Hướng tiếp cận không toàn văn cho bài toán phân lớp tự động bản tin tiếng Việt

Trương Quốc Định
Khoa CNTT&TT
Trường Đại học Cần Thơ
Cần Thơ, Việt Nam
tqdinh@cit.ctu.edu.vn

Trần Thị Thúy
Khoa Kỹ thuật Công nghệ
Trường Đại học Cửu Long
Vĩnh Long, Việt Nam
tranthithuy@mku.edu.vn

Trần Thị Cẩm Tú
Khoa Kỹ thuật Công nghệ
Trường Đại học Cửu Long
Vĩnh Long, Việt Nam
tranthicamtu@mku.edu.vn

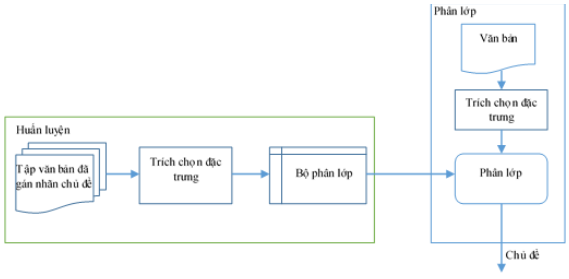
Huỳnh Kim Quýt
Khoa CNTT
Trường Đại học Tiền Giang
Tiền Giang, Việt Nam
huynkimquyt@tgu.edu.vn

***Tóm tắt*—Trong bài báo này chúng tôi trình bày một hướng tiếp cận phân lớp các bản tin tiếng Việt mà không dựa trên nội dung toàn văn của bản tin đó. Chúng tôi đề xuất sử dụng một trong hai thông tin: 1- tóm tắt; 2- từ khóa đại diện, trong đó tóm tắt và từ khóa đại diện được tạo tự động từ nội dung của văn bản, để phân lớp văn bản. Chúng tôi sử dụng tổng cộng 2000 bản tin được tải về từ các trang báo điện tử như vnexpress.net, vietnamnet.vn để kiểm thử giải pháp đề xuất. Kết quả thực nghiệm cho thấy hướng tiếp cận không toàn văn cho bài toán phân lớp văn bản là khả thi và có thể cải tiến để ứng dụng thực tế.**

Từ khóa: phân loại văn bản; tóm tắt tự động; mô hình chủ đề; cây quyết định.

I. GIỚI THIỆU

Bài toán phân loại văn bản (text classification) là bài toán cơ bản của lĩnh vực khai phá văn bản (text mining). Phân loại văn bản chính là gán nhãn (lớp/chủ đề) một cách tự động dựa vào nội dung của văn bản. Phân loại văn bản được ứng dụng trong nhiều lĩnh vực như tìm kiếm thông tin, lọc văn bản, tổng hợp tin tức tự động, thư viện điện tử.



Hình 1. Phân lớp văn bản

Bài toán phân loại văn bản có thể được định nghĩa như sau. Từ một tập các văn bản $D = \{d_1, d_2, \ldots, d_n\}$, được gọi là tập huấn luyện, trong đó các tài liệu d_i được gán nhãn chủ đề c_i với c_i thuộc tập các chủ đề $C = \{c_1, c_2, \ldots, c_n\}$ để xây dựng bộ phân lớp. Nhiệm vụ của bộ phân lớp là gán đúng nhãn chủ đề c_k cho một tài liệu mới d_k bất kỳ, trong đó c_k thuộc vào tập chủ đề C. Hình 1 mô tả bài toán phân lớp văn bản một cách tổng quát.

Bài toán phân lớp văn bản đã thu hút được nhiều nghiên cứu và đạt được nhiều thành công đặc biệt là đối với ngôn ngữ tiếng Anh. Văn bản có thể được phân loại dựa trên nhiều hướng tiếp cận khác nhau ví dụ như kỹ thuật máy học, lý thuyết tập thô hoặc luật kết hợp. Trong số các hướng tiếp cận trên thì hướng tiếp cận sử dụng máy học như là bộ phân lớp thu hút được nhiều nghiên cứu nhất và cho kết quả khả quan. Một số kỹ thuật thường được sử dụng là: naïve bayes, cây quyết định, k láng giềng gần nhất, mạng nơ-ron và máy học vec-tơ hỗ trợ. Phương pháp k láng giềng gần nhất được sử dụng trong nhiều miền ứng dụng vì tính đơn giản trong cài đặt nhưng lại có hiệu năng tốt. [1] đề xuất mô hình k láng giềng hiệu chỉnh trọng số cho bài toán phân lớp văn bản cho kết quả khả quan. Tương tự thì kỹ thuật Naïve bayes cũng được sử dụng nhiều vì tính đơn giản của nó trong tính toán và cài đặt. [2] đã đề xuất 2 độ đo (metric) cho bài toán phân lớp đa chủ đề. Cây quyết định cũng được sử dụng cho bài toán phân lớp văn bản trong đó các nút trong sẽ là các từ và các nút lá sẽ là các nhãn chủ đề. [3] đề xuất một cải tiến của mô hình cây quyết định áp dụng cho bài toán phân lớp trong đó văn bản có thể thuộc vào nhiều chủ đề khác nhau. [4] đề xuất mô hình mạng nơ-ron hồi quy cải tiến (MBPNN) cho bài toán phân lớp văn bản. Máy học vec-tơ hỗ trợ (SVM) ứng dụng cho bài toán phân lớp văn bản được đề xuất lần đầu tiên trong [5]. Bên cạnh đó có thể kết hợp máy học vec-tơ hỗ trợ với xích markov (HMM) để nâng cao hiệu quả của bộ phân lớp. [6] đề xuất sử dụng HMMs cho giai đoạn trích chọn đặc trưng và sau đó các vec-tơ đặc trưng mới sau khi đã chuẩn hóa là đầu vào cho bộ phân lớp SVM.

