USE NATURAL LANGUAGE PROCESSING TECHNIQUE FOR FEATURE SELECTION IN SENTIMENT ANALYSIS ON VIETNAMESE

SỬ DỤNG PHƯƠNG PHÁP XỬ LÝ NGÔN NGỮ TỰ NHIÊN ĐỂ LỰA CHỌN ĐẶC TRƯNG TRONG PHÂN TÍCH TÌNH CẢM

Đậu Mạnh Hoàn
Trường Đại học Quảng Bình

ABSTRACT: *Sentiment analysis is a specialized form of text classification, where a document is categorized to automatically predict its sentiment polarity as positive, negative, or neutral. The sentiment analysis process addresses three fundamental issues: feature selection, identifying sentiment words, and sentiment classification. This study focuses on feature selection, one of the critical steps in the sentiment analysis process. It provides a detailed analysis of various aspects of the natural language processing approach applied to different languages, proposing its application to Vietnamese. Based on the unique characteristics of the Vietnamese language, the study also suggests several specific approaches to improve the efficiency of feature selection and sentiment analysis in Vietnamese.*

Keywords: *Feature selection, feature extraction, sentiment analysis, opinion mining.*

TÓM TẮT: *Phân tích tình cảm là một dạng đặc biệt của phân loại văn bản, trong đó một tài liệu được phân loại để dự đoán tình cảm tự động phân cực là tích cực, tiêu cực hay trung tính. Quá trình phân tích tình cảm cần xử lý ba vấn đề cơ bản đó là lựa chọn đặc trưng, xác định các từ tình cảm và phân loại tình cảm. Trong nghiên cứu này đề cập đến vấn đề lựa chọn đặc trưng, một trong các bước quan trọng của quá trình phân tích tình cảm. Chúng tôi phân tích rõ các khía cạnh trong cách tiếp cận theo hướng xử lý ngôn ngữ tự nhiên trên các ngôn ngữ khác nhau từ đó đề xuất phân loại theo hướng tiếp cận này trên tiếng Việt, ngoài ra từ đặc điểm cơ bản của tiếng Việt, chúng tôi đề xuất một số cách tiếp cận chi tiết để nâng cao hiệu quả của quá trình lựa chọn đặc trưng và phân tích tình cảm trong tiếng Việt.*

Từ khóa: *Lựa chọn đặc trưng, trích xuất đặc trưng, phân tích tình cảm, khai phá ý kiến.*

1. TỔNG QUAN

Công nghệ thông tin ra đời và phát triển, đặc biệt là sự bùng nổ của Internet, các công nghệ và ứng dụng trên nền tảng của Web thì con người ngày càng có xu hướng chia sẻ ý kiến, quan điểm trực tuyến của mình dựa trên văn bản thông qua nhiều kênh khác nhau, chẳng hạn như đánh giá sản phẩm trực tuyến, thể hiện cảm xúc trên mạng xã hội. Các nhà quản lý đã sử dụng hoạt động này để nắm bắt, khai thác các ý kiến, từ đó có thể phân tích suy nghĩ của con người về một sản phẩm, thực thể, chủ đề, dịch vụ hay một nội dung nào đó liên quan đến công việc. Khai phá ý kiến (opinion mining) hay phân tích tình cảm (sentiment analysis) là một kỹ thuật kết hợp giữa tính toán ngôn ngữ học và xử lý ngôn ngữ tự nhiên để rút ra những ý kiến như vậy [1].

Phân tích tình cảm là nhiệm vụ trích xuất thông tin và phân tích tình cảm của con người đối với các thực thể nhất định từ các tài liệu văn bản với mục đích là để hiểu được cảm xúc cá nhân từ nội dung của người sử

dụng thông qua việc đánh giá các nội dung, các sản phẩm trên các ứng dụng và môi trường web. Đây là bài toán đã được nhiều nhà khoa học quan tâm nghiên cứu và ứng dụng để phục vụ các chiến lược kinh doanh, các hệ hỗ trợ ra quyết định, hỗ trợ các dịch vụ và các công việc khác. Ví dụ, cho một câu đánh giá "Khách sạn này gần biển, vị trí đẹp, giá phòng vừa phải" với nhiệm vụ phân tích tình cảm thì kết quả đánh giá sẽ là "tích cực". Ngày nay, khi công cuộc chuyển đổi số diễn ra một cách mạnh mẽ, mọi hoạt động sống của con người luôn gắn với môi trường số thì nhiệm vụ phân tích tình cảm hay khai phá ý kiến người dùng ngày càng trở nên quan trọng hơn.

Trong phân tích tình cảm, mục tiêu là chỉ định các phân cực tình cảm dựa trên nội dung văn bản. Theo B. Liu [1] một ý kiến hay tình cảm được định nghĩa bằng một bộ gồm 5 thành phần (E_i, A_{ij}, S_{ijkl}, H_k, T_l), trong đó:

- E_i: có thể là tên thực thể, chủ đề, sản phẩm, dịch vụ hay một nội dung nào đó được đánh giá hay cho ý kiến.

- A_{ij} là khía cạnh của thực thể đánh giá E_i.

