PHÂN LỚP VĂN BẢN TIẾNG VIỆT TỰ ĐỘNG THEO CHỦ ĐỀ

Mạnh Thiên Lý*, Vũ Văn Vinh, Nguyễn Văn Lễ,
Lâm Thị Họa Mi, Nguyễn Thị Thanh Thủy, Dương Thị Mộng Thùy
Trường Đại học Công nghiệp Thực phẩm TP.HCM
*Email: *lymt@hufi.edu.vn*
Ngày nhận bài: 16/01/2019; Ngày chấp nhận đăng: 06/3/2019

TÓM TẮT

Mạng Internet ngày càng phát triển mạnh mẽ, mang lại nguồn thông tin vô cùng phong phú. Nhu cầu khai thác dữ liệu, phát hiện tri thức cũng ngày càng gia tăng. Phân lớp văn bản đóng vai trò quan trọng trong việc khai thác dữ liệu và phát hiện tri thức. Nhiều kỹ thuật trong học máy được ứng dụng để huấn luyện dữ liệu cho quá trình phân lớp. Hiện nay, có nhiều thuật toán được sử dụng để phân lớp văn bản như Naïve Bayes, K-NN, SVM, Maximum Entropy… Trong bài báo này, nhóm tác giả sử dụng các thuật toán như Naïve Bayes, SVM và K-NN để thực nghiệm phân lớp văn bản tiếng Việt trên 05 bộ dữ liệu thuộc 04 chủ đề khác nhau: Du lịch, Giải trí, Giáo dục và Pháp luật. Các bộ dữ liệu này được rút trích từ Website tin tức VnExpress.net. Một số đặc trưng định danh riêng được đưa vào quá trình xử lý để tăng độ chính xác trong quá trình phân lớp. Kết quả thử nghiệm cho thấy thuật toán SVM cho kết quả phân lớp với độ chính xác cao nhất (trên 90%) và thời gian thử nghiệm mô hình thấp nhất.

Từ khóa: Phân lớp văn bản, Naïve Bayes, K-NN, SVM, thuật toán.

1. TỔNG QUAN VỀ PHÂN LỚP VĂN BẢN

Phân lớp văn bản (Text classification) là quá trình gán nhãn (tên lớp/ nhãn lớp) cho các văn bản ngôn ngữ tự nhiên một cách tự động vào một hoặc nhiều lớp cho trước.

Phân lớp văn bản được xuất hiện từ những năm 1960, nhưng chỉ 15 năm sau đã trở thành lĩnh vực nghiên cứu chính trong hệ thống thông tin bởi sự đa dạng của các ứng dụng. Phân lớp văn bản được sử dụng để hỗ trợ trong quá trình tìm kiếm thông tin (Information retrieval), chiết lọc thông tin (Information extraction), lọc văn bản hoặc tự động dẫn đường cho các văn bản đến những chủ đề xác định trước. Ngoài ra, phân lớp văn bản cũng được ứng dụng trong lĩnh vực hiểu văn bản. Có thể sử dụng phân lớp văn bản để lọc văn bản hoặc một phần văn bản chứa dữ liệu cần tìm mà không làm mất đi tính phức tạp của ngôn ngữ tự nhiên. Phân lớp văn bản có thể thực hiện thủ công hoặc tự động sử dụng các kỹ thuật học máy có giám sát. Tuy nhiên, phân lớp thủ công đôi khi không chính xác vì quyết định phụ thuộc vào sự hiểu biết và động cơ của người thực hiện. Vì vậy, việc xây dựng một bộ phân lớp văn bản tự động là rất quan trọng và cần thiết, đặc biệt khi hầu hết các thông tin được sinh ra và lưu trữ điện tử. Các bài báo khoa học và giải trí là những ví dụ về tập các tài liệu điện tử. Với sự phát triển ngày càng mạnh mẽ của mạng Internet và Intranet đã tạo ra nguồn thông tin vô cùng phong phú. Các kỹ thuật phân lớp văn bản sẽ giúp cho nguồn dữ liệu này được lưu trữ tự động một cách hiệu quả và được tìm kiếm nhanh chóng.

1.1. Định nghĩa phân lớp văn bản

Phân lớp văn bản là nhiệm vụ đặt một giá trị logic cho mỗi cặp $(d_j, c_i) \in D \times C$, trong đó $D = \{d_1, d_2, \dots d_n\}$ là tập các văn bản và $C = \{c_1, c_2, \dots c_m\}$ là tập các lớp cho trước [1].

Giá trị $T(true)$ được gán cho cặp (d_j, c_i) có nghĩa là tài liệu d_j thuộc lớp c_i.

Giá trị $F(false)$ nghĩa là tài liệu d_j không thuộc lớp c_i.

Nói cách khác, phân lớp văn bản là bài toán tìm một hàm Φ: $D \times C \longrightarrow \{T, F\}$, trong đó D là tập các văn bản và C là tập các lớp cho trước, hàm Φ được gọi là bộ phân lớp.

