Xác định đặc điểm tác giả bài viết diễn đàn tiếng Việt dựa trên âm tiết và vần

Syllables and Rhymes for Author Profiling of Vietnamese Forum Posts

Dương Trần Đức, Phạm Bảo Sơn, Tân Hạnh

***Abstract*:** *Author profiling is the task of identifying characteristics of the author just based on a text document. In the previous works, there are a number of linguistic features such as character-based, word-based, grammar-based (often grouped as style-based), and content-based features (content words) have been exploited. The previous results showed that content-based features often achieved better results than style-based features. However, using content-based features is considered as a domain-specific approach, because the content words chosen often have meaning related to the studied domain. In this work, we investigate the use of syllables and rhymes as features for author profiling of Vietnamese text. They are parts of words, but have much less meaning than words, especially the rhymes. Therefore, these features can be considered much less domain-dependent than content words. We experimented on forum post datasets using machine learning approach. With improvement up to 8% compared with baseline results on style-based features, our method shows a new promising approach on author profiling.*

***Keywords**: Author Profiling, Machine Learning, Nature Language Processing*

I. GIỚI THIỆU

Xác định đặc điểm tác giả văn bản (author profiling) là một nhánh nghiên cứu của phân tích tác giả văn bản. Phân tích tác giả văn bản còn có hai nhánh nghiên cứu khác là nhận diện tác giả (authorship attribution) và xác minh tác giả (author verification) [26]. Trong khi việc nhận diện tác giả hoặc xác minh tác giả tiến hành xác định hoặc kiểm chứng một tác giả cụ thể là người tạo nên văn bản và thường áp dụng cho các loại văn bản chính thống như bài báo, tiểu thuyết v.v… xác định đặc điểm tác giả văn bản thường được thực hiện trên các loại văn bản tự do hơn như các loại văn bản trực tuyến (bài viết blog, email, diễn đàn v.v…) [1, 2, 5, 20, 8, 12, 17, 26].

Do đó, các ứng dụng của xác định đặc điểm tác giả văn bản cũng khác so với hai nhánh nghiên cứu còn lại, vốn thường được sử dụng để giải quyết các tranh cãi về quyền tác giả. Ứng dụng chủ yếu của xác định đặc điểm tác giả là trong các lĩnh vực quảng cáo trực tuyến, cá nhân hóa hệ thống web, hỗ trợ điều tra tội phạm trực tuyến v.v… trong đó các đặc điểm cá nhân của tác giả bài viết được dự đoán để hỗ trợ các hoạt động quảng cáo đúng mục đích hoặc điều tra tội phạm.

Cùng với sự phát triển của Internet và các kênh trao đổi thông tin trực tuyến, ứng dụng của việc xác định đặc điểm tác giả văn bản càng trở nên cần thiết và quan trọng hơn. Để dự đoán đặc điểm như giới tính, độ tuổi, trình độ giáo dục v.v... của tác giả một văn bản, các nhà nghiên cứu thường phân tích phong cách viết (writing style) hoặc các từ nội dung (content words) được sử dụng bởi tác giả đó.

Phong cách viết được xem như là một phương pháp độc lập miền và được sử dụng trong nhiều nghiên cứu trước đây về xác định đặc điểm tác giả. Hầu hết các thành phần có tính độc lập nội dung của ngôn ngữ đã được sử dụng làm đặc trưng phong cách như các ký tự, tính chất từ, từ loại, từ công cụ (từ chức năng), các cấu trúc ngữ pháp v.v… Các đặc trưng này thường được tạo ra từ các quy tắc của ngôn ngữ và không phụ thuộc

vào tập dữ liệu hay lĩnh vực cụ thể nào. Ngược lại, các từ nội dung thường được lựa chọn từ chính các tập dữ liệu được sử dụng trong nghiên cứu hoặc được lựa chọn từ các từ ngữ có ngữ nghĩa liên quan đến lĩnh vực cụ thể. Do đó, các từ nội dung được xem là có tính phụ thuộc miền hoặc phụ thuộc dữ liệu ở mức độ nào đó.

Để giảm sự phụ thuộc dữ liệu và phụ thuộc miền của các từ nội dung, nghiên cứu này đề xuất một loại đặc trưng mới cho việc xác định đặc điểm tác giả văn bản tiếng Việt. Đó là các đặc trưng dựa trên các âm tiết và các vần trong tiếng Việt. Một từ trong tiếng Việt có thể gồm nhiều âm tiết, và các vần là một thành phần của âm tiết [7]. Do đó, các âm tiết và vần mang ít ngữ nghĩa hơn nhiều so với một từ hoàn chỉnh. Ví dụ, từ ghép "đồng hồ" được tạo ra bởi hai âm tiết là "đồng" và "hồ" và cả hai âm tiết này đều cần thiết cho việc xác định ngữ nghĩa của từ.

Từ ngữ nghĩa riêng rẽ của một âm tiết, trong nhiều trường hợp, không thể xác định được nghĩa của từ. Ngoài ra, do số lượng các âm tiết và vần trong tiếng Việt không lớn, chúng ta có thể sử dụng toàn bộ âm tiết và vần làm các đặc trưng để loại trừ tính phụ thuộc tập dữ liệu và phụ thuộc miền. Bên cạnh việc mang ít ngữ nghĩa hơn, đây là một khía cạnh quan trọng khác làm cho các âm tiết và vần có sự khác biệt so với các từ nội dung.

Các từ nội dung được lựa chọn từ tập dữ liệu, trong khi các âm tiết và vần có thể được xây dựng từ các quy tắc từ vựng và ngữ pháp mà không cần sử dụng tập dữ liệu, qua đó giảm tính phụ thuộc vào tập dữ liệu và phụ thuộc miền. Về khía cạnh hiệu năng xử lý, với khoảng 6.400 âm tiết và 450 vần được sử dụng làm đặc trưng phân loại, các thuật toán học máy phổ biến hiện nay như máy véc tơ hỗ trợ (Support Vector Machine - SVM) hoàn toàn có thể xử lý mà không gặp nhiều khó khăn.

