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

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

TPHCM - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS NGUYỄN TUẤN ĐĂNG

Phản biện 1: ……

Phản biện 2: ……

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

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

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.

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. Bài toán phân tích cảm xúc có một số phương pháp [7] giải quyết như sau:

- Phương pháp thủ công (dò từ khóa)

- Phương pháp Deep Learning Neural Network [8]:
- Phương pháp kết hợp rule-based và corpus-based [8]

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.

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 dụng 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

Một số vấn đề khái quát về ngôn ngữ tự nhiên

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
- ❖ Đặc điểm từ vựng [1]
- ❖ Đặc điểm ngữ pháp

1.3 Xử lý ngôn ngữ tự nhiên

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.

CHƯƠNG 2
CƠ SỞ LÝ THUYẾT

2.1 Các mô hình mạng neuron dùng trong học sâu

❖ 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].
- Mạng neuron sâu (DNN-Deep neural Network) [34].
- 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

❖ Phương thức hoạt động

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.
- 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.

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.

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]

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

c. Mô hình nghiên cứu tổng quan

Trong nghiên cứu này, trước tiên chúng tôi tiến hành thu thập dữ liệu thô từ 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 ảnh trên 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

a. ***Tách từ*** [27]

b. ***Chuẩn hóa từ*** [24]

c. ***Loại bỏ stopword*** [2]

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

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 chử 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]

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.

❖ 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(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. 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

$$tfidf_i = tf_i * idf_i$$

b. *Phương pháp Neural Embedding*

❖ Word2vec

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

❖ 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.

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.

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. Tù 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,...

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. Trong đó tập huấn luyện chiếm 80%, tập kiểm tra chiếm 20%.

4.1.3 *Tiền xử lý dữ liệu* [31]

Đối với luận văn này, dữ liệu input đầu vào là các phản hồi, đánh giá của khách hàng về sản

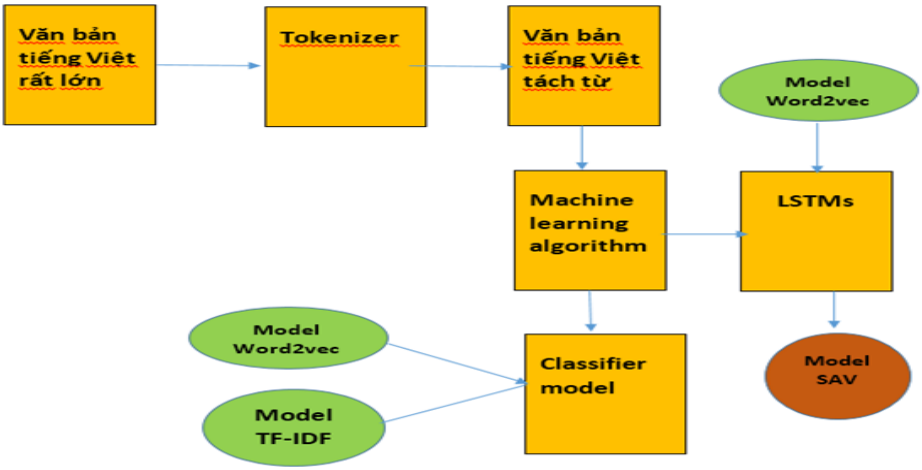
phẩm. Dữ liệu thường không chuẩn, vì thế ta phải tiến hành xử lý dữ liệu:

- 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

Ở đây chúng tôi sẽ áp dụng thuật toán Tokenziner để vec tơ hóa kho ngữ liệu văn bản

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.

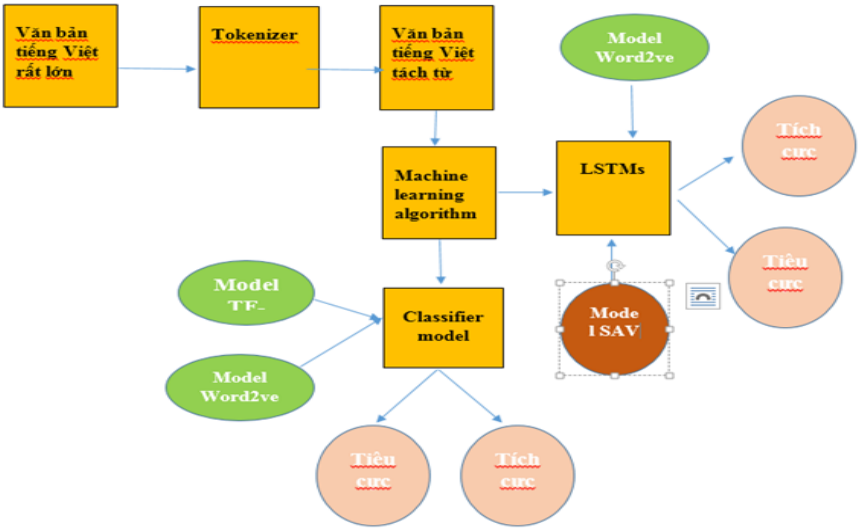
Mô hình Word2Vector là kết quả của quá trình huấn luyện nông dựa trên mô hình Bags of words và TF-IDF để vector hóa từ, hay nói cách khác là đưa từ vào không gian vector

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 lieu 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 nơ ron. 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 đua 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 sẻ được ghi nhận lại để so sánh với nhãn mong đợi ban đầu của dữ liệu, từ đó cho chúng tôi kết quả độ chính xác của mô hình.

4.3 Thực nghiệm và đánh giá kết quả

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

Các thuật toán được sử dụng:

Bảng 4.1. 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]

Kết quả sau khi thực nghiệm với tập dữ liệu:

Bảng 4.2. 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. Kết quả thu được của mô hình với độ chính xác 50,55% và độ mất mát là Nan

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 vectơ hóa dữ liệu TF-IDF kết hợp với phương pháp phân lớp Naïve Bayes đã cho hiệu suất 71%(tính theo F1) 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 mô hình vector hóa dữ liệu 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 các mô hình kết hợp với các phương pháp xử lý dữ liệu 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 them một số phương pháp tiền xử lý dữ liệu và áp dụng them 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