1.2. Phân loại bài toán phân lớp văn bản

Tùy thuộc vào những ràng buộc khác nhau để phân loại bài toán phân lớp văn bản. Nhìn chung, có thể phân loại bài toán phân lớp theo các cách sau:

- Phân lớp văn bản nhị phân/ đa lớp: Bài toán phân lớp văn bản được gọi là nhị phân nếu $|C| = 2$, gọi là đa lớp nếu $|C| > 2$.

- Phân lớp văn bản đơn nhãn/ đa nhãn: Bài toán phân lớp văn bản được gọi là đơn nhãn nếu mỗi tài liệu được gán vào chính xác một lớp. Ngược lại, bài toán phân lớp văn bản được gọi là đa nhãn nếu một tài liệu có thể được gán nhiều hơn một nhãn.

1.3. Quá trình xây dựng bộ phân lớp văn bản

Quá trình phân lớp văn bản thường gồm 2 bước: xây dựng mô hình (tạo bộ phân lớp) và sử dụng mô hình đó để phân lớp văn bản. Các công cụ phân lớp được xây dựng dựa trên một thuật toán phân lớp qua bước học quy nạp. Trong bước học này, hệ thống có tập dữ liệu đầu vào D_t (tập ví dụ) mà thuộc tính lớp của mỗi tài liệu (ví dụ) trong tập đó đã biết. Tại đó, tập dữ liệu ban đầu được chia thành 2 tập dữ liệu rời nhau, một tập được gọi là tập huấn luyện (training set) và một tập được gọi là tập kiểm tra (test set). Thông thường, tập huấn luyện chiếm 2/3 các ví dụ trong D_t, còn tập kiểm tra chiếm 1/3 số lượng ví dụ còn lại. Hệ thống dùng tập huấn luyện để xây dựng mô hình (xác định tham số) phân lớp và dùng tập dữ liệu kiểm tra để đánh giá thuật toán phân lớp vừa được thiết lập.

Quá trình thực hiện cụ thể như sau:

*** Bước 1: Xây dựng mô hình**

Một mô hình sẽ được xây dựng dựa trên phân tích các đối tượng dữ liệu đã được gán nhãn từ trước. Tập các mẫu dữ liệu này còn được gọi là tập huấn luyện. Các nhãn lớp của tập dữ liệu huấn luyện được xác định bởi con người trước khi xây dựng mô hình (học có giám sát).

Ngoài ra, còn phải sử dụng một tập kiểm tra để tính độ chính xác của mô hình. Nếu độ chính xác là chấp nhận được, mô hình sẽ được sử dụng để xác định nhãn lớp cho các dữ liệu khác mới trong tương lai. Trong quá trình kiểm tra lại mô hình, sử dụng các độ đo để đánh giá chất lượng của tập phân lớp, đó là độ hồi tưởng, độ chính xác, độ đo F1, ...

Tùy thuộc vào cách thức xây dựng mô hình phân lớp, nhiều phương pháp được sử dụng để giải quyết bài toán như phương pháp Naïve Bayes, phương pháp K - láng giềng gần nhất (K-NN), phương pháp SVM, phương pháp Maximum Entropy, ...

*** Bước 2: Sử dụng mô hình**

Sử dụng mô hình đã được xây dựng ở bước 1 để phân lớp dữ liệu mới.

Như vậy, thuật toán phân lớp là một ánh xạ từ miền dữ liệu đã có sang một miền giá trị cụ thể của thuộc tính phân lớp dựa vào giá trị các thuộc tính của dữ liệu.

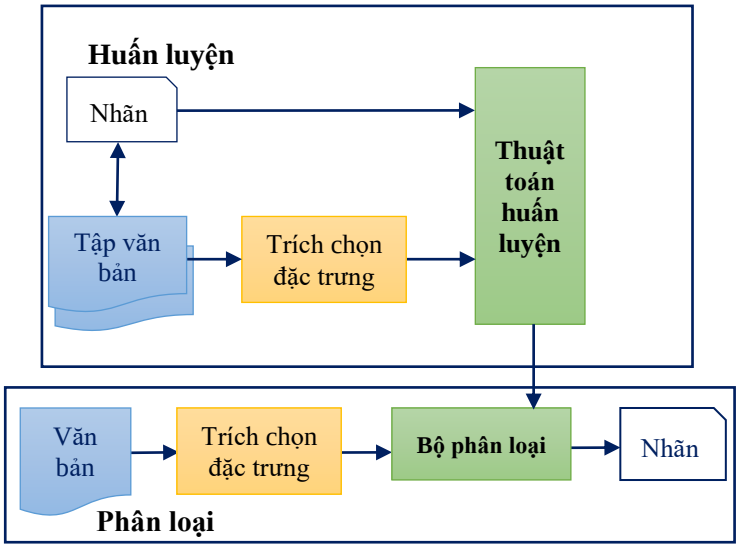
Để xây dựng mô hình trong bước 1 của quá trình phân lớp văn bản, thông thường, được tiến hành theo 2 bước chính sau đây:

- Tiền xử lý dữ liệu: là quá trình biểu diễn văn bản thành một dạng biểu diễn logic mà thuật toán có thể xử lý được (ví dụ: biểu diễn dạng vector của văn bản).
- Học các bộ phân lớp: sử dụng các thuật toán phân lớp để xây dựng mô hình từ dữ liệu đã qua tiền xử lý.