Bài báo có cấu trúc như sau. Phần II trình bày tổng quan về các nghiên cứu liên quan trong lĩnh vực phân tích tác giả văn bản, Phần III đi sâu về âm tiết và vần trong tiếng Việt. Phần IV mô tả phương pháp. Phần V trình bày các kết quả và thảo luận. Cuối cùng, các kết luận sẽ được trình bày trong phần VI của bài báo.

II. TỔNG QUAN VỀ PHÂN TÍCH TÁC GIẢ

Phân tích tác giả văn bản là quá trình phân tích một tài liệu để có thể đưa ra các kết luận về tác giả của nó. Quá trình phân tích tác giả văn bản liên quan đến hai vấn đề chính, đó là kỹ thuật phân tích và tập đặc trưng phân biệt. Kỹ thuật phân tích trong thời kỳ đầu thường sử dụng các kỹ thuật khá đơn giản dựa trên thống kê [22].

Các nghiên cứu gần đây chủ yếu khai thác kỹ thuật học máy để tận dụng khả năng tính toán của máy tính. Mặc dù việc lựa chọn thuật toán học máy phù hợp là một vấn đề quan trọng, nghiên cứu của Koppel [15] cho thấy trong lĩnh vực phân tích tác giả văn bản, việc lựa chọn tập đặc trưng lại có tầm quan trọng cao hơn. Tập đặc trưng có thể được xem như một phương pháp biểu diễn văn bản trên khía cạnh phong cách viết hoặc cách sử dụng từ.

Theo Argamon [2], có hai loại đặc trưng chính được sử dụng trong phân tích tác giả văn bản: đặc trưng về phong cách và đặc trưng dựa trên nội dung. Đặc trưng về phong cách bao gồm các đặc trưng liên quan đến ký tự, tính chất từ (lexical), cách sử dụng các cấu trúc ngữ pháp (syntactic), và các đặc trưng về cấu trúc văn bản. Đặc trưng dựa trên nội dung bao gồm các từ nội dung được sử dụng thường xuyên trong lĩnh vực đó hơn là các lĩnh vực khác. Các từ này thường được chọn theo phương pháp thống kê tần suất xuất hiện trong tập dữ liệu hoặc dựa trên ngữ nghĩa của từ. Phần này trình bày khảo sát về các nghiên cứu trước đây trong lĩnh vực phân tích tác giả có sử dụng các đặc trưng liên quan tới từ vựng như các ký tự, cụm ký tự, các từ v.v...

Các đặc trưng dựa trên các thành phần của hệ thống từ vựng đã được chứng minh là có hữu ích trong việc xác định đặc điểm tác giả văn bản trong nhiều nghiên cứu trước đây. Từ các thành phần cơ bản như các ký tự riêng lẻ [4, 5, 12, 25, 26], các cụm ký tự n-grams [3, 11, 13, 19], đến các đặc điểm của từ như loại từ, mức độ đa dạng của từ vựng [5, 6, 12, 21], các từ công cụ

[2, 6, 9, 12, 14], và các từ nội dung [2, 8, 10, 17, 19, 26] đã được nghiên cứu sử dụng. Trong nghiên cứu đầu tiên được xem là hoàn chỉnh trong lĩnh vực này, Mosteller và Wallace (1964) sử dụng một số từ công cụ để giải quyết vấn đề tranh chấp trong việc xác định tác giả các bài luận liên bang (Federalist Papers).

Sau đó, có rất nhiều các nghiên cứu tiếp theo trong lĩnh vực phân tích tác giả văn bản đã khai thác và xác minh tính hữu ích của các từ công cụ trong lĩnh vực này với số các từ được sử dụng từ 122 đến 645 từ. Các từ công cụ mang ít ngữ nghĩa và được sử dụng để biểu thị mối quan hệ ngữ pháp giữa với các từ khác, do vậy chúng được xem là không liên quan đến nội dung và được xếp loại là dạng đặc trưng dựa theo phong cách.

Các đặc trưng dựa trên ký tự và đặc điểm từ như các ký tự đơn lẻ/cụm ký tự, độ dài từ, loại từ, mức độ đa dạng trong dùng từ cũng được sử dụng phổ biến. De Vel [5] sử dụng các đặc trưng như độ dài từ/câu, loại từ, tần suất các ký tự/loại ký tự, cùng với các đặc trưng ngữ pháp khác để phân biệt 156 emails trong tiếng Anh. Zheng và các đồng tác giả [25] đã sử dụng các đặc trưng dựa trên từ vựng được đề xuất bởi De Vel [5] và bổ sung thêm một số đặc trưng dựa trên nội dung để đề xuất một hệ thống phân tích tác giả văn bản nhằm tự động theo dõi tội phạm mạng dựa trên các bản tin được đăng trên mạng Internet. Các tác giả đánh giá hiệu quả của hệ thống thông qua việc thực hiện các thí nghiệm trên các tập dữ liệu các bản tin trực tuyến tiếng Anh và tiếng Trung Quốc.

Abbasi và Chen [1] sử dụng 79 đặc trưng từ vựng trong tổng số 418 đặc trưng để phân tích tác giả các bài viết diễn đàn tiếng Anh và tiếng Ả rập. Các tác giả sử dụng một tập đặc trưng hiệu quả dựa trên việc khai thác các đặc điểm về hình thái và chính tả tiếng Ả rập (chẳng hạn bổ sung thêm hai đặc trưng về phần kéo dài trong tiếng Arập). Iqbal và các đồng tác giả [12] đã sử dụng 419 đặc trưng bao gồm các đặc trưng dựa trên ký tự, đựa trên đặc điểm từ, đặc trưng ngữ pháp để xây dựng một loại "vân chữ viết" nhằm xác minh các tác giả email hỗ trợ điều tra tội phạm.

Một số nghiên cứu cũng sử dụng các cụm kết hợp ký tự (n-grams) để làm đặc trưng phân loại. Keselj và các đồng tác giả [13] đề xuất một phương pháp nhận diện tác giả dựa trên cụm ký tự. Hiệu quả của phương pháp được kiểm chứng thông qua các thí nghiệm trên các dữ liệu trên các ngôn ngữ tiếng Anh, tiếng Hy Lạp, và tiếng Trung Quốc.

