

XÁC ĐỊNH TÁC GIẢ THEO HƯỚNG TIẾP CẬN PHONG CÁCH TRẮC HỌC

NGUYỄN TUYẾT NHUNG¹

Abstract: In this study, we examine three different stylometric measures for the author identification task, including vocabulary richness, word length, and parts-of-speech frequencies on a 34-opinion article dataset with nearly 3000 word tokens. The findings show that although each of stylometric measures is regarded as a feature representing authorial writing style, it on its own does not have enough discriminating ability. However, if these stylometric measures are combined with Correspondence analysis for parts-of-speech as linguistic variables, we can identify an author of a Vietnamese text successfully.

Key words: *stylometry, author identification, forensic linguistics, corpus.*

1. Giới thiệu

Hiện nay, trên mạng Internet, có rất nhiều văn bản chứa nội dung độc hại không xác định được tác giả vì các bài viết thường ẩn danh hoặc thay đổi danh tính, điều này làm cho quá trình điều tra tội phạm trở nên cực kỳ khó khăn. Một trong những nhiệm vụ chính trong điều tra tội phạm công nghệ cao đó là truy tìm danh tính thực sự của tác giả văn bản. Phương thức truyền thống để giải quyết bài toán này chủ yếu dựa trên cách làm thủ công, điều này gặp nhiều hạn chế khi số lượng văn bản phải xử lý rất lớn. Vì vậy, việc tự động nhận diện tác giả có giá trị rất lớn đối với cơ quan điều tra tội phạm công nghệ cao và ứng dụng của ngành Phong cách trắc học có thể giúp chúng ta nhận diện tác giả của văn bản nghi ngờ.

Phong cách trắc học (Stylometry) là ngành nghiên cứu thuộc lĩnh vực Ngôn ngữ học tội phạm (Forensic Linguistics), giúp chúng ta tìm ra những thành phần then chốt trong phong cách của một tác giả nào đó thông qua việc tham chiếu các tiêu chuẩn định lượng, thường là các đặc trưng từ vựng. Các đặc trưng từ vựng này được gọi là yếu tố phân biệt (discriminators).

Nhận diện tác giả (Author identification), một nhánh trong ngành Phong cách trắc học, gồm ba loại bài toán chính, đó là: Xác định danh tính tác giả (Authorship attribution); Xác minh tác giả (Author verification); và Lí lịch hóa tác giả (Author profiling). Hình 1.1 dưới đây minh họa các bài toán trong lĩnh vực nhận diện tác giả:

Hình 1. Ba bài toán chính trong lĩnh vực nhận diện tác giả



Ứng dụng phổ biến nhất của Phong cách trắc học là xác định tác giả của một văn bản ẩn danh hoặc mạo danh. Trong nghiên cứu này, chúng tôi khảo sát một số phương pháp xác định danh tính tác giả, bao gồm các đại lượng từ đơn giản đến phức tạp như độ phong phú của từ vựng, độ dài từ và tần số từ loại cho các văn bản tiếng Việt.

2. Các công trình nghiên cứu có liên quan

Burrows [3, tr. 91] trước đây đã từng lo ngại về việc “Chưa có ai xác định được một thuộc tính phong cách với tư cách là ‘dấu ấn riêng của con người’ (human fingerprints) mang tính đặc thù và ổn định”. Tuy nhiên, Oakes [9] cho rằng phong cách của một tác giả có thể được mô tả bằng một tập hợp đặc trưng (a set

¹ Trường Đại học An ninh nhân dân.

of features). Cùng quan điểm, Stamatatos [11] cho rằng việc nhận diện tác giả tập trung vào phong cách đặc thù của tác giả, với giả định rằng mỗi tác giả có ‘dấu ấn’ phong cách (stylistic ‘fingerprint’) riêng của mình và ta có thể xác định tác giả của một văn bản chưa rõ nguồn dựa trên dấu ấn phong cách đó.

Tương tự, theo Pinker (2014), phong cách là cách thức biểu đạt ngôn ngữ, được thể hiện qua việc lựa chọn từ ngữ, việc kết hợp các từ ngữ lại với nhau, dấu câu, cấu trúc câu, dạng thức ngữ pháp và tất cả các yếu tố mà một tác giả thích sử dụng. Để xác định một đặc trưng ngôn ngữ ảnh hưởng đến phong cách viết, một số nhà nghiên cứu đề cập đến khái niệm “lựa chọn” (choice) hoặc “tự do” (freedom). Crystal [4, tr. 4] cho rằng khi nói đến phong cách là tôi muốn nói đến “tập hợp các đặc trưng ngôn ngữ mà khi hợp lại có thể nhận diện được người sử dụng ngôn ngữ. Khái niệm này giả định trước là có một sự lựa chọn - tức là một người có thể chọn đặc trưng P thay vì chọn đặc trưng Q hay R, S,...”².

Theo Brezina, những sự khác biệt về mặt phong cách này chỉ có thể được xác định bằng cách khảo sát trên mẫu dữ liệu lớn và phân tích dữ liệu này theo phương thức định lượng [2, tr. 205]. Các nhà nghiên cứu về lĩnh vực Phong cách trắc học cho rằng một số yếu tố ngôn ngữ học có thể được xem là quan trọng hơn những yếu tố khác trong việc phân tích phong cách của một tác giả. Vì vậy, cần phải tìm được những yếu tố đánh dấu phong cách đó và chuyển đổi chúng thành những số liệu định lượng [10, tr. 19].