Các hệ thống phân lớp có thể ứng dụng trong việc phân loại tài liệu của các thư viện điện tử, phân loại văn bản báo chí trên các trang tin điện tử, phân loại văn bản tiếng Việt bằng cách xây dựng mô hình chủ đề, phân loại văn bản theo cảm xúc,... Với những hệ thống tốt, có thể nhận được kết quả khả quan, giúp ích nhiều cho người dùng.

Phân loại văn bản theo cảm xúc, tâm lý, quan điểm hiện đang là một trong những vấn đề được quan tâm nghiên cứu nhiều nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên [2]. Cảm xúc được định nghĩa là phản ứng của con người đối với các sự kiện, hiện tượng (kể cả bên trong hoặc bên ngoài cơ thể) mà có ý nghĩa nào đó đối với con người. Có nhiều quan điểm khác nhau về số lượng các loại cảm xúc. Căn cứ vào tính chất của cảm xúc có thể phân chia cảm xúc thành 3 loại: tích cực (positive), tiêu cực (negative) và trung lập (neutral). Nếu căn cứ vào biểu hiện và nội dung, chúng ta có thể chia cảm xúc thành 6 loại cơ bản: vui, buồn, giận dữ, ngạc nhiên, ghét, sợ hãi. Theo nghiên cứu của W. Gerrod Parrot, từ những cảm xúc cơ bản nhưng dưới sự tác động của các kích thích khác nhau trong những điều kiện, hoàn cảnh khác nhau mà cảm xúc của con người cũng có lúc đan xen, pha lẫn nhiều cảm xúc khác loại nhưng cùng tồn tại trong một thời điểm. Điều này đã tạo ra hàng loạt các cảm xúc khác.

Phân lớp văn bản tiếng Việt bằng cách xây dựng mô hình chủ đề sử dụng cách thức tìm một từ khóa và phát triển để tự sinh ra các từ khác trong chủ đề dựa trên các phương pháp Naïve Bayes, K-NN, SVM. Mỗi loại văn bản (hay còn gọi là lớp – class) tương đương với một chủ đề, ví dụ Giáo dục, Pháp luật, Thời sự, Khoa học, Xe ô tô – Xe máy, Thể thao, Giải trí, Du lịch. Bài toán phân lớp được xây dựng từ một tập các văn bản $D = \{d_1, d_2, \dots d_n\}$, trong đó các tài liệu d_i được gán nhãn c_j - với c_j thuộc tập các chủ đề $C = \{c_1, c_2, \dots c_m\}$, và xác định được mô hình phân lớp.



Hình 1. Mô hình phân lớp văn bản

Việc trích chọn đặc trưng của văn bản đóng vai trò quan trọng với kết quả phân loại văn bản. Nếu lựa chọn đặc trưng phù hợp sẽ giúp cho kết quả bài toán trở nên chính xác hơn. Tuy nhiên, nếu lựa chọn quá nhiều đặc trưng sẽ làm cho quá trình huấn luyện cũng như quá trình phân loại mất nhiều thời gian hơn. Do đó, vấn đề của lựa chọn đặc trưng là chọn một tập con nhỏ từ tập các đặc trưng mà vẫn đảm bảo tính chính xác của quá trình phân loại. Để tăng tính chính xác khi phân lớp, nhóm tác giả đưa thêm đặc trưng về định danh tên riêng trong quá trình xử lý. Ví dụ: một văn bản nếu có từ "Công_Phượng", "Quang_Hải" thì xác xuất cao được phân loại vào lĩnh vực Thể thao, văn bản có chứa từ "Mỹ_Tâm" thì xác suất cao được phân loại vào lĩnh vực Giải trí.

Trong phạm vi bài báo này, nhóm tác giả tập trung nghiên cứu phương pháp phân lớp văn bản tiếng Việt bằng cách xây dựng mô hình chủ đề. Phần còn lại của bài báo giới thiệu một số phương pháp phân lớp văn bản, trình bày kết quả thực nghiệm và kết luận.

2. MỘT SỐ PHƯƠNG PHÁP PHÂN LỚP VĂN BẢN

2.1. Thuật toán Naïve Bayes

Naïve Bayes là kỹ thuật phân loại phổ biến trong học máy có giám sát. Ý tưởng chính của kỹ thuật này dựa vào xác suất có điều kiện giữa từ hay cụm từ và nhãn phân loại để dự đoán văn bản mới cần phần loại thuộc lớp nào. Naïve Bayes được ứng dụng nhiều trong giải quyết các bài toán phân loại văn bản, xây dựng bộ lọc thư rác tự động, hay trong bài toán khai phá quan điểm bởi tính dễ hiểu, dễ triển khai cũng như độ chính xác tốt [3-8].

Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa các đặc trưng và nhãn để dự đoán xác suất nhãn của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các đặc trưng trong văn bản đều độc lập với nhau. Giả định đó làm cho việc tính toán Naïve Bayes hiệu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các đặc trưng để đưa ra phán đoán nhãn. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng…

Thuật toán Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Trong đó:

$P(X|Y)$ là xác suất xảy ra của một sự kiện ngẫu nhiên X khi biết sự kiện liên quan Y đã xảy ra.