- S_{ijkl} là ý kiến cảm xúc về khía cạnh A_{ij} của thực thể đánh giá E_i xác định bởi chủ thể H_k tại thời điểm T_l, ý kiến đó có thể là tích cực, tiêu cực hay trung lập.

- H_k là chủ thể thể hiện ý kiến.

- T_l là thời gian thể hiện ý kiến của H_k.

Dựa trên định nghĩa đó thì việc phân tích tình cảm hay quan điểm sẽ liên quan đến việc nắm bắt và phân tích các thành phần của bộ cảm xúc, tuỳ vào từng trường hợp cụ thể để khai thác một hay một số hay toàn bộ thành phần của bộ cảm xúc.

Phân tích tình cảm là một trường hợp đặc biệt của phân loại văn bản thuộc lĩnh vực nghiên cứu khai phá văn bản, xử lý ngôn ngữ tự nhiên (Natural Language Processing: NLP) và tính toán thông minh. Phân tích tình cảm được tiếp cận với ba dạng cơ bản đó là phân tích tình cảm xác định trên mức độ từ, cụm từ và câu; phân tích cảm xúc mức văn bản và phân tích tình cảm mức khía cạnh. Trong quá trình phân tích tình cảm cần xử lý ba vấn đề cơ bản đó là lựa chọn đặc trưng, xác định các từ tình cảm và phân loại tình cảm [2].

Phân tích tình cảm đã được thực hiện rộng rãi trên nhiều ngôn ngữ khác nhau như tiếng Nhật, Pháp, Romania, Trung Quốc và phổ biến nhất là tiếng Anh. Đối với tiếng Việt, trước đây bài toán này chưa được nghiên cứu rộng rãi do hạn chế về ngữ liệu và thực tế vấn đề phân tích tình cảm người sử dụng từ các trang web kinh doanh trực tuyến của Việt Nam chưa được chú ý nhiều, mặt khác xử lý ngôn ngữ tự nhiên tiếng Việt là một bài toán phức tạp và phải thực hiện qua nhiều giai đoạn. Hiện nay, bài toán phân tích tình cảm trong tiếng Việt đã được nhiều tác giả nghiên cứu và đạt được các kết quả phong phú và đa dạng. Điển hình như các tác giả Dang Van Thin và các cộng sự [3] đã chứng minh hiệu quả của mô hình PhoBERT đơn ngữ và đa ngữ trên tập dữ liệu SemEval-2016 bằng các ngôn ngữ khác nhau trong đó có tập dữ liệu tiếng Việt cho lĩnh vực nhà hàng và khách sạn, kết quả cho thấy cao hơn các mô hình khác. Các tác giả [4] phân tích tình cảm mức văn bản bằng việc biểu diễn các đặc trưng và ảnh hưởng của nó đến hiệu quả phân tích tình cảm, kết quả nghiên cứu đã chỉ ra rằng việc lựa chọn các đặc trưng phù hợp sẽ mang lại hiệu quả cao trong phân tích cảm xúc đối với phân lớp nhị phân hay phân lớp nhiều cấp độ.

Trong nghiên cứu này chúng tôi tiến

hành tìm hiểu vấn đề lựa chọn đặc trưng bằng nhóm phương pháp xử lý ngôn ngữ tự nhiên và tiến hành đề xuất một số cách tiếp cận trong tiếng Việt thông qua quá trình nghiên cứu của mình.

2. PHÂN TÍCH TÌNH CẢM

Theo [1] phân tích tình cảm hay khai phá ý kiến được định nghĩa là quá trình nghiên cứu phân tích ý kiến, bình luận, suy nghĩ, thái độ và đánh giá về cảm xúc của con người đối với các thực thể, các sản phẩm, dịch vụ, các tổ chức, chính trị, các vấn đề hiện tại, sự kiện và con người, cũng như các đặc điểm tồn tại trong các thực thể đó. Tiến trình phân tích tình cảm được thực hiện qua 6 bước cơ bản sau [5]:

Bước 1. Tiền xử lý văn bản (Document preprocessing)

Tiền xử lý văn bản là quá trình loại bỏ các từ lóng không có nghĩa hoặc không cần thiết hoặc gây rối văn bản từ đó giúp cải thiện và nâng cao độ chính xác của việc tìm kiếm các từ quan trọng trong mỗi câu đồng thời sẽ xác định các từ đặc trưng chính xác hơn. Các phương pháp để tiền xử lý dữ liệu dựa trên văn bản bao gồm loại bỏ các từ dừng, phân tách, tách câu, từ gốc, từ viết sai chính tả và gắn thẻ loại từ (PoST).

Bước 2. Chuyển đổi dữ liệu (Data transformation)

Là quá trình chuyển đổi dữ liệu văn bản thành các vectơ đặc trưng vì việc phân tích tình cảm dựa vào các đặc trưng này. Tất cả các văn bản cần phân tích không thể tiến hành phân loại một cách trực tiếp nên các văn bản phải được chuyển đổi thành định dạng mà máy tính có thể xác định được mà phương pháp phổ biến và hiệu quả để thực hiện việc này đó là chuyển thành mô hình không gian vectơ. Mô hình này sẽ chuyển đổi một tài liệu thành một vectơ đa chiều và các đặc trưng được chọn từ tập dữ liệu là các chiều của vectơ này. Các vectơ đặc trưng biểu diễn các đối tượng dữ liệu trong không gian đặc trưng.