Những công trình nghiên cứu trước đây sử dụng một đại lượng đơn, với hy vọng đại lượng này không biến đổi trong những tác phẩm của một tác giả, ví dụ như độ phong phú từ vựng, độ dài của từ, v.v. Có nhiều phép đo độ phong phú từ vựng (vocabulary richness), hay còn gọi là độ đa dạng từ vựng, được một số nhà nghiên cứu đề xuất (Baayen, 2008; [5]) để chỉ biên độ của từ vựng mà một tác giả sử dụng. Nhìn chung, từ vựng được xem là phong phú nếu nhiều từ mới xuất hiện trong một phần văn bản có độ dài nhất định và từ vựng được xem là nghèo nàn nếu có ít từ khác biệt xuất hiện trong một phần văn bản có độ dài tương tự.

Độ dài của từ cũng là một yếu tố đánh dấu độ phức tạp của văn bản. Hay nói cách khác, văn bản có tỉ lệ từ có nhiều kí tự càng cao cho thấy phong cách của nó phức tạp, có tính khoa học và khó hiểu hơn. Lakoff và Wehling chỉ ra mối quan hệ giữa độ dài từ và độ phức tạp của từ được người tiếp nhận thông tin phân tích: “...từ ngữ có tác động lớn nhất lên não bộ chúng ta khi chúng là những từ đơn giản... Những từ thuộc cấp độ cơ bản khiến ta dễ nhớ hơn; những văn bản sử dụng ngôn ngữ cấp độ cơ bản là dễ nhớ nhất” [7, tr. 41]. Dựa trên kết quả này, tần số của những từ gồm nhiều kí tự cho thấy tác giả sử dụng từ vựng phong phú và có độ khó cao (Tausczik và Pennebaker, 2010). Một điều cần lưu ý đó là, việc chỉ đơn giản đếm số lượng của các kí tự để tìm độ phức tạp của từ cần phải hết sức cẩn trọng vì kí tự không phải là thành tố trực tiếp để tạo nên từ, mà vai trò này là của các âm tiết hoặc hình vị [10, tr. 234].

Những thành tựu phát triển hiện nay đã giúp sức cho những tiến bộ lớn lao trong Phong cách trắc học, giúp cho việc tính toán được thực hiện nhanh chóng và chính xác, giúp lưu trữ các khối ngữ liệu có kích cỡ lớn. Đáng chú ý nhất là PAN (viết tắt của Plagiarism analysis, Authorship identification, and Near - duplicate detection), chuỗi sự kiện khoa học về khảo nghiệm văn bản số và phong cách trắc học. PAN cung cấp các nguồn tài nguyên đánh giá bao gồm các khối ngữ liệu quy mô lớn, các phương pháp thực hiện và các dịch vụ trực tuyến để có thể thực hiện việc đánh giá chính xác. Mục tiêu chính đó là cung cấp những phương thức đánh giá mang tính bền vững và có thể tái tạo lại kết quả.

Ở Việt Nam, lĩnh vực Ngôn ngữ học máy tính đã có những bước phát triển đáng kể so với nhiều năm trước đây. Tuy nhiên, tài nguyên ngôn ngữ định dạng số đang ở những bước đi đầu tiên, chưa đủ để tiến hành khai thác văn bản tự động với quy mô lớn. Do đó, xác định danh tính tác giả chưa thực sự đạt được nhiều thành tựu như các công tác điều tra khác. Ngoài ra, theo hiểu biết của chúng tôi, hiện nay các công trình nghiên cứu về xác định danh tính tác giả trên dữ liệu tiếng Việt với các thuật toán học máy là chủ yếu và chỉ có rất ít công trình nghiên cứu tập trung vào việc sử dụng các phương pháp thống kê định lượng.

² “By style I mean the set of linguistic features that, taken together, uniquely identify a language user. The notion presupposes that there has been a choice—that someone has opted for Feature P rather than Feature Q (or R, S,...)” [4]

Một nghiên cứu mới được công bố gần đây của Hồ Ngọc Lâm và cộng sự [6] sử dụng các đặc trưng ngôn ngữ cơ bản, đó là từ có tần số cao nhất và độ dài của từ trong văn bản. Nhóm tác giả này đã thực nghiệm trên 1.304 văn bản số (bài đăng Facebook và bài báo trực tuyến) được viết bởi 10 tác giả. Bằng việc sử dụng ba phép đo phong cách là đường cong đặc trưng của Mendenhall, phương pháp Chi-squared và phương pháp Delta, nhóm nghiên cứu đã giải quyết bài toán xác định danh tính tác giả theo hướng tiếp cận thống kê định lượng. Trong đó, phương pháp Chi-squared, được thực nghiệm trên 500 từ có tần số cao nhất trong tập dữ liệu, đạt kết quả chính xác tuyệt đối. Tuy nhiên, khi áp dụng phương pháp Delta, độ chính xác chỉ đạt 40% với 30 từ có tần số cao nhất và đạt 50% với 20 từ có tần số cao nhất. Như vậy, mặc dù công trình nghiên cứu này đã áp dụng ba phương pháp khác nhau và đạt được kết quả rất cao với phương pháp Chi-squared (Chi bình phương), nhưng chỉ mới thực nghiệm trên hai nhóm đặc trưng ngôn ngữ, đó là từ có tần số cao và độ dài của từ.

Theo Savoy [10, tr.8], những phương pháp và kỹ thuật trắc học phong cách thường cũng hiệu quả đối với không chỉ riêng với tiếng Anh mà với tất cả các ngôn ngữ tự nhiên khác. Do đó, chúng tôi cũng áp dụng các đặc trưng đạt hiệu quả cao trên tập dữ liệu tiếng Anh cho các văn bản tiếng Việt, nhất là khi việc lựa chọn thể loại, chủ đề và mục đích giao tiếp của văn bản có những sự tương đồng nhất định. Trong phần tiếp theo, chúng tôi sẽ mô tả quy trình lựa chọn và xử lý khối ngữ liệu báo chí ý kiến tiếng Việt.