$P(Y|X)$ là xác suất xảy ra Y khi biết X xảy ra.

$P(X)$ là xác suất xảy ra của riêng X mà không quan tâm đến Y.

$P(Y)$ là xác suất xảy ra của riêng Y mà không quan tâm đến X.

Áp dụng trong bài toán phân loại, các dữ kiện gồm có:

D: tập dữ liệu huấn luyện đã được vectơ hóa dưới dạng $\vec{x} = (x_1, x_2, \dots, x_n)$

C_i: phân loại i, với $i = \{1,2, \dots, m\}$.

Các thuộc tính độc lập điều kiện đôi một với nhau.

Theo định lý Bayes:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Theo tính chất độc lập điều kiện:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Trong đó:

$P(C_i|X)$ là xác suất thuộc phân loại i khi biết trước mẫu X.

$P(C_i)$ xác suất là phân loại i.

$P(x_k|C_i)$ xác suất thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân loại i.

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính $P(C_i)$ và $P(x_k|C_i)$.

Bước 2: Phân loại $X^{new} = (x_1, x_2, \dots, x_n)$, ta cần tính xác suất thuộc từng phân loại khi đã biết trước X^{new}. X^{new} được gán vào lớp có xác suất lớn nhất theo công thức:

$$\max_{C_i \in C} \left(P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \right)$$

Ví dụ 2.1: Xét bài toán phân loại email là thư rác (spam) hay không phải thư rác (non-spam). Để đánh giá một email, bước đầu tiên phải chuyển email sang vector $x = (x_1, x_2 \dots x_n)$ với $x_1, x_2 \dots x_n$ là giá trị các thuộc tính $X_1, X_2 \dots X_n$ trong không gian vector đặc trưng X. Mỗi thuộc tính được thể hiện bởi một token đơn. Theo phương pháp đơn giản nhất ta có thể lập ra một từ điển chứa các token. Sau đó với mỗi token trong email nếu nó xuất hiện trong từ điển thì giá trị thuộc tính sẽ là 1, ngược lại thì là 0. Tuy nhiên, trên thực tế, tập huấn luyện không thường là một bộ từ điển như vậy. Thay vào đó, tập huấn luyện lúc này sẽ gồm 2 kho ngữ liệu. Kho ngữ liệu thư rác sẽ chứa một danh sách các email đã được xác định là thư rác trước đó, và tương tự với kho ngữ liệu không thư rác sẽ chứa các email hợp lệ.

Như vậy, nếu vẫn để giá trị các thuộc tính là 0 hoặc 1 thì sẽ rất khó đánh giá được một email là spam hay không. Đặc biệt, nếu email nhận được là dài, khi đó nếu ta vẫn sử dụng giá trị thuộc tính là 0 hoặc 1 thì sự xuất hiện của một token 100 lần cũng tương đương với việc xuất hiện chỉ 1 lần.

2.2. Thuật toán K-Nearest Neighbors

K-Nearest Neighbors (K-NN) là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp và tất cả các đối tượng trong tập dữ liệu huấn luyện.

Một đối tượng được phân lớp dựa vào K láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Khoảng cách Euclid thường được dùng để tính khoảng cách giữa các đối tượng [9-11].

Các bước của thuật toán

1. Xác định giá trị tham số K (số láng giềng gần nhất).

2. Tính khoảng cách giữa đối tượng cần phân lớp với tất cả các đối tượng trong tập dữ liệu huấn luyện.

3. Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với đối tượng cần được phân lớp.

4. Lấy tất cả các lớp của K láng giềng gần nhất đã xác định.

5. Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho đối tượng.

Ví dụ 2.2:

Xét tập tài liệu huấn luyện {TL1, TL2, TL3, TL4} và tập các từ vựng {doanh thu, cáo buộc, thuế, điện ảnh, diễn viên, ca sĩ, nghi phạm, kinh doanh} có được sau khi thực hiện các bước tiền xử lý dữ liệu. Mỗi tài liệu thuộc một lớp chủ đề được xác định trước như: TL1 và TL2 thuộc lớp chủ đề Kinh doanh, TL3 thuộc lớp chủ đề Giải trí, TL4 thuộc lớp chủ đề Pháp luật. Các tài liệu này được mô hình hóa thành các vector nhiều chiều. Giá trị mỗi chiều là tần suất xuất hiện của từ vựng tương ứng trong tài liệu.

