HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THANH HUY

NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT BẰNG MÔ HÌNH MÁY HỌC

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

TP.HỒ CHÍ MINH - NĂM 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT BẰNG MÔ HÌNH MÁY HỌC

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN
MÃ SỐ: 8.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS.TS NGUYỄN TUẤN ĐĂNG

TP.HỒ CHÍ MINH - NĂM 2022

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn thạc sĩ công nghệ thông tin " ***Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình máy học***" là do tôi nghiên cứu, tổng hợp và thực hiện dưới sự hướng dẫn của **PGS.TS Nguyễn Tuấn Đăng.**

Toàn bộ nội dung luận văn, những điều được trình bày là của chính cá nhân tôi hoặc là được tham khảo, tổng hợp từ nhiều nguồn tại liệu khác nhau. Tất cả các tài liệu tham khảo, tổng hợp đều được trích xuất nguồn gốc rỏ ràng. Các số liệu, kết quả được nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác

TP.HCM, ngày 25 tháng 01 năm 2022
Học viên thực hiện luận văn

Nguyễn Thanh Huy

LỜI CẢM ƠN

Trước hết, em xin bày tỏ tình cảm và lòng biết ơn của em tới Thầy **PGS.TS Nguyễn Tuấn Đăng**. Người đã từng bước hướng dẫn, giúp đỡ em trong quá trình thực hiện luận văn tốt nghiệp của mình.

Em xin chân thành cảm ơn Thầy Cô của Học viện Bưu Chính Công Nghệ Bưu Chính Viễn thông cơ sở tại TP.HCM đã dìu dắt, dạy dỗ em cả về kiến thức chuyên môn và tinh thần học tập để em có được những kiến thức thực hiện đồ án tốt nghiệp của mình.

Em xin chân thành cảm ơn Thầy **TS. Tân Hạnh** – Phó giám đốc Học viện Công Nghệ Bưu Chính Viễn Thông cơ sở tại TP.HCM, các phòng ban và quý Thầy Cô đã giúp đỡ tạo điều kiện tốt nhất cho em trong suốt thời gian học tập tại trường.

Tuy có nhiều cố gắng trong quá trình học tập, cũng như trong quá trình làm luận văn tốt nghiệp không thể tránh khỏi những thiếu sót, em rất mong được sự góp ý quý báu của tất cả các thầy cô giáo cũng như tất cả các anh chị để kết quả của em được hoàn thiện hơn.

Một lần nữa em xin chân thành cảm ơn.

TP.HCM, ngày 25 tháng 01 năm 2022
Học viên thực hiện luận văn

Nguyễn Thanh Huy

MỤC LỤC

DANH SÁCH HÌNH VẼ

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Diễn giải
NLP	Natural Language Processing
CNN	Convolutional Neural Network
SVM	Support Vector Machine
TF- IDF	Term Frequency – Inverse Document Frequency
CBOW	Continous Bag Of Words
DNN	Deep Neural Network
BOW	Bag Of Words
LSTMs	Long Short Term Memory Neural Network
PP	Phương pháp
SAV	Sentiment Analysis Vietnamese

MỞ ĐẦU

1. Lý do chọn đề tài

Với sự phát triển không ngừng các lĩnh vực công nghệ, việc nhận diện cảm xúc trong văn bản tiếng Việt được ứng dụng trong nhiều lĩnh vực như: quản trị doanh nghiệp, quản trị thương hiệu sản phẩm, quản trị quan hệ khách hàng, khảo sát ý kiến khách hàng hay dễ hiểu hơn là phân tích đánh giá, ý kiến phản hồi của khách hàng về một sản phẩm, …. Việc dự đoán là vô cùng quan trọng vì ý kiến, đánh giá của khách hàng ngày càng trở nên có giá trị thiết thực hơn. Do đó, vấn đề này được các doanh nghiệp quan tâm nhiều hơn. Họ cần xây dựng một hệ thống để phân tích ý kiến phản hồi của khách hàng về sản phẩm một cách tự động để qua đó nắm bắt được cảm nhận và thị hiếu của khách hàng, từ đó họ sẽ có chiến lược để nâng cao khả năng cạnh tranh với đối thủ và thích ứng được với sự biến động không ngừng của thị trường. Những thông tin này giúp hỗ trợ doanh nghiệp trong việc nhận biết các vấn đề để xây dựng và phát triển sản phẩm.

Còn trong nghiên cứu, việc xây dựng hệ thống nhận diện cảm xúc trong văn bản tiếng Việt là một bước tiến lớn trong xử lý ngôn ngữ tự nhiên, giúp giải quyết được nhiều vấn đề đang mắc phải. Xây dựng mô hình giải quyết bài toán phân tích cảm xúc người dùng. Cụ thể chúng tôi chia cảm xúc của khách hàng qua các ý kiến đánh giá, phản hồi ra thành hai trạng thái cảm xúc riêng biệt. Từ đó, chúng tôi xây dựng bài toán nhận diện cảm xúc người dùng bằng phương pháp phân lớp. Trong đó, mỗi ý kiến đánh giá, phản hồi diễn đạt cảm xúc từ khách hàng được biểu diễn thành một vector để đưa vào huấn luyện mô hình phân lớp.

2. Tổng quan về vấn đề nghiên cứu

Trong những năm gần đây, phân tích và nhận diện cảm xúc ngày càng trở nên phổ biến để xử lý dữ liệu truyền thông xã hội trên các cộng đồng trực tuyến, blog, wiki, nền tảng tiểu blog và các phương tiện cộng tác trực tuyến khác. Phân tích nhận diện cảm xúc là một nhánh của nghiên cứu điện toán sinh thái nhằm phân loại văn bản (nhưng đôi khi cả âm thanh và video) thành tích cực hoặc tiêu cực. Đây là một lĩnh

vực liên quan đến truy xuất thông tin và tổng hợp thông tin vì nó yêu cầu dữ liệu phải được thu thập, tích hợp và phân loại. Hầu hết các tài liệu về ngôn ngữ tiếng Anh nhưng gần đây ngày càng có nhiều ấn phẩm đề cập đến vấn đề đa ngôn ngữ. Hệ thống phân tích nhận diện cảm xúc có thể được phân loại rộng rãi thành dựa trên tri thức và dựa trên thống kê. Trong khi hầu hết các công việc áp dụng nó như là một bài toán phân loại đơn giản, phân tích cảm xúc là một bài toán nghiên cứu đòi hỏi phải giải quyết nhiều nhiệm vụ NLP (Natural Language Processing), bao gồm nhận dạng thực thể được đặt tên [3], trích xuất khái niệm [4], phát hiện châm biếm[5], trích xuất khía cạnh và phát hiện tính chủ quan [6]. Phát hiện tính chủ quan là một nhiệm vụ cần thiết của phân tích cảm xúc vì hầu hết các công cụ phát hiện cảm tính đều được tối ưu hóa để phân biệt giữa văn bản tích cực và tiêu cực

Hiện tại thì cộng đồng khoa học mới chỉ giải quyết tốt bài toán phân tích và nhận diện cảm xúc trong văn bản tiếng Việt ở cấp độ đơn giản, tức là phân tích cảm xúc với hai lớp cảm xúc tiêu cực và tích cực với độ chính xác hơn 85%. Bài toán phân tích cảm xúc có một số phương pháp [7] giải quyết như sau:

- Phương pháp thủ công (dò từ khóa): việc dự đoán cảm xúc dựa vào việc tìm kiếm các từ cảm xúc riêng lẻ, xác định điểm số cho các từ tích cực, xác định điểm số cho các từ tiêu cực và sau đó là tổng hợp các điểm số này lại theo một độ đo xác định để quyết định xem văn bản mang màu sắc cảm xúc gì. Điểm hạn chế của phương pháp này là quan tâm đến thứ tự các từ và sẽ bỏ qua các từ quan trọng. Độ chính xác của mô hình phụ thuộc vào độ tốt của bộ từ điển các từ cảm xúc.Ưu điểm của phương pháp này là dễ thực hiện, tính toán nhanh, chỉ tốn công sức cho việc xây dựng bộ từ điển dữ liệu của các từ cảm xúc thôi.
- Phương pháp Deep Learning Neural Network [8]: phương pháp phân tích nhận diện cảm xúc đã được giải quyết bằng mô hình học Recurrent Neural Network với một phương pháp được dùng phổ biến hiện nay là Long Short Term Memory Neural Network (LSTMs), kết hợp với phương pháp mô hình vector hóa từ Word2Vector với kiến trúc được sử dụng là Continuous Bag-

of-Words (CBOW).

- Phương pháp kết hợp rule-based và corpus-based [8]: Phương pháp này kết hợp sử dụng mô hình Deep Learning Recursive Neural Network với hệ tri thức chuyên gia được sử dụng trong xử lý ngôn ngữ tự nhiên được gọi là Sentiment Treebank. Sentiment Tree là một mô hình cây phân tích cú pháp của một câu văn, trong đó ở mỗi nút trong cây được kèm theo bộ trọng số cảm xúc lần lượt là: rất tiêu cực, tiêu cực, trung tính, tích cực và rất tích cực.

3. Mục đích nghiên cứu

Tìm hiểu các lí thuyết cần thiết để xây dựng được mô hình giải quyết bài toán nhận diện cảm xúc người dùng tiếng Việt qua các ý kiến đánh giá, phản hồi ... với cảm xúc mong đợi ở hai dạng định tính:

- Nhận diện tính tích cực – tiêu cực của văn bản.
- Xác định tính chủ quan – khách quan của văn bản.

Bên cạnh đó, mô hình giải quyết bài toán nhận diên cảm xúc trong văn bản tiếng việt phải được tối ưu về độ chính xác, hiệu suất thời gian thực hiện, giúp giải quyết các vấn đề còn mắc phải trong nhận diện cảm xúc khách hàng nói riêng và xử lý ngôn ngữ tự nhiên ở Việt Nam nói chung.

4. Đối tượng nghiên cứu

Đối tượng nghiên cứu: Nhận diện cảm xúc cho văn bản tiếng việt theo văn bản và đặc trưng của văn bản. Từ kết quả nhận diện cảm xúc, xây dựng mô hình nhận diện cảm xúc cho văn bản tiếng Việt

Phạm vi nghiên cứu: Nhận diện cảm xúc trong văn bản tiếng Việt với các phản hồi, ý kiến đánh giá sản phẩm trên website bán hàng shopee.vn, Lazada.vn

5. Phương pháp nghiên cứu

Trong luận văn này chúng tôi sử dụng phương pháp nghiên cứu lý thuyết kết hợp với xây dựng mô hình ứng dụng thực nghiệm:

- Thu thập các tài liệu, các nghiên cứu liên quan đến đề tài
- Về mặt lý thuyết, luận án tìm hiểu tổng quan về cảm xúc trong văn bản tiếng Việt, các phương pháp nhận dạng cảm xúc, đồng thời cũng trình bày một số mô hình nhận diện cảm xúc được tổng hợp từ các tài liệu, bài báo khoa học.
- Về mặt thực nghiệm, chúng tôi sử dụng các bộ công cụ để tính toán, phân tích, thống kê và đánh giá các tham số đặc trưng, tiến hành nghiên cứu và thực hiện các thực nghiệm để nhận diện cảm xúc dựa trên các mô hình với hai loại cảm xúc tích cực, tiêu cực, từ đó đánh giá kết quả đạt được để xác nhận giá trị của các mô hình và các tham số sử dụng.

CHƯƠNG 1
TỔNG QUAN TÀI LIỆU

1.1 Ngôn ngữ tự nhiên

Trong ngôn ngữ học, ngôn ngữ tự nhiên là ngôn ngữ nào phát sinh, không suy nghĩ trước trong não bộ của con người. Một số ngôn ngữ điển hình mà con người được sử dụng để giao tiếp với nhau, có thể ngôn ngữ âm thanh, ngôn ngữ ký hiệu, các ký hiệu xúc giác hay chữ viết [2]. Hiểu một cách đơn giản, ngôn ngữ tự nhiên (Natural Language) là ngôn ngữ mà con người dùng để giao tiếp với nhau như tiếng Việt, tiếng Anh,… và khác với ngôn ngữ nhân tạo như ngôn ngữ máy tính (Pascal, C, Python, …) hay mã Morse, Braille, ….

Theo thống kê, trên thế giới có khoảng 5600 ngôn ngữ, được phân bố rất không đồng đều và chỉ có một số ít các ngôn ngữ là có chữ viết.

- Đặc điểm

  Một số đặc điểm của ngôn ngữ tự nhiên [2]:

  - Ngôn ngữ tự nhiên là một hiện tượng xã hội đặc biệt.
  - Ngôn ngữ tự nhiên là một trong những phương tiện giao tiếp quan trọng nhất của con người, các phương tiện khác cũng được diễn giải qua ngôn ngữ tự nhiên.
  - Ngôn ngữ tự nhiên là một hệ thống các tín hiệu đặc biệt.

- Phân loại [8]

  - Phân loại ngôn ngữ theo nguồn gốc lịch sử
  - Phân loại ngôn ngữ theo trật tự từ
  - Phân loại ngôn ngữ theo loại hình: được nhiều người sử dụng nhất.