Các nghiên cứu trong nước về phân loại văn bản tiếng Việt cũng có được nhiều kết quả khả quan trong

đó có thể liệt kê một số công trình như sau, chủ yếu tập trung vào hướng tiếp cận sử dụng nội dung toàn văn của văn bản. Các hướng tiếp cận chủ yếu là học không giám sát và chỉ mục [15], sử dụng lý thuyết tập thô [16] hoặc cách tiếp cận thống kê [17]. Thời gian gần đây, các nghiên cứu về phân loại văn bản tiếng Việt tập trung vào các kỹ thuật cải tiến để phù hợp với ngữ cảnh ngôn ngữ tiếng Việt. [18] đề xuất sử dụng mô hình từ khóa chủ đề kết hợp naïve bayes cho mục tiêu giảm số lượng đặc trưng và phân lớp hiệu quả. [7] đề xuất giải pháp biểu diễn văn bản tiếng Việt dựa trên âm tiết. Phương pháp biểu diễn mới này được thực nghiệm với 6 thuật toán phân lớp để kiểm chứng tính khả thi và đều cho kết quả khả quan. [8] đề xuất cách đánh trọng số normalize($tf.rf_{max}$) cho từ chỉ mục trong ngữ cảnh bài toán phân lớp văn bản. Thực nghiệm cho thấy kết quả phân lớp được nâng cao tối đa là 5% so với các mô hình đánh trọng số truyền thống. [9] đề xuất sử dụng kỹ thuật SVM và Naïve bayes để xây dựng bộ phân lớp áp dụng cho bài toán phân lớp tự động các bản tin trên các trang tin điện tử. Kết quả thực nghiệm trên hơn 1000 bản tin cho thấy giải pháp đề xuất là khả thi.

Đối với bài toán phân loại đối tượng nói chung và bài toán phân loại văn bản nói riêng, giai đoạn trích chọn đặc trưng là quan trọng. Đại đa số các công trình vừa nêu sử dụng toàn văn nội dung của văn bản cho giai đoạn trích chọn đặc trưng, điều này có thể là nguyên nhân của 2 hạn chế: (1) số lượng đặc trưng lớn sẽ dẫn đến độ phức tạp cao, (2) khi số lượng đặc trưng quá lớn có thể sẽ chứa nhiễu dẫn đến độ chính xác của giai đoạn phân lớp bị hạn chế.

Trong phạm vi của nghiên cứu này, chúng tôi đề xuất giảm số chiều của đặc trưng bằng 2 giải pháp: (1) tạo tóm tắt tự động cho văn bản, (2) rút trích danh sách từ khóa đại diện cho văn bản. Với mỗi giải pháp chúng tôi đối chiếu kết quả với giải pháp truyền thống (không giảm chiều đặc trưng) và sử dụng cây quyết định cho bộ phân lớp. Nội dung còn lại của bài báo được tổ chức như sau: phần 2 giới thiệu các kỹ thuật có liên quan để giải quyết bài toán phân lớp theo hướng tiếp cận đề xuất, phần 3 trình bày giải pháp thực nghiệm và thảo luận, phần cuối là kết luận và đề xuất hướng nghiên cứu tiếp theo.

II. MÔ HÌNH ĐỀ XUẤT

A. Biểu diễn văn bản

Văn bản đầu vào cho việc huấn luyện và phân lớp có cấu trúc plain text. Chúng tôi sử dụng mô hình túi từ (BoW - Bag of Words) để biểu diễn văn bản. Mô hình này chỉ quan tâm đến trọng số một từ chỉ mục nào đó trong văn bản mà không quan tâm đến vị trí xuất hiện của từ chỉ mục đó. Đối với mô hình túi từ, hai công việc cần phải giải quyết đó là tách từ và gán trọng số.