Houvardas và Stamatatos [11] nghiên cứu phương pháp sử dụng các cụm ký tự có độ dài biến đổi để giải quyết vấn đề nhận diện tác giả trên các bản tin Reuters của 50 tác giả khác nhau. Ý tưởng chính của phương pháp này là so sánh mỗi cụm ký tự với các cụm ký tự tương đồng và giữ lại các cụm ký tự nổi trội hơn. Peersman và các đồng tác giả [19] dự đoán tuổi và giới tính của người dùng chat dựa trên các đoạn chat thu thập từ mạng xã hội Netlog tại Bỉ. Tác giả sử dụng các cụm ký tự và từ làm đặc trưng phân loại. Các cụm 1 từ, 2 từ, 3 từ, 4 từ và các cụm 2 ký tự, 3 ký tự, 4 ký tự được trích từ tập dữ liệu và sau đó được chọn lọc bởi thuật toán lựa chọn đặc trưng χ^2 (chi-square). Xét về khía cạnh phụ thuộc nội dung hoặc phụ thuộc miền, các cụm ký tự và cụm từ được tạo thành theo phương pháp trích chọn từ tập dữ liệu cũng có tính chất này do chúng được lấy ra từ dữ liệu và kể cả cụm ký tự vẫn có thể chứa nội dung nếu nó bao hàm một từ.

Bên cạnh các đặc trưng dựa trên các ký tự và từ không liên quan đến nội dung, các từ mang ngữ nghĩa cũng được khai thác và sử dụng làm đặc trưng phân loại trong lĩnh vực phân tích tác giả văn bản. Như đã nói ở trên, các đặc trưng dựa trên nội dung này thường mang lại kết quả tốt hơn so với các đặc trưng dựa trên phong cách. Tuy nhiên, loại đặc trưng này được xem là có tính đặc thù miền và có thể cho kết quả kém hơn khi áp dụng vào lĩnh vực khác. Mặc dù vậy, các đặc trưng này vẫn có rất nhiều ý nghĩa, chẳng hạn trong trường hợp áp dụng mô hình trong cùng lĩnh vực/lĩnh vực tương tự hoặc có thể được tổng quát hóa thông qua các phương pháp xử lý loại bỏ tính đặc thù miền.

Zheng et al. [26] sử dụng 11 từ khóa mang nội dung như "deal" (mặc cả), "sale" (bán hàng), "check" (kiểm tra), v.v... trong tổng số 270 đặc trưng để nhận

diện tác giả các bản tin trực tuyến tiếng Anh và tiếng Trung Quốc. Goswani và đồng tác giả [10] thực hiện trích các từ không có trong từ điển trong tập dữ liệu blog và sử dụng làm đặc trưng để phân loại các bài viết blogs theo các nhóm giới tính và độ tuổi.

Argamon và đồng tác giả [2] trích khoảng 1.000 từ nội dung có tần suất cao trong tập dữ liệu và có khả năng phân biệt các lớp tốt nhất, được xác định bằng độ đo độ lợi thông tin (Information Gain). Các từ này sau đó được sử dụng làm các đặc trưng để xác định giới tính, độ tuổi, ngôn ngữ gốc, và các mặt tính cách như tính cởi mở, hướng ngoại, dễ bị kích động, tận tâm, dễ hòa hợp của các tác giả bài viết blogs và bài luận của sinh viên. Iqbal [12] sử dụng 13 từ đặc thù trong lĩnh vực như các từ "agreement" (thỏa thuận), "team" (nhóm), "section" (phần), v.v... làm đặc trưng để khai phá ra các "vân chữ viết" của các tác giả email vô danh phục vụ cho việc điều tra tội phạm mạng.

Peersman [19] sử dụng các cụm từ đơn, ghép đôi, ghép ba để xác định đặc điểm tác giả các đoạn chat và các đặc trưng này có thể xem là đặc trưng dựa theo nội dung. Nguyen [17] sử dụng các từ đơn xuất hiện ít nhất 10 lần trong tài liệu huấn luyện làm đặc trưng phân loại trong nghiên cứu dự đoán tuổi của người dùng mạng xã hội Twitter sử dụng hồi quy tuyến tính. Duong [8] nghiên cứu việc sử dụng các từ nội dung để làm đặc trưng trong việc xác định đặc điểm tác giả bài viết diễn đàn tiếng Việt, qua đó dự đoán các đặc điểm như giới tính, độ tuổi, vùng miền, và nghề nghiệp. Các từ nội dùng xuất hiện tần suất cao trong mỗi lớp được trích ra từ tập dữ liệu và áp dụng phương pháp lựa chọn đặc trưng dựa trên độ lợi thông tin để chọn lọc ra các từ có thứ hạng cao nhất làm đặc trưng dựa trên nội dung.

Nhìn chung, hầu hết các nghiên cứu trước đây đều khai thác các đặc trưng thuần phong cách hoặc các đặc trưng mang nhiều nội dung. Trong nghiên cứu này, chúng tôi đề xuất phương pháp sử dụng các âm tiết và vần trong tiếng Việt làm đặc trưng nhận diện. Đây là các đặc trưng có mức độ ngữ nghĩa cao hơn các ký tự hoặc cụm ký tự ghép ngẫu nhiên, tuy nhiên mang ngữ nghĩa ít hơn nhiều so với các từ nội dung. Các đặc trưng này có thể coi như các cụm ký tự n-grams nhưng được kết hợp theo quy tắc từ vựng và ngôn ngữ thay vì kết hợp một cách ngẫu nhiên. Theo khảo sát của chúng tôi, đến nay chưa có nghiên cứu nào trong lĩnh vực phân tích tác giả văn bản được thực hiện trên loại đặc trưng này.

III. ÂM TIẾT VÀ VẦN TRONG TIẾNG VIỆT

Tiếng Việt là ngôn ngữ thuộc hệ ngôn ngữ Nam Á (Austroasiatic). Trong quá khứ, tiếng Việt sử dụng chữ viết Trung Quốc và có cải tiến cho phù hợp với đặc thù Việt Nam. Tuy nhiên, tiếng Việt hiện đại sử dụng chữ viết dựa trên bảng chữ cái Latin gọi là chữ Quốc ngữ. Chữ Quốc ngữ được phát minh bởi các nhà truyền giáo châu Âu khi với Việt nam để truyền đạo Thiên chúa giáo vào cuối thế kỷ 19.