Phân loại các ngôn ngữ tự nhiên theo loại hình là cách phân loại ngôn ngữ tự nhiên theo cấu trúc và chức năng của ngôn ngữ tự nhiên. Từ việc phân loại người ta thu được các loại hình ngôn ngữ. Loại hình ngôn ngữ tự nhiên là một tập hợp các

ngôn ngữ tự nhiên. Trong mỗi ngôn ngữ thì có ba nhóm thuộc tính: thuộc tính phổ quát (thuộc tính chung), thuộc tính riêng biệt, thuộc tính loại hình. Trong đó thuộc tính loại hình được dùng làm tiêu chuẩn khi phân loại ngôn ngữ.

1.2 Ngôn ngữ tiếng Việt

Tiếng Việt là ngôn ngữ đơn lập, nghĩa là trong mỗi âm tiết đều được phát âm tách rời nhau và được biểu diễn bằng một chữ viết cụ thể. Đặc điểm này được thể hiện ở tất cả các mặt như về ngữ âm, từ vựng, ngữ pháp.

❖ Đặc điểm ngữ âm

Trong ngôn ngữ tiếng Việt thì 'tiếng' là một loại đơn vị đặc biêt. Về mặt ngữ âm, mỗi tiếng của tiếng Việt là một âm tiết. Hệ thống âm vị trong ngôn ngữ tiếng Việt thì rất phong phú và có tính cân đối. Trong ngôn ngữ tiếng Việt có rất nhiều từ được dùng để gợi hình, tượng thanh có giá trị gợi tả đặc sắc. Khi chúng ta viết câu, viết lời trong tiếng Việt thì phải chú ý đến sự hài hoà về ngữ âm, đến ngữ điệu của câu văn [1].

❖ Đặc điểm từ vựng [1]

Trong tiếng Việt. mỗi tiếng đều là một yếu tố có nghĩa. Tiếng là một đơn vị cơ sở trong hệ thống các đơn vị có nghĩa của ngôn ngữ tiếng Việt. Từ tiếng, người ta có thể tạo ra rất nhiều đơn vị từ vựng khác nhau để định danh cho sự vật, hiện tượng,… và chủ yếu được tạo ra bằng các phương thức ghép và phương thức láy.

Việc tạo ra các đơn vị trong ngôn ngữ tiếng Việt ở phương thức ghép chịu sự chi phối của quy luật kết hợp về ngữ nghĩa, ví dụ: đất nước. xe lửa, nhà lầu xe hơi, dậu đổ bìm leo,…. Theo phương thức này, tiếng Việt sử dụng các yếu tố cấu tạo từ thuần Việt hay được vay mượn từ các ngôn ngữ khác nhau để tạo ra các từ ngữ mới, ví dụ: nhân viên, karaoke, thư điện tử (e-mail), hộp thư thoại (voice mail), phiên bản (version), xa lộ thông tin, văn bản siêu liên kết, truy cập ngẫu nhiên, ...

Việc tạo ra các đơn vị từ vựng bằng phương thức láy thì quy luật phối hợp ngữ âm chi phối chủ yếu việc tạo ra các đơn vị từ vựng, chẳng hạn: đom đóm, bơ vơ,long

lanh, ầm ầm, lấm tấm, …

❖ Đặc điểm ngữ pháp

Từ của tiếng Việt đặc trưng là không biến đổi hình thái. Khi kết hợp các từ thành các kết cấu như ngữ, câu, phương thức trật tự từ và hư từ [2] rất quan trọng.

Việc sắp xếp các từ trong tiếng Việt theo một trật tự nhất định sẽ mạng ý nghĩa khác nhau qua đó biểu thị các quan hệ cú pháp. Trong tiếng Việt khi nói "Mùa xuân lại đến" là khác với "Lại đến mùa xuân".Nhờ kết hợp trật tự của từ mà ngữ nghĩa của chúng cũng khác nhau. Trong tiếng Việt thì trật tự kết cấu câu chủ ngữ đứng trước, vị ngữ đúng sau là trật tự phổ biến nhất.

Phương thức hư từ cũng là một trong những phương thức ngữ pháp chủ yếu được sử dụng trong ngôn ngữ tiếng Việt. Nhờ hư từ mà tổ hợp các từ khác nhau có nghĩa khác nhau. Hư từ kết hợp với trật tự từ cho phép tiêng Việt tạo ra các câu về hình thức và nội dung cơ bản giống nhau nhưng khác nhau hoàn toàn về sắc thái biểu cảm. Ví dụ, so sánh các câu sau đây:

- Bạn ấy không uống nước ngọt.
- Nước ngọt, bạn ấy không uống.
- Nước ngọt, bạn ấy cũng không uống.

1.3 Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên [2] là một phạm trù trong xử lý thông tin với dữ liệu đầu vào là các văn bản hay là tiếng nói. Ngày nay, các dữ liệu dạng này ngày càng trở thành là một trong những kiểu dữ liệu chính và lưu dưới dạng điện tử. Đặc điểm chung của dữ liệu là không có cấu trúc, hoặc nửa cấu trúc và không thể lưu lại dưới dạng bảng biểu. Do đó chúng ta cần phải xử lý để chuyển từ không thể hiểu thành có thể hiểu được.

Xử lý ngôn ngữ tự nhiên (Natural Language Processing) [2] là một lĩnh vực khoa học máy tính kết hợp giữa Trí tuệ nhân tạo & Ngôn ngữ học tính toán chủ yếu tập trung các xử lý tương tác giữa con người và máy tính sao cho máy tính có thể hiểu được ngôn

ngữ của con người.

Xử lý ngôn ngữ tự nhiên là hướng dẫn máy tính thay thế và giúp đỡ con người thực hiện các công việc về xử lý ngôn ngữ như: dịch thuật, phân tích dữ liệu văn bản, nhận dạng tiếng nói, tìm kiếm thông tin, tóm tắt văn bản,…

❖ Một số bài toán về xử lý ngôn ngữ tiêu biểu

Nhận dạng tiếng nói [12] (Speech recognization): phổ biến trong các hệ thống trợ lý ảo như Siri của Apple, Cortana của Microsoft, Google Assistant của Google, Alexa của Amazon, ….

Tổng hợp tiếng nói (Speech Synthesis) [32] : từ dữ liệu đầu vào là một văn bản, phân tích và chuyển thành tiếng nói. Hiện tại có rất nhiều các hãng công nghệ lớn như IBM và Amazon đều có dịch vụ Text to Speech chất lượng tốt nhưng chưa được hỗ trợ ngôn ngữ tiếng Việt..

Nhận dạng ký tự quang học [14] (Optical Character Recognition): từ một văn bản được in trên giấy máy tính sẽ chuyển thành một tệp văn bản và lưu được.

Tổng hợp tiếng nói (Speech synthesis hoặc Text to Speech – TTS) [33]: chuyển đổi ngôn ngữ từ dạng văn bản sang tiếng nói, thường được dùng trong đọc các văn bản tự động.

Truy xuất thông tin (Information Retrieval) [19]: hệ thống thực hiện xử lý và tìm các tài liệu dưới dạng không có cấu trúc (thường là văn bản) đáp ứng nhu cầu về thông tin từ những nguồn dữ liệu lớn. Các hệ thống được sử dụng phổ biến nhất hiện nay như các công cụ tìm kiếm như Google, Yahoo, hoặc Bing search.

Trích chọn thông tin (Information Extraction – IE): nhận diện loại thực thể, mối quan hệ giữa các thực thể và các sự kiện trong văn bản ngôn ngữ tự nhiên. Trích chọn thông tin sẽ trả về thông tin mà người dùng mong muốn.

Trả lời câu hỏi(Question Answering) [33]: Hệ thống có khả năng tự động trả lời câu hỏi của người dùng bằng phương thức là truy xuất thông tin từ một tập hợp các tài liệu.

Tóm tắt văn bản tự động (Automatic Text Summarization): là ứng dụng mà đầu vào là một văn bản và đầu ra là một văn bản được tóm tắt nội dung ngắn gọn mà vẫn giữ được nội dung chính của văn bản.

Chatbot: là một chương trình mà máy tính có khả năng trò chuyện (chat), hỏi đáp với con người qua hình thức hội thoại dưới dạng văn bản (text).

Dịch máy (Machine Translation): là việc con người sử dụng máy tính để dịch từ một ngôn ngữ này sang ngôn ngữ khác một cách tự động hóa toàn bộ .

Kiểm lỗi chính tả tự động: là việc sử dụng máy tính để tự động phát hiện các lỗi chính tả trong văn bản (lỗi từ vựng, lỗi ngữ pháp, lỗi ngữ nghĩa) và đưa ra gợi ý cách thức chỉnh sửa lỗi.

Tìm kiếm và truy xuất thông tin (Information Retrieval): từ một nguồn có rất nhiều file thông tin, tìm ra những file có liên quan đến câu hỏi cần tìm. Điển hình như Google Search, Yahoo Search, Bing, … hay một công cụ thuần Việt là Tìm kiếm Cốc Cốc.

Rút trích thông tin văn bản (Information Extraction) [26]: tìm ra những đoạn bên trong của văn bản chứa nội dung ta cần biết.

Khai phá dữ liệu (Data Mining) [17]: là quá trình phân tích dữ liệu từ một tập dữ liệu lớn để tìm ra các mẫu. Data Mining rất hữu ích trong việc tăng doanh thu và cắt giảm chi phí cho lĩnh vực kinh doanh. Đây là một hướng đi rất tiềm năng ở Việt Nam.

❖ Tình hình và những vấn đề chính trong xử lý ngôn ngữ tiếng Việt

Về xử lý tiếng nói và tiếng Việt, theo chúng tôi biết, hiện nay có rất nhiều nghiên cứu đã phân tích, nhận dạng và xử lý ngôn ngữ tự nhiên. Bên ngoài Việt Nam, cũng có nhiều nghiên cứu về xử lý ngôn ngữ tiếng Việt và có những thành tựu nhất định

CHƯƠNG 2
CƠ SỞ LÝ THUYẾT

Bài toán nhận diện cảm xúc [17] thuộc dạng bài toán phân tích ngữ nghĩa văn bản. Vì vậy, chúng tôi sẽ xây dựng một mô hình để phân tích và hiểu được ý nghĩa của câu văn, đoạn văn để quyết định xem câu văn đó hay đoạn văn đó mang ý nghĩa sắc thái cảm xúc nào. Về cơ bản, chúng ta có thể chia cảm xúc con người thành nhiều loại và việc này tương ứng với các bài toán phân lớp dữ liệu trong khai thác dữ liệu. Do đó, chúng tôi xây dựng ứng dụng nhận diện cảm xúc người dùng bằng phương pháp phân lớp dữ liệu.

Chúng tôi mô tả khái quát mô hình phân tích cảm xúc của người dùng. Dữ liệu đầu vào của bài toán là một câu văn, đoạn văn hay tổng quát hơn là một văn bản, còn kết quả đầu ra mong muốn là loại cảm xúc nào. Tùy vào mức độ chi tiết của việc phân tích mà ta phân chia thành số lượng loại cảm xúc. Chẳng hạn với bài toán đánh giá sản phẩm tiêu dùng, ta có thể phân loại cảm xúc người dùng ở hai mức độ có tính chất định tính: tích cực và tiêu cực.

2.1 Các mô hình mạng neuron dùng trong học sâu

❖ Định nghĩa

Có một số cách để mô tả học sâu. Học sâu [23] là một lớp của các thuật toán máy học mà:

- Sử dụng một tầng (cascade) nhiều lớp để trích tách các đặc điểm với các đơn vị xử lý phi tuyến. Mỗi lớp sau thì dùng đầu ra từ lớp trước để làm đầu vào.
- Học nhiều cấp độ ứng với các mức độ trừu tượng khác nhau, ở mỗi mức độ thì hình thành một hệ thống phân cấp của các khái niệm.

❖ Các mạng nơ ron nhân tạo

Một trong những phương pháp học sâu thành công nhất là mạng nơron nhân tạo [34].

Phương pháp mạng bộ nhớ dài ngắn hạn (LSTM) [34] của Hochreiter & Schmidhuber (1997). Trong năm 2009, các mạng LSTM đa chiều sâu đã có những thành công nhất định trong năm 2009 với nghiên cứu nhận dạng chữ viết tay.

Các phương pháp sử dụng đào tạo trước không có giám sát để tạo ra một mạng nơ ron. Sau đó mạng nơ ron này được đào tạo tiếp tục bằng cách truyền ngược có giám sát để tiến hành phân loại dữ liệu và có dán nhãn.