Bước 3. Lựa chọn đặc trưng (Feature selection)

Đặc trưng của văn bản là những hạng (một nhóm từ) trong văn bản. Sau chuyển đổi dữ liệu chúng ta thu được các vectơ đặc trưng biểu diễn các đối tượng dữ liệu trong không gian đặc trưng, nhiệm vụ tiếp theo là lựa chọn các đặc trưng phù hợp với bài toán và văn bản đang xem xét. Bước này xác định và loại bỏ các đặc trưng dư thừa và không liên quan khỏi danh sách các đặc trưng đã chọn để giảm kích thước cho không gian đặc trưng đồng thời nâng cao độ chính xác của quá trình phân loại. Lựa chọn đặc trưng là khâu quan trọng và có ý nghĩa quyết định đến hiệu quả của quá trình phân tích tình cảm. Bởi vì nếu lựa chọn tập đặc trưng lớn quá sẽ khó xác định nội dung và phân tích được văn bản, nếu chọn tập đặc trưng nhỏ quá sẽ không đủ các thông tin để phân tích văn bản. Trong khi đó, vấn đề đặt ra là chọn tập đặc trưng vừa đủ, không lớn quá cũng không nhỏ quá để biểu diễn chính xác các đặc trưng của các đối tượng trong văn bản mà người dùng bình luận. Chúng tôi sẽ trình bày chi tiết vấn đề này ở các phần tiếp theo.

Bước 4. Nhận dạng từ ngữ tình cảm (Sentiment word identification)

Nhận dạng từ ngữ tình cảm có mối quan hệ mật thiết với quá trình lựa chọn đặc trưng, từ những đặc trưng đã trích xuất được, chúng ta sử dụng chúng để nhận dạng các từ ngữ tình cảm có liên kết với một đặc trưng trong câu văn bản. Việc xác định mối quan hệ giữa đặc trưng và từ ngữ tình cảm là một quá trình quan trọng vì mối quan hệ này

trong thực tế có thể là rất phức tạp nếu câu chứa nhiều hơn một đặc trưng hoặc từ chứa tình cảm.

Bước 5. Phân loại tình cảm (Sentiment classification)

Phân tích tình cảm có thể được xem là một trường hợp đặc biệt của phân loại văn bản, trong đó các tình cảm có thể phân thành các loại tích cực, tiêu cực và trung tính. Quá trình phân loại là một quá trình xem xét cấp độ tài liệu, câu hoặc đặc điểm để xác định tài liệu, câu hoặc đặc trưng nào thể hiện tình cảm tích cực, tiêu cực hay trung tính.

Các giải thuật học máy, đặc biệt là các phương pháp học sâu [2, 3] đã được chứng minh là những giải thuật phân lớp tốt nhất hiện nay cho vấn đề phân loại văn bản. Các giải thuật học máy phù hợp với bài toán phân loại tình cảm vì chúng có khả năng đáp ứng được không gian đầu vào có số chiều rất lớn, các đặc trưng rời rạc, ít liên hệ lẫn nhau, các vectơ tài liệu là thưa và các vấn đề phân lớp trong văn bản là có thể chia cắt được.

Bước 6. Kiểm tra và Đánh giá (Testing and Evaluation)

Sau quá trình xác định mối quan hệ giữa đặc trưng và quá trình nhận dạng từ ngữ tình cảm thì chúng ta cần tiến hành đánh giá để kiểm tra độ chính xác của mối quan hệ giữa đặc trưng và nhận dạng từ ngữ tình cảm đó, đồng thời đánh giá mức độ chính xác của các loại tình cảm khác nhau đối với các bình luận tích cực, tiêu cực hay trung tính trong văn bản đang xem xét.

3. LỰA CHỌN ĐẶC TRƯNG TRONG PHÂN TÍCH TÌNH CẢM

Lựa chọn đặc trưng là bước trọng tâm và quyết định trong tiến trình phân loại tình cảm. Trong phân tích tình cảm mỗi đặc trưng là một khía cạnh mà người dùng sử dụng để nhận xét các đối tượng, sản phẩm, dịch vụ, tổ chức, sự kiện hay cá nhân. Khi xử lý dữ liệu ở dạng văn bản, các đặc trưng được biểu diễn bằng vectơ đặc trưng, là một nhóm từ hay còn gọi là phương pháp tập hợp từ.

