5 cho chúng ta thấy kết quả Accuracy khi chạy kho dữ liệu trên 5 fold. Fold 5 cho điểm tốt nhất với 69,24%.

Bảng 4-7 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 4-8 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**

4.5 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.

4.6 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.