Bảng 1. Tần suất của các từ vựng trong văn bản

Tài liệu	Doanh thu	Cáo buộc	Thuế	Điện ảnh	Diễn viên	Ca sĩ	Nghi phạm	Kinh doanh	Lớp chủ đề
TL1	2	0	1	0	0	0	0	3	Kinh doanh
TL2	1	0	0	1	2	1	0	0	Giải trí
TL3	1	0	0	3	1	2	0	0	Giải trí
TL4	0	4	0	0	0	0	2	0	Pháp luật

Xét tài liệu cần phân loại có nội dung như sau:

"Khi nói đến những ca sĩ thành danh trên mặt trận điện ảnh Hollywood, chắc chắn không thể bỏ qua Jennifer Lopez. Cô đã tham gia đóng phim và lồng tiếng cho 31 bộ phim đình đám. Có thể nói, trong điện ảnh Jennifer Lopez có khả năng diễn xuất đa năng khi cô vừa có thể diễn những bộ phim tình cảm hài nhẹ nhàng cho đến những tác phẩm điện ảnh tội phạm hình sự. Lopez từng được đề cử giải Quả cầu vàng cho "Vai nữ diễn viên chính xuất sắc nhất - phim hành động hoặc hài" năm 1998".

Tài liệu này được biểu diễn thành vectơ nhiều chiều $V = (0,0,0,3,1,1,0,0)$. Sau đó sử dụng độ đo Euclid để tính khoảng cách đến tất cả các tài liệu trong tập huấn luyện, sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với đối tượng cần được phân lớp.

Bảng 2. Khoảng cách từ tài liệu đang xét đến các tài liệu khác

Tài liệu	Khoảng cách	Lớp chủ đề
TL3	1,4	Giải trí
TL2	2,4	Giải trí
TL1	4,3	Kinh doanh
TL4	5,5	Pháp luật

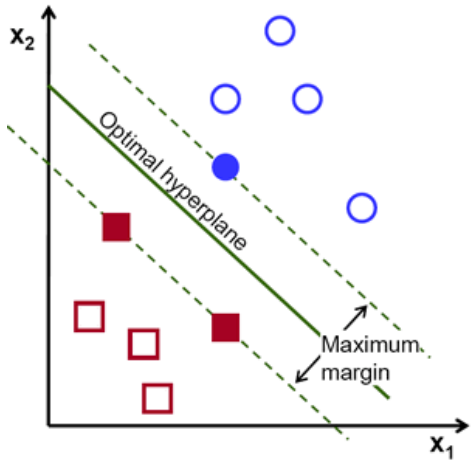
Trường hợp $K = 2$, chọn 2 tài liệu có khoảng cách ngắn nhất (láng giềng) lần lượt là TL3 và TL2. Cả 2 tài liệu láng giềng này đều thuộc lớp chủ đề giải trí nên tài liệu cần phân loại thuộc chủ đề giải trí.

Trường hợp $K = 3$, chọn 3 tài liệu có khoảng cách ngắn nhất (láng giềng) là TL3, TL2 và TL1. Trong đó có 2 tài liệu thuộc chủ đề giải trí, 1 tài liệu thuộc chủ đề kinh doanh. Nên tài liệu cần phân loại sẽ thuộc lớp chủ đề phổ biến hơn đó là chủ đề giải trí.

2.3. Thuật toán Support Vector Machine

Support Vector Machine (SVM) là một giải thuật máy học dựa trên lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng.

Bài toán cơ bản của SVM là bài toán phân loại 2 lớp: Cho trước n điểm trong không gian n chiều, mỗi điểm thuộc vào một lớp kí hiệu là $+1$ hoặc -1, mục đích của giải thuật SVM là tìm một siêu phẳng $f(x)$ (hyperplane) phân hoạch tối ưu cho phép chia các điểm này thành 2 phần sao cho các điểm cùng một lớp nằm về một phía với siêu phẳng này. Tất cả các điểm x_+ được gán nhãn 1 thuộc về phía dương của siêu phẳng, các điểm $x-$ được gán nhãn -1 thuộc về phía âm của siêu phẳng. Một siêu phẳng phân chia dữ liệu được gọi là "tốt nhất", nếu khoảng cách từ điểm dữ liệu gần nhất đến siêu phẳng (margin) là lớn nhất [12].



Hình 2. Phân lớp với SVM trong mặt phẳng

Thuật toán tìm siêu phẳng:

Bộ phân lớp tuyến tính được xác định bằng siêu phẳng: $\{x: f(x) = w^T + w_0\}$

Trong đó $w \in R^m$ và $w_0 \in R$ đóng vai trò là tham số của mô hình. Hàm phân lớp nhị phân b: $R^n \rightarrow \{0,1\}$ có thể thu được bằng cách xác định dấu của $f(x)$.

Rosenblatt đã đưa ra một thuật toán đơn giản để xác định siêu phẳng:

1. $w \leftarrow 0$
2. $w_0 \leftarrow 0$
3. repeat
4. $e \leftarrow 0$
5. for $i \leftarrow 1, \ldots, n$ do
6. $s \leftarrow sign(y_i(w^T x_i + w_0)$
7. if $s < 0$ then $w \leftarrow w + y_i x_i$
9. $w_0 \leftarrow w_0 + y_i x_i$
10. $e \leftarrow e + 1$
11. until $e = 0$
12. return (w, w_0).

Việc tìm siêu phẳng tối ưu có thể mở rộng trong trường hợp dữ liệu không thể tách rời tuyến tính bằng cách ánh xạ dữ liệu vào một không gian có số chiều lớn hơn bằng cách sử dụng một hàm nhân K (Kernel).

