NGHIÊN CỨU XÂY DỰNG TIẾN TRÌNH PHÂN LOẠI TÌNH CẢM TRÊN TIẾNG VIỆT

Đậu Mạnh Hoàn
Trường Đại học Quảng Bình

Tóm tắt. *Phân loại văn bản là quá trình phân tích nội dung văn bản và sau đó đưa ra quyết định văn bản đó có thể thuộc về một nhóm, nhiều nhóm hoặc không thuộc vào nhóm tài liệu văn bản nào được định nghĩa trước. Phân loại 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 (tích cực hay tiêu cực). Trên thế giới đã có nhiều nghiên cứu có hiệu quả về vấn đề này, đặc biệt là trên các tài liệu văn bản bằng tiếng Anh. Tuy nhiên, rất ít nghiên cứu về tài liệu văn bản tiếng Việt. Hơn nữa, các kết quả nghiên cứu và ứng dụng vẫn còn hạn chế một phần do các đặc điểm đặc trưng của ngôn ngữ tiếng Việt trong cấu trúc từ, câu và có nhiều từ đa nghĩa trong nhiều bối cảnh khác nhau. Trong bài viết này, tác giả tiến hành xây dựng một mô hình tiến trình nhằm phục vụ quá trình phân loại tình cảm trên tiếng Việt và đề xuất kỹ thuật lựa chọn đặc trưng cho tiến trình đó.*

Từ khóa: khai phá văn bản, phân loại tình cảm, lựa chọn đặc trưng

1. TỔNG QUAN

Phân tích tâm lý, khai phá ý kiến là các lĩnh vực nghiên cứu chuyên sâu trong lĩnh vực nghiên cứu khai phá văn bản. Phân tích tâm lý là nghiên cứu tính toán các ý kiến của con người, thái độ, cảm xúc và tình cảm của người đó đối với một thực thể. Các thực thể có thể đại diện cho các cá nhân, tổ chức, sự kiện hoặc chủ đề. Các vấn đề đưa ra xem xét sẽ được bao phủ bởi các ý kiến nhiều nhất. Trong thực tế, khai phá ý kiến là trích xuất các thông tin của con người về một thực thể sau đó xử lý các ý kiến của con người về thực thể đó, trong khi phân tích tâm lý lại xác định xu hướng ý kiến thể hiện trong một văn bản, sau đó phân tích nó. Do đó, mục tiêu của phân tích tâm lý là tìm xu hướng ý kiến, xác định những tình cảm mà họ thể hiện theo khuynh hướng nào, từ đó phân loại các ý kiến phân cực của họ đối với thực thể đó. Vì vậy, phân tích tâm lý có thể được xem là một quá trình phân loại đặc biệt trong phân loại văn bản và thường được gọi là phân loại tình cảm.

Phân loại văn bản là bài toán được sử dụng nhiều trong khai phá văn bản. Để thực hiện phân loại, người ta ánh xạ một văn bản vào một chủ đề đã biết trong một tập hữu hạn các chủ đề dựa trên ngữ nghĩa của văn bản [1]. Ý nghĩa của phân loại văn bản sẽ giúp cho việc sắp xếp, lưu trữ, tìm kiếm và truy xuất tài liệu dễ dàng hơn, hiệu quả hơn. Một điều đáng chú ý khi giải quyết bài toán phân loại văn bản đó là sự đa dạng của chủ đề văn bản và tính đa chủ đề của văn bản. Ảnh hưởng của đa chủ đề trong văn bản làm cho sự phân loại chỉ mang tính tương đối và có thể không rõ ràng khi thực hiện phân loại tự động. Trên phương diện cấu trúc tổ chức văn bản thì văn bản bao gồm một tập hợp các từ ngữ có liên quan với nhau tạo nên nội dung ngữ nghĩa cho văn bản. Từ ngữ của mỗi văn bản rất đa dạng và phong phú do đặc điểm của từ đa nghĩa và đa dạng trong

ngôn ngữ. Trong thực tế, có thể một văn bản có số lượng từ không nhiều, nhưng số lượng từ cần xét là rất nhiều, bởi vì nó phải bao hàm tất cả các từ của ngôn ngữ đang xét đó. Do đó, có thể xem việc giải quyết bài toán là đọc nội dung và phân tích nội dung đọc được, sau đó chọn thuật toán để phân loại văn bản.