3. Phương pháp nghiên cứu

3.1. Dữ liệu nghiên cứu

Trong nghiên cứu này, trọng tâm được đặt vào việc xác định phong cách viết cá nhân của tác giả những bài viết trên báo trực tuyến và cụ thể đó là các bài viết ý kiến trên trang báo trực tuyến tiếng Việt VnExpress, một trang web báo trực tuyến có lượng truy cập cao nhất hiện nay. Các văn bản trong tập dữ liệu là những bài viết ý kiến được đăng trên VnExpress trong khoảng thời gian từ năm 2014 đến năm 2018. Chúng tôi lựa chọn thể loại báo chí ý kiến trực tuyến bởi vì:

(1) Các bài báo trực tuyến được đăng công khai trên mạng Internet, rất thuận lợi trong việc truy cập;

(2) Chúng thường là những văn bản có chất lượng cao, có rất ít lỗi chính tả;

(3) Các bài báo thường dễ đọc và dễ hiểu hơn các văn bản khoa học chuyên sâu;

(4) Việc tác giả nêu lên những ý kiến, quan điểm cá nhân sẽ giúp bộc lộ phong cách viết cá nhân rõ hơn so với những thể loại văn bản có cấu trúc khác. Đây là đặc điểm phù hợp để làm dữ liệu khảo sát cho các phương pháp phong cách trắc học khác nhau, nhất là cho nhiệm vụ xác định danh tính tác giả.

Trong nghiên cứu này, vì phương pháp thực nghiệm được sử dụng để xác định danh tính tác giả chỉ đòi hỏi phải có đủ dữ liệu phân tích cho mỗi tác giả ứng viên, nên chúng tôi chỉ giới hạn cho các tác giả có từ ba bài trở lên, trong đó mỗi bài có ít nhất 800 từ. Vì vậy, chúng tôi chọn 10 tác giả đảm bảo các tiêu chí này. Bảng 1 dưới đây liệt kê thông tin cơ bản của các tác giả trong khối ngữ liệu con VVC_AA.

STT	Tác giả	Mã tác giả	Số lượng văn bản trong VVC	Giới tính	Nghề nghiệp	Mã năm sinh
1	F1	49	17	Female	Journalist	Unknown
2	F2	129	14	Female	Journalist	Unknown
3	F3	414	24	Female	Journalist	Unknown
4	F4	465	13	Female	Teacher	Unknown
5	F5	890	18	Female	Journalist	Unknown
6	M1	59	24	Male	Doctor	1960
7	M2	83	15	Male	Teacher	1977
8	M3	97	29	Male	Researcher	1988
9	M4	203	29	Male	Journalist	Unknown
10	M5	1050	29	Male	Teacher	Unknown

Bảng 1. Thông tin tác giả trong khối ngữ liệu con VVC_AA

3.2. Định lượng phong cách viết bằng các đặc trưng ngôn ngữ cấp độ từ vựng

Các đặc trưng phong cách của một tác giả hiện diện ở nhiều cấp độ khác nhau, nên khó có thể khảo sát tất cả các đặc trưng này cùng một lúc. Vì vậy, trong nghiên cứu này, chúng tôi chỉ lựa chọn khảo sát các đặc trưng ngôn ngữ thuộc cấp độ từ vựng, cụ thể là độ phong phú từ vựng, độ dài từ và tần số từ loại. Bốn công thức được sử dụng để đo lường độ phong phú từ vựng bao gồm:

(1) Số R của Honoré: Số R của Honoré (1979) phụ thuộc vào hapax legomena.

(2) Số S của Sichel: Trong Sichel's S, mật độ của dis legomena được sử dụng.

(3) Số K của Yule: Số K của Yule phụ thuộc vào từ với mọi tần số, nó được coi là phép tính độ phong phú từ vựng ổn định hơn vì nó độc lập với độ dài của văn bản.

(4) Số W của Brunet: W có ưu điểm là không bị ảnh hưởng bởi độ dài văn bản và nó mang tính đặc trưng cho từng tác giả.

3.3. Xác định tác giả bằng phương pháp Phân tích tương ứng

Trong nghiên cứu này, chúng tôi sẽ thực nghiệm độ chính xác của việc áp dụng đặc trưng cấp độ từ vựng cho nhiệm vụ xác định danh tính tác giả. Để thực hiện điều này, chúng tôi lựa chọn một hướng tiếp cận mang tính khám phá, đó là phương pháp Phân tích tương ứng.

Phân tích tương ứng là một phương pháp mô tả đa biến trong phân tích dữ liệu, giúp trực quan hóa mối quan hệ giữa các tác giả và mối quan hệ giữa các biến ngôn ngữ [2, tr. 200]. Chúng tôi đếm số lần xuất hiện của tám nhóm từ loại lớn trong khối ngữ liệu và khảo sát sự tương tác giữa các biến này và tác giả. Đặc biệt là quá trình thực hiện Phân tích tương ứng bằng Ngôn ngữ lập trình R sẽ giúp chúng ta trực quan hóa việc sử dụng từ loại của các tác giả trên đồ thị tương ứng hai chiều (2D correspondence plot). Đồ thị hai chiều là sự biểu diễn hữu hiệu nhất cho tính phức tạp bởi vì nó giúp giảm thiểu số chiều của sự khác biệt xuống chỉ còn hai chiều được biểu diễn bởi trục tung và trục hoành. Điểm đặc biệt của đồ thị tương ứng chính là nó có thể bao quát cả thông tin ở cột lẫn thông tin ở hàng trong bảng dữ liệu trên cùng một không gian [2]. Lý do chính cho việc lựa chọn phương pháp nghiên cứu này đó là, không giống như phân tích dữ liệu khẳng định, các phương pháp khám phá không đòi hỏi phải có trước các giả thuyết đã được kiểm định. Thay vào đó, các phương pháp khám phá bắt đầu bằng việc quan sát sự tương tác diễn ra giữa các biến trong dữ liệu [2].