Bảng 3. Một số hàm nhân thường dùng

Kiểu hàm nhân	Công thức
Linear kernel	$K(x,y) = x.y$
Polynomial kernel	$K(x,y) = (x.y + 1)^d$
Radial basis function (Gaussian) kernel	$K(x,y) = e^{\frac{-\lvert x-y\rvert^2}{2\sigma^2}}$
Hyperbolic tangent kernel	$K(x,y) = tanh(a.x.y - b)$

Ví dụ 2.3:

Để kiểm tra một văn bản bất kỳ d_i nào đó thuộc hay không thuộc một phân loại c_j cho trước? Nếu $d_i \in c_j$ thì d_i được gán nhãn là 1, ngược lại thì d_i được gán nhãn là –1.

Giả sử lựa chọn được tập các đặc trưng là $T = \{t_1, t_2, \dots, t_n\}$, thì mỗi văn bản d_i sẽ được biểu diễn bằng một vector dữ liệu $x_i = (w_{i1}, w_{i2}, \dots, w_{in})$, $w_{ij} \in R$ là trọng số của từ t_j trong văn bản d_i. Như vậy, tọa độ của mỗi vector dữ liệu x_i tương ứng với tọa độ của một điểm trong không gian R^n.

Dữ liệu huấn luyện là tập các văn bản đã được gán nhãn trước $T_r = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$, trong đó, x_i là vector dữ liệu biểu diễn văn bản d_i ($x_i \in R^n$), $y_i \in \{+1, -1\}$, cặp (x_i, y_i) được hiểu là vector x_i được gán nhãn là y_i.

Việc xác định một văn bản x có thuộc phân loại c hay không, tương ứng với việc xét dấu của $f(x)$, nếu $f(x) > 0$ thì x thuộc c, nếu $f(x) \leq 0$ thì x không thuộc c.

3. KẾT QUẢ THỰC NGHIỆM

Để phân lớp văn bản theo chủ đề, nhóm tác giả tiến hành thực nghiệm trên máy tính Macbook Pro x64, Core i7 3.30GHz, 4 CPUs, 16GB RAM. Dữ liệu trên các trang báo điện tử có vốn từ ngữ và nội dung rất phong phú, dữ liệu đa dạng thuộc các lĩnh vực trong đời sống xã hội như: Kinh tế, Chính trị, Văn hóa, Giáo dục, Thể thao,… Nội dung các bài báo được đăng trên các trang báo điện tử uy tín đã được kiểm duyệt phù hợp với từng chủ đề. Vì vậy, việc thu thập dữ liệu từ các trang báo điện tử uy tín làm tập dữ liệu huấn luyện có độ chính xác cao, đáng tin cậy. Thực nghiệm được tiến hành trên tập dữ liệu tin tức tiếng Việt được trích xuất từ website VnExpress.net gồm 05 bộ dữ liệu với số lượng lần lượt là 400, 800, 1200, 1600 và 2000 tập tin văn bản thuộc 4 chủ đề như: Du lịch, Giải trí, Giáo dục và Pháp luật. Trong mỗi bộ dữ liệu thì số lượng các các tập tin ở các chủ đề là như nhau. Các tập tin dữ liệu này được xử lý tách từ bằng công cụ vnTokenizer [13], sau đó sử dụng công cụ Weka (phần mềm mã nguồn mở hỗ trợ xây dựng mô hình huấn luyện cho các bài toán về phân lớp dữ liệu) [14] để biểu diễn văn bản thành dạng vector, đồng thời loại bỏ những từ ngữ không có ý nghĩa (Stop words). Các vector văn bản này được sử dụng làm dữ liệu huấn luyện và dữ liệu kiểm tra.

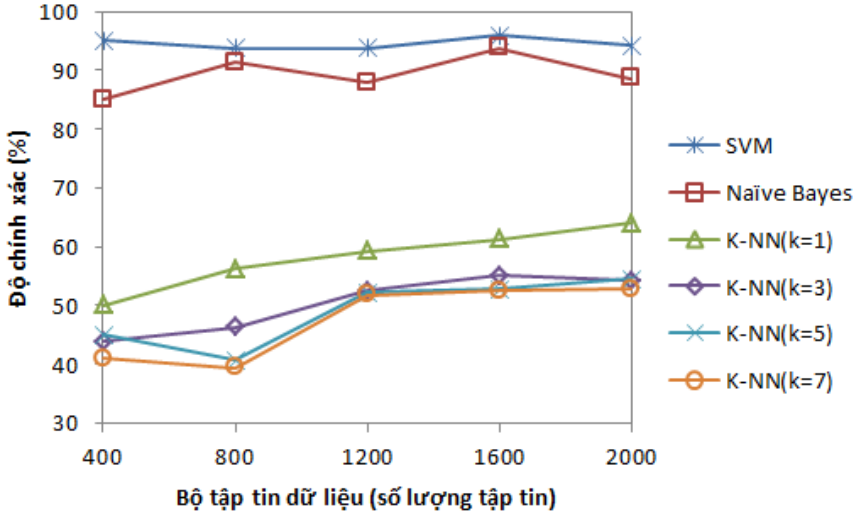
Trong bài báo này, nhóm tác giả đã chạy thực nghiệm 03 thuật toán là Naïve Bayes, SVM và K-NN trên cùng bộ dữ liệu huấn luyện. Trong đó, mỗi bộ dữ liệu có 80% dữ liệu dùng để huấn luyện và 20% dữ liệu còn lại dùng để thử nghiệm phân lớp. Bảng 4 trình bày kết quả thử nghiệm, so sánh độ chính xác giữa các thuật toán dựa trên giá trị trung bình của các tham số khi chạy thử nghiệm trên 05 bộ dữ liệu. Các tham số gồm: tỷ lệ văn bản được phân loại đúng (TP Rate), tỷ lệ văn bản phân loại sai (FP Rate), độ chính xác (Precision), độ bao phủ (Recall) và độ trung bình điều hòa (F-Measure).