Tiếng Việt có đặc điểm là từ có thể là từ đơn hoặc từ ghép vì thế khoảng trắng không còn là dấu hiệu nhận biết các từ. Việc phân tách một câu thành tập hợp đúng các từ có nghĩa là hết sức quan trọng đối với các bài toán thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Chúng tôi sử dụng thư viện vnTokenizer [10] cho giai đoạn tách từ với độ chính xác tách đúng từ theo công bố của tác giả là trong khoảng từ 96% đến 98%. Ví dụ sau đây minh họa kết quả của giai đoạn tách từ:

- Văn bản nguồn: "*Để có thể thực hiện rút trích tự động tóm tắt cũng như phân lớp văn bản với máy học vectơ hỗ trợ thì văn bản cần được biểu diễn dưới dạng thích hợp*".
- Văn bản sau giai đoạn tách từ: "*Để có_thể thực_hiện rút trích tự_động tóm_tắt cũng_như phân lớp văn_bản với máy học vectơ hỗ_trợ thì văn_bản cần được biểu_diễn dưới dạng thích_hợp*". Trong đó các từ có dấu "_" kết nối là các từ ghép.

Việc giảm chiều đặc trưng với giải pháp tạo tóm tắt tự động và rút trích danh sách từ khóa đại diện được thực hiện trên đơn văn bản vì thế giải pháp khả thi cho trọng số của từ trong văn bản là tần suất xuất hiện của từ trong văn bản đó.

B. Tóm tắt văn bản tiếng Việt tự động

Giải pháp tóm tắt tự động văn bản tiếng Việt được chúng tôi đề xuất trong [19] dựa trên khái niệm độ tương tự giữa các câu. Giá trị thông tin của mỗi câu trong văn bản được tính dựa trên giải thuật PageRank cải tiến. Các câu có giá trị thông tin cao là các câu được đưa vào tóm tắt, số lượng câu của tóm tắt do người dùng quyết định. Các bước thực hiện chính như sau:

- Biểu diễn câu trong không gian vec-tơ các từ chỉ mục.
- Xây dựng đồ thị trong đó mỗi đỉnh của đồ thị tương ứng với một câu của văn bản. Cung nối giữa hai đỉnh có trọng số là độ tương tự giữa hai câu.
- Thuật toán PageRank cải tiến được sử dụng để tính giá trị thông tin của mỗi đỉnh.
- Các câu được sắp xếp theo thứ tự giảm dần của giá trị thông tin.
- Một tỷ lệ nhất định (tham số đầu vào) các câu có giá trị thông tin cao nhất được trả về như tóm tắt.

Ví dụ sau đây minh họa kết quả là tóm tắt của một bản tin được hệ thống tạo tự động: "*Nhiều nhân viên bán hàng bảo hiểm tại Nhật Bản sẽ được chuyển từ máy tính cũ lên tablet chạy Windows 8 để tương tác tốt hơn với khách hàng. Microsoft tại Nhật Bản hôm nay thông báo đang giúp một công ty bảo hiểm lớn của Nhật Bản là Meiji Yasuda nhằm nâng cấp hàng loạt máy tính chạy hệ điều hành sắp tròn 12 tuổi.*

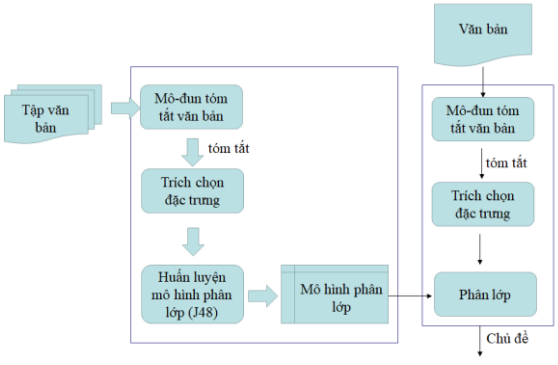
Trước đây, đội ngũ bán hàng sẽ chuẩn bị các đề xuất trên máy tính chạy Windows XP và sau đó in ra để chia sẻ với các khách hàng. Tuy nhiên, hệ thống thiết bị mới sẽ giúp chấm dứt các bước làm phiền toái này, thông báo của Microsoft có đoạn."

C. *Rút trích danh sách từ khóa đại diện cho nội dung của văn bản*

Rút trích từ khóa tự động (keywords extraction) từ một văn bản được ứng dụng trong nhiều lĩnh vực khác nhau như: tìm kiếm văn bản, tìm kiếm web, gom nhóm văn bản … Nhiều nghiên cứu đã khẳng định danh sách các từ khóa thích hợp có thể đại diện được cho thông tin cốt lõi của văn bản [11]. Trong phạm vi của nghiên cứu này, chúng tôi dựa trên phương pháp được đề xuất trong [12] và điều chỉnh một số bước để phù hợp với ngữ cảnh văn bản tiếng Việt. Các bước thực hiện chính bao gồm:

- Tiền xử lý: Sử dụng vnTokenizer để tách từ, loại bỏ các từ dừng (stop words).
- Giữ lại tất cả các từ được sinh ra ở bước 1 (nội dung bản tin ngắn nên mỗi một từ không có tần suất xuất hiện lớn).
- Gom cụm: với mỗi một 2 từ bất kỳ, tính giá trị khoảng cách Jensen-Shannon $J(w_1, w_2)$, giá trị xuất hiện cùng nhau $M(w_1, w_2)$. 2 từ w_1, w_2 sẽ thuộc cùng một nhóm nếu như $J(w_2, w_2) >= (0.95 \times \log2)$ hoặc $M(w_1, w_2) >= \log2$.
- Với mỗi từ w, tính $\chi^2(w)$, giá trị này thể hiện rằng từ w là quan trọng như thế nào với các từ thuộc cùng nhóm và phân biệt như thế nào với các nhóm từ khác.
- Trả về N từ có giá trị χ^2 cao nhất như là các từ đại diện.

D. *Phân lớp văn bản dựa trên tóm tắt của văn bản*

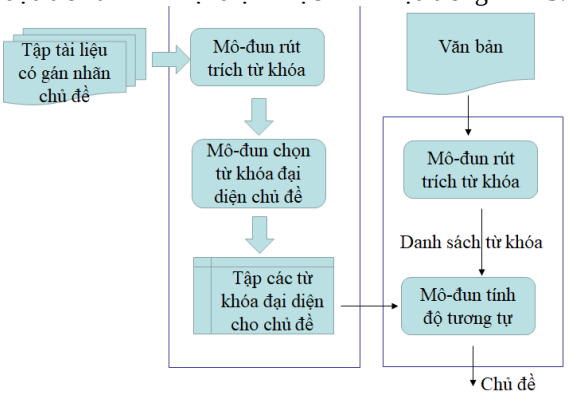


Hình 2. Mô hình phân lớp văn bản dựa trên tóm tắt

Mô hình tổng quan cho hệ thống phân lớp văn bản dựa trên tóm tắt được minh họa trong hình 2. Hệ thống đề xuất bao gồm hai thành phần chính: thành phần huấn luyện và thành phần phân lớp. Văn bản đầu vào được đưa qua mô-đun tạo tóm tắt trước khi đưa vào thành phần huấn luyện mô hình phân lớp. Thành phần phân lớp cũng có cách xử lý tương tự với kết quả là chủ đề của văn bản cần phân lớp. Trong phạm vi nghiên cứu này chúng tôi sử dụng cây quyết định J48 cung cấp bởi công cụ WEKA [13] để xây dựng bộ phân lớp. Số lượng câu được chọn đưa vào tóm tắt là 15% tổng số câu của văn bản.

E. *Phân lớp văn bản dựa trên từ khóa đại diện*

Mô hình tổng quan cho hệ thống phân lớp văn bản dựa trên từ khóa đại diện được minh họa trong hình 3.



Hình 3. Phân lớp văn bản dựa trên rút trích từ khóa đại diện

Mô-đun rút trích từ khóa sẽ nhận vào là một văn bản và trả về kết quả là N (trong đó N là tham số) từ khóa đại diện cho nội dung của văn bản đó. Tập tài liệu có gán nhãn chủ đề sẽ được sử dụng để tạo tập từ khóa đại diện cho mỗi chủ đề. Quá trình tạo lập từ khóa đại diện cho một chủ đề được tóm lược qua các bước chính như sau:

- Duyệt qua các tập tin văn bản có nhãn chủ đề là chủ đề cần tạo tập từ khóa đại diện. Với mỗi văn bản, rút trích đúng N từ khóa đại diện cho văn bản đó với N là số lượng từ khóa đại diện cho chủ đề.
- Tổng hợp danh sách các từ khóa được trả về ở bước 1, trong đó mỗi từ khóa sẽ có thêm thông tin đó là số lượt mà từ khóa đó được trả về. Các từ khóa được xếp theo thứ tự giảm dần của số lượt trả về.
- Trả về N từ khóa đầu danh sách có được ở bước 2 như là N từ khóa đại diện cho chủ đề.

Như vậy với mỗi chủ đề, chúng tôi xác định được N từ khóa đại diện. Các từ khóa này có trọng số giống nhau và có thể là các từ loại khác nhau.