Xét với một vector đặc trưng đầu vào ngẫu nhiên $F = \{f_1, f_2, ..., f_d\}$ và X là giá trị đầu ra có thể dự đoán từ vector đặc trưng F. Nhiệm vụ lựa chọn đặc trưng chính là việc tìm ra các đặc trưng f_i có liên quan nhất đến dự đoán giá trị X. Mỗi văn bản đưa ra phân tích có thể xem như là một tập hợp các đặc trưng. Quá trình phân tích tình cảm của người dùng sẽ dựa trên tập đặc trưng này. Trong thực tế, số đặc trưng của một văn bản là lớn và không gian các đặc trưng của tất cả các văn bản đang xem xét là rất lớn, về nguyên tắc, nó bao gồm tất cả các từ trong một ngôn ngữ. Trên thực tế, người ta không thể xem xét tất cả các từ của ngôn ngữ mà là dùng tập hợp các từ được rút ra từ một tập đủ lớn các văn bản đang xét và thường được gọi là tập ngữ liệu. Đầu tiên chúng ta thực hiện trích xuất các đặc trưng, đó là quá trình nhận diện các đặc trưng của văn bản theo nhận xét của người dùng. Ngoài ra, trích xuất đặc trưng cũng có thể thực hiện bằng cách chọn các đặc trưng đại diện từ các cụm danh từ trong câu và trích xuất các ý kiến liên quan dưới dạng thông tin. Bước tiếp theo là lựa chọn đặc trưng. Như đã phân tích ở trên, số đặc trưng của một văn bản là lớn và không gian các đặc trưng của tất cả các văn bản đang xem xét là rất lớn, do đó cần phải lựa chọn đặc trưng nhằm rút ngắn số chiều của không gian đặc trưng. Mục đích của lựa chọn đặc trưng là cải thiện độ chính xác của dự đoán hoặc giảm kích thước không gian đặc trưng bằng cách chọn một

tập hợp con đặc trưng mà không làm giảm đáng kể độ chính xác của phân tích. Quá trình lựa chọn đặc trưng là làm giảm số chiều của vector đặc trưng bằng cách bỏ đi những thành phần đặc trưng không quan trọng nhưng vẫn đảm bảo tính chính xác của nội dung văn bản. Như vậy, lựa chọn đặc trưng là quá trình lập một tập hợp các danh sách đặc trưng từ đó hệ thống phân loại và chọn ra một tập hợp con các đặc trưng tiêu biểu biểu diễn từ không gian đặc trưng gốc có giá trị tốt nhất, kết quả của việc lựa chọn đặc trưng sẽ tìm ra một tập các đặc trưng nhỏ nhất, có giá trị nhất và giúp cải thiện độ chính xác của phân loại cảm tính [6].

4. LỰA CHỌN ĐẶC TRƯNG SỬ DỤNG PHƯƠNG PHÁP XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Phân tích tình cảm là một lĩnh vực nghiên cứu của phân loại văn bản vì vậy quá trình xử lý và phân tích tình cảm phần lớn sử dụng phương pháp xử lý ngôn ngữ tự nhiên hoặc sử dụng NLP kết hợp với các phương pháp khác. Cách tiếp cận đầu tiên đó là sử dụng phương pháp túi từ (Bag of Words: BoW), đây là một trong những cách đơn giản và phổ biến nhất hiện nay để biểu diễn một câu hay một phần văn bản. Biểu diễn vectơ BoW là một vectơ đặc trưng trong đó mỗi phần tử biểu diễn một từ. Phương pháp này được gọi là "túi từ" vì tiến hành tập hợp các từ lại với nhau mà bỏ qua cấu trúc của văn bản, là một phương pháp để trích xuất các đặc điểm từ các dữ liệu văn bản. Nó tạo ra một kho từ vựng chứa tất cả các từ duy nhất có trong tất cả các dữ liệu văn bản, BoW hầu như không quan tâm đến thứ tự xuất hiện của các từ đó. Đối với phương pháp biểu diễn BoW, mỗi phần tử x_i trong vectơ $\boldsymbol{x}_j$ biểu diễn một từ duy nhất từ vốn từ vựng. Vectơ BoW đơn giản nhất là một vectơ nhị phân trong đó mỗi phần tử $x_{i,j}$ trong vectơ $\boldsymbol{x}_j$ được đặt bằng 1 nếu từ tương ứng xuất hiện trong R_j và bằng 0 nếu không, có thể xem xét tần suất của các từ bằng cách đặt mỗi phần tử $x_{i,j}$ bằng với số lần từ tương ứng xuất hiện trong R_j [7]. Tuy nhiên, khi sử dụng phương pháp này sẽ xuất hiện tình huống một số từ được sử dụng thường xuyên, trong khi một số từ lại ít được sử dụng. Điều này có thể gây ra sự thiên vị đối với một số từ nhất định trong số lượng từ. Với bản chất mỗi đặc trưng là một khía cạnh mà người dùng bình luận trong văn bản được chứa trong một nhóm câu ở định dạng văn bản. Các đặc trưng được phân loại thành hai loại chính đó là các đặc trưng là nhóm danh sách các từ trong văn bản, chúng có thể ở dạng unigram, bi-gram hoặc tri-gram và loại thứ hai đó là các thẻ PoST, được sử dụng để xác định từng từ trong văn bản như danh từ, tính từ, động từ, trạng từ, giới từ hay từ hạn định [8].