4. Kết quả và phân tích

4.1. Độ phong phú từ vựng

Bảng 2 dưới đây thể hiện kết quả các phép tính độ phong phú của từ vựng cho các tác giả bằng cách chia trung bình cho một chuỗi các phân đoạn không trùng lặp nhau, trong đó mỗi phân đoạn gồm 500 lượt từ.

Authors	Vocabulary richness measures				
	Honore's R	Sichel's S	Yule's K	Brunet's W	Trung bình
X4	529,1529704	0,169812310	35,88648311	11,91732125	135,2676
F465	529,0820666	0,165573582	34,07328754	11,84654120	135,2322
M1050	526,7175637	0,148486720	20,48425685	12,02908288	134,6867
M59	526,078574	0,150732555	36,29369829	12,14602697	134,5562
F414	526,7885069	0,156328583	22,43141628	11,16685955	134,4888
F890	525,2088235	0,162448889	33,21334744	11,47489157	134,1709
M97	521,1671892	0,152062150	32,71557939	11,34017417	133,1268
X3	518,1671892	0,144611520	32,45368133	11,62516291	132,4481
F49	517,1144371	0,147841365	30,05771796	11,49132991	132,1514
X1	513,2243981	0,158937312	26,06362281	12,98846331	131,5532
X2	512,7609112	0,153334583	25,83718182	12,16378211	131,2312
M203	513,1529704	0,149678058	26,85352151	10,57280108	130,9314
M83	510,8669892	0,153301498	41,89841835	12,76191804	130,9072
F129	512,9516512	0,151809494	24,53136221	10,13593955	130,7719

Bảng 2. Độ phong phú từ vựng của mười tác giả đã biết và bốn văn bản ẩn danh (giá trị trung bình xếp theo thứ tự giảm dần)

Theo kết quả ở Bảng 2, xét về trị trung bình, trong mười tác giả đã biết thì các giá trị cao nhất thuộc về hai nữ tác giả F465 và M1050, lần lượt là 135,2322 và 134,6867, điều này cho thấy rằng hai tác giả này có kho từ vựng lớn nhất. Ngược lại, tác giả nam M83 (130,9072) và tác giả nữ F129 (130,7719) có giá trị thấp nhất, cho thấy hai tác giả này sử dụng từ ít đa dạng hơn so với các tác giả còn lại. Trong bốn văn bản ẩn danh, văn bản X4 có giá trị độ phong phú từ vựng cao nhất (135,2676), đồng thời nó tương đồng nhất với tác giả F465 và M1050. Còn giá trị thấp nhất là X2 (131,2312), văn bản này tương đồng với tác giả M203 và đặc biệt là X2 còn tương đồng với X1 (131,5532). Văn bản X3 với giá trị 132,4481 có độ tương đồng với M97 và F49.

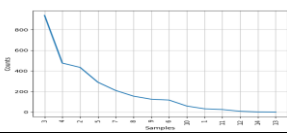
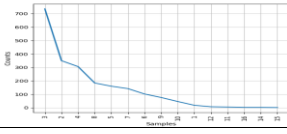
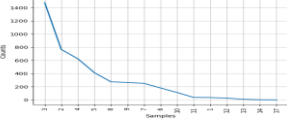
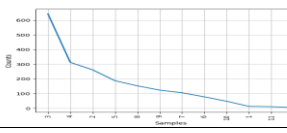
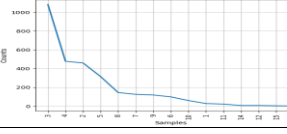
Tuy nhiên, theo phép tính số R của Honoré và số S của Sichel, các giá trị cao nhất thuộc về tác giả M1289, điều này cho thấy rằng tác giả này có kho từ vựng lớn nhất. Trong khi đó, các phép tính số K của Yule và số W của Brunet lại nghiêng về tác giả M83.

Như vậy, nhìn chung ta có thể sử dụng giá trị độ phong phú từ vựng với tư cách là một phương pháp xác định danh tính tác giả vì mỗi tác giả có một giá trị độ phong phú từ vựng riêng của mình. Tuy nhiên, như chúng tôi đã đề cập từ trước, không có phép đo độ phong phú từ vựng nào trong số này chứng minh được rằng độ chính xác của riêng mình nó thỏa mãn bài toán xác định danh tính tác giả (Hoover, 2003; Tweedie, 1998), một phần là do sự phân bố của từ, bao gồm cả bigram và trigram, bị kiểm soát bởi một lượng lớn những từ có tần suất thấp (Baayen, 2001). Do đó, cần phải kết hợp chúng với những phương pháp khác để tăng độ chính xác của nhiệm vụ xác định danh tính tác giả.