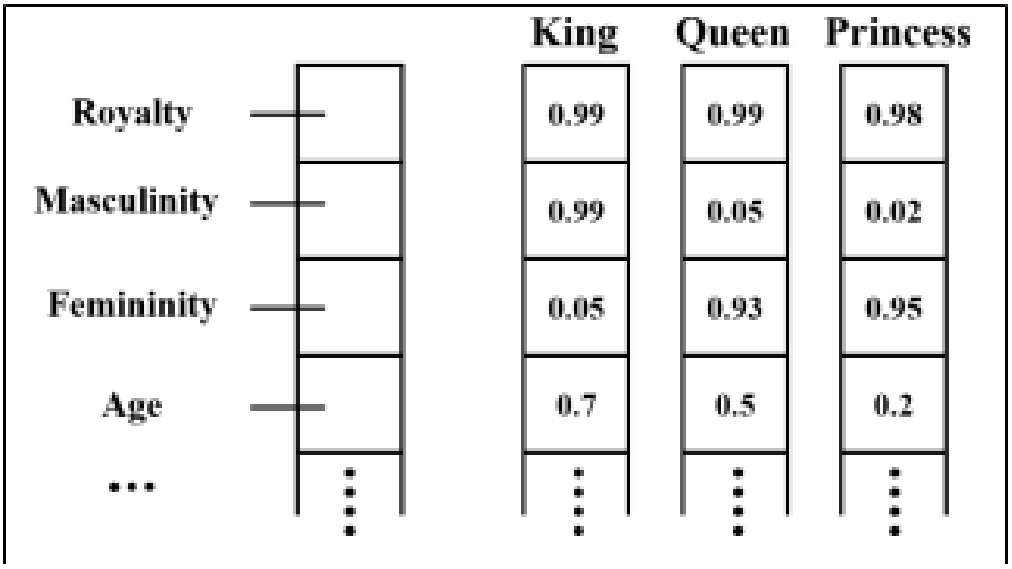
Mạng neuron sâu (DNN-Deep neural Network) [34] là một mạng neuron nhân tạo (ANN) với nhiều đơn vị lớp ẩn giữa các lớp đầu vào và các lớp đầu ra. Các kiến trúc DNN này được thể hiện như một thành phần được xếp lớp của các hình ảnh nguyên thủy.

Các mạng neuron sâu tích chập (CNN) [26] được sử dụng thành công trong lĩnh vực thị giác máy tính.

2.2 Word2Vec Text Embedding

❖ Khái niệm

Word2Vec [30] là biểu diễn các từ (word) dưới dạng một phân bố quan hệ với những từ còn lại (distributed representation) [8]. Mỗi từ thì được biểu diễn bằng một vector mang giá trị là biểu diễn phân bố quan hệ của từ này đối với các từ khác có trong từ điển. Như thế thay vì chỉ có kết nối one-to-one giữa các phần từ trong vector và một từ, biểu diễn từ sẽ là sự dàn trải của tất cả các thành phần liên quan của vector và mỗi phần tử trong một vector sẽ góp phần định nghĩa cho nhiều từ khác.



Hình 2.1: Cách biểu diễn các từ trên Word2Vec [23]

Với cách biểu diễn như vậy, chúng ta phát hiên ra rằng các vector mang lại cho ta cả cú pháp và ngữ nghĩa ở mức độ nào đó để máy tính có thể hiểu được

❖ Phương thức hoạt động

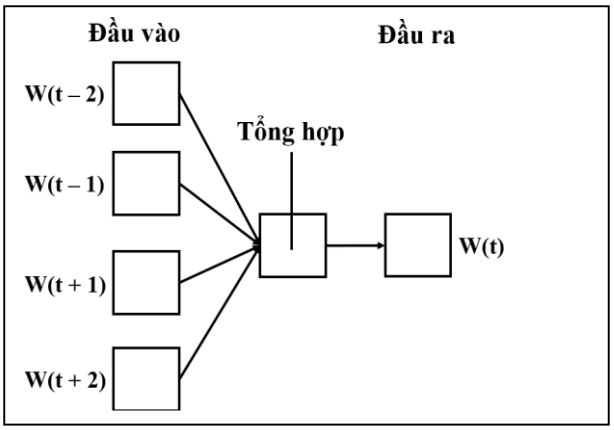
Công cụ Word2Vec sẽ lấy các bộ ngữ liệu của văn bản (Corpus) như là một dữ liệu đầu vào và từ đó tạo ra các dữ liệu đầu ra là Word Vector [23]. Đầu tiên, nó sẽ xây dựng một bộ từ vựng (Vocabulary) từ các văn bản dữ liệu sau khi đã được huấn luyện, sau đó nó sẽ học cách biểu diễn từ của Vector. Kết quả chúng ta thu được là một file Word Vector có thể được sử dụng trong các ứng dụng của xử lý ngôn ngữ tự nhiên và các ứng dụng học máy.

Có hai dạng mô hình chính trong Word2Vec: Continuous Bag of Words với Continuous Skip-Gram và có hai thuật toán chính được sử dụng trong Word2Vec là Hierarchical Softmax và Negative Sampling [21].

Về mô hình:

- Continuous Bag of Words: Ý tưởng của mô hình CBOW là mô hình dự đoán của từ hiện tại dựa trên các từ xung quanh hay các từ trong cùng một ngữ cảnh. Ngữ cảnh ở đây có thể là một câu, một đoạn văn hay một

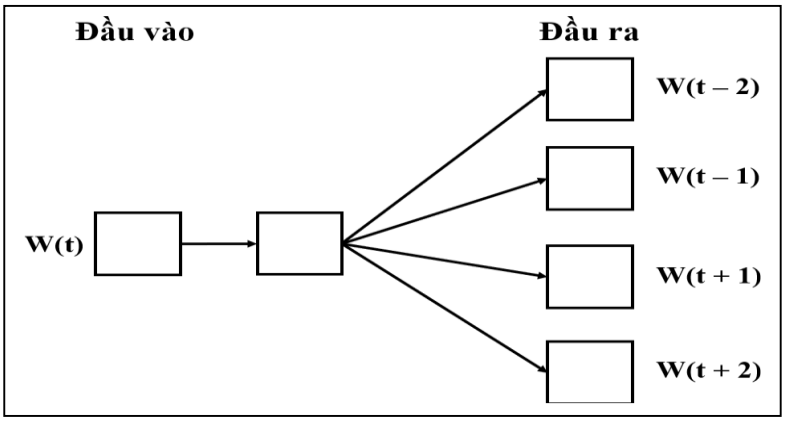
tập các từ đứng cạnh nhau [23]. Đầu vào của mô hình CBOW sẽ là tập hợp tất cả các ngữ cảnh và đầu ra là từ hiện tại mà chúng ta cần dự đoán. Mô hình CBOW sử dụng tầng chiếu chung cho tất cả các từ trong toàn văn bản, do đó tất cả các từ sẽ được chiếu vào các vị trí giống nhau. Ngoài việc sử dụng các từ đứng trước, CBOW còn sử dụng các từ đứng sau từ hiện tại để phân loại chính xác các từ hiện tại dựa trên việc xây dựng bộ phân loại Log-Linear cho các từ đứng trước và từ đứng sau. Trong mô hình này, thứ tự của các từ không làm ảnh hưởng đến kết quả dự đoán.



Hình 2.2: Mô hình Continous Bag of Words [23]

- Continuous Skip-gram: Kiến trúc của Continuous Skip-gram giống với Continuos Bag of Word, tuy nhiên thay vì dự đoán từ hiện tại dựa trên ngữ cảnh, mô hình này sẽ tập trung vào việc tối ưu hóa việc phân loại của một từ dựa trên các từ khác trong cùng một câu. Phương pháp này sử dụng tần chiếu liên tục và dự đoán các từ theo một dải phía trước và phía sau từ hiện tại. Việc tăng kích thước dải từ sẽ cải thiện chất lượng của Vector từ đầu ra, tuy nhiên nó cũng làm tăng độ phức tạp của việc tính toán. Vì những từ càng xa nhau thì sẽ thường ít liên quan đến các từ hiện tại hơn là những từ gần với nó, do đó chúng ta có thể đánh số trọng số cho những từ ở xa nhỏ hơn để khắc phục vấn đề này. Không giống với các kiến trúc mạng nơ-ron được sử dụng trước đó để học Vector từ, việc chúng ta xây dựng đào tạo mô hình Skip-gram không sử dụng đến các phép nhân ma

trận dày đặc. Điều này sẽ giúp chúng ta thực hiện việc đào tạo trở nên vô cùng hiệu quả, một máy đơn đã được tối ưu thì có thể đào tạo hơn 100 tỉ từ trong một ngày. Một mở rộng đáng ngạc nhiên của phương pháp này đó là việc áp dụng các phép tính cộng/trừ đại số cho các Vector chúng ta có thể thu được các kết quả khả quan về ngữ nghĩa [23].



Hình 2.3: Mô hình Continuous Skip-gram [23]

Về thuật toán:

- Phương pháp này để biểu diễn tất cả các từ có trong từ điển thì chúng tôi sử dụng cây nhị phân. Ứng với mỗi từ sẽ được biểu diễn là một lá trong cây. Với mỗi lá thì sẽ tồn tại duy nhất một đường đi từ gốc tới lá, từ đó đường này sẽ được sử dụng để ước lượng xác suất mỗi từ biểu diễn bởi lá .
- Negative Sampling chỉ đơn giản là chúng ta chỉ cập nhật mẫu đầu ra của từ ở mỗi vòng lặp . Từ đầu ra đó mục tiêu sẽ được giữ trong mẫu và được cập nhật và chúng ta sẽ thêm một vài từ như mẫu âm tính .

2.3 GloVe Vectors Text Embedding

GloVe (Vectơ toàn cầu cho đại diện từ) [22] là một trong những phương pháp được dùng thay thế để tạo nhúng từ. Phương pháp này được dựa trên kỹ thuật là phân tích nhân tử ma trận trên các ma trận ngữ cảnh của từ. Một ma trận lớn về

thông tin sẽ đồng xuất hiện được xây dựng và chúng ta đếm từng "từ" và tần suất xuất hiện của từ, chúng ta thấy các từ này trong một số "ngữ cảnh" trong một kho ngữ liệu lớn. Chúng tôi tiến hành quét kho dữ liệu của mình theo cách sau: đối với mỗi thuật ngữ, chúng tôi sẽ tìm kiếm với các thuật ngữ ngữ cảnh trong một số khu vực thì được xác định bởi kích thước cửa sổ trước của thuật ngữ và kích thước cửa sổ sau của thuật ngữ. Ngoài ra, chúng tôi đưa ra ít trọng lượng hơn cho các từ xa hơn.

Tất nhiên, số lượng "ngữ cảnh" là rất lớn, vì nó có kích thước của một tổ hợp. Vì vậy, sau đó chúng tôi phân tích nhân tử của ma trận này để tạo ra một ma trận với số chiều thấp hơn, khi đó mỗi hàng được biểu diễn một vectơ cho mỗi từ.

Trong luận văn này, chúng tôi đã sử dụng cả GloVe và Word2Vec để chuyển đổi văn bản của chúng tôi thành các bản nhúng và cả hai phương pháp này đều thể hiện các hiệu suất tương đương.

2.4 Các mô hình nhận diện cảm xúc trong văn bản

a. *Phân tích cảm xúc tiếp cận theo xử lý ngôn ngữ tự nhiên* [2]

Các ý kiến đánh giá, bình luận của khách hàng trên các website là dạng ngôn ngữ tự nhiên được viết ra. Do đó, việc chuẩn bị tập dữ liệu để phân tích, ở đây là dữ liệu văn bản là các nội dung bình luận, ý kiến đánh giá, phản hồi của khách hàng để lại sau khi sử dụng những sản phẩm của các cửa hàng, có thể trên website, trên các trang mạng xã hội. Tiếp theo là tiền xử lý, chúng ta tiến hành làm sạch dữ liệu, loại bỏ các kí tự đặc biệt, các loại dữ liệu rác, các dữ liệu không chuẩn hóa, sau đó chuẩn hóa dữ liệu về dạng ngữ pháp ngữ nghĩa. Ở giai đoạn khảo sát phân tích nghiên cứu sẽ phân tích được khái quát tính chất, nội dung, số lượng của tập dữ liệu thu được. Lựa chọn các yếu tố đầu vào để phân tích và dữ liệu đầu vào sẽ có rất nhiều chiều. Lựa chọn chiều nào thích hợp nhất để phân tích là môt việc rất quan trọng. Các chiều đầu vào có độ chính xác càng cao thì kết quả phân tích sẽ có độ chính xác càng cao. Bước cuối cùng là đánh giá kết quả thực nghiệm và tiến hành triển khai dự án.

b. Phân tích cảm xúc tiếp cận theo phương pháp học máy

Phân tích cảm xúc đã được định nghĩa là tính toán nghiên cứu các ý kiến đánh giá, tình cảm và cảm xúc được thể hiện trong văn bản (Liu, 2012) [16]. Hay nói cách khác, khai thác ý kiến là một phương pháp trích xuất ý kiến của người dùng để tạo ra một tài liệu cụ thể, đã trở thành mối quan tâm nghiên cứu lớn nhất trong mạng xã hội (Pang & Lee, 2008) [18]. Đặc biệt, trong thời đại phát triển kỹ thuật số, chúng ta hiện có một khối lượng dữ liệu rất lớn được lưu lại dưới dạng văn bản dùng để phân tích.