Theo [7], tiếng Việt có ba loại âm vị là thanh điệu, phụ âm, và nguyên âm, trong khi hầu hết các ngôn ngữ châu Âu như tiếng Anh đều không sử dụng các thanh điệu. Thanh điệu được xem như là âm vị trong tiếng Việt vì việc thay đổi một thanh điệu có thể làm thay đổi nghĩa của từ. Có 6 thanh điệu trong tiếng Việt, đó là không dấu, huyền, ngã, hỏi, sắc, nặng.

Hệ thống ngữ âm tiếng Việt có 23 âm vị phụ âm đầu: b, ph, v, m, t, đ, th, x, d, n, l, tr, s, (gi, r), ch, nh, (c, k, q), g, kh, ng, h, p, r, sáu âm vị phụ âm cuối: p, t, (c/ch), m, n, (ng/nh), và hai bán âm vị nguyên âm cuối: (i/y), (o/u). Tiếng Việt cũng có 11 âm vị nguyên âm đơn: i, ê, e, ư, ơ, â, a, ă, u, ô, o và 3 nguyên âm đôi: (iê/ia), (ươ/ưa), (uô/ua). Các âm vị này kết hợp với nhau tạo thành các âm tiết. Các âm tiết trong tiếng Việt có ba thành phần: thanh điệu, âm đầu, và vần. Âm đầu là một phụ âm đơn và vần lại bao gồm một nguyên âm trung, một nguyên âm chính, và một âm cuối, trong đó chỉ có nguyên âm chính là bắt buộc phải có trong một âm tiết [7].

Theo phân tích của Tang [23], cấu trúc này khác với cấu trúc âm tiết của các ngôn ngữ châu Âu như tiếng Anh, theo đó các âm tiết của các ngôn ngữ này được mô tả là chuỗi các phụ âm (C) và nguyên âm (V) đan xen. Do đó, có nhiều cấu trúc âm tiết trong các

ngôn ngữ đó như CV, CVC, CCVC, CCCVC, v.v... Nghiên cứu của Tang [23] cũng chỉ ra rằng tiếng Việt có sự đa dạng hơn về khả năng kết hợp nguyên âm và phụ âm, nhưng có số phụ âm ít hơn so với các ngôn ngữ châu Âu như tiếng Anh.

Từ cấu trúc âm tiết này và các quy tắc từ vựng, ngữ pháp có thể tạo ra danh sách khoảng 6.700 âm tiết và 480 vần trong tiếng Việt. Sau khi loại bỏ đi một số âm tiết hiếm khi được sử dụng và gần như không có ảnh hưởng tới quá trình nhận diện, có 6.400 âm tiết và 480 vần được sử dụng làm đặc trưng trong nghiên cứu này.

Bảng 1. Cấu trúc âm tiết tiếng Việt

Thanh điệu			
Âm đầu	Vần		
	Nguyên âm trung	Nguyên âm chính	Âm cuối

(nguồn [7])

IV. PHƯƠNG PHÁP

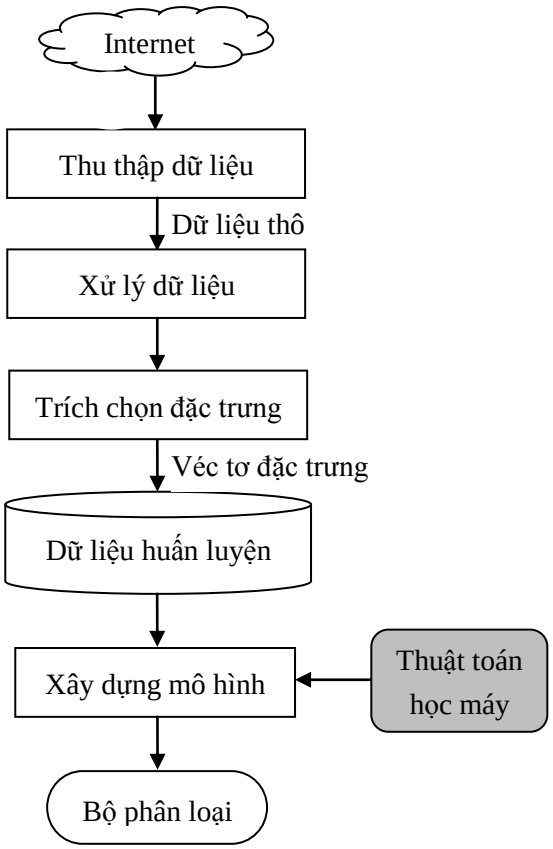
IV.1. Tổng quan về phương pháp

Mô hình tổng thể của phương pháp bao gồm các bước thu thập và xử lý dữ liệu, trích chọn đặc trưng, và xây dựng mô hình phân loại. Theo đó, các bài viết diễn đàn tiếng Việt đã có thông tin về đặc điểm người viết được thu thập từ Internet thông qua bước thu thập dữ liệu. Bước xử lý dữ liệu tiến hành các thao tác tiền xử lý trước khi thực hiện trích chọn đặc trưng và tạo các tập dữ liệu huấn luyện. Cuối cùng, các bộ phân loại sẽ được xây dựng bằng các thuật toán học máy trên các tập dữ liệu tạo được. Hình 1 cho thấy mô hình tổng quát của phương pháp.

IV.2. Thu thập và xử lý dữ liệu

Trong bước này, dữ liệu được thu thập từ các diễn đàn qua bộ thu thập dữ liệu tự động trên Web (Web crawler) sẽ được xử lý làm sạch và phân nhóm theo đặc điểm tác giả.

Hoạt động xử lý làm sạch tiến hành loại bỏ các nội dung không phải do người dùng tạo ra (ví dụ các đoạn trích từ các bài viết khác), hoặc các nội dung không phù hợp (ví dụ các đoạn chứa quá nhiều ký tự đặc biệt) v.v... Hoạt động phân nhóm sẽ nhóm các bài viết theo các lớp đặc điểm của người dùng. Các đặc điểm giới tính và vùng miền được phân làm hai nhóm (nam/nữ và bắc/nam), trong khi các đặc điểm độ tuổi và nghề nghiệp được phân làm ba nhóm. Độ tuổi được phân chia làm ba lớp theo các giai đoạn trong cuộc đời (học sinh, sinh viên/người mới đi làm/người trung niên).