Phân loại văn bản tự động [2] thường được thực hiện theo nhiều cách tiếp cận như dựa vào từ khóa, dựa vào thống kê tần số xuất hiện của các từ trong văn bản... Với cách tiếp cận như thế, một quá trình quy nạp tổng quát sẽ tự động xây dựng một trật tự phân lớp cho phân lớp d_i bằng cách ghi nhận những đặc trưng có được của tài liệu thuộc lớp d_i và những tài liệu không thuộc phân lớp d_i. Từ những đặc trưng này, quá trình thu thập có tính chất quy nạp sẽ dự đoán các đặc trưng sẽ phải có đối với những tài liệu thuộc phân lớp d_i. Hầu hết các phương pháp máy học áp dụng cho bài toán phân loại văn bản đều sử dụng cách biểu diễn văn bản dưới dạng vectơ đặc trưng. Lựa chọn đặc trưng có tầm quan trọng rất lớn trong thực tế và nhất là cho vấn đề phân loại văn bản.

Phân loại tình cảm người sử dụng là một lĩnh vực nghiên cứu mới trong khai phá văn bản, trong xử lý ngôn ngữ tự nhiên và ngôn ngữ học. Phân loại tình cảm người sử dụng là trường hợp đặc biệt của phân loại văn bản, mục đích chính của phân tích tình cảm người sử dụng là xác định sự phân cực của văn bản trong xử lý ngôn ngữ tự nhiên. Các nhà nghiên cứu trên thế giới đã tập trung nghiên cứu nhiều lĩnh vực liên quan đến nó như kỹ thuật phân loại, kỹ thuật lựa chọn đặc trưng cho văn bản trong phân loại, xử lý các vấn đề của ngôn ngữ tự nhiên thông qua học máy. Các nhà nghiên cứu đã khám phá việc sử dụng các kỹ thuật học máy để liên kết các loại tài liệu tự động bằng cách sử dụng một tập huấn luyện ban đầu để thích ứng với các phân loại theo tính năng thiết lập của các tài liệu cụ thể. Quá trình học máy được bắt đầu bằng việc kiểm tra các văn bản mẫu để xác định các tính năng thiết lập tối thiểu từ đó sản xuất ra các kết quả phân loại dự kiến. Hầu hết các kỹ thuật liên quan được tìm thấy trên hai ngôn ngữ tự nhiên chủ yếu là tiếng Anh và tiếng Trung Quốc. Đối với tiếng Việt đây là một bài toán rất mới. Tiếng Việt được xếp vào loại hình đơn lập, tức là mỗi một tiếng (âm tiết) được phát âm tách rời nhau, không biến hình, đơn tiết và được thể hiện bằng một chữ viết. Sự khác nhau cơ bản giữa tiếng Anh và tiếng Việt là về loại hình (biến cách và đơn lập) nên từ tiếng Việt và từ tiếng Anh khác nhau cả về mặt từ vựng hóa và hình thái học. Đặc biệt ranh giới từ không được xác định mặc nhiên bằng khoảng trắng, vì vậy tách từ là nhiệm vụ quan trọng trước khi đi vào xử lý. Xử lý ngôn ngữ tự nhiên tiếng Việt là một bài toán phức tạp, qua nhiều giai đoạn. Trong nghiên cứu này chúng tôi chỉ tập trung xây dựng mô hình cho tiến trình phân loại tình cảm trên tiếng Việt và đề xuất các phương pháp lựa chọn đặc trưng cho tiến trình phân loại.

2. BÀI TOÁN PHÂN LOẠI TÌNH CẢM

Theo định nghĩa của Jiliang Tang, Salem Alelyani và Huan Liu (2014) [3] phân loại là một quá trình mà chia các đối tượng được nhận ra, được phân biệt và hiểu được. Quá trình phân loại các đối tượng được nhóm thành các bộ phân loại, một bộ phân loại mô tả mối quan hệ giữa các chủ thể và đối tượng tri thức. Có hai cách cơ bản để tiếp cận

phân loại đó là phân loại được huấn luyện trước và phân loại không được huấn luyện trước. Như đã phân tích trong phần 1, bài toán phân loại tình cảm là một trường hợp đặc biệt của bài toán phân loại văn bản mà giá trị phân loại nhận được theo tính phân cực, nó được sử dụng trong các lĩnh vực khác nhau như kinh doanh, chính trị và tâm lý học.