Học máy là một ứng dụng của Trí tuệ nhân tạo, là một lĩnh vực giúp hệ thống tự động hiểu được dữ liệu từ dữ liệu được đào tạo mà chúng ta không cần lập trình cụ thể. Học máy chia làm 4 phần (Das, Dey, Pal, & Roy, 2015) [19]: học có giám sát, học bán giám sát, học không giám sát và học củng cố.

Máy học có giám sát là thuật toán được dùng để dự đoán tập dữ liệu đầu ra dựa vào tập dữ liệu đã được huấn luyện. Phương pháp phân loại và hồi quy là hai loại của máy học có giám sát. Phân loại là chia dữ liệu theo từng nhóm rồi đưa ra kết quả dự đoán, hồi quy thì cho ra kết quả dự đoán là một số thực cụ thể.

Máy học không giám sát là thuật toán dự đoán dữ liệu đầu ra dựa vào duy nhất tập dữ liệu đầu vào, dữ liệu đầu vào sẽ không được dán nhãn hoặc kết quả đầu ra. Máy học không giám sát bao gồm phân nhóm và tích hợp. Thuật toán phân nhóm là phân tập dữ liệu thành các nhóm nhỏ dựa vào các liên quan của dữ liệu trong nhóm. Thuật toán tích hợp sẽ tìm ra một số quy luật trên tập dữ liệu để tiến hành khai phá dữ liệu.

Học bán giám sát [10] là thuật toán kết hợp của cả hai thuật toán có giám sát và không giám sát. Dữ liệu chia một phần được gán nhãn, phần còn lại thì không được gán nhãn.

Trong nghiên cứu này, chúng tôi chọn phương pháp học có giám sát để áp dụng cho bài toán nhận diện cảm xúc khách hàng dựa trên các ý kiến đánh giá, bình luận.

c. Mô hình nghiên cứu tổng quan

Trong nghiên cứu này, trước tiên chúng tôi tiến hành thu thập dữ liệu thô từ các trang web shopee.vn, lazada.vn. Sau đó dữ liệu thô được tiền xử lý và gán nhãn trước khi tiến hành học máy. Dữ liệu được chia thành hai nhóm: tập dữ liệu huấn luyện (training data), tập dữ liệu kiểm tra (test data).

Giai đoạn huấn luyện: là giai đoạn học tập trên tập dữ liệu huấn luyện của mô hình phân loại cảm xúc trong văn bản. Ở bước này, mô hình sẽ học từ dữ liệu có nhãn (trong nghiên cứu này nhãn là Tích cực, Tiêu cực). Dữ liệu văn bản sẽ được số hóa thông qua bộ trích xuất đặc trưng để mỗi mẫu dữ liệu trong tập huấn luyện trở thành 1 vector nhiều chiều. Thuật toán máy học sẽ học và tối ưu các tham số để đạt được kết quả tốt trên tập dữ liệu này. Nhãn của dữ liệu được dùng để đánh giá việc mô hình học tốt không và dựa vào đó để tối ưu.

Giai đoạn kiểm tra: là giai đoạn sử dụng mô hình học máy sau khi nó đã học xong. Ở giai đoạn này, dữ liệu trên tập dữ liệu kiểm tra cần dự đoán cũng vẫn thực hiện các bước trích xuất đặc trưng. Mô hình đã học sau đó nhận đầu vào là đặc trưng đó và đưa ra kết quả dự đoán.

CHƯƠNG 3
NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT

3.1 Tiền xử lý ngữ liệu

Dữ liệu được thu thập về sẽ có dạng thô, do chưa được xử lý nên có thể dữ liệu bị rỗng, dữ liệu không đúng chính tả, dữ liệu quá ngắn, quá dài hoặc chứa các biểu tượng icon. Điều này sẽ làm ảnh hưởng đến kết quả thu được sao cùng của mô hình, vì vậy chúng ta cần làm sạch dữ liệu trước khi phân tích.

Xóa các icon, các kí tự đặc biệt: các kí tự đặc biệt không mang ý nghĩa trong việc phân loại mà nó sẽ gây nhiễu trong quá trình phân tích. Chuyển tất cả về chữ thường, mỗi số, các ký tự đặc biệt. Chữ in hoa và chữ thường sẽ có mã unicode khác nhau, về mặt ngữ nghĩa thì giống nhau nhưng máy tính sẽ không thể phân biệt dữ liệu đầu vào, dẫn đến kết quả dự đoán có thể bị ảnh hưởng. Vì vậy, việc chuyển toàn bộ chữ về chữ thường là công việc hợp lý và cần thiết cho hệ thống phân tích và dự đoán.

Chuyển dạng từ rõ nghĩa: việc chuyển dạng từ rõ nghĩa là rất cần thiết cho bước tiền xử lý dữ liệu khi phân tích. Các bình luận hay các ý kiến trên các website do người dùng bình luận bằng tiếng Việt nên việc viết tắt, viết sai chính tả là hiển nhiên. Chẳng hạn như: từ ko đúng (không đúng), vs (với), 100k (100.000),… hay dữ liệu sẽ không đồng bộ, không được chuẩn hóa. Việc này sẽ làm ảnh hưởng gây nhiều đến kết quả phân tích.

Xóa dòng dữ liệu: tập dữ liệu thu về sẽ có rất nhiều dữ liệu bị trống, dữ liệu trống sẽ không có ý nghĩa trong quá trình phân tích, gây hao tốn bộ nhớ lưu trữ.

a. ***Tách từ*** [27]

Trong tiếng Việt, dấu cách (space) sẽ không được sử dụng như một dạng kí hiệu dùng để phân tách từ, nó chỉ có ý nghĩa khi phân tách các âm tiết với nhau. Vì thế, khi chúng tôi tiến hành xử lý tiếng Việt, công đoạn tách từ (word segmentation) là một trong những công việc đầu tiên cơ bản và quan trọng nhất.

Ví dụ : từ “đất nước” được tạo ra từ 2 âm tiết “đất” và “nước”, cả 2 âm tiết này

đều có ý nghĩa riêng khi chúng đứng độc lập, nhưng khi ghép lại sẽ mang một ý nghĩa khác. Vì vậy, bài toán tách từ trở thành một trong những bài toán tiền đề cho việc xây dựng các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, tóm tắt văn bản, máy dịch tự động,….

Về mặt biểu hiện, tách từ là gom nhóm các từ đơn liền kề thành một cụm từ có ý nghĩa. Ví dụ: "Cách tách từ cho tiếng Việt." sau khi tách từ thì thành "Cách tách từ cho tiếng_Việt." Về hình thức, các từ đơn được gom nhóm với nhau bằng cách nối với nhau bằng ký tự gạch dưới "_", trong trường hợp này là từ “tiếng_Việt”. Sau khi thực hiện tách từ thì mỗi từ (Token) trong câu được cách nhau bởi một khoảng trắng, trong trường hợp này: "tiếng Việt " thì từ "tiếng_Việt" cách dấu "." bởi một khoảng trắng. Đây là quy ước chung cho tất cả các ngôn ngữ của bài toán tách từ trong XLNNTN [2].

Về mặt ngữ nghĩa, việc tách từ văn bản đầu vào trước khi đưa vào huấn luyện mô hình máy học là để giải quyết các bài toán liên quan đến ngữ nghĩa của văn bản, tức là kết quả đầu ra mang tính suy luận dựa trên việc hiểu ý nghĩa của văn bản đầu vào. Ví dụ như các dạng bài toán: phát hiện đạo văn, tóm tắt văn bản, hỏi đáp tự động, hỗ trợ khách hàng tự động, phân tích cảm xúc văn bản, dịch máy, và trợ lý ảo.

Mục tiêu của việc tách từ văn bản đầu vào là để khử tính nhập nhằng về ngữ nghĩa của văn bản. Tùy vào từng loại ngôn ngữ có những đặc điểm khác nhau mà việc tách từ văn bản cũng có độ khó khăn khác nhau. Với ngôn ngữ hòa kết như tiếng Anh, thì việc tách từ khá đơn giản vì ranh giới từ được nhận diện bằng khoảng trắng và dấu câu. Với ngôn ngữ tiếng Việt, thuộc loại hình đơn lập, mang đặc điểm là từ tiếng Việt không biến đổi hình thái, ranh giới từ không được xác định mặc nhiên bằng khoảng trắng. Tiếng Việt có đặc điểm là ý nghĩa ngữ pháp nằm ở ngoài từ, phương thức ngữ pháp chủ yếu là trật tự từ và hư từ.

Cho nên có trường hợp một câu có thể có nhiều ngữ nghĩa khác nhau tuỳ vào cách chúng ta tách từ như thế nào, dẫn đến gây nhập nhằng về ngữ nghĩa của câu. Ví dụ câu "Cam phun thuốc sâu không ăn." có thể tách với ý nghĩa hoàn toàn khác như sau:

- Cam / phun thuốc / sâu / không / ăn.

- Cam / phun / thuốc sâu / không / ăn.

Với câu "Ăn cơm không được uống rượu", có thể được tách từ như sau:

- Ăn / cơm / không / được / uống / rượu.
- Ăn / cơm không / được / uống / rượu.

Bài toán tách từ có ba hướng tiếp cận:

- Tiếp cận dựa vào từ điển cố định: Ý tưởng của hướng tiếp cận này là những cụm từ được tách ra từ văn bản phải được so khớp với các từ trong từ điển
- Tiếp cận dựa vào thống kê thuần túy: Dựa trên các thông tin như tần số xuất hiện của từ trong tập huấn luyện ban đầu
- Tiếp cận dựa trên cả hai hướng trên: Với mục đích kết hợp các hướng tiếp cận khác nhau để thừa hưởng được các ưu điểm của nhiều kỹ thuật và các hướng tiếp cận khác nhau nhằm nâng cao kết quả. Hướng tiếp cận này thường kết hợp giữa hướng dựa trên thống kê và dựa trên từ điển nhằm tận dụng các mặt mạnh của các phương pháp này. Tuy nhiên hướng tiếp cận này lại mất nhiều thời gian xử lý, không gian đĩa và đòi hỏi nhiều chi phí.
- Hướng tiếp cận dựa trên ký tự: Hướng tiếp cận này đơn thuần là rút trích ra một số lượng nhất định các tiếng trong văn bản như rút trích từ 1 ký tự (unigram) hay nhiều ký tự (n- gram). Ưu điểm nổi bật của hướng tiếp cận này là tính đơn giản, dể ứng dụng, ngoài ra còn có thuận lợi là ít tốn chi phí cho các thao tác tạo chỉ mục và xử lý câu truy vấn.

Tách từ chính xác là công việc rất quan trọng, nếu không chính xác sẽ làm cho ý nghĩa của câu sẽ không đúng, ảnh hưởng trực tiếp đến tính chính xác kết quả của chương trình.

Để thực việc tách từ trong văn bản tiếng Việt trong luận văn này chúng tôi sử dụng bộ công cụ Tokenizer.

b. ***Chuẩn hóa từ*** [24]

Mục đích của việc chuẩn hóa từ là đưa văn bản không đồng nhất về cùng một dạng. Chuẩn hóa cũng giúp tối ưu bộ nhớ lưu trữ và tăng tính chính xác là rất quan trọng.

Ví dụ: U.S.A = USA

Ví dụ trong từ điển, dữ liệu huấn luyện của chúng ta không có U.S.A, chỉ có USA, thì việc chuyển đổi những từ như U.S.A về USA là điều cần thiết để thực hiện các bước xử lý tiếp theo được chính xác.

Có rất nhiều cách viết, với mỗi cách viết khi chúng ta lưu trữ sẽ làm tốn lượng bộ nhớ khác nhau, như một nữa kích thước tập tin chỉ tốn 1/2 dung lượng so với toàn bộ tập tin nên tuỳ theo nhu cầu cũng như tình hình thực tế, chúng ta sẽ đưa văn bản về một dạng đồng nhất.

Ngoài ra trong một vài trường hợp, nếu các ký tự số không mang lại lợi ích gì khi phân tích thì cũng sẽ tiến hành loại bỏ các ký tự số đó, nếu cứ để nguyên rất có thể các ký tự số sẽ trở thành tiếng ồn, ảnh hưởng đến tính chính xác của mô hình sau này.

c. ***Loại bỏ stopword*** [2]

StopWords là những từ trong ngôn ngữ tự nhiên xuất hiện rất nhiều, tuy nhiên nó lại không mang nhiều ý nghĩa mà chủ yếu đóng vai trò trong ngữ pháp. Ở tiếng việt StopWords là những từ như: để, đó, này, kia,....

Có rất nhiều cách để loại bỏ StopWords nhưng thường được sữ dụng có 2 cách chính là:

- Dùng từ điển

  Cách này đơn giản nhất, chúng ta tiến hành lọc văn bản, loại bỏ những từ mà chúng xuất hiện trong từ điển StopWords: cậu, của, cứ, dù, nọ, phóc, này, kia, để,….

- Dựa theo tần suất xuất hiện của từ

  Với cách này, chúng ta tiến hành đếm xem số lần xuất hiện của từng từ trong

toàn bộ dữ liệu sau đó sẽ tiến hành loại bỏ những từ xuất hiện nhiều lần.

d. ***Xóa HTML code trong dữ liệu*** [2]

Dữ liệu được chúng ta thu thập từ các website thường vẫn còn sót lại các đoạn mã HTML. Các mã HTML code này là rác, chẳng những chúng không có tác dụng cho việc phân loại mà còn làm cho kết quả phân loại văn bản bị giảm đi.