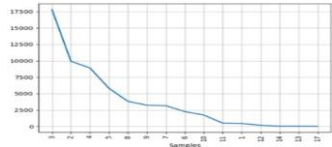
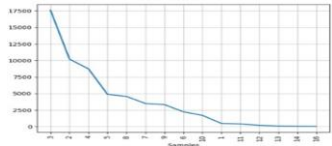
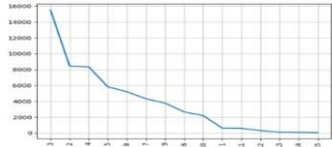
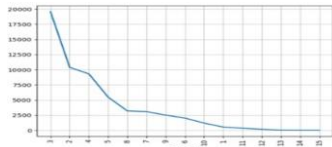
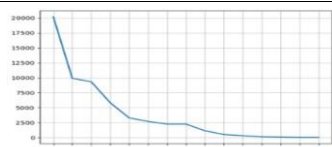
4.2. Độ dài của từ

Kết quả đường cong Mendenhall của các tác giả bài viết thuộc chuyên mục *Góc nhìn* được trình bày trong Hình 2 và 3 dưới đây, sử dụng các bài viết thuộc khối ngữ liệu VVC.

Hình 2. Đường cong Mendenhall của các tác giả nữ

Tác giả	Đường cong Mendenhall	Thông tin đường cong
F49		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 13 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-4-2-5-7 kí tự - Từ dài nhất: 13 kí tự - Từ ngắn nhất: 1 kí tự
F129		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 15 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-8-5 kí tự - Từ dài nhất: 16 kí tự - Từ ngắn nhất: 1 kí tự
F414		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 17 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 17 kí tự - Từ ngắn nhất: 1 kí tự
F465		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 12 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 12 kí tự - Từ ngắn nhất: 1 kí tự
F890		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 18 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-4-2-5-8 kí tự - Từ dài nhất: 18 kí tự - Từ ngắn nhất: 1 kí tự

Hình 3. Đường cong Mendenhall của các tác giả nam

Authors	Mendenhall curves	Information
M59		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 17 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 17 kí tự - Từ ngắn nhất: 1 kí tự
M83		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 16 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 16 kí tự - Từ ngắn nhất: 1 kí tự
M97		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 15 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 15 kí tự - Từ ngắn nhất: 1 kí tự
M203		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 15 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 15 kí tự - Từ ngắn nhất: 1 kí tự
M1050		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 15 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-2-4-5-8 kí tự - Từ dài nhất: 15 kí tự - Từ ngắn nhất: 1 kí tự

Xét đồ thị đường cong Mendenhall của các tác giả đã biết, ta có thể thấy rằng:

(1) Tất cả các tác giả đều sử dụng từ có 3 kí tự với tỉ lệ cao nhất và từ có 1 kí tự là từ ngắn nhất. Vì vậy việc xét riêng hai đặc trưng này không giúp phân biệt phong cách tác giả.

(2) Tuy nhiên, khi kết hợp đặc trưng chiếm tỉ lệ cao nhất với bốn đặc trưng có thứ hạng cao kế tiếp, gọi là top 5 chiếm tỉ lệ cao, ta có thể rút ra được kết luận như sau:

- Các tác giả có kiểu dùng từ chiếm tỉ lệ cao nhất với lần lượt 3-4-2-5-7 kí tự đó là tác giả F49 và F1083.

Nhưng nếu bạn có thể nhận được tiền từ bố_mẹ và sử_dụng chúng để thay_đổi cuộc_sống của người khác, thì đó là những gì tôi gọi là “thế_hệ thứ hai giàu_có”. (Tác giả F49)

- Các tác giả có kiểu dùng từ chiếm tỉ lệ cao nhất với lần lượt 3-4-2-5-8 kí tự đó là tác giả F890 và F1160.

Nếu vẫn không có chế_tài để kiểm_soát dầu_thải nói_riêng và chất_thải nói_chung từ nơi sinh ra nó, thì việc ô_nhiễm nguồn nước luôn là sự đã rồi, bất_kể bạn có nghĩ thấy chúng hay không. (Tác giả F890)

- Các tác giả có kiểu dùng từ chiếm tỉ lệ cao nhất với lần lượt 3-2-4-5-8 kí tự đó là tác giả F169, F414, F465 và F1193.

- Chỉ có tác giả F1083 có kiểu dùng từ chiếm tỉ lệ cao nhất với lần lượt 3-2-4-5-7 kí tự và duy nhất tác giả F129 có kiểu dùng từ chiếm tỉ lệ cao nhất với lần lượt 3-2-4-8-5 kí tự.

Chúng_ta cần trả lại cho điểm 9, điểm 10 vương_miền hào_quang của nó.... (Tác giả F129)

(3) Khi xét về từ có độ dài lớn nhất thì giữa các tác giả có sự khác biệt lớn:

- Tác giả M59 sử dụng từ độ dài lớn nhất gồm 17 kí tự.

- Tác giả M83 sử dụng từ độ dài lớn nhất gồm 16 kí tự.

- Tác giả M97 và M1050 sử dụng từ độ dài lớn nhất gồm 15 kí tự.

Lần đầu tiên Kiên đi ô tô tới tận văn phòng đón tôi ăn trưa. “Sao mà xịn thế”, tôi soi từ mái tóc xuống đôi giày mũi nhọn hoắt của thằng bạn thân hồi đại học. (Tác giả M97)

Bạn có muốn để lại cho con cháu mình một thế giới, nơi những bạn bè thân yêu nhất bị cướp đi, hành hạ và ăn thịt? (Tác giả M1050)

Hình 4. Đường cong đặc trưng Mendenhall của các văn bản ẩn danh.