Phương pháp phổ biến thường được sử dụng để lựa chọn đặc trưng trong bài toán phân loại tình cảm là n-gram, bi-gram và tri-gram. Đây là kỹ thuật có nhiều ưu điểm mà các nhà nghiên cứu đã sử dụng để đánh giá trên nhiều miền dữ liệu khác nhau và mang lại hiệu suất cao khi sử dụng chúng để phân tích tình cảm. Do đặc điểm không gian đặc trưng luôn có kích thước lớn nên đòi hỏi phải sử dụng kỹ thuật lựa chọn đặc trưng phù hợp và n-gram chính là kỹ thuật thích hợp để giải quyết vấn đề đó khi n-gram có thể thực hiện cố định và biến đổi. Trong đó n-gram cố định là một dãy xuất hiện ở cấp độ ký tự hoặc ở cấp độ mã thông báo, trong khi n-gram biến đổi là một trích xuất mẫu biểu diễn theo ngữ cảnh cao hơn [8].

Đối với các sử dụng loại thẻ PoST (Parts of Speech Tagging), một trong

những nhiệm vụ cốt lõi trong xử lý ngôn ngữ tự nhiên là gắn thẻ từ loại, nghĩa là gán cho mỗi từ trong văn bản một danh mục ngữ pháp, chẳng hạn như danh từ, động từ, tính từ và trạng từ. Việc đánh dấu là một loại phân loại bằng cách định nghĩa và tự động gán mô tả cho các mã thông báo là các thẻ từ loại nói trên, nó thường được gọi là gắn thẻ PoST. Như vậy, gắn thẻ PoST là nhiệm vụ gắn nhãn cho từng từ trong câu bằng loại từ thích hợp của nó. Hầu hết các loại gắn thẻ PoST đều thuộc loại gắn thẻ PoST theo quy tắc, gắn thẻ PoST ngẫu nhiên và gắn thẻ dựa trên chuyển đổi. Trong quá trình phân tích tình cảm và truy xuất thông tin thì việc gắn thẻ PoST là nhiệm vụ quan trọng và cần thiết. Gắn thẻ PoST đóng vai trò là mối liên kết giữa ngôn ngữ và sự hiểu biết của máy, cho phép tạo ra các hệ thống xử lý ngôn ngữ phức tạp và đóng vai trò là nền tảng cho phân tích ngôn ở mức cao hơn. Trong phân tích tình cảm việc đánh dấu các phần của văn bản là đánh dấu ngữ pháp hoặc giải nghĩa từ loại. PoST là quá trình đánh dấu một từ trong văn bản (ngữ liệu) tương đương với một phần cụ thể của các bình luận. Trong PoST, mỗi thuật ngữ trong tài liệu hoặc trong câu được phân bổ một thẻ hoặc nhãn biểu thị vị trí của nó trong ngữ cảnh ngữ pháp. Các danh từ và tính từ trong câu được đưa ra phân tích và được xác định từ quá trình đánh dấu này, sau đó chúng được sử dụng để lựa chọn các đặc trưng và làm cơ sở để phân tích tình cảm. Trong phương pháp NLP đã được áp dụng để xác định các đặc trưng trong các câu bình luận của người sử dụng đối với các dịch vụ , đối tượng hay chủ thể. PoST được sử dụng để đánh giá từng câu và các từ trong câu được gắn thẻ và được xác định là danh từ hoặc cụm danh từ hay các từ loại khác. Các đặc trưng văn bản trong phân tích tình cảm thường được chia thành loại phổ biến và không phổ biến. Để có thể phân tích có hiệu quả thì sẽ sử dụng tập đặc trưng phổ biến bởi vì nó biểu thị các tập từ tình cảm gần nhất và khi đó sẽ phân tích tình cảm có hiệu quả hơn [9].

Một nhóm phương pháp lựa chọn các đặc trưng khác là sử dụng mô hình hóa chủ đề, là phương pháp học máy không giám sát, chức năng của nó là xác định sự kết hợp của các chủ đề trong mỗi tài liệu. Trong phương pháp này, một mô hình xác suất được tạo ra để sử dụng và phân phối các từ vựng để phát hiện ra các chủ đề từ một bộ sưu tập lớn các tài liệu văn bản. Đầu ra của phương pháp này là một nhóm từ và chính nhóm từ này dùng để xác định danh sách các đặc trưng trong phân tích tình cảm của văn bản đang xét [10].