Hình 1. Mô hình tổng quát của phương pháp

Để tránh sự nhập nhằng trong dự đoán do thực tế có nhiều tác giả tham gia viết bài trong nhiều năm liền, một số nghiên cứu trước đây sử dụng các nhóm tuổi không liên tục [2, 19, 21]. Trong nghiên cứu này, chúng tôi cũng sử dụng cách chia nhóm tuổi không liên tục (16-21, 24-27, 33-47). Đặc điểm nghề nghiệp được phân thành ba nhóm nghề phổ biến ở Việt Nam hiện nay là kinh doanh, bán hàng/kỹ thuật, công nghệ/giáo dục, y tế.

Ngoài ra, để có thể trích chọn được các đặc trưng liên quan đến từ vựng và ngữ pháp, cần có thêm các thao tác xử lý về mặt ngôn ngữ. Đó là các tác vụ phân chia văn bản thành các câu hoặc các từ và việc gán nhãn loại từ. Đây là những tác vụ quan trọng trong việc trích xuất các từ và các đặc điểm ngữ nghĩa trong bước kế tiếp, đặc biệt khi cấu tạo từ trong tiếng Việt phức tạp hơn các ngôn ngữ khác như tiếng Anh (có nhiều loại từ như từ ghép đôi, ghép ba, v.v…). Trong nghiên cứu này, chúng tôi sử dụng các công cụ VnTokenizer và VnTagger được mô tả trong [16].

IV.3. Các đặc trưng phân loại

Trong nghiên cứu này, chúng tôi đề xuất một loại đặc trưng mới có mức độ ngữ nghĩa cao hơn các ký tự nhưng ở mức thấp hơn so với các từ nội dung. Số lượng các đặc trưng cũng phải ở mức chấp nhận được. Các âm tiết và vần trong tiếng Việt thỏa mãn các yêu cầu này. Trong tiếng Việt, một từ có thể là từ đơn (chứa một âm tiết) hoặc từ ghép (chứa từ hai âm tiết trở lên). Mỗi âm tiết là một cụm ký tự riêng rẽ trong câu.

Do đó, các âm tiết mang ít ngữ nghĩa hơn so với các từ. Một số âm tiết có thể mang đầy đủ ngữ nghĩa nếu nó hình thành một từ đơn (ví dụ từ "ghế"), nhưng nó cũng có thể mang ít hoặc thậm chí không rõ ngữ nghĩa nếu là thành phần của một từ ghép. Ví dụ, từ "đồng hồ" có hai âm tiết là "đồng" và "hồ". Nghĩa của hai âm tiết này đứng riêng rẽ không có liên quan gì tới từ ghép. Vần là một phần của âm tiết, do vậy nó mang rất ít hoặc không mang ngữ nghĩa. Chẳng hạn "ế" là vần của từ "ghế" hay "ồng" và "ồ" là vần của từ "đồng hồ" nhưng hầu như không mang ngữ nghĩa liên quan đến các âm tiết mà nó thuộc vào.

Một tính chất quan trọng nữa là các âm tiết và vần được sử dụng làm đặc trưng không được chọn từ tập dữ liệu hoặc có liên quan về ngữ nghĩa với lĩnh vực nghiên cứu như các từ nội dung. Điều này giúp loại bỏ được tính phụ thuộc dữ liệu và phụ thuộc lĩnh vực của các đặc trưng này.

Âm tiết và vần có thể được hình thành sử dụng một số luật trong tiếng Việt. Chẳng hạn, một âm đầu có thể được lựa chọn từ 23 âm vị phụ âm đầu và 6 nguyên âm cùng với 2 bán nguyên âm có thể sử dụng làm âm cuối. Nguyên âm có thể là đơn lập hoặc kết hợp thành các nguyên âm đôi hoặc nguyên âm đa. Cụ thể, có 11 nguyên âm đơn, 3 nguyên âm đôi, và 20 nguyên âm đa trong tiếng Việt.

Từ các luật này, chúng ta có thể dễ dàng xây dựng danh sách các âm tiết và vần trong tiếng Việt và sử dụng chúng như các đặc trưng nhận diện trong xác định đặc điểm tác giả văn bản. Bên cạnh các đặc trưng này, chúng tôi cũng thực hiện thí nghiệm trên các đặc trưng đơn thuần theo phong cách và đơn thuần theo nội dung để tạo cơ sở so sánh và đánh giá kết quả. Bảng 2 cho thấy danh sách và số lượng các đặc trưng được sử dụng trong nghiên cứu.

Bảng 2. Các đặc trưng

Loại đặc trưng	**Số lượng**
Âm tiết	6.400
Vần	480
Các đặc trưng theo phong cách	333
Ký tự và tính chất từ	*90*
Ngữ pháp	*26*
Từ công cụ (từ chức năng)	*212*
Cấu trúc	*5*
Đặc trưng nội dung (từ nội dung)	2.400
Tổng cộng	**9.613**

IV.4. Xây dựng mô hình phân loại

Vấn đề xác định đặc điểm tác giả bài viết được chuyển thành bài toán phân loại bài viết theo các đặc điểm trên. Một bộ phân loại sẽ khớp các tài liệu với các nhãn đặc điểm người viết dựa trên các đặc trưng trích chọn được. Bộ phân loại sẽ được xây dựng từ các tài liệu đã gắn nhãn, sử dụng phương pháp học máy, với các đặc trưng của tài liệu là đầu vào và đặc điểm tác giả là đầu ra của thuật toán. Bên cạnh thuật toán học máy, một số kỹ thuật hỗ trợ khác cũng được áp dụng để nâng cao độ chính xác phân loại và giảm độ phức tạp mô hình như các thuật toán tối ưu tham số và lựa chọn đặc trưng.