Authors	Mendenhall curves	Information
X1		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 12 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-4-2-5-8 kí tự - Từ dài nhất: 12 kí tự - Từ ngắn nhất: 1 kí tự
X2		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 1 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-4-5-2-7 kí tự - Từ dài nhất: 10 kí tự - Từ ngắn nhất: 1 kí tự
X3		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 12 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-4-2-5-9 kí tự - Từ dài nhất: 15 kí tự - Từ ngắn nhất: 1 kí tự
X4		- Chiếm tỉ lệ cao nhất: từ có 3 kí tự - Chiếm tỉ lệ thấp nhất: từ có 11 kí tự - Chiếm tỉ lệ cao nhất: từ có lần lượt 3-4-2-8-5 kí tự - Từ dài nhất: 11 kí tự - Từ ngắn nhất: 2 kí tự

Kết quả ở Biểu đồ 4 cho thấy:

(i) Xét về nhóm từ chiếm tỉ lệ thấp nhất, văn bản X2 có thông số khác hẳn so với các tác giả đã biết, với nhóm từ gồm 1 kí tự là nhóm chiếm tỉ lệ thấp nhất.

(ii) Xét về từ chiếm tỉ lệ cao nhất, ta thấy rằng tất cả các tác giả đều sử dụng từ có 3 kí tự với tỉ lệ cao nhất.

(iii) Xét về độ dài từ, ngoài văn bản ẩn danh X4 có từ ngắn nhất là từ có 2 kí tự, thì cả ba văn bản ẩn danh còn lại đều sử dụng từ có 1 kí tự là từ ngắn nhất. Văn bản có từ dài nhất là văn bản X3, từ gồm 15 kí tự.

(iv) Xét về nhóm từ chiếm tỉ lệ cao nhất, văn bản X3 có những thông số khác biệt lớn so với các tác giả đã biết, với thứ hạng giảm dần tương ứng là 3-4-2-5-9.

Khi so sánh đường cong đặc trưng Mendenhall của bốn văn bản ẩn danh này với mười tác giả đã biết, chúng ta thấy rằng:

(v) X1 tương đồng với tác giả F890 về hình dáng đồ thị, tương đồng với tác giả F465 và F49 về số lượng kí tự trong từ dài nhất.

(vi) X2 tương đồng với tác giả F49 về hình dáng đồ thị, tương đồng với tác giả F465 về số lượng kí tự trong từ dài nhất.

(vii) X3 có rất ít điểm tương đồng về hình dáng đồ thị với mười tác giả đã biết, nhưng lại có độ tương đồng về số lượng kí tự trong từ dài nhất với các tác giả nam M97, M203 và M1050.

(viii) X4 tương đồng với các tác giả F465 và F49 về số lượng kí tự trong từ dài nhất.

4.3. Tần số từ loại

Các Bảng 3, 4 và 5 dưới đây lần lượt biểu diễn tần số từ loại trong các văn bản của 5 tác giả nữ, 5 tác giả nam và 4 tác giả ẩn danh.

Sample_code	Sample	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions	Other classes
F_49_2956939	F1_1	194	325	23	98	55	43	51	11
F_49_2958414	F1_2	195	324	42	66	78	38	40	17
F_49_3230024	F1_3	173	326	47	83	62	40	43	26
F_129_3072820	F2_1	169	284	68	72	82	42	65	18
F_129_3707810	F2_2	173	292	42	74	82	45	79	13
F_129_3772628	F2_3	203	286	29	94	65	47	66	10
F_414_3292585	F3_1	201	298	72	54	59	43	59	14
F_414_3665604	F3_2	214	301	78	47	66	27	46	21
F_414_3745372	F3_3	204	317	60	53	67	41	35	23
F_465_3806059	F4_1	213	297	50	66	49	55	61	9
F_465_3819460	F4_2	173	301	48	83	72	55	61	7
F_465_3913373	F4_3	190	299	51	74	76	35	64	11
F_890_3777816	F5_1	184	382	18	50	53	59	40	14
F_890_3908942	F5_2	177	405	20	54	50	50	37	7
F_890_3986445	F5_3	165	398	28	55	60	42	42	10

Bảng 3. Từ loại trong bài viết của các tác giả nữ trong VVC_AA

Sample_code	Sample	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions	Other classes
M_59_3467751	M1_1	175	315	50	68	72	57	54	9
M_59_3564512	M1_2	175	325	50	56	71	68	41	14
M_59_3715054	M1_3	176	346	54	50	67	47	42	18
M_83_3581012	M2_1	185	247	82	51	73	60	84	18
M_83_3599855	M2_2	184	249	81	71	78	50	75	12
M_83_3815138	M2_3	186	225	82	81	91	49	79	7
M_97_3788559	M3_1	171	317	43	89	62	50	50	18
M_97_3802665	M3_2	159	337	43	90	47	54	60	10
M_97_3862241	M3_3	175	326	40	88	43	51	65	12
M_203_3536291	M4_1	153	377	40	77	62	35	35	21
M_203_3561690	M4_2	140	331	33	89	61	48	62	36
M_203_3594835	M4_3	150	361	38	68	65	59	46	13
M_1050_3742860	M5_1	206	237	72	77	96	40	59	13
M_1050_3812920	M5_2	192	234	83	78	79	47	56	31
M_1050_3898207	M5_3	183	270	83	77	77	41	50	19

Bảng 4. Từ loại trong bài viết của các tác giả nam trong VVC_AA

Sample_code	Sample	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions	Other classes
	X1	200	315	45	69	67	34	56	14
	X2	190	270	57	87	88	26	57	25
	X3	192	313	41	86	64	51	41	12
	X4	219	293	42	60	72	41	66	7

Bảng 5. Từ loại trong các văn bản X1, X2, X3 và X4

Kết quả cho thấy rằng chỉ riêng tần số từ chưa đủ độ phân biệt để trở thành một đặc trưng đại diện cho phong cách viết của các tác giả trong khối ngữ liệu VVC, do là độ chính xác đạt được từ chỉ từ hơn 40% đến hơn 57% cho 10 tác giả. Kết quả này thấp hơn so với nghiên cứu của Grieve (2007) trên dữ liệu tiếng Anh cùng thể loại văn bản báo chí ý kiến, với kết quả của tần số từ khá cao, từ gần 50% đến hơn 60%.