Bảng 4. Giá trị trung bình các tham số theo phân lớp chủ đề với 05 bộ dữ liệu

Thuật toán	Tỷ lệ đúng (TP Rate)	Tỷ lệ sai (FP Rate)	Độ chính xác (Precision)	Độ bao phủ (Recall)	Độ trung bình điều hòa (F-Measure)
SVM	0,946	0,018	0,946	0,946	0,945
NaiveBayes	0,893	0,036	0,896	0,893	0,892
K-NN (k = 1)	0,582	0,144	0,645	0,582	0,580
K-NN (k = 3)	0,504	0,169	0,630	0,504	0,483
K-NN (k = 5)	0,500	0,162	0,677	0,500	0,481
K-NN (k = 7)	0,491	0,163	0,704	0,491	0,471

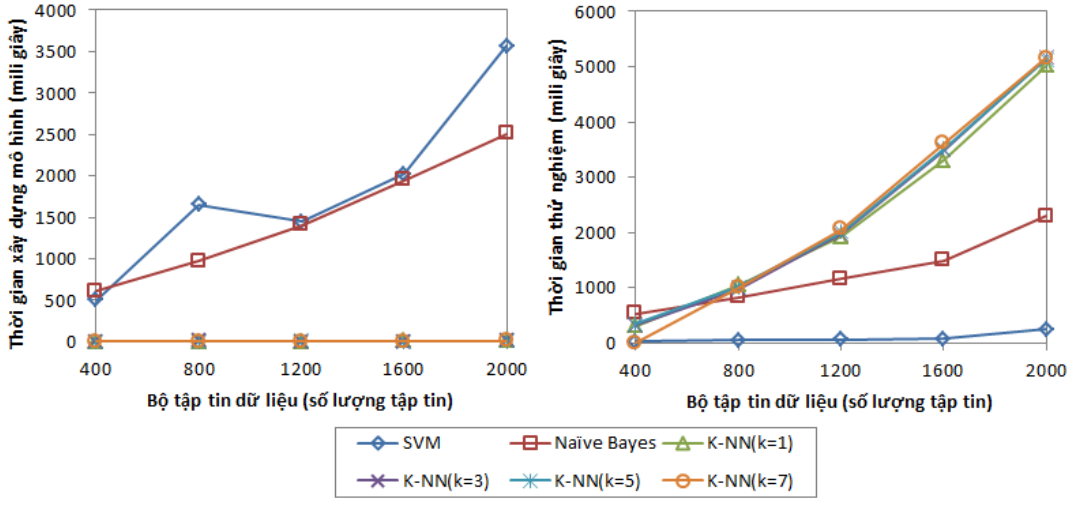
Hình 3 so sánh độ chính xác (%) của các thuật toán khi phân loại văn bản theo 4 chủ đề trên 05 bộ tập tin dữ liệu văn bản tiếng Việt. Độ chính xác của thuật toán K-NN phụ thuộc vào

việc chọn giá trị cho tham số k. Kết quả cho thấy giá trị của k càng nhỏ thì độ chính xác càng cao (độ chính xác cao nhất khi k = 1). Thuật toán SVM cho kết quả phân loại văn bản với độ chính xác cao nhất (trên 90%), tiếp đến là Naïve Bayes và cuối cùng là thuật toán K-NN.



Hình 3. So sánh độ chính xác các thuật toán

Hình 4 so sánh thời gian xây dựng mô hình huấn luyện và thời gian thử nghiệm của các thuật toán. Kết quả cho thấy thời gian xây dựng mô hình huấn luyện của thuật toán K-NN thấp nhất (gần bằng 0), trong khi thuật toán Naïve Bayes và SVM có thời gian xây dựng mô hình tăng tuyến tính theo độ lớn của bộ dữ liệu huấn luyện. Thuật toán SVM mất nhiều thời gian nhất để xây dựng mô hình huấn luyện. Tuy nhiên, thời gian thử nghiệm phân loại văn bản trên mô hình đã huấn luyện thì thuật toán SVM cho kết quả với thời gian thực hiện thấp nhất, kế đến là Naïve Bayes và cao nhất là K-NN.