V. THỰC NGHIỆM

V.1. Dữ liệu

Trong nghiên cứu này, chúng tôi sử dụng tập dữ liệu của nghiên cứu trước đây về nhận diện đặc điểm tác giả bài viết diễn đàn [8] để tiện so sánh kết quả. Tập dữ liệu này được thu thập bằng cách sử dụng bộ thu thập dữ liệu tự động (crawler) để thu thập các bài viết từ các diễn đàn phổ biến ở Việt Nam như otofun.net.vn, webtretho.com, tinhte.vn. Do các bài viết diễn đàn được viết khá tự do và chứa nhiều nội dung nhiễu, các phương pháp lọc và làm sạch dữ liệu đã được thực hiện như đã nói ở trên.

Sau bước xử lý và làm sạch, tập dữ liệu thu thập được bao gồm có 6.831 bài viết từ 104 người dùng. Tổng cộng có 736.252 từ và trung bình 107 từ/bài. Các bài viết được lựa chọn là các bài có ít nhất một thông tin về đặc điểm người viết, có thể dùng làm dữ liệu huấn luyện cho hệ thống. Độ dài của các bài viết cũng được giới hạn trong khoảng từ 250 đến 1.500 ký tự để loại bỏ các bài viết quá ngắn hoặc quá dài (bài viết quá dài có thể chứa các đoạn văn bản sao chép từ các nguồn khác). Bảng 3 cho thấy các thông số thống kê về tập ngữ liệu huấn luyện.

Bảng 3. Thống kê về tập ngữ liệu huấn luyện

Đặc điểm tác giả	**Số bài viết**	**Lớp đặc điểm**	**Tỷ lệ trong tập dữ liệu**
Giới tính	4.474	Nam	54%
		Nữ	46%
Độ tuổi	3.017	Ít hơn 22	21%
		Từ 24 đến 27	27%
		Nhiều hơn 32	52%
Vùng miền	3.960	Bắc	57%
		Nam	43%
Nghề nghiệp	3.453	Kinh doanh, bán hàng	36%
		Kỹ thuật, công nghệ	31%
		Giáo dục, y tế	33%

V.2. Thuật toán và phương pháp đánh giá

SVM là phương pháp học máy được lựa chọn trong nghiên cứu này để xây dựng mô hình phân loại do đây là phương pháp đã chứng minh được tính hiệu quả trong rất nhiều nghiên cứu về phân tích tác giả trước đây. SVM có ưu điểm là có thể xử lý số lượng lớn các đặc trưng phân loại và không cần đến việc giảm bớt số lượng đặc trưng nhằm tránh vấn đề quá khớp (over-fitting). Đặc điểm này rất hữu ích khi xử lý các vấn đề có số chiều lớn thường gặp trong các lĩnh vực như phân tích văn bản [5].

Kỹ thuật tối ưu tham số sử dụng Grid Search để tiến hành rà soát các giá trị của cặp tham số thuật toán và thử nghiệm từng cặp để chọn ra tham số tốt nhất. Qua thực nghiệm cho thấy SVM đạt kết quả tốt nhất với nhân đa thức (PolyKernel) cho bài toán này, do vậy, thuật toán tìm kiếm lưới (Grid Search) sẽ được thực hiện trên hai tham số là *c* và *exp* (bậc đa thức). Thuật toán lựa chọn đặc trưng dựa trên độ lợi thông tin (Information Gain) được sử dụng làm phương pháp lựa chọn đặc trưng nhằm loại bỏ bớt các đặc trưng không liên quan, qua đó giảm độ phức tạp và tăng độ chính xác của mô hình. Information Gain sử dụng cách đo độ quan trọng của mỗi đặc trưng trong việc phân biệt các lớp phân loại (dựa trên mức độ giảm entropy ứng với đặc trưng) và đã được ứng dụng trong nhiều nghiên cứu trước đây và cho kết quả tốt.

Các thực nghiệm của nghiên cứu được tiến hành trên công cụ Weka [24] với phương pháp kiểm chứng chéo 10 phần (10-fold cross-valiadion) và độ đo chính xác.

Độ đo chính xác (accuracy) được định nghĩa là tổng số mẫu được phân loại đúng trên tổng số mẫu trong tập dữ liệu kiểm tra. Đây là độ đo được sử dụng phổ biến để đánh giá độ chính xác tổng quát của một mô hình học máy và được sử dụng trong nhiều nghiên cứu trước đây về phân tích tác giả văn bản.

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \quad (1)$$

Trong đó *tp (true positive)* là số các mẫu mang nhãn "dương" được phân đúng vào lớp "dương", *tn (true nagative)* là số các mẫu mang nhãn "âm" được phân đúng vào lớp "âm", *fp (false positives)* là số các mẫu mang nhãn "âm" được phân sai vào lớp "dương", và *fn (false negative)* là số các mẫu mang nhãn "dương" được phân sai vào lớp "âm".

V.3. Kết quả và đánh giá

Chúng tôi thực hiện các thực nghiệm về xác định đặc điểm tác giả trên các tập con đặc trưng khác nhau để kiểm chứng hiệu quả của các đặc trưng âm tiết và vần. Bảng 4 cho thấy kết quả xác định đặc điểm tác giả bài viết diễn đàn tiếng Việt trên 9 tập con đặc trưng.

Các tập con đặc trưng được xây dựng theo nguyên tắc như sau: xem tập đặc trưng dựa theo phong cách làm cơ sở, mỗi loại đặc trưng khác được thử nghiệm riêng rẽ và kết hợp với tập đặc trưng theo phong cách. Cuối cùng, tập đặc trưng kết hợp tất cả các loại đặc trưng trên được thử nghiệm. Do số lượng âm tiết và từ nội dung có số lượng lớn, các tập con đặc trưng có chứa các loại đặc trưng này sẽ được thực hiện lựa chọn bằng thuật toán Information Gain trước khi thực hiện nhận diện bằng thuật toán SVM như đã nói ở trên.

Cụ thể, các tập con đặc trưng âm tiết, từ nội dung, kết hợp phong cách và âm tiết, kết hợp phong cách và nội dung, kết hợp tất cả đặc trưng là các trường hợp sẽ được thực hiện lựa chọn đặc trưng trước khi thực hiện nhận diện. Ngoài ra, do nghiên cứu này áp dụng thêm thuật toán Grid Search để tối ưu tham số cho thuật toán phân loại SVM nên các kết quả trên các tập đặc trưng theo phong cách và nội dung có sự cải tiến so với các kết quả được trình bày trong [8].