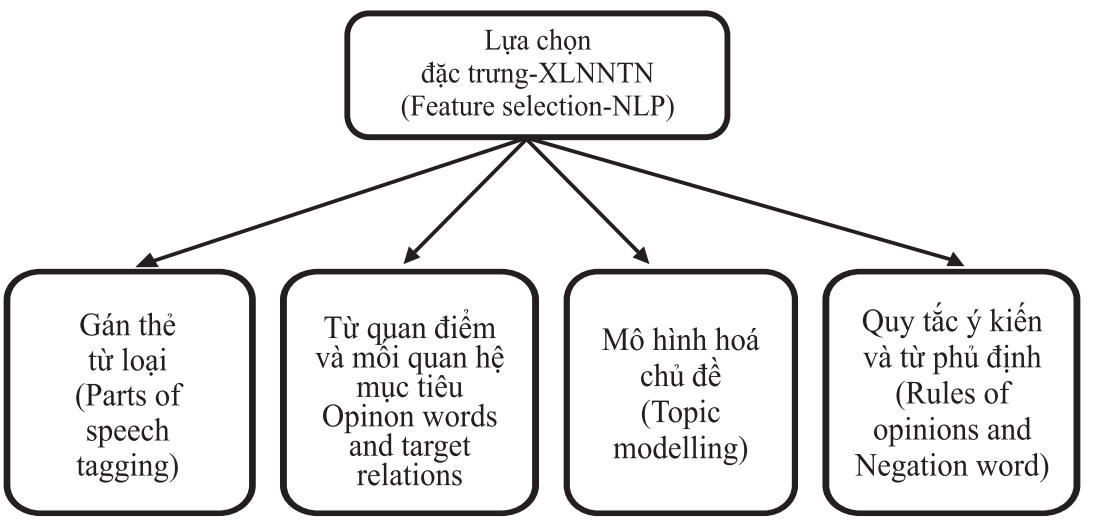
Sử dụng từ ngữ ý kiến và mối quan hệ giữa đặc trưng và từ ngữ ý kiến đó để trích xuất các đặc trưng cũng là một hướng tiếp cận khác để lựa chọn đặc trưng. Quá trình khai thác các quy tắc liên kết sẽ xác định được các đặc trưng phổ biến, đó là tập hợp các từ hoặc cụm từ thường được sử dụng nhất. Mối quan hệ này có thể được sử dụng để xác định các đặc trưng được chọn. Liên quan đến phương pháp này chúng ta cũng có thể sử dụng và khai thác quy tắc ý kiến. Đó là cách chọn lựa và sử dụng các thuật ngữ hoặc cấu trúc ngôn ngữ mà chúng có thể được sử dụng để mô tả tình cảm và ý kiến trong văn bản. Trong nhóm phương pháp này đặc biệt chú ý đến từ phủ định, ý nghĩa của đặc trưng này là một đặc trưng quan trọng vì nghĩa của nó có thể làm thay đổi hướng của tình cảm. Ví dụ, “không tốt”

tương ứng với "xấu" [11].

5. MỘT SỐ CÁCH TIẾP CẬN ĐỐI VỚI TIẾNG VIỆT

5.1. Một số đặc điểm cơ bản của tiếng Việt

Tiếng Việt được xếp vào loại ngôn ngữ biệt lập, nghĩa là mỗi âm tiết được phát âm riêng biệt, không có sự biến đổi hình thái, ngôn ngữ là đơn âm tiết và mỗi âm tiết xuất hiện như một từ (viết). Những đặc điểm này thể hiện rõ ràng trong mọi khía cạnh của tiếng Việt: âm vị học, từ vựng và ngữ pháp. Nhìn chung, trong tiếng Việt mỗi âm tiết là một yếu tố có nghĩa. Âm tiết là đơn vị cơ bản trong hệ thống các đơn vị có nghĩa của tiếng Việt. Các đơn vị từ vựng khác từ âm tiết được tạo ra để đặt tên cho các đối tượng, sự kiện, hiện tượng,... chủ yếu thông qua ghép và lặp lại. Sự hình thành các đơn vị từ vựng thông qua ghép luôn được chi phối bởi quy tắc kết hợp ngữ nghĩa, ví dụ "xe" kết hợp với "đạp" tạo thành "xe đạp" (bicycle)... Đây là phương thức chính để xây dựng các đơn vị từ vựng, trong đó tiếng Việt tận dụng tối đa các yếu tố tạo thành từ kho từ tiếng Việt.



Hình 1. Danh mục các hướng lựa chọn đặc trưng sử dụng NLP

Từ vựng tối thiểu của tiếng Việt chủ yếu bao gồm các từ đơn âm tiết. Ranh giới từ không được xác định bằng khoảng trắng như trong các ngôn ngữ chuyển đổi khác. Điều này làm cho việc phân tích hình thái (phân đoạn từ) của tiếng Việt trở nên khó khăn. Ví dụ: "Học sinh học sinh học". Ngoài ra có những loại từ đặc biệt gọi là "từ loại" đứng trước danh từ như chiếc/a (chiếc xe/a car), cái/an (cái dù/an umbrella),…vv.

Về ngữ pháp, từ tiếng Việt không bị biến đổi về mặt hình thái. Đặc điểm này chi phối các đặc điểm ngữ pháp khác. Nghĩa ngữ pháp nằm ngoài từ, khi các từ tiếng Việt được kết hợp thành các cấu trúc như mệnh đề và câu, trọng tâm sẽ chủ yếu là trật tự từ. Ví dụ: "lên xe và xe lên". Việc sắp xếp các từ theo một thứ tự nhất định là cách chính để biểu thị các mối quan hệ cú pháp. Trong tiếng Việt, khi chúng ta nói, "Anh ta lại đến" thì điều này sẽ có nghĩa hoàn toàn khác với "Lại đến anh ta" mặc dù số từ của hai câu này là như nhau. Khi các từ cùng loại được kết hợp với nhau bằng phép hạ

ngữ, từ đầu tiên sẽ đóng vai trò cốt lõi trong khi từ hoặc các từ theo sau sẽ là thành phần hạ ngữ. Trật tự từ phổ biến trong cấu trúc câu tiếng Việt là chủ ngữ đi trước, theo sau là vị ngữ và các thành phần bổ ngữ khác. Trong một văn bản, ngữ điệu thường được phản ánh thông qua các dấu câu [12].