Hình 4. Thời gian xây dựng mô hình và thời gian thử nghiệm của các thuật toán

Thực nghiệm chứng tỏ thuật toán SVM cho kết quả phân loại văn bản theo chủ đề tốt hơn Naïve Bayes và K-NN ở cả 2 khía cạnh là độ chính xác cao nhất và thời gian phân loại thử nghiệm trên mô hình thấp nhất. Mặc dù SVM tốn nhiều thời gian hơn để xây dựng mô hình huấn luyện nhưng có thể cải thiện điều này dễ dàng khi được huấn luyện trên các hệ thống máy tính tốc độ cao.

4. KẾT LUẬN

Trong bài báo này, nhóm tác giả đã trình bày vấn đề tiền xử lý văn bản, phương pháp phân lớp và thực hiện phân lớp văn bản tiếng Việt tự động theo chủ đề bằng cách sử dụng 3 thuật toán Naïve Bayes, K-NN và SVM. Thực nghiệm cho thấy thuật toán SVM cho kết quả phân lớp với độ chính xác cao nhất (trên 90%) và thời gian phân loại thấp nhất ở cả 05 bộ dữ liệu có số tập tin lần lượt là 400, 800, 1200, 1600 và 2000. Kết quả này cho thấy việc sử dụng thuật toán SVM để phân lớp văn bản tiếng Việt theo chủ đề là sự lựa chọn phù hợp trong các ứng dụng về phân lớp văn bản.

Kết quả nghiên cứu này là cơ sở cho nghiên cứu tiếp theo về ứng dụng phân loại văn bản theo hướng tích cực, tiêu cực và trung lập để xây dựng ứng dụng phát hiện và phân loại cảm xúc: tích cực (positive), tiêu cực (negative) và trung lập (neutral) của con người dựa trên nội dung các bài viết có trên Internet về một chủ đề cần quan tâm.

TÀI LIỆU THAM KHẢO

1. Sebastiani F. - Machine learning in automated text categorization, ACM Computing Surveys (CSUR) **34** (1) (2002) 1-47.
2. Ezhilarasi R. and Minu R. I. - Automatic emotion recognition and classification, Procedia Engineering **38** (2012) 21-26.
3. Rennie J. D. M. - Improving multi-class text classification with Naive Bayes, Massachusetts Institute of Technology, Cambridge (2001).
4. Dai W., Xue G., Yang Q., and Yu Y. - Transferring Naive Bayes classifiers for text classification, In Association for the Advancement of Artificial Intelligence (AAAI), (2007) 540-545.
5. Frank E. and Bouckaert R. R. - Naive Bayes for text classification with unbalanced classes, In European Conference on Principles of Data Mining and Knowledge Discovery (2006) 503–510.
6. Hovold J. - Naive Bayes spam filtering using word-position-based attributes, The Common European Asylum System (CEAS) (2005).
7. Soelistio Y. E., Raditia M., and Surendra S. - Simple text mining for sentiment analysis of political figure using naive bayes classifier method, arXiv preprint arXiv, (2015) 99–104.
8. Pang B. and Lee L. - A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics (2004) 271.
9. Cunningham P. and Delany S. J. - k-Nearest Neighbour Classifiers, Multiple Classifier Systems (2007) 1–17.
10. Zhang M. and Zhou Z. - A k-Nearest Neighbor based algorithm for Multi-label classification, Granular Computing (GrC) (2005) 718–721.
11. Dharmadhikari S. C., Ingle Maya, and Kulkarni P. - Empirical Studies on machine learning based text classification algorithms, Advanced Computing (2011) 161–169.
12. Campbell C., Ying Y. - Learning with support vector machines, Synthesis lectures on artificial intelligence and machine learning (2011) 1–95.

13. Lê Hồng Phương - Vietnamese Word Tokenizer, 2018 (http://mim.hus.vnu.edu.vn/dsl/tools/tokenizer).

14. Hall M., Frank E., Holmes G., Pfahringer B., and Reutemann P. - The WEKA data mining software: An Update, ACM SIGKDD explorations Newsletter (2009) 11-17.

ABSTRACT

AUTOMATICALLY VIETNAMESE TEXT CLASSIFICATION BY TOPIC

Manh Thien Ly*, Vu Van Vinh, Nguyen Van Le,
Lam Thi Hoa Mi, Nguyen Thi Thanh Thuy, Duong Thi Mong Thuy
Ho Chi Minh City University of Food Industry
*Email: *lymt@hufi.edu.vn*

The Internet is strongly growing every day with a huge amount of information. The need of data mining and knowledge discovery is also increasing, in which the text classification plays an important role. Many techniques in machine learning are applied in classification process and achieved good results. Nowadays, there are many algorithms used for text classification such as: Naïve Bayes, K-NN, SVM, Maximum Entropy, etc. In this paper, Naïve Bayes, SVM and K-NN algorithms were used to experiment on Vietnamese text classification with 05 datasets belonging to 4 different topics: Tourism, Entertainment, Education and the Law. These datasets were extracted from vnexpress.net website. Some unique identifiers were applied during processing to increase the classification accuracy. The results show that SVM algorithm has the highest accuracy (over 90%) and the lowest amount of execution time.

Keywords: Text classification, Naïve Bayes, K-NN, SVM, algorithm.