Để xác định chủ đề của văn bản mới, văn bản này cũng được rút trích N từ khóa đại diện cho nội dung. Việc tiếp theo đó là xác định sự tương đồng giữa tập từ khóa đại diện văn bản mới với mỗi một tập từ khóa đại diện cho các chủ đề. Giá trị tương đồng lớn nhất ứng với tập từ khóa của chủ đề nào thì văn bản mới

thuộc vào chủ đề đó. Trong phạm vi của nghiên cứu này, chúng tôi đề xuất sử dụng độ đo Jaccard [14] để xác định độ tương đồng giữa hai tập hợp. Lý do chúng tôi chọn độ đo Jaccard là vì tập từ khóa đại diện cho văn bản và tập từ khóa đại diện cho chủ đề đơn thuần chỉ là tập các phần tử không có trọng số. Độ đo Jaccard được định nghĩa như sau: $J(A, B)=\frac{(A\cap B)}{(A\cup B)}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$

III. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

A. Dữ liệu thực nghiệm

Theo hiểu biết của chúng tôi thì với lĩnh vực phân lớp văn bản tiếng Việt chưa có bất kỳ một tập tài liệu kiểm thử chuẩn nào được công bố. Để đánh giá cho giải pháp đề xuất, chúng tôi đã tải về 2000 bản tin từ các trang báo điện tử vnexpress.net và vietnamnet.vn. Các tài liệu này được chia đều trong 10 chủ đề, cụ thể như bảng I.

BẢNG I. TẬP DỮ LIỆU KIỂM THỬ

Chủ đề	Số lượng tài liệu	Kích thước (MB)
Vi tính	200	6.69
Kinh doanh	200	7.62
Làm đẹp	200	6.80
Giáo dục	200	7.34
Sức khỏe	200	7.00
Thể thao	200	7.06
Khoa học	200	6.94
Du lịch	200	7.25
Gia đình	200	7.70
Ẩm thực	200	7.06

Đối với phương pháp mà chúng tôi đề xuất thì số lượng đặc trưng sử dụng cho bộ phân lớp chắc chắn sẽ giảm rất nhiều so với cách sử dụng nội dung toàn văn. Tuy nhiên thời gian thực hiện phân lớp cũng là một vấn đề cần được quan tâm vì đây là giai đoạn thực hiện online. Thông tin về số lượng đặc trưng trung bình, thời gian tạo tóm tắt trung bình, thời gian rút trích từ khóa đại diện trung bình được cho ở bảng II. Các thông số này được ghi nhận khi thực nghiệm trên máy tính cá nhân Asus X202E, CORE i3, 4GB RAM, WINDOWS 8.1.

BẢNG II. ĐẶC TÍNH CÁC GIẢI PHÁP ĐỀ XUẤT

Số lượng đặc trưng trung bình			Thời gian thực hiện trung bình (giây)	
Toàn văn	*Tóm tắt*	*Từ khóa*	*Tóm tắt*	*Từ khóa*
462	123	30	1.4	1.2

B. Đánh giá kết quả

Dù là hướng tiếp cận nào đi nữa thì chúng tôi cũng sử đụng 2/3 tập tài liệu cho giai đoạn huấn luyện và 1/3 tập dữ liệu còn lại cho kiểm thử.

Đối với giải pháp rút trích từ khóa đại điện, trong phạm vi của nghiên cứu này, với mỗi chủ đề chúng tôi chỉ chọn 30 từ khóa làm đại diện, một trong số các lý do là vì các bản tin tải về từ các trang báo điện tử có nội dung không nhiều. Bảng III sau đây mô tả danh sách 30 từ khóa đại diện cho mỗi chủ đề.