So với bài toán phân loại tình cảm tiếng Anh, bài toán phân loại tình cảm tiếng Việt có nhiều đặc điểm riêng, đó chính là các đặc điểm của ngôn ngữ tiếng Việt. Một bước quan trọng trong xử lý văn bản tiếng Việt là thực hiện tách từ sau quá trình tiền xử lý văn bản, bởi vì các từ có nghĩa trong tiếng Việt không được phân biệt mặc nhiên bằng khoảng trống mà còn phải dựa vào các yếu tố khác. Các giai đoạn còn lại khi xử lý tiếng Việt được thực hiện như ngôn ngữ tiếng Anh. Đầu tiên văn bản được thu thập, tiếp theo được tiến hành tiền xử lý, rồi tiến hành tách từ, khởi tạo số điểm cho các từ ban đầu, tiến hành xử lý tiền tố, hậu tố, sau đó sử dụng từ điển để kiểm tra từ đó có trong từ điển hay không, tương ứng với quá trình kiểm tra sẽ cập nhật số điểm, lặp lại cho đến khi kết thúc sẽ có tổng số điểm cuối cùng, dựa vào số điểm này để xác định từ đó là tiêu cực hay tích cực. Việc gán số điểm phân cực cho danh sách các từ để tiến hành phân loại các ý kiến dựa trên các chủ đề và dữ liệu văn bản liên quan. Ví dụ nhập vào một câu "Điện thoại iphone 6 thiết kế bắt mắt, ưa nhìn, có đầy đủ tính năng". Đầu tiên thực hiện tách từ ta được các từ sau: #Điện thoại#iphone 6#thiết kế#bắt mắt#ưa nhìn#có#đầy đủ#tính năng#. Sau đó tiến hành xác định từ đánh giá tình cảm và các cụm từ, chẳng hạn từ chỉ "**kiểu dáng"** là "*bắt mắt, ưa nhìn*"; từ chỉ "**tính năng"** là "đầy đủ*"*. Bước cuối cùng là đánh giá sản phẩm thông qua các từ chỉ ý kiến, chẳng hạn "kiểu dáng" được gán điểm tích cực là 1, điểm tiêu cực là 0; "**tính năng**" gán điểm tích cực là 1, điểm tiêu cực là 0. Kết thúc đánh giá ta có giá trị trọng số của "kiểu dáng" và "tính năng" là 1 và 1, tổng bằng 2 điểm, trọng số này chỉ ra rằng ý kiến tình cảm của người dùng là tích cực đối với sản phẩm Iphone 6.

Thông thường giả sử rằng các bộ phân loại chỉ là những nhãn ký hiệu. Các thuộc tính của các tài liệu liên quan đến bộ phân loại được nhận ra dựa trên nội dung cơ bản của tài liệu. Phụ thuộc vào từng ứng dụng cụ thể mà phân loại văn bản có thể chia thành các loại khác nhau, ở đây chúng tôi dựa vào Fabrizio Sebastiani (2002) [4].

3. ĐẶC TRƯNG VÀ QUÁ TRÌNH LỰA CHỌN ĐẶC TRƯNG

3.1. Khái niệm đặc trưng

Đặc trưng của văn bản là những hạng trong văn bản [5]. Người ta sử dụng các thuật toán để biểu diễn không gian đặc trưng trong quá trình phân loại. Lựa chọn đặc trưng nhằm mục đích chọn lựa ra một tập con các đặc trưng tiêu biểu biểu diễn từ không gian đặc trưng gốc.