Việc xóa các HTML code này cũng khá đơn giản, chúng tôi đã sử dụng regex trong Python để xóa đi một cách đơn giản

e. ***Xóa các ký tự không cần thiết*** [2]

Tiền xử lý dữ liệu bao gồm việc loại bỏ các dữ liệu không có tác dụng cho việc phân loại văn bản. Việc này giúp:

- Giảm thiểu tối đa số chiều đặc trưng, làm tăng tốc độ học và xử lý dữ liệu
- Tránh làm ảnh hưởng xấu tới kết quả của mô hình khi phân loại dữ liệu

Trong ngôn ngữ tiếng Việt, các ký tự không cần thiết thường gặp rất nhiều như: các dấu ngắt câu, số đếm và các ký tự đặc biệt, các ký tự này không giúp chúng ta phân loại một văn bản thuộc loại cảm xúc nào. Do đó, chúng ta nên loại bỏ nó đi.

Riêng với số đếm, ngày tháng, email, chúng ta đưa nó về các token chung như: <number>, <date>, <email>,.… Việc này có thể không giúp ích cho mô hình học tốt hơn nhưng sẽ giúp ích trong việc giữ được mạch của dữ liệu.

3.2 Chuẩn hóa các đặc trưng văn bản

Mục tiêu của chuẩn hóa:

Khi chúng ta chuẩn hóa một tài nguyên ngôn ngữ tự nhiên [23], chúng ta cần giảm bớt tính ngẫu nhiên trong đó, đưa chúng về gần hơn với các "tiêu chuẩn" đã được xác định trước. Khi chuẩn hóa cần cố gắng đưa mọi thứ gần hơn với "phân phối chuẩn" nghĩa là chúng ta tìm cách làm cho mọi thứ "hoạt động như mong đợi" theo hình dạng tốt và có thể đoán được

Đầu tiên, bằng cách chúng ta làm giảm biến thể, tức là làm ít đi các biến đầu vào

để việc xử lý được dễ dàng và làm tăng hiệu xuất tổng thể của mô hình

Thứ hai, việc chuẩn hóa sẽ làm giảm kích thước đầu vào, khi chúng ta sử dụng các cấu trúc như BoW và TF-IDF sẽ làm giảm số lượng các xử lý để tạo bản nhúng.

Thứ ba, việc chuẩn hóa giúp xử lý các đầu vào vi phạm mã, nhằm đảm bảo rằng dữ liệu đầu vào sẽ được tuân theo quy định cụ thể.

Cuối cùng, việc chuẩn hóa dữ liệu nếu được thực hiện đúng cách giúp cho việc trích xuất thông kê được chính xác và đáng tin cậy

Khi thực hiện việc chuẩn hóa thì chúng tôi quan tâm nhất hai điều là cấu trúc câu và từ dựng. Khi chuẩn hóa cần giải quyết các vấn đề sau:

- Loại bỏ những khoảng trắng và dấu câu bị trùng lắp.
- Loại bỏ các chữ in hoa.
- Xóa hoặc thay thế các ký tự đặc biệt và các biểu tượng cảm xúc. Ví dụ: xóa các thẻ bắt đầu bằng dấu $,#,@ ,...
- Chuyển các chữ số thành số. Ví dụ: ‘năm mươi lăm’ thành ’55’.
- Thay thế các giá trị cho loại của chúng. Ví dụ ‘$100’ -> ‘money’.
- Chuẩn hóa các từ viết tắt. Ví dụ : ‘VN’ -> ‘Việt Nam’.
- Chuẩn hóa định dạng ngày tháng.
- Sữa lỗi chính tả: khi viết các bình luận người dùng thường viết sai chính tả rất nhiều cho nên làm giảm biến thể của từ dựng.
- Thay thế cho các từ hiếm gặp thành các từ đồng nghĩa được thông dụng hơn

3.3 Vector hóa văn bản [24]

Word embedding là một trong những bước quan trọng khi xây dựng bài toán phân tích cảm xúc trong văn bản tiếng Việt bằng mô hình máy học. Một lý do cơ bản mà chúng ta cần phải vector hóa văn bản là máy tính không thể hiểu được nghĩa của các từ. Như vậy để xử lý ngôn ngữ tự nhiên chúng ta cần có một phương pháp để biểu

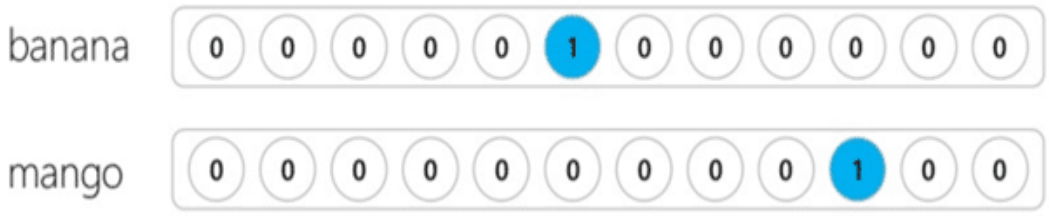
diễn văn bản dưới dạng mà máy tính có thể hiểu được. Phương pháp tiêu chuẩn để biểu diễn các văn bản thành các vector. Khi đó các từ hay các cụm từ được ánh xạ thành những vector trong không gian số thực.

Khi vector hóa văn bản, chúng tôi sử dụng 2 phương pháp chính sau đây: Phương pháp Word Embedding cổ điển và Neural Embedding (Vector hóa văn bản theo phương pháp mạng nơ-ron).

a. Phương pháp word embedding cổ điển

❖ Bag of words(BoW)

BoW [24] là một phương pháp biểu diễn vector cổ điển được sử dụng nhiều nhất. Khi đó mỗi từ sẽ được biểu diễn thành một vector có số chiều bằng đúng với số từ trong bộ từ vựng và ứng với vị trí của từ đó trong túi từ, phần tử đó sẽ được đánh dấu là 1, còn các vị trí còn lại đánh dấu là 0.



Hình 3.1: Mô hình BoW [24]

Trong phương pháp BoW này, các từ giống nhau sẽ có trọng số như nhau. Phương pháp này không quan tâm đến tần suất xuất hiện của từ hay ngữ cảnh của từ. Trong thực tế, khi phân tích từ chúng ta cần hiểu rõ nghĩa của từ, chúng ta cần xét nghĩa của từ trong toàn văn bản hơn là xét nghĩ độc lập

❖ TF_IDF [24]

TF-IDF là một phương pháp thống kê nhằm giúp phản ánh được độ quan trọng của từ đối với văn bản trên toàn bộ dữ liệu đầu vào. TF-IDF thể hiện được trọng số của mỗi từ theo ngữ cảnh trong toàn văn bản. TF-IDF chuyển đổi dạng

biểu diễn văn bản thành dạng không gian vector (VSM), hoặc thành những vector thưa thớt. Phương pháp này giúp làm tăng tỷ lệ thuận với số lần xuất hiện của từ trong văn bản và số các văn bản mà có chứa các từ đó trên toàn tập liệu đầu vào.

TF(Term frequency) : Tần suất xuất hiện của một từ trong một đoạn văn bản.

IDF(Invert Document Frequency) : Được dùng để đánh giá mức độ quan trọng của một từ trong văn bản.

Khi tính TF thì mức độ quan trọng của các từ là như nhau. Tuy nhiên trong văn bản thường xuất hiện nhiều từ không quan trọng với tần xuất cao. Do đó ta cần làm giảm đi mức độ quan trọng của từ đó bằng IDF. Cách tính TF-IDF được cho bởi công thức sau:

$$tf_i = n_i/N_i$$

Trong đó:

- i: 1….D
- n_i: Tần số xuất hiện của từ trong văn bản i.
- N_i : Tổng số từ trong văn bản i.

$$Idf_i = \log_2 D/d$$

Trong đó:

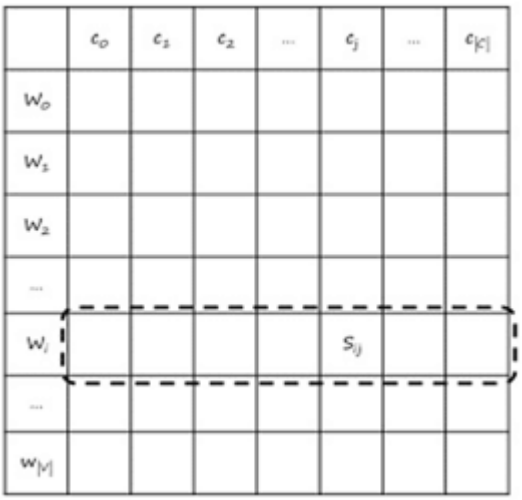
- D : Tổng số document trong tập dữ liệu.
- d : Số lương document có sự xuất hiện của từ.

$$tfidf_i = tf_i * idf_i$$

❖ Distributional Embedding

Phương pháp Dítributional Embedding giúp chúng ta có thể xem xét tổng quan toàn bộ ngữ cảnh của từ. Ở phương pháp này mỗi từ sẽ được biểu diễn

trên các thông tin tương hỗ với các từ khác trong tập dữ liệu đầu vào và dưới dạng tần suất xuất hiện trong ma trận trên toàn bộ tập dữ liệu hay xem xét trong tập dữ liệu lân cận hoặc xét trên giới hạn của các từ xung quanh.



Hình 3.2: Ví dụ ma trận thuật toán Distributional Embedding [24]

Phương pháp Dítributional Embedding ra đời trước phương pháp Neural Embedding. Phương pháp này giúp chúng ta quan sát được tầm quan trọng của mỗi từ tốt hơn.

b. Phương pháp Neural Embedding

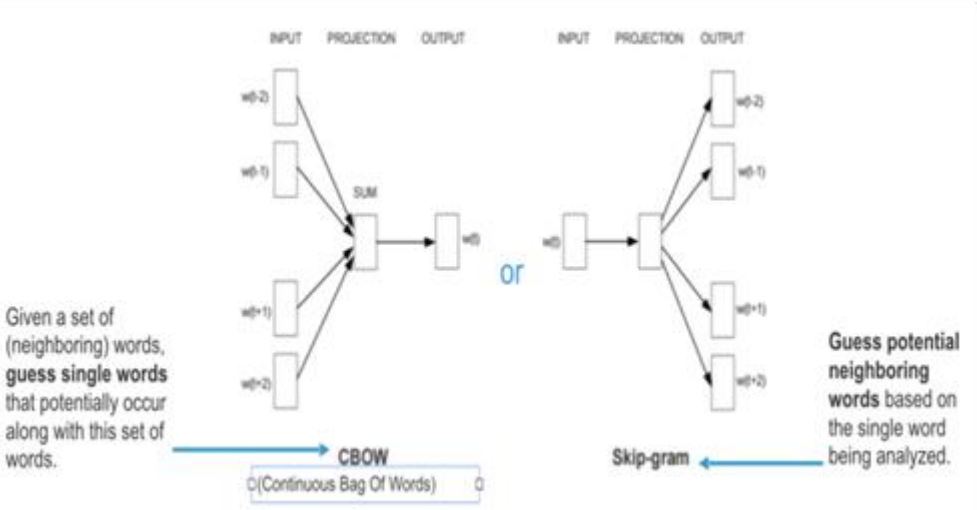
❖ Word2vec

Phương pháp Word2vec [30] là một mô hình đơn giản và nổi tiếng giúp tạo ra các biểu diễn embedding của từ trong một không gian có số chiều thấp hơn nhiều lần so với số từ trong từ điển. Mô hình dự đoán sẽ biểu diễn vector từ thông qua các từ, ngữ cảnh xung quanh nhằm tăng khả năng dự đoán được nghĩa của từ

Có 2 cách xây dựng mô hình Word2vec dùng để biểu diễn phân tán của từ trong không gian vector:

- Sử dụng ngữ cảnh để dự đoán mục tiêu (CBOW).

- Sử dụng một từ để chúng ta dự đoán ngữ cảnh mục tiêu (Continuous skip-gram) xem xét các từ ngữ cảnh xung quanh sẽ được đánh giá tốt hơn so với các từ trong ngữ cảnh nhưng ở vị trí xa hơn.



Hình 3.3: Mô hình CBOW và Skip-gram [30]

Trong hai thuật toán trên, thuật toán CBOW khi thực thi sẽ ít tốn thời gian để huấn luyện mô hình hơn Skip-gram. Tuy nhiên, thuật toán Skip-gram có độ chính xác cao hơn và có chứa cả những từ ít xuất hiện.

❖ Glove

Thuật toán GloVe [26] dựa trên sự tương phản có lợi với cùng dự đoán của ma trận đồng xuất hiện được sử dụng trong thuật toán Distributional Embedding, nhưng sử dụng phương pháp Neural Embedding để phân tích ma trận đồng xuất hiện thành những vector có ý nghĩa và có tỷ trọng hơn.