BẢNG III. DANH SÁCH CÁC TỪ KHÓA ĐẠI DIỆN CHỦ ĐỀ

Chủ đề	Từ khóa
Vi tính	dùng; sản phẩm; máy; triệu; điện thoại; màn hình; bán; chip; thiết bị; việc làm; apple; samsung; giá; máy tính; microsoft; usd; hãng; chạy; công nghệ; gb; đồng; hd; tablet; mỹ; công ty; lõi; thế giới; smartphone; so sánh.
Kinh doanh	giá; đồng; công ty; triệu; giảm; tháng; mức; usd; tăng; tỷ giá; bán; việc làm; cao; số; doanh nghiệp; lớn; khoảng; thị trường; cho biết; đầu tư; nước; việt nam; đây; chưa; ngân hàng; tới; hà nội; hàng hóa; thế giới; sáng.
Làm đẹp	da; làm; giúp; công nghệ; vùng; phương pháp; điều trị; hiệu quả; làn da; mỡ; hay; cơ thể; sử dụng; gây mê; sản phẩm; bác sĩ; giảm; việc làm; cần; quá trình; đau; nhỏ; tạo; lông; phẫu thuật; rf; khoảng; ánh sáng; triệt; đẹp.
Giáo dục	thi; thí sinh; thpt; học sinh; gd&đt; trường; sở; bắc giang; tốt nghiệp; clip; làm; sinh viên; môn; thanh tra; ném; phòng; quay; hội đồng; việc; tỉnh; giải; tổ chức; đại học; kỳ; tỷ lệ; điểm; nói; chưa; xếp loại; cho biết.
Sức khỏe	bệnh; bác sĩ; phát hiện; cho biết; đây; nghiên cứu; cao; y tế; bé; chị; điều trị; bệnh viện; khám; giảm; thấy; việc; trẻ; nguy cơ; giúp; đốt; cơ thể; tp hcm; bệnh nhân; trung quốc; phòng khám; tuổi; sở; kiểm tra; loại; tăng.
Thể thao	trận; đấu; cầu thủ; đội; thắng; hlv; việt nam; anh; bóng; tuyển; chơi; tới; giải; sân; euro; bảng; nhà; phút; tốt; đội tuyển; tây ban nha; mùa; thua; chiến thắng; phan thanh hùng; vòng; lần; qua; việc; nói.
Khoa học	khả năng; tới; mỹ; sử dụng; nghiên cứu; công nghệ; đưa; loại; thiết bị; nhóm; tạo; công ty; chế tạo; robot; sản xuất; điện tử; điện; hoạt động; thử nghiệm; đại học; giúp; pin; đường; bay; cao; chuyên gia; cơ thể; tin; máy bay; chống.
Du lịch	du khách; du lịch; phòng; giá; khu vực; khách; đồng; biển; đây; chương trình; việt nam; nước; thành phố; hay; hà nội; khách sạn; đà nẵng; đêm; tp hcm; hạ long; nơi; giữa; qua; điểm; thế giới; thư giãn; dịch vụ; nghỉ ngơi; thiên nhiên; vé.
Gia đình	mình; làm; chồng; biết; nhà; vợ; thấy; anh; gia đình; mẹ; lần; gì; nói; việc; em; lúc; chị; bà; trẻ; tuổi; chuyện; cần; muốn; cách; khác; vợ chồng; học; đàn ông; bố mẹ; con cái.
Ẩm thực	món; ăn; nhà hàng; ngon; thịt; thực khách; nướng; loại; nước; chế biến; thưởng thức; hương vị; thơm; mang; vừa; dùng; đây; tươi; gia vị; thành; đồng; việt nam; vị; bếp; làm; buffet; màu; khoảng; nguyên liệu; phong cách.

Bảng IV cho thấy giải pháp mà chúng tôi đề xuất là khả thi, đặc biệt là giải pháp dựa trên tóm tắt.

BẢNG IV. KẾT QUẢ THỰC NGHIỆM TRÊN 10 CHỦ ĐỀ

Chủ đề	Phân lớp dựa trên tóm tắt (J48)	Phân lớp dựa trên từ khóa	Phân lớp dùng nội dung toàn văn (J48)
Vi tính	84.5%	84%	79%
Kinh doanh	72.9%	88%	66.5%
Làm đẹp	83.5%	94%	65%
Giáo dục	85.9%	82%	86.5%
Sức khỏe	77.5%	62%	63.5%
Thể thao	92%	82%	83.5%
Khoa học	84.5%	78%	70.9%
Du lịch	83%	72%	62%
Gia đình	75.5%	60%	74.7%
Ẩm thực	85%	86%	84%
Trung bình	**82.4%**	**79%**	**73.6%**

Chúng ta có thể dễ dàng nhận thấy rằng về độ chính xác trung bình thì cả 2 giải pháp mà chúng tôi đề xuất đều vượt trội so với phương pháp truyền thống. Nếu xét từng chủ đề thì giải pháp mà chúng tôi đề xuất chỉ thua giải pháp truyền thống ở chủ đề giáo dục, sức khỏe, thể thao, gia đình cho trường hợp đề xuất dựa trên từ khóa trong khi đó giải pháp dựa trên tóm tắt đều vượt so với giải pháp truyền thống.

IV. KẾT LUẬN

Trong bài báo này chúng tôi giới thiệu mô hình phân lớp văn bản không dựa trên nội dung toàn văn của văn bản. Đây là một hướng tiếp cận mới và chưa có nhiều nghiên cứu trên thế giới cũng như ở Việt Nam vì đại bộ phận đều cho rằng khi thực hiện tóm tắt văn bản thì thông tin dùng cho phân lớp đã mất đi khá nhiều. Kết quả thực nghiệm cho thấy giải pháp mà chúng tôi đề xuất có thể giảm đáng kể số đặc trưng cho bộ phân lớp từ đó có thể giảm được độ phức tạp của hệ thống phân lớp. Kết quả mà chúng tôi thu được từ nghiên cứu này là hết sức khả quan và thiết nghĩ là hoàn toàn khả thi khi ứng dụng vào thực tế.