3.2. Quá trình lựa chọn đặc trưng

Văn bản có thể xem như là một tập hợp các đặc trưng. Việc phân loại văn bản sẽ dựa trên các đặc trưng này. Do 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. Do đó, cần phải lựa chọn đặc trưng nhằm rút ngắn số

chiều của không gian đặc trưng. Thực chất của 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. Lựa chọn đặc trưng sẽ tìm ra một tập nhỏ các đặc trưng có giá trị nhất.

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. Trên thực tế, người ta không thể xem xét tất cả các từ của ngôn ngữ mà 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.

4. PHƯƠNG PHÁP LỰA CHỌN ĐẶC TRƯNG

Có nhiều phương pháp lựa chọn đặc trưng khác nhau, mỗi phương pháp có các ưu điểm riêng biệt. Ở đây, chúng tôi giới thiệu phương pháp phân tích giá trị riêng, là phương pháp được đánh giá tốt hơn các phương pháp cổ điển nhờ có ưu điểm vượt trội để rút gọn không gian đặc trưng trong quá trình lựa chọn đặc trưng. Hiệu quả từ việc rút gọn không gian đặc trưng đã làm tăng hiệu quả phân loại và giảm bớt các tính toán được minh chứng trong công trình nghiên cứu của tác giả [6] và phương pháp Optimal Orthogonal Centroid Feature Selection của nhóm Microsoft Asia đề xuất năm 2004 [7].

4.1. Phương pháp phân tích giá trị riêng

Phương pháp phân tích giá trị riêng [8] (Singular value decomposition: SVD) là một dạng khai triển của ma trận, phương pháp này dựa trên nền tảng trong kỹ thuật chỉ mục ngữ nghĩa tiềm ẩn (LSI: Latent Semantic Indexing) có rất nhiều ứng dụng trong nghịch đảo, số hóa các dữ liệu, tìm kiếm, truy hồi thông tin dạng văn bản, xử lý tín hiệu số, tính các giá trị xấp xỉ trong kĩ thuật và được ứng dụng nhiều trong các công cụ tìm kiếm trên các website. Ý tưởng chính của phương pháp [6, 9] như sau:

Với mọi ma trận $A_{m\times n}$ bất kỳ đều có thể phân tích thành A = U. Σ. V^T (4.1)

trong đó:

- U là ma trận trực giao $m\times m$ có các cột là các vectơ riêng bên trái của A.

- ∑ là ma trận $m\times n$ có đường chéo chứa các giá trị riêng, không âm có thứ tự giảm dần: $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_{min(m,n)} \geq 0$. Ma trận ∑ được xây dựng:

$$\sum_{m\times n} = \begin{bmatrix} D_{r\times r} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \qquad \text{với D} = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}$$

- V^T là ma trận trực giao $n\times n$ có các cột là các vectơ riêng bên phải của A.

Hạng của ma trận A là số các số khác 0 trên đường chéo chính của ma trận ∑. Thông thường A là một ma trận thưa có kích thước lớn. Để giảm số chiều của ma trận người ta thường tìm cách xấp xỉ ma trận A có hạng r bằng một ma trận A_k có hạng là k nhỏ hơn r rất nhiều. Ma trận xấp xỉ của A theo kỹ thuật này chính là: $A_k = U_k\sum_k V_k^T$ (4.2)

trong đó:

- U_k là ma trận trực giao $m\times k$ có các cột là k cột đầu của ma trận U.

- $\sum_k$ là ma trận đường chéo $k \times k$ chứa k phần tử đầu tiên δ_1 , δ_2 , δ_3 , δ_k trên đường chéo chính.

- V_k là ma trận trực giao $n \times k$ có các cột là k cột đầu của ma trận V.

Mục đích của quá trình thực hiện này là để chuyển không gian đang xét r chiều về không gian k chiều, với k nhỏ hơn rất nhiều so với r. Quá trình thực hiện chuyển đổi như sau: đầu tiên văn bản sẽ được mô hình hóa thành một vectơ cột trong không gian xác định bởi ma trận $A_{m \times n}$; sau khi chuyển đổi ma trận $A_{m \times n}$ về A_k, tất cả các vectơ đang xét đều được chiếu lên không gian A_k để có số chiều k theo công thức:

$$\text{Proj}(x) = x^T.U_k.\sum_k^{-1} \quad (4.3)$$

4.2. Phương pháp Optimal Orthogonal Centroid Feature Selection (OCFS)

Phương pháp OCFS là phương pháp được các tác giả tại Trung tâm nghiên cứu Microsoft Asia đề xuất năm 2004, phương pháp này được đánh giá là tốt hơn các phương pháp tìm ra trước đây [7]. Phương pháp OCFS dựa trên nền tảng của thuật toán Orthogonal Centroid (OC**)**. Thuật toán Orthogonal Centroid được sử dụng để rút trích đặc trưng có giám sát bằng cách tận dụng phép biến đổi trực giao trên trọng tâm [10, 11]. Thuật toán này đã được chứng minh là rất hiệu quả với các vấn đề phân lớp dữ liệu dưới dạng văn bản và nó dựa trên phép tính toán không gian véctơ trong đại số tuyến tính. Tư tưởng chính của phương pháp OCFS để tìm ra kỹ thuật lựa chọn đặc trưng tối ưu bằng cách tối ưu *J(W)* trong không gian H^{dxp} dựa trên tiêu chuẩn *J(W)* trong thuật toán Orthogonal Centroid. Độ phức tạp của thuật toán OCFS là O(cd), thuật toán OCFS dễ cài đặt và có thời gian tính toán nhanh hơn các phương pháp khác [7].

Phương pháp lựa chọn số đặc trưng *k*: giả sử tất cả các đặc trưng đã được tính điểm và sắp xếp theo thứ tự giảm dần $s(k_1) >= s(k_2) >= \ldots >= s(k_d)$, ta tính hàm

$$E(p) = \frac{\sum_{j=1}^{p} s(k_j)}{\sum_{i=1}^{d} s(i)} \quad (4.4)$$

p đặc trưng được chọn phải thỏa mãn p = arg min E(p) sao cho E(p)> = T, với T > = 80%.

Thuật toán OCFS

Thay vì tìm ma trận W ta tìm ma trận $\tilde{W}$ với tiêu chuẩn tối ưu $J(\tilde{W})$ như sau:

$$\arg\max J(\tilde{W}) = \arg\max trace(\tilde{W}^T S_b \tilde{W}), \quad \tilde{W} \in H^{dxp}. \quad (4.5)$$

Trong đó các thành phần trong công thức được định nghĩa như trong phần OC, chỉ khác là ma trận $\tilde{W}$ là ma trận nhị phân mà mỗi cột chỉ duy nhất có một phần tử khác 0. Định nghĩa K = {k_i, $1 \le k_i \le d$, i= 1, 2, ... p} là tập các chỉ mục của đặc trưng, ta có:

$$trace(\tilde{W}^T S_b \tilde{W}) = \sum_{i=1}^{p} \tilde{w}_i^T S_b \tilde{w}_i = \sum_{i=1}^{p} \sum_{j=1}^{c} \frac{n_j}{n} (m_j^{k_i} - m^{k_i})^2 \quad (4.6)$$

Quá trình thực hiện của thuật toán OCFS là tìm tập K ở trên để làm cực đại:

$$\sum_{i=1}^{p}\sum_{j=1}^{c}\frac{n_j}{n}(m_j^{k_i}-m^{k_i})^2 \quad (4.7)$$

Từ đó thuật toán OCFS được xây dựng như sau:

+ Đầu vào: Tập dữ liệu huấn luyện (ngữ liệu)

+ Phương pháp:

- Bước 1: Tính centroid m_i i=1, 2, …, c của mỗi lớp cho dữ liệu huấn luyện
- Bước 2: Tính centroid m của tất cả các mẫu huấn luyện
- Bước 3: Tính điểm cho từng đặc trưng i theo công thức