5.2. Một số cách tiếp cận trong vấn đề lựa chọn đặc trưng cho tiếng Việt

Trên cơ sở các phân tích ở mục 3, 4 và các đặc điểm của tiếng Việt ở mục 5.1 chúng tôi đề xuất phân nhóm hướng tiếp cận lựa chọn đặc trưng trong tiếng Việt theo hướng xử lý ngôn ngữ tự nhiên theo và mô tả ở hình 1, đồng thời đề xuất một số cách tiếp cận chi tiết công việc phân tích tình cảm trong tiếng Việt, trong đó có bước lựa chọn đặc trưng như sau:

- Phân biệt quá trình trích xuất đặc trưng và lựa chọn đặc trưng. Phân biệt giữa trích xuất đặc trưng và lựa chọn đặc trưng trong phân tích tình cảm và quá trình lựa chọn đặc trưng trong học máy.

- Do đặc điểm tiếng Việt rất phong phú và đa dạng nên khi tiến hành phân tích tình cảm chú ý không xét các trường hợp không chuẩn của văn bản tiếng Việt mà chỉ giải quyết các vấn đề chính quy.

- Điều chỉnh các bước trong tiến trình phân loại tình cảm trên tiếng Việt để phù hợp với đặc điểm ngôn ngữ.

- Trong bước tiền xử lý dữ liệu chú ý với các nhiệm vụ tách đoạn, tách câu, chuẩn hóa chính tả, chuẩn hóa dấu chấm câu, đồng nhất chữ hoa, chữ thường, xử lý các từ viết tắt, loại bỏ các con số, các phụ từ, hư từ như "là", "của", "nhất là",... các từ nối, từ chỉ số lượng "và", "các", "những, "mỗi",…vì chúng không mang tính phân biệt trong khi phân loại, loại bỏ chúng để tăng hiệu năng cũng như giảm bớt số lượng đặc trưng vốn đã rất lớn trong các mô hình phân loại văn bản.

- Không giống như tiếng Anh, trong tiếng Việt ranh giới từ không phải là những khoảng trắng. Từ là đơn vị cơ bản để phân tích cấu trúc của ngôn ngữ, do vậy việc xác định ranh giới từ là rất quan trọng. Tách từ trong tiếng Việt quyết định quá trình phân loại đúng hay sai, hiệu quả cao hay thấp.

- Việc sử dụng các đặc trưng dựa trên từ sẽ đạt được kết quả tốt hơn so với sử dụng các đặc trưng dựa trên âm tiết.

- Trích xuất các đặc trưng dựa trên các từ là rất quan trọng, bao gồm danh từ riêng, danh từ chung, động từ, tính từ, trạng từ và liên từ phụ thuộc có hiệu quả hơn trích xuất các đặc trưng sử dụng tất cả các từ.

- Điểm số chung rất quan trọng để dự đoán chính xác tình cảm của các câu trong tiếng Việt.

- Khi sử dụng cách tiếp cận n-gram chú ý đến số đặc trưng phải phù hợp tức là không lớn quá cũng không bé quá, thường thì sử dụng unigram có hiệu quả hơn. Ngoài ra, áp dụng các kỹ thuật làm mịn (smoothing) sẽ nâng cao hiệu quả của quá trình lựa chọn đặc trưng trong tiếng Việt.

- Sử dụng các phương pháp rút gọn số chiều của không gian đặc trưng văn bản đã được áp dụng thành công cho bài toán phân loại văn bản tiếng Việt vì thế chúng cũng khả quan để áp dụng cho tiến trình phân loại tình cảm tiếng Việt.

- Đối với quá trình gắn thẻ PoST, đây là bước cần thiết trong xử lý ngôn ngữ tự nhiên để gán từ loại cho từng từ trong câu dưới dạng danh từ, động từ, tính từ, đại từ. Cần nâng cao hiệu quả của quá trình này trong tiếng Việt, dựa trên đặc điểm của ngôn ngữ, điều này quyết định việc tăng ngữ nghĩa trong văn bản cần phân tích.

- Như đã phân tích ở trên, nếu kết hợp quá trình trích xuất đặc trưng và lựa chọn đặc trưng trong bước lựa chọn đặc trưng và quá trình lựa chọn đặc trưng trong học máy thành một quá trình thì sẽ tiết kiệm được thời gian và nâng cao hiệu suất, đồng nhất được số đặc trưng cho mỗi bài toán từ đó nâng cao hiệu quả công việc.

- Cuối cùng là hoàn thiện các nguồn tài nguyên quan trọng cho bài toán phân tích tình cảm đó là dữ liệu nhận xét, mô hình từ nhúng (word embedding) được huấn luyện sẵn và bộ từ điển cảm xúc tiếng Việt. Vấn đề này đã được nhóm tác giả [3] minh chứng trong nghiên cứu của mình bằng việc thử nghiệm trên mô hình học sâu PhoBERT đơn ngữ cho kết quả cao nhất là 84,4 %, cao hơn các mô hình khác trên hai tập dữ liệu chuẩn tiếng Việt trong lĩnh vực nhà hàng và khách sạn, đồng thời cũng chứng minh rằng việc kết hợp tập dữ liệu từ các ngôn ngữ khác bằng cách sử dụng các mô hình đa ngôn ngữ có thể cải thiện hiệu suất trên ngôn ngữ đích.