Bảng 4. Kết quả xác định đặc điểm tác giả trên các tập đặc điểm khác nhau

Tập đặc trưng	**Giới tính**	**Độ tuổi**	**Nghề nghiệp**	**Vùng miền**
Theo phong cách	83.47	62.76	52.46	71.22
Các vần	84.13	58.26	50.22	72.80
Âm tiết	89.98	66.24	57.43	80.38
Từ nội dung	90.01	70.05	60.99	82.98
Kết hợp phong cách và vần	86.56	60.90	54.30	75.70
Kết hợp phong cách và âm tiết	91.33	69.23	58.70	81.07
Kết hợp phong cách và nội dung	90.55	70.70	61.04	83.13
Kết hợp tất cả	91.72	71.26	61.43	84.28

Từ kết quả trong Bảng 4, sử dụng kết quả nhận diện khi dùng đặc trưng theo phong cách làm cơ sở, có thể thấy các kết quả khi sử dụng đặc trưng vần làm tăng độ chính xác lên khoảng 1-2%, trong khi sử dụng các đặc trưng âm tiết làm tăng khoảng 7%. Việc kết hợp đặc trưng theo phong cách và đặc trưng vần cũng như kết hợp đặc trưng phong cách và đặc trưng âm tiết làm tăng hiệu quả nhận diện lên 4%-8% tương ứng.

So sánh với các từ nội dung, các đặc trưng theo âm tiết mặc dù mang ít ngữ nghĩa hơn và có tính độc lập dữ liệu hơn nhưng có kết quả nhận diện gần tương đương với các từ nội dung (đặc biệt ở đặc điểm giới tính và vùng miền).

Mặc dù vẫn còn một số ngoại lệ, như các đặc trưng vần cho kết quả không tốt khi nhận diện các đặc điểm về độ tuổi và nghề nghiệp so với đặc trưng phong cách, hoặc các đặc trưng âm tiết cho kết quả kém khi nhận diện độ tuổi, có thể kết luận rằng các đặc trưng âm tiết và vần mang lại kết quả khả quan và tốt hơn đặc trưng phong cách và tiệm cận với các đặc trưng nội dung. Hơn nữa, việc kết hợp tất cả các loại đặc trưng cho kết quả cao nhất chứng tỏ việc sử dụng các âm tiết và vần đã có những ảnh hưởng tích cực tới kết quả nhận diện kể cả khi các từ nội dung được sử dụng. Mặc dù vần và âm tiết vẫn còn mang nội dung và chưa được coi là hoàn toàn không phụ thuộc nội dung như các đặc trưng phong cách khác, các kết quả trên vẫn rất khả quan vì các lý do sau:

- Tiếng Việt là một ngôn ngữ đa âm tiết, trong đó một từ có thể chứa nhiều âm tiết. Theo [7], 80% các từ tiếng Việt có từ hai âm tiết trở lên. Vần chỉ là một bộ phận cấu thành âm tiết. Do vậy, sử dụng âm tiết và vần làm đặc trưng nhận diện sẽ làm giảm tính phụ thuộc nội dung hơn rất nhiều so với sử dụng các từ nội dung. Đặc biệt, các vần có thể được xem như đặc trưng phong cách và không có tính phụ thuộc nội dung.
- Mặc dù các âm tiết vẫn mang một phần ngữ nghĩa, việc xây dựng danh sách các âm tiết và vần làm đặc trưng không phụ thuộc vào tập dữ liệu mà dựa vào các quy tắc từ vựng và ngữ pháp. Do đó, tập các âm tiết và vần đại diện cho hệ thống từ vựng của toàn bộ ngôn ngữ và không bị ảnh hưởng bởi tập

dữ liệu hay một lĩnh vực cụ thể nào. Điều này làm cho các âm tiết và vần có tính độc lập hơn nhiều so với các từ nội dung.

VI. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề xuất một phương pháp xác định đặc điểm tác giả văn bản tiếng Việt mới dựa trên các đặc trưng về âm tiết và vần tiếng Việt. Âm tiết và vần là các thành tố cấu tạo nên từ, do vậy chúng mang ít ngữ nghĩa nội dung hơn so với các từ nội dung. Hơn nữa, các âm tiết và vần được xây dựng từ các quy tắc từ vựng mà không chọn lọc từ tập dữ liệu. Do đó, sử dụng các đặc trưng này sẽ làm giảm đi tính đặc thù dữ liệu và đặc thù miền trong phân tích tác giả văn bản. Các kết quả thực nghiệm cho thấy độ chính xác nhận diện khi sử dụng các đặc trưng này có cải tiến đáng kể so với các đặc trưng dựa theo phong cách, đồng thời làm tăng kết quả khi sử dụng kết hợp với các đặc trưng nội dung khác.

Hướng phát triển tiếp theo của nghiên cứu có thể là tiến hành thực nghiệm trên các tập dữ liệu thuộc các lĩnh vực khác nhau để kiểm chứng tính tổng quát của phương pháp. Ngoài ra, chúng tôi cũng có kế hoạch khai thác thêm các đặc trưng của tiếng Việt để nâng cao kết quả nhận diện, chẳng hạn như các đặc trưng về thanh điệu, hình vị, v.v...

TÀI LIỆU THAM KHẢO

[1] AHMED ABBASI, HSINCHUN CHEN. *Applying Authorship Analysis to Extremist-Group Web Forum Messages*, IEEE Intelligent Systems, v.20 n.5, p.67-75, 2005.

[2] S. ARGAMON, M. KOPPEL, J. W. PENNEBAKER, J. SCHLER. *Automatically profiling the author of an anonymous text*, Communications of the ACM, v.52 n.2, 2009.

[3] R. CLEMENT, D. SHARP. *Ngram and Bayesian classification of documents for topic and authorship*. Literary and Linguistic Computing, 18(4), pp: 423—447, 2003.

[4] M. CORNEY, O. DE VEL, A. ANDERSON, G. MOHAY. *Gender-preferential text mining of e-mail discourse*, In ACSAC'02: Proc. of the 18th Annual Computer Security Applications Conference, Washington, DC, pp : 21-27, 2002.