4.4. Thực nghiệm xác định tác giả

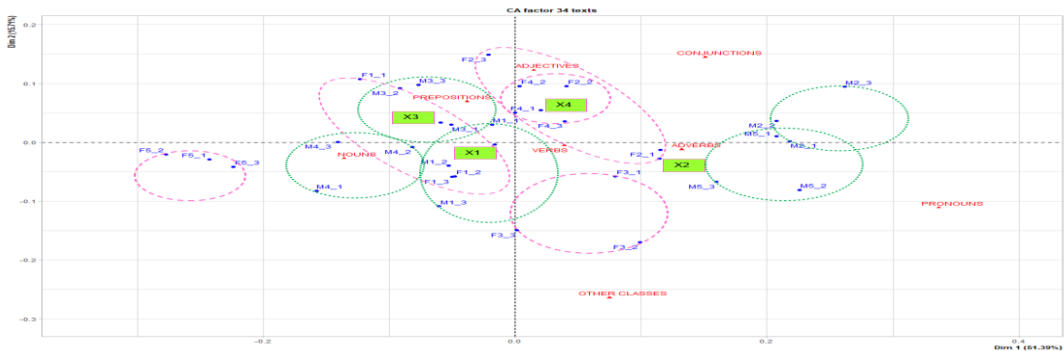
Bối cảnh thực nghiệm: Dữ liệu đầu vào: Khối ngữ liệu con VVC_AA gồm 30 bài báo ý kiến, với ba mẫu văn bản có trên 800 từ được chọn cho mỗi người trong số 10 tác giả. Cho 4 bài báo ý kiến bị ẩn danh X1, X2, X3 và X4, hãy xác định tác giả của văn bản này trong số danh sách 10 tác giả ứng viên.

Với phương pháp Phân tích tương ứng, các văn bản được khảo sát tần số tám nhóm từ loại: động từ, danh từ, đại từ, tính từ, phó từ, giới từ, liên từ và nhóm các từ loại khác. Phân tích này khảo sát tỉ lệ của

các từ loại khác nhau trong các mẫu văn bản này với vai trò là các biến ngôn ngữ. Cả văn bản của tác giả lẫn từ loại đều được hiển thị trên đồ thị.

Đồ thị tương ứng ở Hình 5 minh họa cho 14 tác giả, được gom nhóm theo cách sử dụng các từ loại khác nhau trong các bài viết của họ.

Hình 5. Đồ thị tương ứng: Từ loại trong các văn bản của 14 tác giả



Theo như Hình 5, ta có thể quan sát thấy rằng: X1 gần với M1 và F1, X2 gần với F2 và F3, X3 gần với M3, X4 gần với F4. Còn các tác giả F5, M2, M4 và M5 nằm cách xa bốn văn bản ẩn danh này. Như vậy, rất có khả năng các tác giả còn lại là F5, M2, M4 và M5 không phải là tác giả của các văn bản ẩn danh này. Ngoài ra, ta có thể quan sát rõ ràng nhất đó chính là các mẫu văn bản thuộc tác giả F3, bao gồm văn bản F3_1, F3_2 và F1_3 được gom nhóm ở phía trên hơi lệch về bên trái, trong khi các mẫu văn bản của tác giả M3, bao gồm M3_1, M3_2 và M3_3 được gom nhóm ở góc dưới bên phải. Và cả sáu mẫu văn bản này đều gần vùng trung tâm hơn so với các văn bản thuộc các tác giả M2, M5 và F5.

Xét văn bản X4: trong đồ thị này, văn bản ẩn danh X4 có sự gần kề nhất với các mẫu văn bản của tác giả F4 và nằm cách xa các mẫu văn bản của các tác khác. Tuy nhiên, để khẳng định văn bản ẩn danh là do ai viết, ta kết hợp thêm với các kết quả của các phép đo định lượng độ dài từ và tần số từ loại ở các tiêu mục 4.1 và 4.2 ở trên. Với một xác suất cao như vậy, từ đây ta có thể khẳng định rằng văn bản ẩn danh X4 có khả năng thuộc về tác giả F4 nhất. Lối viết của F4 có đặc điểm là sử dụng nhiều động từ, tính từ và giới từ; trong khi lại sử dụng tương đối ít đại từ so với các tác giả khác.

Bảng 5. Từ loại trong bài viết của tác giả F465 và văn bản X4

Sample_code	Sample	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions	Other classes
F_465_3806059	F4_1	213	297	50	66	49	55	61	9
F_465_3819460	F4_2	173	301	48	83	72	55	61	7
F_465_3913373	F4_3	190	299	51	74	76	35	64	11
	X4	219	293	42	60	72	41	66	7

Kết hợp những kết quả thu được, chúng tôi khẳng định rằng trong số 10 tác giả được khảo sát, văn bản X4 có nhiều khả năng là do tác giả F465 viết. Khi truy tìm trong khối ngữ liệu con VVC_X, chúng tôi tìm được tác giả của văn bản X4 là Đỗ Sông Hương, cũng chính là tác giả có mã số F465 trong khối ngữ liệu VVC.