Kết quả khả quan của mô hình dựa trên tóm tắt có thể được lý giải bởi nhiều nguyên nhân: 1- Tóm tắt của một văn bản về lý thuyết sẽ tóm lược được nội dung cốt lõi truyền tải bởi văn bản. Một khi đã tóm lược được nội dung chính thì chủ đề của văn bản hoàn toàn có thể xác định được. 2- Cách thức biểu diễn văn bản đã thể hiện tốt nội dung, ngữ nghĩa của văn bản. Thật vậy, trong nghiên cứu của mình, chúng tôi dựa trên "mô hình túi từ - bag of words" để biểu diễn nội dung văn bản, phương pháp này có ưu điểm là cài đặt đơn giản nhưng có hạn chế lớn là làm mất đi ngữ nghĩa của văn bản vì không quan tâm đến vị trí của từ mà chỉ quan tâm đến tần suất xuất hiện của từ. Việc sử dụng thư viện vnTokenizer có khả năng nhận biết chính xác từ đơn và từ ghép đồng thời việc tạo tóm tắt được thực hiện trên mức câu nên đã giúp giữ lại phần nào ngữ nghĩa của văn bản; 3- Mô hình tóm tắt tự động văn bản mà chúng tôi đề xuất trong nghiên cứu trước đây thật sự là khả thi. Điểm mấu chốt của bài toán tóm tắt là tính độ tương tự giữa các câu và tính điểm xếp hạng các câu dựa trên mô hình đồ thị. Độ tương tự giữa các câu được tính thông qua độ đo Jaccard có chú trọng đến mối tương quan về độ dài của các câu. Thuật toán PageRank dùng để tính điểm xếp hạng các câu đưa vào tóm tắt là thuật toán xếp hạng các trang web và đã chứng tỏ được tính khả thi khi được ứng dụng thành công trong các bộ máy tìm kiếm thông tin web. Một ưu điểm khác của mô hình tóm tắt tự động đó là quá trình tóm tắt không cần tập ngữ liệu huấn luyện, cũng như không cần xem xét tính ngữ nghĩa và cấu trúc ngữ pháp của câu và việc tóm tắt được áp dụng trên từng văn bản đơn.

Với mô hình phân lớp dựa trên từ khóa thì kết quả bước đầu cũng thể hiện tính khả thi của giải pháp đề xuất, tuy nhiên cũng bộ lộ một số điểm cần cải tiến. Trước tiên đối với mô-đun rút trích từ khóa đại diện cho văn bản, trong phạm vi nghiên cứu này chúng tôi chưa quan tâm đến từ loại của từ khóa mà chỉ xử lý loại bỏ từ dừng (stop words) ở giai đoạn tiền xử lý, điều này có thể dẫn đến nhiễu khi xây dựng tập từ khóa đại diện cho chủ đề. Thiết nghĩ các từ loại có thể dùng để đại diện cho chủ đề đó là danh từ, động từ và tính từ. Bên cạnh đó, khi xây dựng tập từ khóa đại diện cho chủ đề, chúng tôi đã sử dụng đồng nhất một trọng số cho tất cả các từ và vì thế khi so khớp sự trùng lắp giữa hai tập từ khóa, chúng tôi chỉ có thể sử dụng độ đo Jaccard, điều này dẫn đến xác định sai chủ đề cho văn bản khi từ khóa đại diện cho một văn bản có có thể thuộc vào cùng lúc nhiều chủ đề (số từ khóa đại diện cho chủ đề nhỏ, chỉ là 30, và chưa được gán trọng số). Tập dữ liệu dùng cho huấn luyện và kiểm thử chưa đủ lớn và có thời gian xuất bản nằm trong khoảng thời gian ngắn nên chưa có tính đại diện. Nhãn chủ đề bản tin là chủ đề của các trang báo điện tử vì thế đôi khi cũng không thật chính xác dẫn đến nhiễu trong việc xây dựng từ khóa đại diện cho mỗi chủ đề.

Mặc dù kết quả nghiên cứu bước đầu đã khẳng định mô hình đề xuất phân lớp văn bản không dựa vào nội dung toàn văn là hoàn toàn khả thi và hoàn toàn có thể áp dụng vào thực tế, tuy nhiên kết quả ấy cũng chỉ được thực nghiệm trên một tập chưa đủ lớn các tài liệu và cũng chỉ mới kiểm thử với phương pháp phân lớp là cây quyết định. Chúng tôi thiết nghĩ giải pháp dựa trên từ khóa có thể có kết quả tốt hơn nếu như chỉ giữ lại các loại từ là danh từ, động từ và tính từ. Hơn nữa thay vì đồng hóa trọng số cho tất cả các từ khóa thì sẽ tốt hơn nếu mỗi từ khóa biểu diễn cho một chủ đề với trọng số khác nhau. Khi đó các độ đo tương đồng khác có tính đến trọng số của các phần tử (ví dụ như cosine) sẽ là phù hợp hơn so với độ đo Jaccard.

Một giải pháp khả dĩ cần được kiểm chứng trong nghiên cứu tiếp theo đó là kết hợp rút trích từ khóa đại diện trên tóm tắt của văn bản để giảm tiểu nhiễu đến mức tối thiểu. Và để kết quả nghiên cứu có tính thuyết phục hơn thì tập dữ liệu thực nghiệm cần có kích thước lớn hơn nữa (số lượng văn bản cũng như nội dung của mỗi văn bản).