Thuật toán Glove nhanh hơn thuật toán CBOW, nhưng cả hai thuật toán đều không hiển thị để cung cấp kết quả đầu ra tốt và rỏ ràng hơn thay vì dùng cả hai

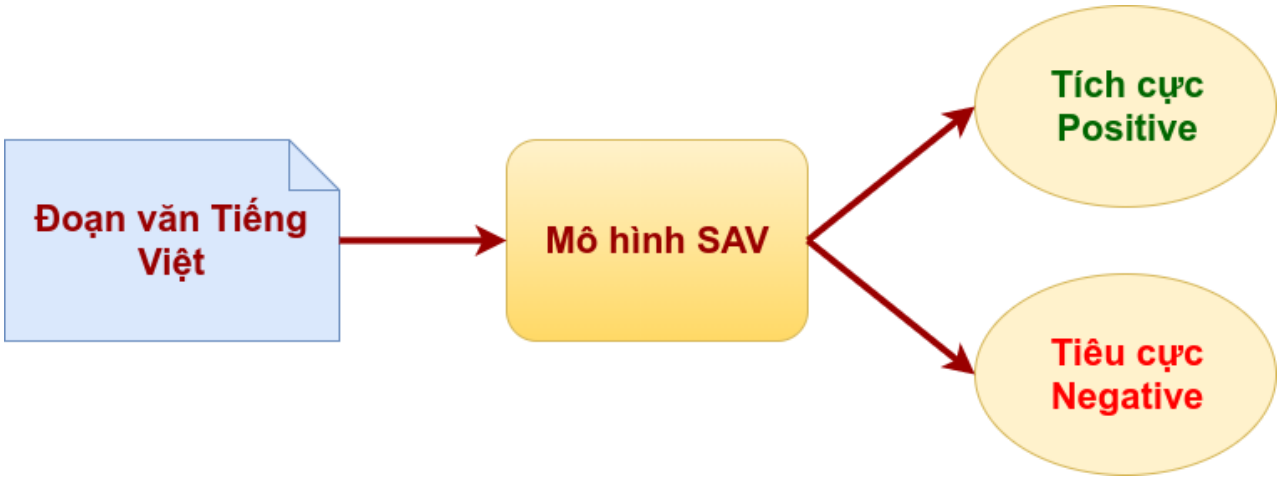
❖ FastTest

Thuật toán FastText [26] được xây dựng trên Word2Vec bằng cách học các

biểu diễn vectơ cho mỗi từ và n-gram được tìm thấy trong mỗi từ. Các giá trị của các biểu diễn sau đó được tính trung bình thành một vectơ ở mỗi bước đào tạo. Việc này sẽ giúp rất nhiều trong tính toán để bổ sung cho việc đào tạo, và nó cho phép chúng ta nhúng từ để mã hóa thông tin các từ phụ.

3.4 Mô hình nhận diện cảm xúc sử dụng học sâu

Bài toán nhận diện cảm xúc được giải quyết bằng mô hình học sâu recurrent neural network với phương pháp được sử dụng là mô hình học máy không giám sát , mô hình máy học có giám sát và mô hình Naïve Bayes, được kết hợp với mô hình vector hóa từ Word2vector với kiến trúc Continuous Bag of Words và mô hình vector hóaTF-IDF.



Hình 3.4: Mô hình SAV [29]

Để thực hiện được mô hình này thì đòi hỏi chúng ta phải có được một tập dữ liệu càng lớn càng tốt để tạo Word2Vec CBOW và TF-IDF đạt được chất lượng tốt và dữ liệu đươc gán nhãn đủ lớn để tao tập huấn luyện và tập kiểm tra bằng mô hình máy học có giám sát. Từ đó chúng tôi sẽ đánh giá được độ chính xác thông qua mô hình.

CHƯƠNG 4
THỰC NGHIỆM

4.1 Xây dựng ngữ liệu

4.1.1 Cơ sở lý thuyết của bộ dữ liệu

Với mục tiêu xây dựng một hệ thống nhận diện cảm xúc trong văn bản tiếng Việt, luận văn tập trung vào khía cạnh phân tích cảm xúc trong các bình luận, đánh giá sản phẩm trên website Shopee.vn, Lazada.vn,... Với chủ đề này, luận văn tập trung nghiên cứu xoay quanh phân loại cảm xúc trong các phản hồi của khách hàng

4.1.2 Xây dựng bộ dữ liệu

Với nội dung đã tìm hiểu về chủ đề phản hồi, đánh giá của khách hàng, bộ dữ liệu của chúng tôi được thu thập từ các trang bán hàng trực tuyến và được phân tích sẵn thành tập huấn luyện và tập kiểm tra.

Bộ dữ liệu được xây dựng thành 1 file data.csv. Mỗi dòng dữ liệu có 2 thành phần: review, sentiment.

- Review: là các phản hồi, đánh giá sản phẩm của khách hàng.
- Sentiment: Phân loại cảm xúc của khách hàng được phân loại như sau: đánh giá thuộc 4, 5 sao là ‘Tích cực’, các đánh giá 1, 2 sao là ‘Tiêu cực’.

Với cấu trúc như trên, luận văn đã xây dựng được 1 bộ dữ liệu với 4063 đánh giá, trong đó có 2030 đánh giá ‘Tích cực’, 2033 đánh giá ‘Tiêu cực'.

	review	sentiment
0	Chất lượng cần tốt hơn. Mình đã mua mấy lần ng...	Tiêu cực
1	Áo đẹp lắm nha ❤.chất vải mịn .mỗi tội mk thấy...	Tích cực
2	Áo xinh lắm mọi người ơii,mọi người nên mua nh...	Tích cực
3	Áo đẹp giao hàng nhanh □□□□□□□□□□□□□□□□□□□□□□...	Tích cực
4	Màu áo rất xấu, không giống như trong hình	Tiêu cực
...	...	...
95	Không bám tường cho lắm Mới mang vô nhà tắm te...	Tiêu cực
96	Chuyên nghiệp, thân thiện. Đẹp như mô tả. Giá ...	Tích cực
97	Chuyên nghiệp, thân thiện. Đẹp như mô tả. Đóng...	Tích cực
98	Chuyên nghiệp, thân thiện. Đóng gói kỹ lưỡng. ...	Tích cực
99	Mới mở ra xài đc. Hết pin xạc ko vô. Đã vậy cắ...	Tiêu cực

Hình 4.1: Mô tả bộ dữ liệu

Thông kê dữ liệu được trình bày trong bảng sau:

Bảng 4.1: Bộ dữ liệu

Bộ dữ liệu	Tích cực	Tiêu cực
Huấn luyện	1,608	1,592
Kiểm tra	422	441

Sau khi chúng tôi thực hiện việc tiền xử lý dữ liệu gồm tách từ và loại bỏ các hư từ cũng như các dấu câu không cần thiết. Chúng tôi thu được dữ liệu như trong bảng sau:

Bảng 4.2: Bộ dữ liệu đã tiền xử lý

Bộ dữ liệu	Tích cực	Tiêu cực
Huấn luyện	1,608	1,592
Kiểm tra	422	441

4.1.3 *Tiền xử lý dữ liệu* [31]

Tiền xử lý dữ liệu là một trong những bước quan trọng nhất khi giải quyết bất kỳ bài toán nào trong lĩnh vực Học máy. Để mô hình có thể đưa ra kết quả có độ chính xác cao thì bộ dữ liệu luôn cần được xử lý, làm sạch và biến đổi trước khi trở thành dữ liệu huấn luyện cho mô hình học máy.

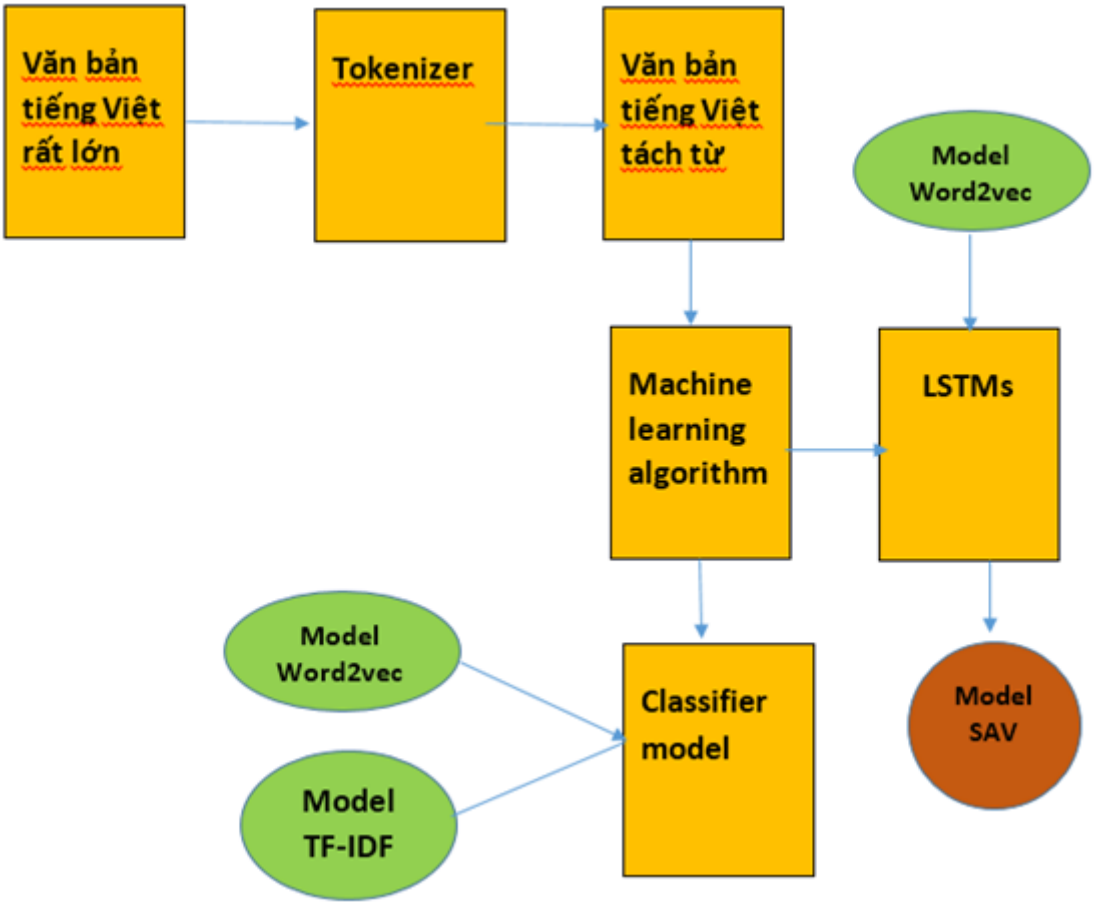
Đối với luận văn này, dữ liệu input đầu vào là các phản hồi, đánh giá của khách hàng về sản phẩm. Dữ liệu thường không chuẩn, vì thế ta phải tiến hành xử lý dữ liệu:

- Loại bỏ các dãy html.
- Loại bỏ các dấu ngoặc vuông.
- Loại bỏ văn bản nhiễu.
- Loại bỏ các ký tự đặc biệt.
- Đưa các từ trong văn bản về từ gốc.
- Loại bỏ các từ dừng trong tiếng Việt.

Tiếp theo, chúng tôi tiến hành xử lý bộ dữ liệu. Ở đây chúng tôi sẽ áp dụng thuật toán Tokenziner đây là một nhánh con trong tập xử lý ngôn ngữ tự nhiên. Tokenziner cho phép ta vectơ hóa một kho ngữ liệu văn bản, bằng cách biến mỗi văn bản thành một chuỗi các số nguyên hoặc thành một vectơ trong đó hệ số cho mỗi mã thông báo có thể là nhị phân, dựa trên số từ, dựa trên tf-idf ...

4.2 Huấn luyện mô hình

Sơ đồ huấn luyện:



Hình 4.2: Mô hình huấn luyện

Theo sơ đồ trên, chúng tôi sử dụng đầu vào của mô hình học có giám sát LSTMs (Long short-term memory) là các tập tin đã gán nhãn, chứa các đoạn văn bản đã được xử lý tách từ bằng công cụ Tokenizer và mô hình Word2Vector.