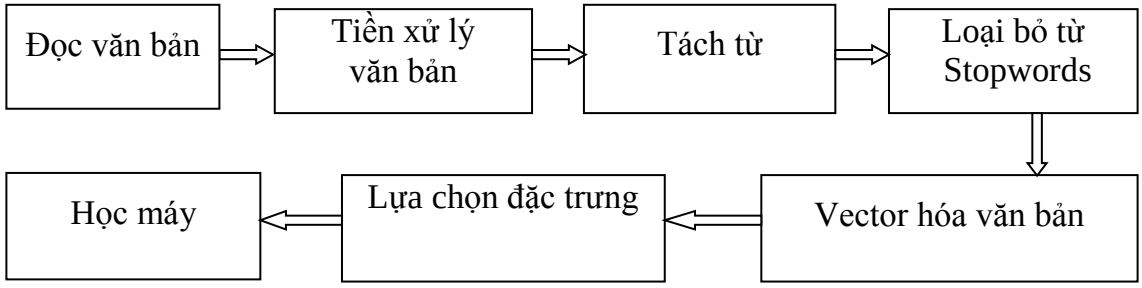
$$s(i)=\sum_{j=1}^{c}\frac{n_j}{n}(m_j^i-m^i)^2$$

- Bước 4: Chọn *k* đặc trưng có điểm cao nhất trong tập S={s(i)| 1≤i ≤ d}

+ Đầu ra: giá trị chính xác và F_1 của tập dữ liệu.

5. TIẾN TRÌNH PHÂN LOẠI TÌNH CẢM

Mô hình tiến trình phân loại tình cảm thực hiện như sau:



Hình 1. Mô hình tiến trình phân loại tình cảm tiếng Việt

5.1. Tiền xử lý văn bản

Do đặc điểm tiếng Việt rất phong phú và đa dạng nên không xét các trường hợp không chuẩn của văn bản tiếng Việt mà chỉ giải quyết các vấn đề chính quy. Dữ liệu văn bản được tiến hành tách đoạn, tách câu, chuẩn hóa chính tả, chuẩn hóa dấu chấm câu.

5.2. Tách từ

Để tiến hành phân loại văn bản, tài liệu văn bản được biến đổi thành dạng vectơ đặc trưng, trong đó đặc trưng là một từ. Không giống như tiếng Anh là ngôn ngữ thuộc loại hình hòa kết, tiếng Việt là ngôn ngữ thuộc loại hình đơn lập, không biến hình, các ký tự được dựa trên hệ chữ cái Latin. Trong tiếng Việt ranh giới từ không phải là những khoảng trắng. Một từ có thể được tạo bởi một hoặc nhiều hình vị và mỗi hình vị phân tách nhau bởi các khoảng trắng. Từ là đơn vị cơ bản để phân tích cấu trúc của ngôn ngữ, do vậy việc xác định ranh giới từ rất quan trọng và cũng có nhiều xử lý phức tạp. Tách từ là vấn đề quan trọng, nó quyết định quá trình phân loại đúng hay sai, hiệu quả cao hay thấp. Tất cả các tài liệu qua bước này đều được xử lý thành các từ là đầu vào cho bước xử lý tiếp theo.

5.3. Loại bỏ từ Stopwords

Từ Stopwords là các từ chức năng hay các phụ từ, hư từ chẳng hạn như "là", "của", "nhất là", …vv, các từ như từ nối, từ chỉ số lượng "và", "các", "những, "mỗi",… chúng không mang tính phân biệt trong khi phân loại. Ngoài ra, còn có rất nhiều từ khác cũng không có giá trị phân loại, ví dụ như từ xuất hiện hầu hết khắp các văn bản hay dùng không phổ biến trong văn bản, những từ gọi là stopword này sẽ được lược bỏ để tăng hiệu năng cũng như giảm bớt số lượng đặc trưng vốn đã rất lớn trong các mô hình phân loại văn bản.

5.4. Trọng số hóa đặc trưng

Trọng số (Weight) là một giá trị đặc trưng cho hạng, giá trị này thường là số thực. Công thức tính toán giá trị là TF_IDF (Terms Frequency Inverse Document Frequency) và các dạng mở rộng của nó là logTF_IDF và TF_IWF (Terms Frequency Inverse Word Frequency) [10].

5.5. Chọn lựa đặc trưng

Đặc trưng của văn bản là những hạng trong văn bản. Phương pháp lựa chọn đặc trưng được chúng tôi giới thiệu trong mục 3 và 4.