TÀI LIỆU THAM KHẢO

[1] Fang Lu Qingyuan Bai, "A Refined Weighted K-Nearest Neighbours Algorithm for Text Categorization", IEEE 2010.

[2] Jingnian Chen, Houkuan Huang, Shengfeng Tian, Youli Qua, "Feature selection for text classification with Naïve", China Expert Systems with Applications, vol. 36, p. 5432–5435, 2009.

[3] Peerapon Vateekul and Miroslav Kubat, "Fast Induction of Multiple Decision Trees in Text Categorization From Large Scale,Imbalanced, and Multi-label Data", IEEE International Conference on Data Mining, 2009.

[4] Cheng Hua Li , Soon Choel Park "An efficient document classification model using an improved back propagation neural network and singular value decomposition", Expert Systems with Applications, 3208–3215, 2009.

[5] Joachims, T. "Text categorization with support vector machines: learning with many relevant features". In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE), pp. 137–142 1998.

[6] Chen donghui, Liu zhijing, "A new text categorization method based on HMM and SVM", 2010 2nd Int. Conf. Comput. Eng. Technol., IEEE (2010).

[7] Giang-Son Nguyen, Xiaoying Gao, and Peter Andreae, "Vietnamese Document Representation and Classification". In Proceedings of the 22nd Australasian Joint Conference on Advances in Artificial Intelligence (AI '09), Ann Nicholson and Xiaodong Li (Eds.). Springer-Verlag, Berlin, Heidelberg, 577-586. DOI=http://dx.doi.org/10.1007/978-3-642-10439-8_58

[8] Vu Thanh Nguyen, Nguyen Tri Hai, Nguyen Hoang Nghia, and Tuan Dinh Le, "A Term Weighting Scheme Approach for Vietnamese Text Classification", In Proceedings of the Second International Conference on Future Data and Security Engineering - Volume 9446 (FDSE 2015), Tran Khanh Dang, Roland Wagner, Josef Küng, Nam Thoai, Makoto Takizawa, and Erich Neuhold (Eds.), Vol. 9446. Springer-Verlag New York, Inc., New York, NY, USA, 46-53. DOI: http://dx.doi.org/10.1007/978-3-319-26135-5_4

[9] Phan Thi Ha, Nguyen Quynh Chi, "Automatic Classification for Vietnamese News", Advances in Computer Science: an International Journal, Vol. 4, No. 4, p.126-135, 2015.

[10] Le Hong Phuong, Nguyen Thi Minh Huyen, Azim Roussanaly, Ho Tuong Vinh, "A Hybrid Approach to Word Segmentation of Vietnamese Texts", Language and Automata Theory and Applications: Second International Conference, LATA 2008, Tarragona, Spain, March 13-19, 2008.

[11] Blei, D., and Lafferty, J. 2009. "Topic models". In Srivastava, A., and Sahami, M., eds., Text Mining: Theory and Applications. Taylor and Francis.

[12] Matsuo, Y., Ishizuka, M., "Keyword extraction from a single document using word co-occurrence statistical information", Int. Journal on AI Tools 13(1), 157-169 (2004).

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009), "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1.

[14] Jaccard P., "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Société Vaudoise des Sciences Naturelles 37: 547–579.

[15] Huỳnh Quyết Thắng, Đinh Thị Phương Thu, "Tiếp cận phương pháp học không giám sát trong học có giám sát với bài toán phân lớp văn bản tiếng Việt và đề xuất cải tiến công thức tính độ liên quan giữa hai văn bản trong mô hình vector", Kỷ yếu Hội thảo ICT.rda'04, trang 251-261, Hà Nội 2005.

[16] Nguyễn Ngọc Bình, "Dùng lý thuyết tập thô và các kỹ thuật khác để phân loại, phân cụm văn bản tiếng Việt", Kỷ yếu hội thảo ICT.rda'04. Hà nội 2004.

[17] Nguyễn Linh Giang, Nguyễn Duy Hải, "Mô hình thống kê hình vị tiếng Việt và ứng dụng", Chuyên san "Các công trình nghiên cứu, triển khai Công nghệ Thông tin và Viễn thông, Tạp chí Bưu chính Viễn thông, số 1, tháng 7-1999, trang 61-67. 1999.

[18] Bùi Khánh Linh, Nguyễn Quỳnh Anh, Nguyễn Nhật An, Nguyễn Thị Thu Hà, Đào Thanh Tĩnh, "Phân loại văn bản tiếng việt dựa trên mô hình chủ đề và lý thuyết Naive Bayes", Tạp chí Nghiên cứu Khoa học Công nghệ quân sự, Số 37, tập 2, trang 89-95, 2015.

[19] Trương Quốc Định, Nguyễn Quang Dũng, "Một giải pháp tóm tắt văn bản tiếng Việt tự động", Kỷ yếu hội thảo khoa học quốc gia lần thứ XV, trang 233-238, Nhà xuất bản Khoa học.