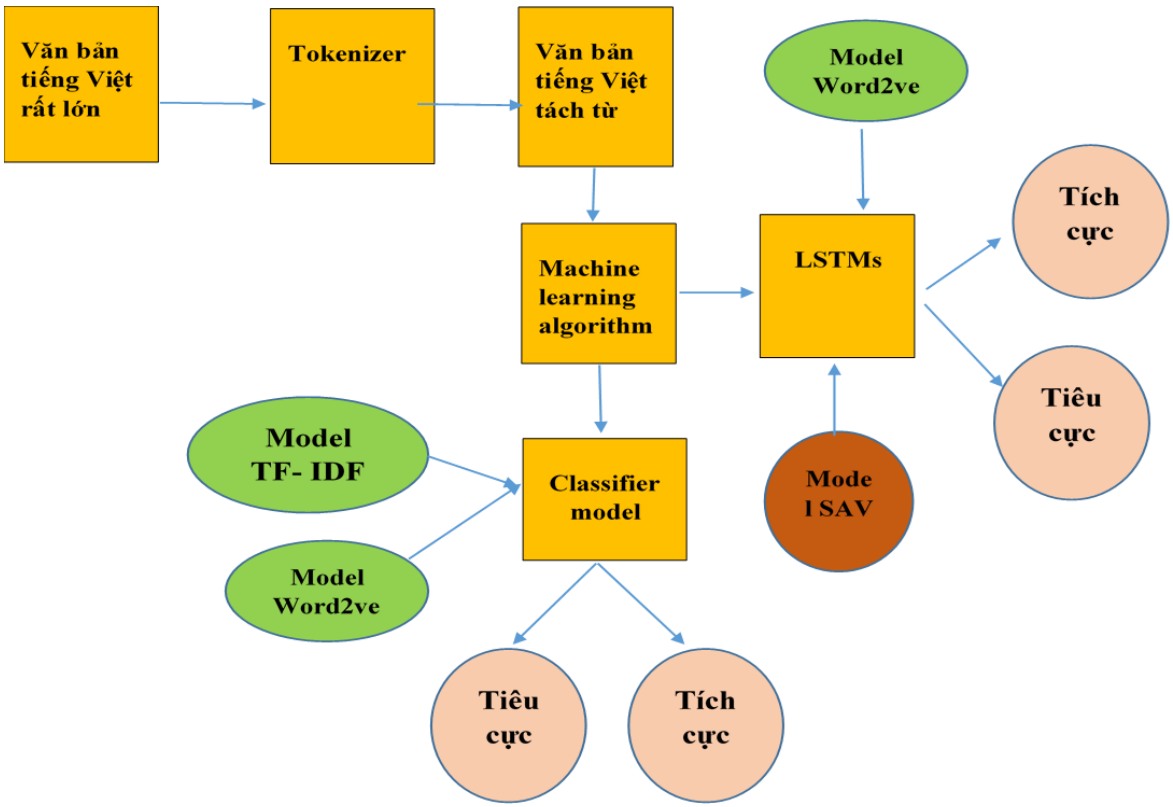
Mô hình Word2Vector là kết quả của quá trình huấn luyện nông dựa trên mô hình Bags of words và TF-IDF để vector hóa từ, hay nói cách khác là đưa từ vào không gian vector

Kết quả của quá trình huấn luyện ta thu được:

Xây dựng được mô hình phân lớp để khi có dữ liệu mới thì có thể xác định dữ liệu đó được phân lớp nào.

Một bộ trọng số của mạng nơron LSTMs [28] được lưu xuống file cùng với các siêu tham số cấu hình mạng LSTMs mà chúng tôi đã thiết lập. Hai tập tin này sẽ được tải vào mạng LSTMs để kiểm tra, vận hành hoặc có thể tiếp tục huấn luyện sau này.

Sơ đồ kiểm tra:



Hình 4.3: Mô hình kiểm tra

Ở giai đoạn kiểm tra:

- Mô hình LSTMs [28] sẻ tải lên các file cấu hình và file lưu bộ trọng số của mạng nơron. Đồng thời chúng tôi sử dụng đến mô hình Word2Vector và mô hình TF-IDF với vai trò là hệ tri thức từ vựng.
- Mô hình Classifier: dữ liệu ở tập kiểm tra được đưa vào mô hình để tiến hành phân lớp.

Trong quá trình kiểm tra, chúng tôi đưa vào bộ dữ liệu bao gồm các tập tin chứa các đoạn văn được gán nhãn đã tách từ bằng công cụ Tokenizer trước đó. Kết quả phân lớp đầu ra sẽ được ghi nhận lại để so sánh với nhãn mong đợi ban đầu của dữ liệu, từ

đó cho chúng tôi thu được kết quả độ chính xác của mô hình.

Nếu sau quá trình kiểm tra, độ chính xác của mô hình đạt được ở một mức độ chấp nhận được thì ta sử dụng kết quả mô hình này vào vận hành thực tế.

4.3 Thực nghiệm và đánh giá kết quả

Sau giai đoạn tiền xử lý, dữ liệu sẽ được thực thi theo 2 mô hình vectơ hóa Word2vec và TF-IDF được giới thiệu ở trên và phân lớp theo 3 phương pháp phân lớp khác nhau. Mỗi lần chạy, chúng tôi sẽ tiến hành đánh giá hiệu suất, tốc độ huấn luyện cùng với tính chính xác và đưa ra hướng điều chỉnh thích hợp cho chúng trong các nghiên cứu sau.

Toàn bộ quá trình chạy thực nghiệm được tiến hành trên cấu hành máy và IDE với cấu hình như sau:

- Mã máy: HP Elitebook 2540p
- CPU: Core i7-640LM
- SSD: 120GB
- RAM 6GB, DDR3 1333Mhz (PC3-10666)
- Ngôn ngữ : Python
- Thực thi: https://colab.research.google.com/drive

Bảng 4.2: Kết hợp mô hình vector hóa dữ liệu với các phương pháp phân lớp

Tên	Mô hình vector hóa	Phương pháp phân lớp
1	BoW	Logistic Regression
2	BoW	Linear SVM
3	BoW	Naive Bayes
4	TF-IDF	Logistic Regression
5	TF-IDF	Linear SVM
6	TF-IDF	Naive Bayes
7	CNN	Tensorflow

Thực nghiệm để phân lớp đánh giá [31]

Đầu tiên, chúng tôi phân loại các văn bản tiếng Việt để nhận biết các văn bản có tính chủ quan và tính khách quan của văn bản. Cụ thể là, khi tiến hành thực nghiệm phân lớp đánh giá chủ quan và đánh giá khách quan với số lần lặp tối đa 500 lần, kết quả thu được như sau:

Đối với phương pháp hồi quy logistic, chúng tôi sử dụng mô hình vector hóa BOW và TF-IDF để xây dựng tập huấn luyện thu được : đối với mô hình BOW độ chính xác là 0,6998, đối với mô hình TF-IDF độ chính xác là 0,7056

```
#Accuracy score for bag of words
lr_bow_score=accuracy_score(test_sentiments,lr_bow_predict)
print("lr_bow_score :",lr_bow_score)
#Accuracy score for tfidf features
lr_tfidf_score=accuracy_score(test_sentiments,lr_tfidf_predict)
print("lr_tfidf_score :",lr_tfidf_score)
```

```
lr_bow_score : 0.6998841251448435
lr_tfidf_score : 0.7056778679026651
```

Hình 4.4: Điểm quyết định cho phương pháp Logistic Regression

Sau đó, chúng tôi tiến hành xây đựng báo cáo trên tập dữ liệu kiểm tra về các chỉ số phân loại chính.

```
                precision    recall  f1-score   support

      Tích cực       0.74      0.63      0.68       441
      Tiêu cực       0.67      0.77      0.72       422

      accuracy                           0.70       863
     macro avg       0.70      0.70      0.70       863
  weighted avg       0.71      0.70      0.70       863

                precision    recall  f1-score   support

      Tích cực       0.76      0.61      0.68       441
      Tiêu cực       0.67      0.80      0.73       422

      accuracy                           0.71       863
     macro avg       0.71      0.71      0.70       863
  weighted avg       0.72      0.71      0.70       863
```

Hình 4.5: Báo cáo trên tập dữ liệu kiểm tra với PP Logistic Regression

Kết quả thu được: đối với điểm precision lớp 'Tích cực' mô hình BOW thấp hơn mô hình TF-IDF lần lượt là 0,74 và 0,76, lớp 'Tiêu cực' thì tương đồng nhau. Đối với điểm Recall, ở mô hình BOW số điểm có nhãn là 'Tiêu cực' được mô hình nhận ra cao hơn nhãn ' Tích cực' lần lượt là 0,63 và 0,77, tương tự với mô hình TF-IDF là 0,61 và 0,8. Đối với điểm f1-score, ở cả 2 mô hình đều khá tương đồng nhau lần lượt lá 0,70 và 0,71.

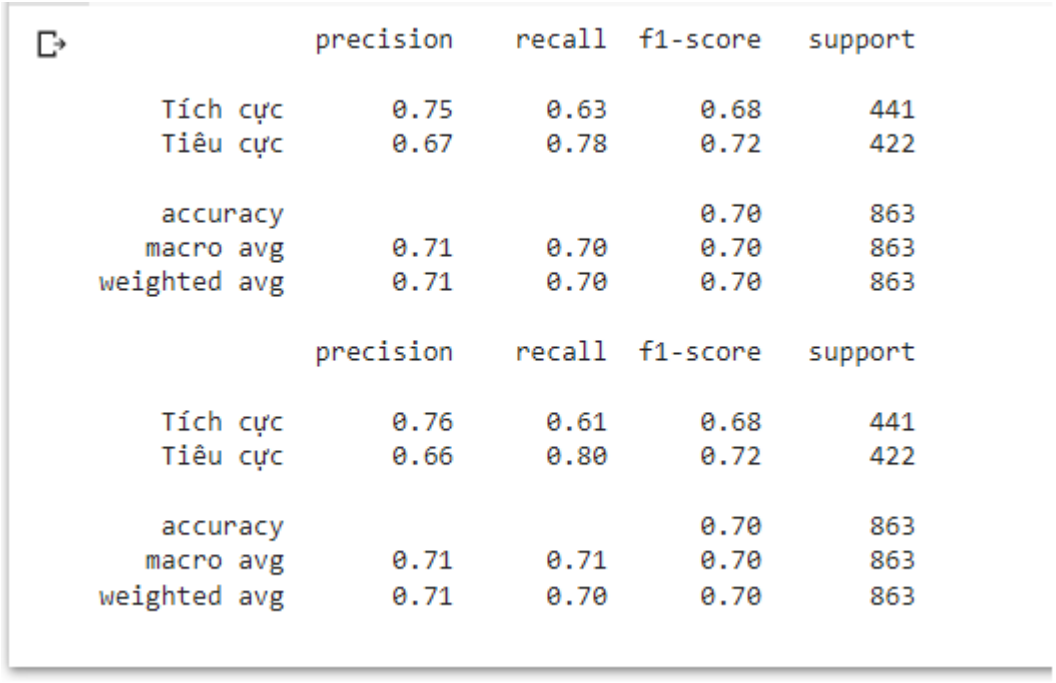
Đối với phương pháp Linear SVM, điểm chính xác khi xây dựng tập huấn luyện của 2 mô hình tương đồng nhau

```
#Accuracy score for bag of words
svm_bow_score=accuracy_score(test_sentiments,svm_bow_predict)
print("svm_bow_score :",svm_bow_score)
#Accuracy score for tfidf features
svm_tfidf_score=accuracy_score(test_sentiments,svm_tfidf_predict)
print("svm_tfidf_score :",svm_tfidf_score)
```

```
svm_bow_score : 0.7022016222479722
svm_tfidf_score : 0.7033603707995365
```

Hình 4.6: Điểm quyết định cho phương pháp Linear SVM

Kết quả báo cáo phân lớp của 2 mô hình



	precision	recall	f1-score	support
Tích cực	0.75	0.63	0.68	441
Tiêu cực	0.67	0.78	0.72	422
accuracy			0.70	863
macro avg	0.71	0.70	0.70	863
weighted avg	0.71	0.70	0.70	863

	precision	recall	f1-score	support
Tích cực	0.76	0.61	0.68	441
Tiêu cực	0.66	0.80	0.72	422
accuracy			0.70	863
macro avg	0.71	0.71	0.70	863
weighted avg	0.71	0.70	0.70	863

Hình 4.7: Báo cáo trên tập dữ liệu kiểm tra với phương pháp Linear SVM

Theo kết quả thu được thực hiện trên tập dữ dữ liệu kiểm tra là tương đương nhau trên cả 2 mô hình: các điểm accuracy, precision, recall , f1-score .