5. Kết luận, một số hạn chế và hướng phát triển

5.1. Kết luận

Xác định danh tính tác giả có nhiều ứng dụng thực tiễn trong lĩnh vực Ngôn ngữ học tội phạm, đặc biệt là nhằm tìm ra tác giả của một văn bản điện tử, ví dụ như thư điện tử, bài đăng trên mạng xã hội, bài báo trực tuyến,... liên quan đến một vụ án. Trong nghiên cứu này, chúng tôi khảo sát một số phương pháp xác định danh tính tác giả hiệu quả nhất được sử dụng cho văn bản tiếng Việt. Như kết quả phân tích ở trên, ngay cả các công cụ tính toán đơn giản như sự thay đổi về tần số tương đối của một từ cũng có thể

cung cấp thông tin hữu ích về phong cách viết. Chúng tôi còn áp dụng thêm những mô hình phong cách trắc học phức tạp hơn để chỉ ra được các dạng thức ẩn trong các tập dữ liệu như bốn phép tính độ phong phú từ vựng R, S, K và W. Để đánh giá tính hiệu quả của các phương pháp định lượng này, chúng tôi tiến hành nghiên cứu thực nghiệm trên khối ngữ liệu bài viết ý kiến trên báo trực tuyến tiếng Việt VnExpress.

Kết quả tìm được cho thấy mặc dù từng đại lượng riêng lẻ chưa đủ độ phân biệt để trở thành một đặc trưng đại diện cho phong cách viết của các tác giả; tuy nhiên, nếu kết hợp phương pháp Phân tích tương ứng với biến ngôn ngữ là các từ loại cơ bản như danh từ, động từ, tính từ,..., cùng với các đại lượng còn lại, ta có thể xác định danh tính tác giả của văn bản tiếng Việt.

5.2. Một số hạn chế và hướng phát triển

Việc sử dụng các công cụ phân tích ngôn ngữ từ các thứ tiếng khác vào áp dụng cho tiếng Việt không phải lúc nào cũng phù hợp. Ngoài ra, khung nghiên cứu thông dụng nhất cho việc kiểm định các thuật toán là nhiệm vụ xác định danh tính tác giả trong tập ứng viên đồng, tức là cho trước một mẫu nghiên cứu gồm các văn bản tham chiếu của một nhóm nhỏ xác định gồm các tác giả ứng viên, nhiệm vụ đặt ra là xác định người có nhiều khả năng nhất là tác giả của một văn bản chưa biết nguồn tác giả. Nhiệm vụ này gặp nhiều thử thách trong các trường hợp khác lĩnh vực (cross-domain), trong đó các văn bản đã biết và chưa biết nguồn tác giả thuộc các lĩnh vực khác nhau, chẳng hạn như chủ đề, thể loại văn bản, v.v. Ngoài ra, thường thì trong thực tế, tác giả thực sự của một văn bản gây tranh cãi không phải lúc nào cũng nằm trong danh sách tác giả ứng viên. Nếu khắc phục được những hạn chế trên, bài toán tự động xác định danh tính tác giả trên văn bản tiếng Việt sẽ có nhiều bước phát triển mạnh mẽ.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Nguyễn Tuyết Nhung, Võ Diệp Như, Hồ Ngọc Lâm, Đinh Điền, *Gán nhãn siêu dữ liệu cho khối ngữ liệu VnExpress Viewpoint Corpus (VVC)*, Kỷ yếu Hội thảo Ngôn ngữ học và những lĩnh vực ứng dụng - lần I, 12, 2020.

Tiếng Anh

2. Brezina, V., *Statistics in corpus linguistics: A practical guide*, Cambridge University Press, Cambridge, 2018.

3. Burrows, J., *Not unless you ask nicely: The interpretative nexus between analysis and information*, Literary and Linguistic Computing 7(2): 91-109, 1992. . DOI: 10.1093/llc/7.2.91

4. Crystal, D., *English as a global language*, Cambridge University Press, 2003.

5. Holmes, D.I., *The Federalist revisited: new directions in authorship attribution*, Literary and Linguistic Computing 10(1), 111-127, 1995.

6. Hồ Ngọc Lâm, Võ Diệp Như, Nguyễn Tuyết Nhung, Đinh Điền, *Identifying authors based on stylometric measures of Vietnamese texts*, Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, pp. 447-452, 2020.

7. Lakoff, G., Wehling, E., *The little blue book: The essential guide to thinking and talking democratic*, Free Press, New York, 2012.

8. Nguyen, T. N., Do, T. A. D., & Dinh, D., *Applying the text stylometry in detecting the gender of authors in Vietnamese texts*, The International Workshop on Vietnamese studies and Vietnamese linguistics, Hue, 2018.

9. Oakes, M. P., *Literary detective work on the computer* (Natural Language Processing, ISSN 1567-8202), 2014.

10. Savoy, J., *Machine learning methods for stylometry*, Springer Nature Switzerland AG, 2020. DOI: 10.1007/978-3-030-53360-1

11. Stamatatos, E. et al., *Overview of PAN 2018: author identification, author profiling, and author obfuscation*, In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol.11018, pp. 267-285, 2018. Springer, Cham. DOI: 10.1007/978-3-319-98932-7 25

12. Tweedie, F. & Baayen, R. H., *How variable may a constant be? Measures of lexical richness in perspective*, Computers and the Humanities, 32, 323-52, 1998.