6. KẾT LUẬN

Bài toán phân tích tình cảm thuộc loại đặc biệt của phân loại văn bản, vì vậy có số đặc trưng rất nhiều, nâng cao hiệu quả phân loại phụ thuộc vào quá trình lựa chọn đặc trưng này. Các nhiệm vụ liên quan của quá trình phân tích tình cảm đều thuộc lĩnh vực nghiên cứu xử lý ngôn ngữ tự nhiên. Do đó nhóm các phương pháp lựa chọn đặc trưng đều thuộc cách tiếp cận xử lý ngôn ngữ tự nhiên mà chúng tôi đã chỉ ra trong quá trình nghiên cứu là hợp lý. Từ cách tiếp cận chúng tôi đã phân loại các nhóm phương pháp theo cách tiếp cận này cho tiếng Việt. Ngoài ra, phân loại tình cảm là một hướng nghiên cứu mới đối với tiếng Việt, ngôn ngữ này có những đặc điểm riêng biệt và đa dạng. Trên cơ sở những đặc điểm đó, chúng tôi đã đề xuất một số cách tiếp cận chi tiết khi tiến hành thực hiện phân loại tình cảm trên tiếng Việt. Trong đó có việc đề xuất bước hợp nhất của hai quá trình trích xuất đặc trưng và lựa chọn đặc trưng. Trong bước lựa chọn đặc trưng và quá trình lựa chọn đặc trưng trong học máy thành một quá trình thì sẽ tiết kiệm được thời gian và nâng cao hiệu suất. Tuy nhiên, thách thức lớn nhất của hướng tiếp cận này là hiện tại chưa có một nghiên cứu nào cho các ngôn ngữ được triển khai, trong thời gian tới chúng tôi sẽ tiếp tục nghiên cứu về vấn đề này.

TÀI LIỆU THAM KHẢO

[1] Opinion Mining, Synth. Lect. Hum. Lang. Technol (2012), 5: 1:1-167, doi: 10.2200/S00416ED1V01Y201204HLT01. May 2012.

[2] Trần Khải Thiện, Tiếu Phùng Mai Sương (2020), Một số khái niệm và hướng tiếp cận phân tích cảm xúc áp dụng cho tiếng Việt. Tạp chí Khoa học HUFLIT (HJS), Tập 6 Số 1.https://hjs.huflit.edu.vn/index.php/hjs/article/view/94/36.

[3] Dang Van Thin, Lac Si Le, Vu Xuan Hoang, Ngan Luu-Thuy Nguyen (2020), Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. December 2022.DOI:10.1109/RIVF55975.2022.10013792. Conference: 2022 RIVF International Conference on Computing and Communication Technologies (RIVF).

[4] T. K. Tran, T. T. Phan (2016), Multi-Class Opinion Classification for Vietnamese Hotel Reviews, Int. J. Intell. Technol.

Appl. Stat., 9:1:7-18, doi: 10.6148/IJITAS.2016.0901.02, Mar.

[5] G. Vinodhini and R. Chandrasekaran (2012), Sentiment analysis and opinion mining: A survey, International Journal of Advanced Research in Computer Science and Software Engineering 2, 282-292.

[6] Siti Ahmad, Muhammad Zakwan, Nurlaila Syafira, Nurhafizah Moziyana (2019), A Review of Feature Selection and Sentiment Analysis Technique in Issues of Propaganda, January 2019 International Journal of Advanced Computer Science and Applications 10(11). DOI:10.14569/IJACSA.2019.0101132.

[7] Gerard. Salton (1968), Automatic Information Organization and Retrieval. McGraw Hill Text

[8] Abbasi, S. France, Z. Zhang and H. Chen (2010), Selecting attributes for sentiment classification using feature relation networks, IEEE Transactions on Knowledge and Data Engineering 23 (2011), 447-462. doi: 10.1109/TKDE. 110.

[9] Le Hong Phuong, Huyen Thi Minh Nguyen (2013), Part-of-Speech Induction for Vietnamese, Conference: The Fifth International Conference on Knowledge and Systems Engineering (KSE), DOI:10.1007/978-3-319-02821-7_24.

[10] B. Liu (2012), Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

[11] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen (2011), Opinion Word Expansion and Target Extraction through Double Propagation. March Computational Linguistics 37(1):9-27. DOI:10.1162/coli_a_00034.

[12] Đinh Đien, Ho Bao Quoc (2002), Word boundaries in English- Vietnamese bilingual corpus. Seminar report "Protecting and Developing Vietnamese". Linguistic Institute, HCMC Linguistic Association, University of Social Sciences and Humanities - HCMC, 12-2002, p.70 - 78.

Liên hệ:

TS. Đậu Mạnh Hoàn

Khoa Công nghệ thông tin, Trường Đại học Quảng Bình

Địa chỉ: 18 Nguyễn Văn Linh, Đồng Hới, Quảng Bình

Email: daumanhhoan@yahoo.com

Ngày nhận bài: 30/7/2024

Ngày gửi phản biện: 05/8/2024

Ngày duyệt đăng: 23/10/2024