Đối với phương pháp Naïve Bayes điểm chính xác khi xây dựng tập huấn luyện là tương đương nhau

```
[ ]  #Accuracy score for bag of words
     mnb_bow_score=accuracy_score(test_sentiments,mnb_bow_predict)
     print("mnb_bow_score :",mnb_bow_score)
     #Accuracy score for tfidf features
     mnb_tfidf_score=accuracy_score(test_sentiments,mnb_tfidf_predict)
     print("mnb_tfidf_score :",mnb_tfidf_score)

     mnb_bow_score : 0.712630359212051
     mnb_tfidf_score : 0.7103128621089224
```

Hình 4.8: Điểm quyết định cho phương pháp Naive Bayes

Kết quả báo cáo phân lớp cho tập dữ liệu kiểm tra

	precision	recall	f1-score	support
Tích cực	0.78	0.61	0.69	441
Tiêu cực	0.67	0.82	0.74	422
accuracy			0.71	863
macro avg	0.72	0.71	0.71	863
weighted avg	0.72	0.71	0.71	863

	precision	recall	f1-score	support
Tích cực	0.78	0.61	0.68	441
Tiêu cực	0.67	0.82	0.73	422
accuracy			0.71	863
macro avg	0.72	0.71	0.71	863
weighted avg	0.72	0.71	0.71	863

Hình 4.9: Báo cáo trên tập dữ liệu kiểm tra với phương pháp Naive Bayes

Từ kết quả cho thấy các điểm precision, recall, cho cả 2 mô hình BOW và TF-IDF là bằng nhau, riêng điểm f1-score ở các nhãn của mô hình BOW cao hơn 0,01 so với mô hình TF-IDF

Đầu tiên ta có các quy ước sau: số điểm true Tích cực (TP), những điểm được phân loại là Tích cực (FP), những điểm thực sự là Tích cực (FN)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2\frac{\text{Pr}ecison * \text{Re}call}{\text{Pr}ecison + \text{Re}call}$$

Kết quả sau khi thực nghiệm với tập dữ liệu:

Bảng 4.3: Hiệu suất của các phương pháp phân lớp cảm xúc (đo bằng F1)

Tên	Tích cực			Tiêu cực			Average F1
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	
1	74	63	68	67	77	72	70
2	75	63	68	67	78	72	70
3	78	61	69	67	82	74	71
4	76	61	68	67	80	73	70
5	76	61	68	66	80	72	70
6	78	61	68	67	82	73	71

Ngoài ra, trong luận văn này chúng tôi còn thực nghiệm trên mạng nơ ron nhân tạo với phương pháp Tensorflow [26] được sử dụng . Chúng tôi thực hiện huấn luyện mô hình với epochs=100.

```
Epoch 1/100
102/102 [==============================] - 1s 8ms/step - loss: nan - accuracy: 0.4991 - val_loss: nan - val_accuracy: 0.5055
Epoch 2/100
102/102 [==============================] - 1s 8ms/step - loss: nan - accuracy: 0.4991 - val_loss: nan - val_accuracy: 0.5055
Epoch 3/100
 97/102 [===========================>..] - ETA: 0s - loss: nan - accuracy: 0.4968Restoring model weights from the end of the best epoch: 1.
102/102 [==============================] - 1s 7ms/step - loss: nan - accuracy: 0.4991 - val_loss: nan - val_accuracy: 0.5055
Epoch 00003: early stopping
<keras.callbacks.History at 0x7f10eb2d1c90>
```

Hình 4.10: Kết quả huấn luyện với phương pháp Tensorflow

Sau khi chạy huấn luyện cho mô hình, ta quan sát thấy độ chính xác đạt mức trung bình chỉ 50,55% .

Sau khi xây dựng mô hình huấn luyện, chúng tôi tiến hành thực nghiệm trên tập dữ liệu kiểm tra

```
loss, accuracy = model.evaluate(test_data,test_label)

print(f"Accuracy : {accuracy}")
print(f"Loss     : {loss}")
```

```
26/26 [==============================] - 0s 2ms/step - loss: nan - accuracy: 0.5055
Accuracy : 0.5055350661277771
Loss     : nan
```

Hình 4.11: Kết quả trên tập dữ liệu kiểm tra với phương pháp Tensorflow

Kết quả thu được của mô hình với độ chính xác 50,55%

KẾT LUẬN VÀ KIẾN NGHỊ

1. Các kết quả đạt được của luận văn

Sau một thời gian tìm hiểu và nghiên cứu, chúng tôi đã áp dụng mô hình giải quyết bài toán gồm các bước: Tiền xử lý dữ liệu, vector hóa dữ liệu và phân loại cảm xúc bằng mô hình nhận diện cảm xúc sử dụng học sâu đã đạt được kết quả khả quan. Sau khi huấn luyện và kiểm tra trên cùng một tập dữ liệu ban đầu thì phương pháp xử lý dữ liệu TF-IDF và BoW kết hợp với phương pháp phân lớp Naïve Bayes đã cho độ chính xác 71% là tốt nhất

Để làm được điều đó, chúng tôi đã hoàn tất những việc như sau:

- Tìm hiểu về các đặc điểm của ngôn ngữ tiếng Việt, về xử lý ngôn ngữ tự nhiên và xử lý ngôn ngữ tiếng Việt. Tìm hiểu, phân tích và xây dựng thành công mô hình giải quyết bài toán phân lớp cảm xúc người dùng với định tính "Xác định tính tích cực – tiêu cực của văn bản".
- Nghiên cứu và áp dụng phương pháp vector hóa dữ liệu Word2Vec, TF-IDF và CNN.
- Nghiên cứu các phương pháp tiền xử lý tiếng Việt nhằm cải thiện hiệu suất khi tiến hành huấn luyện.
- Nghiên cứu và áp dụng các phương pháp phân lớp và kết hợp với ba phương pháp xử lý văn bản tiếng Việt kể trên để chọn ra được phương pháp máy học tốt nhất cho phân lớp cảm xúc người dùng.
- Áp dụng kết hợp các phương pháp xử lý văn bản tiếng Việt và các thuật toán phân lớp để đánh giá trên bộ dữ liệu
- Xây dựng và gán nhãn cho bộ dữ liệu (Dataset)

2. Nhận xét, đề xuất, khuyến nghị

2.1 Nhận xét

Tất cả mô hình kết hợp với các phương pháp xử lý dữ liệu văn bản tiếng Việt

đã sử dụng thì đều cần một lượng lớn dữ liệu đầu vào. Nếu dữ liệu ít hoặc thiếu cân bằng, độ chính xác khi tiến hành các phương pháp phân lớp sẽ bị ảnh hưởng và không ổn định.

2.2 Đề xuất

Luận văn có thể áp dụng thêm một số phương pháp tiền xử lý dữ liệu và áp dụng thêm các thuật toán phân lớp hay tối ưu các thuật toán phân lớp hiện có để mô hình giải quyết bài toán nhận diên cảm xúc trong văn bản tiếng Việt được tốt hơn.

2.3 Kiến nghị

Phân tích cảm xúc nói riêng và xử lý ngôn ngữ tự nhiên nói chung là một trong những nhánh nghiên cứu phức tạp nhưng lợi ích mà nó mang lại trong cuộc Cánh mạng công nghiệp 4.0 tai Việt Nam là rất lớn. Nếu đề tài được đầu tư và phát triển tốt có thể được áp dụng rộng rãi trong các lĩnh vực như giáo dục, y tế, kinh doanh, giải trí, ….. Vì tất cả các lĩnh vực này đều cần một mô hình để xây dựng phân lớp và nhận diện cảm xúc của người dùng hiệu quả như đề tài.

3. Hướng nghiên cứu tiếp theo

Trong những nghiên cứu tiếp theo, chúng tôi sẽ tiếp tục nghiên cứu để cải thiện hiệu suất phân loại và nhận diện cảm xúc trong văn bản tiếng Việt. Kể tiếp, chúng tôi cũng sẽ tiến hành thu thập thêm dữ liệu thực nghiệm để ổn định hiệu suất của mô hình. Cùng với đó, chúng tôi cũng tiến hành thực nghiệm trên bộ dữ liệu phong phú hơn về số lượng, khía cạnh, ý kiến của người dùng.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1]. N.V. Xtankêvich, Các loại hình ngôn ngữ, Nhà xuất bản Đại học và trung học chuyên nghiệp Hà Nội.

[2]. Giáo trình "Xử lý ngôn ngữ tự nhiên", Đinh Điền, NXB Đại học Quốc gia – HCM, năm 2006.

[3]. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.

[4]. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21

[5]. Binh Thanh Kieu, Son Bao Pham. "Sentiment Analysis for Vietnamese." Proc.

KSE'10, tr. 152-157, 2010.

[6]. Montoyo Andre´s, Martı´nez-Barco Patricio, Balahur Alexandra. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675–9.

[7]. Zhang Wenhao, Hua Xu, Wan Wei. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Syst Appl 2012;39:10283–91.

[8]. Socher, C. D. Manning, and A. Y. Ng. 2010. Learning continuous phrase representations and syntactic pars- ing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Fea- ture Learning Workshop.*

[9]. Daniel Jurafsky and James H. Martin. *"Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition."* 1st edition, 2000.

[10]. Binh Thanh Kieu, Son Bao Pham. *"Sentiment Analysis for Vietnamese."* Proc.

KSE'10, tr. 152-157, 2010.

[11]. Nguyen Thi Duyen, Ngo Xuan Bach, and Tu Minh Phuong, *"An Empirical*

Study on Sentiment Analysis for Vietnamese." Proc. ATC'14, tr. 309-314, 2014.

[12]. Bo Pang và Lillian Lee. *"A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts."* Proc. ACL'04, tr. 271- 278, 2008.

[13]. Le Anh Cuong, Nguyen Thi Minh Huyen, and Nguyen Viet Hung. *"Report on Sentiment Analysis Evaluation Campaign: Data and Systems."* Proc. VLSP 2016, 2016.

[14]. Vi Ngo Van, Minh Hoang Van, and Tam Nguyen Thanh. *"Sentiment Analysis for Vietnamese using Support Vector Machines with application to Facebook comments."* Proc. VLSP 2016.

[15]. Thien Khai Tran and Tuoi Thi Phan. *"Computing Sentiment Scores of Verb Phrases for Vietnamese."* Proc. ROCLING'16, tr. 204-213, 2016.

[16]. Liu, B. (2012). Sentiment analysis and opinion mining. New York, NY: Morgan & Claypool Publishers.

[17]. Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information Fusion, 36(2017), 10-25.

[18]. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1/2), 1-135.

[19]. Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: Review and prospect. International Journal of Computer Applications, 115(9), 31-41.

[20]. Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. MIS Quarterly, 35(3), 553-572.

[21]. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *"Efficient Estimation of Word Representations in Vector Space."* CoRR 2013

[22]. http://www.vietlex.com, truy cập 18-06-2018.

[23]. https://streetcodevn.com, truy cập 18-06-2018.

[24]. https://trituenhantao.io

[25]. https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/

[26]. https://studymachineleaning.com

[27]. https://text.123docz.net

[28]. https://www.tensorflow.org/text/guide/word_embeddings

[29]. https://streetcodevn.com

[30]. https://www.kaggle.com/ nitin194/ twitter-sentiment-analysis-word2vec

[31]. https://www.kaggle.com/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews

[32]. http://123doc.org

[33]. https://text.123doc.org

[34]. https://vi.wikipedia.org/wiki/Học sâu

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung luận văn qua phần mềm DoIT một cách trung thực và đạt kết quả mức độ tương đồng 13% toàn bộ nội dung luận văn. Bản luận văn kiểm tra qua phần mềm là bản cứng luận văn đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của học viện.

TP.HCM, Ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Thanh Huy

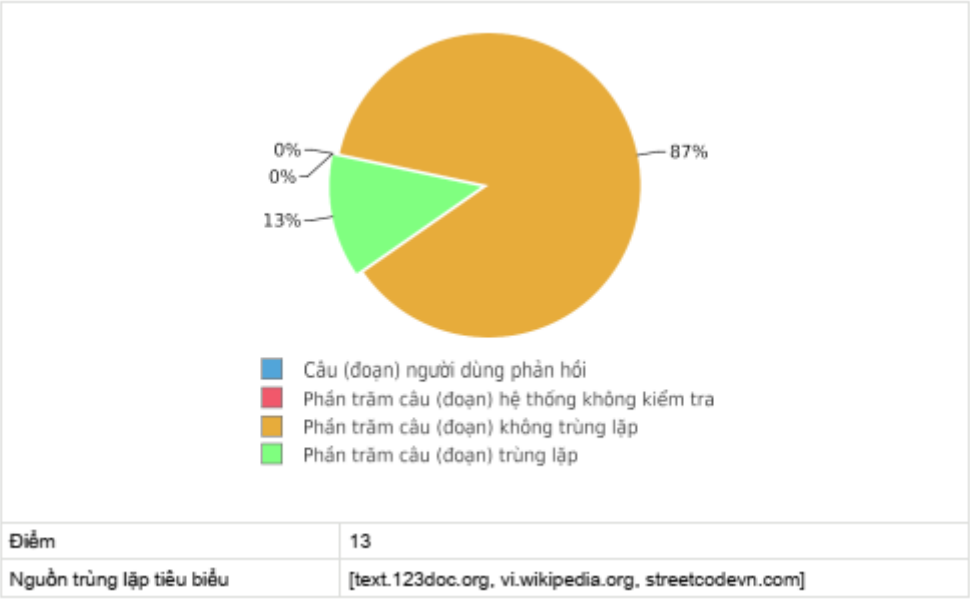
Hệ thống hỗ trợ nâng cao chất lượng tài liệu

KẾT QUẢ KIỂM TRA TRÙNG LẶP TÀI LIỆU

THÔNG TIN TÀI LIỆU

Tác giả	Nguyễn Thanh Huy
Tên tài liệu	NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT BẰNG MÔ HÌNH MÁY HỌC
Thời gian kiểm tra	25-01-2022, 14:15:36
Thời gian tạo báo cáo	25-01-2022, 14:20:51

KẾT QUẢ KIỂM TRA TRÙNG LẶP



Điểm	13
Nguồn trùng lặp tiêu biểu	[text.123doc.org, vi.wikipedia.org, streetcodevn.com]

Học viên thực hiện luận văn — Người hướng dẫn khoa học

Nguyễn Thanh Huy — **PGS.TS Nguyễn Tuấn Đăng**