[5] O. DE VEL, A. ANDERSON, M. CORNEY, G. MOHAY. *Mining e-mail content for author identification forensics*. SIGMOD Record 30(4), pp. 55-64, 2001.

[6] J. DIEDERICH, J. KINDERMANN, E. LEOPOLD, G. PAASS. *Authorship Attribution with Support Vector Machines*, Applied Intelligence, v.19 n.1-2, p.109-123, 2003.

[7] D. L. THU, N. V. HUE. *Cơ cấu ngữ âm tiếng Việt*, Vietnam Education Publishing, 1998.

[8] T. D. Duong, S. B. Pham, H. Tan. *Using Content-based Features for Author Profiling of Vietnamese Forum Posts*, In: Recent Developments in Intelligent Information and Database Systems, pp. 287–296. Springer International Publishing, Berlin, 2016.

[9] MICHAEL GAMON. *Linguistic correlates of style: authorship classification with deep linguistic analysis features*, Proceedings of the 20th international conference on Computational Linguistics, p.611-es, 2004

[10] S. GOSWANI, S. SARKAR, M. RUSTAGI. *Stylometric analysis of bloggers' age and gender*, In Proceedings of the Third International ICWSM Conference, San Jose, USA, 2009.

[11] J. HOUVARDAS, E. STAMATATOS. *N-Gram feature selection for authorship identification*, Proceedings of the 12th international conference on Artificial Intelligence: methodology, Systems, and Applications, Varna, Bulgaria, 2006.

[12] F. IQBAL, H. BINSALLEEH, B. C. M. FUNG, M. DEBBABI. *Mining writeprints from anonymous e-mails for forensic investigation*, Digital Investigation: The International Journal of Digital Forensics & Incident Response, v.7 n.1-2, p.56-64, 2010.

[13] V. KESELJ, F. PENG, N. CERCONE, C. THOMAS. *N-gram-based author profiles for authorship attribution*. In: Pasific Association for Computational Linguistics, pp. 256–264, 2003.

[14] M. KOPPEL, J. SCHLER, K. ZIGDON. *Determining an author's native language by mining a text for errors*, Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery in data mining, USA, 2005.

[15] M. KOPPEL, J. SCHLER, S. ARGAMON. *Computational methods in authorship attribution*. Journal of the American Society for

information Science and Technology, 60(1), p.9-26, 2009.

[16] H. P. LE, A. ROUSSANALY, T. M. H. NGUYEN, M. ROSSIGNOL. *An empirical study of maximum entropy approach for part-of-speech tagging of vietnamese texts*, In Traitement Automatique des Langues Naturelles-TALN, page 12, 2010.

[17] D. NGUYEN, R. GRAVEL, D. TRIESCHNIGG, T. MEDER. *"How old do you think I am?" a study of language and age in Twitter*. In ICWSM, 2013.

[18] D. H. NGUYEN. *Vietnamese*, Amsterdam: John Benjamins Publishing Company, 1997.

[19] C. PEERSMAN, W. DAELEMANS, L. V. VAERENBERGH. *Predicting age and gender in online social networks*, In Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, pages 37–44, New York, NY, USA, 2011.

[20] D. D. PHAM, G. B. TRAN, S. B. PHAM, *Author Profiling for Vietnamese Blogs*, Proceedings of the 2009 International Conference on Asian Language Processing, p.190-194, 2009.

[21] F. RANGEL, P. ROSSO. *Use of language and author profiling: Identification of gender and age*. In Natural Language Processing and Cognitive Science, p. 177, 2013.

[22] E. STAMATATOS. *A survey of modern authorship attribution methods*, Journal of the American Society for information Science and Technology, 60(3), pp.538-556, 2009.

[23] G. TANG. *Cross-linguistic analysis of Vietnamese and English with implications for Vietnamese language acquisition and maintenance in the United States*, Journal of Southeast Asian-American Education & Advancement, 2, 1–33, 2006.

[24] I. H. WITTEN, E. FRANK. *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, second edition, 2005.

[25] R. ZHENG, H. CHEN, Z. HUANG, Y. QIN. *Authorship Analysis in Cybercrime Investigation* (Eds.): ISI 2003, LNCS 2665, pp : 59-73, 2003.

[26] R. ZHENG, J. LI, H. CHEN, Z. HUANG. *A framework for authorship identification of online messages: Writing-style features and classification techniques*, Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 378–393, 2006.

Nhận bài ngày: 09/02/2017

SƠ LƯỢC VỀ CÁC TÁC GIẢ

DƯƠNG TRẦN ĐỨC

Sinh ngày 28/02/1978.

Tốt nghiệp trường ĐH Khoa học Tự nhiên, ĐH Quốc gia Hà Nội ngành CNTT năm 1999. Tốt nghiệp Thạc sỹ chuyên ngành Hệ thống thông tin tại ĐH Tổng hợp Leeds, Vương Quốc Anh năm 2004.

Hiện đang công tác tại Khoa CNTT, Học viện Công nghệ Bưu chính Viễn thông.

Hướng nghiên cứu chính: Học máy, dữ liệu lớn

Email: ducdt@ptit.edu.vn

PHẠM BẢO SƠN

Sinh năm 1977.

Tốt nghiệp ĐH Tổng hợp New South Wales năm 1999. Tốt nghiệp Thạc sỹ và sau đó nhận bằng Tiến sĩ chuyên ngành Khoa học máy tính tại ĐH Tổng hợp New South Wales năm 2007.

Hiện đang công tác tại trường ĐH Công nghệ, ĐH Quốc gia Hà Nội.

Hướng nghiên cứu chính: Học máy, xử lý ngôn ngữ tự nhiên.

Email: sonpb@vnu.edu.vn

TÂN HẠNH

Sinh năm 1966.

Nhận bằng Tiến sĩ chuyên ngành Khoa học máy tính tại Viện Công nghệ Grenoble, Pháp.

Hiện đang công tác tại Học viện Công nghệ Bưu chính Viễn thông.

Hướng nghiên cứu chính: Học máy, xử lý tín hiệu.

Email: tanhanh@ptit.edu.vn