5.6. Học máy

Các giải thuật học máy 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 văn bản 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. Học máy là một lĩnh vực có liên quan đến việc nghiên cứu các thuật toán và kỹ thuật cho phép các máy tính để "học hỏi" tự động từ kinh nghiệm. Máy học tập dựa trên các khái niệm và kỹ thuật từ nhiều lĩnh vực, bao gồm cả số liệu thống kê, lý thuyết thông tin, trí tuệ nhân tạo, sinh học, triết học và xử lý tri thức thông minh.

6. KẾT LUẬN

Bài toán phân loại văn bản là bài toán có số lượng đặc trưng rất nhiều, nâng cao hiệu quả phân loại văn bản là mục đích mà nhiều nhà nghiên cứu hướng đến. Phân loại tình cảm là một hướng nghiên cứu mới, đặc biệt đối với tiếng Việt. Tiếng Việt có những đặc điểm riêng biệt và đa dạng. Chúng tôi đã đề xuất mô hình để sử dụng cho việc thực hiện phân loại tình cảm trên tiếng Việt và đề xuất sử dụng phương pháp rút gọn số chiều của không gian đặc trưng văn bản bằng phương pháp phân tích giá trị riêng và phương pháp OCFS áp dụng cho tiếng Việt. Các phương pháp này 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 trên tiếng Việt. Trong tương lai chúng tôi sẽ tiếp tục nghiên cứu và thực nghiệm cho quá trình phân loại này.

TÀI LIỆU THAM KHẢO

[1] Feldman, R., Sanger, J. (2007), The Text Mining Handbook: *Advanced Approaches in Analyzing Unstructured Data* (Cambridge University Press, Cambridge, UK.

[2] Maron, M. E, (1961), "Automatic Indexing: An Experimental Inquiry", *Journal of the Association for Computing Machinery*, 8(3): 404–417.

[3] Jiliang Tang, Salem Alelyani and Huan Liu (2014) *Feature selection for classification: A review*. In: Data Classification: Algorithms and Applications. CRC Press, p. 37

[4] Fabrizio Sebastiani (2002), *Machine Learning in Automated Text Categorization, ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47.

[5] Girish Chandrashekar, Ferat Sahin. *A survey on feature selection methods*. Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA. Available online 6 December 2013.

[6] Hoan Manh Dau, Ning Xu (2014), *Text Document Classification Using Support Vector Machine with Feature Selection Using Singular value Decomposition*, AMEE, Vol. 905, pp 528-532.

[7] Jun Yan-Ning Liu-Benyu Zhang-Shuicheng Yan (2005), *OCFS: Optimal Orthogonal Centroid Feature Selection for Text Categorization,* Microsoft Research Asia, China.

[8] Golub, G.H., Loan, C.F.V.(1996), *Matrix Computations,* third ed. Johns Hopkins University Press, pp. 48–80.

[9] T. Letsche, M. Berry (2001), "Large-scale Information Retrieval with Laten Semantic Analysis", SIGIR , pp. 19-25.

[10] Yang, Y. and Pedersen , J,O. (1997), *A comparative Study On Feature Selection in Text Categorization*, In Proceedings of the 14th International Conference on Machine Learning(ICML), 412-420.

[11] Tao Liu, Zheng Chen, Benyu Zhang, Wei-ying Ma, Gongti Wu (2004), *Improving Text Classificaion using Local Latent Semantic Indexing,* Data Mining, ICDM 2004. Proceedings, Fourth IEEE International Conference.

RESEARCH OF CONSTRUCTING THE SENTIMENT CLASSIFICATION PROCESS ON VIETNAMESE TEXTS

Abstract. *Text classification is the process of analyzing text content and then giving decision whether this text could belong to one group, many groups or it does not belong to the text group which is defined before. Sentiment classification is a special kind of text classification in which a document is classified to predict automatically sentiment polarity (positive or negative). In all over the world, there have been many effective researches on this problem, especially on texts in English. However, there have been few researches on Vietnamese texts. Moreover, these researching results and applications are still limited partly due to the typical characteristics of Vietnamese language in term of words and sentences and there are many words with many meanings in many different contexts. In this research, the author constructs a model to serve the process of sentiment classification on Vietnamese texts and suggests techniques feature selection for that process.*

Key words: text mining, sentiment classification, features selection.