Identifying Authors Based on Stylometric measures of Vietnamese texts

Ho Ngoc Lam
Ho Chi Minh City University of Education
ngoclam0706@gmail.com

Vo Diep Nhu
VNUHCM-University of Science
vodiepnhu@gmail.com

Dinh Dien
VNUHCM-University of Science
ddien@fit.hcmus.edu.vn

Nguyen Tuyet Nhung
VNUHCM-University of Social Sciences and Humanities
velvetsnow.nguyen@gmail.com

Abstract

Author identification has many applications in investigating or resolving authorship disputes. Research on author identification has been conducted in many high resource languages, such as English, Chinese, Spanish, etc. However, for Vietnamese, studies are limited because of the lack of relevant language resources. This paper represents the topic of author identification with the application of stylometric methods: Mendenhall's characteristic curve, Kilgariff's squared method (Kilgariff's Chi-Squared), the Delta method of John Burrows. The study applied three different methods based on a corpus extracted from Vietnamese online newspapers, categorized by each author and achieved results from 50% to 100% depending on the method and number of linguistic features.

1 Introduction

International integration, along with the exponential growth of the Internet, has led to an increase in plagiarism, imitation of celebrities' writing style, and copyright disputes.

Due to the enormous amount of information, looking for the style and characteristics of written works in order to identify the author's style is a huge challenge. Globally, there have been numerous studies which find out models to identify the author's style in many languages. However, there are very few studies in natural language processing applying writing style in Vietnamese to attribute authorship.

Stylometry, beginning with attempts to settle authorship disputes, was first developed by Augustus De Morgan in 1851 based on word length. By the late 1880s, Thomas C. Mendenhall had analyzed the word length distribution for works written by Bacon, Marlowe, and Shakespeare to determine the true author of plays supposedly written by Shakespeare. In 1932, George Kingsley Zipf discovered the connection between ranking and the frequency of words, later stated in Zipf's law. In 1944, George Yule created a way to measure frequency of words, used to analyze vocabulary richness, namely Yule's characteristic. In the early 1960s, most research papers refer to Mosteller and Wallace's works on the Federalist Papers, which was considered as a basis of using computation in stylometry. In the next several decades, with the increasing number of digital texts, as well as the growth of the Internet, machine learning techniques, and neural networks, accessing information led to the development of natural language processing tools. Semantics continued to grow in the 21st century, and due to the overwhelming amount of information, copying texts also became more popular, leading to the growth of stylometry which is used in plagiarism detection, author identitication, author profiling, etc.

In this paper, we use a corpus of Vietnamese online texts to attribute authorship using the following measures: Mendenhall's characteristic curves, Kilgariff's Chi-Squared, John Burrows's Delta measure.

2 Related work

Broadly, there are two categories of stylometry:

Adversarial Stylometry: When translated, a piece of writing has its style imitated, and going through many translators makes its characteristics less distinct. These changes make detecting the original style more difficult.

Detecting stylistic similarities includes the following tasks:

Stylochronometry: In time, an author may change his/her writing style due to changes in vocabulary, lifestyle, environment, age, etc. Studies have sharp distinction because they depend on a language in a specific time period and on a particular author. Author Profiling: extracting the characteristics of a text to gain information about an author such as gender, age, region, time of writing.

Authorship Verification: Based on characteristics readily available in the training data, determining whether two texts were written by the same author. Authorship Attribution: an individual or group of authors has characteristic styles that are developed subconsciously. Based on these distinctions, we will identify the true author(s) of texts in a corpus.

3 Experimentation

In authorship identification using corpus-based approach, we use the NLTK Python package to process the corpus in order to execute the methods of author attribution. Due to limitations in the number of the texts per author, we will choose only 10 authors whose texts contain appropriate number of sentences and words and closely similar in size. Depending on methods and characteristic numbers, our results vary between 50% and 100%.

3.1 Corpus

We use a corpus of Vietnamese online texts, including 1304 texts extracted from several Vietnamese online newspapers (largely from VnExpress), Facebook, and blogs. These texts are written by 10 authors, who give their own opinion or share their own experiences on social issues. The corpus was pre-processed to eliminate links, images, captions, and tokenized semiautomatically. The process of tokenization was carried out with CLC toolkit, an automatic tool developed by Computational Linguistics Center (VNUHCM-University of Science). Then we manually checked the whole corpus and correct the mistakes. The number of texts and tokens of each author are displayed in Table 1 below.

No.	Authors	Corpus			Test set
		Texts	Tokens (Include punctuations)	Tokens (Exclude punctuations)	Tokens (Exclude punctuations)
1	Author59	162	118,148	101,375	43,287
2	Author83	45	84,010	72,613	14,525
3	Author88	151	110,930	96,404	38,316
4	Author97	91	99,910	85,781	27,693
5	Author203	108	94,452	79,747	21,659
6	Author1028	152	87,341	75,029	16,941
7	Author1035	184	122,028	101,938	43,850
8	Author1050	133	120,337	102,653	44,565
9	Author1262	121	173,865	148,085	89,997
10	Author1289	157	118,267	102,940	44,852
Total		**1304**	**1,129,288**	**966,565**	**385,685**

Table 1. Information of the corpus and test set

3.2 Stylometric measures

Measure 1: Mendenhall's characteristic curves

Mendenhall once wrote that an author's "stylistic signature" could be found by measuring the frequency with which he or she used words of different lengths. These characteristic curves give results quickly and visually, allowing the researcher to draw a conclusion on the author's style. Applying this method, our group worked on our dataset of works by ten chosen authors. To standardize the size of the text while applying this method, we made the token number in works from each author's bibliography 58,088 token (punctuations removed). On each author's bibliography, we sequentially did the following: calculating the length of each token, calculating the frequency of calculated length in the bibliography, and visualize the data. Besides the visualized data, we use Caroll's index R to measure each author's lexical diversity to have an overview of style:

$$\boldsymbol{R = \frac{V}{N}}$$

V: vocabulary size (number of word types)

N: text size (number of word tokens)

Figure 1. Equation for lexical diversity

Measure 2: Kilgariff's Chi-squared

In the dataset whose authors are known, namely

Known: let denote the file of i^{th} candidate author K_i (i = 1, 2, … 10)

Let denote the unknown author's file U.

Calculate Chi-squared for each of the ten candidate authors.

1. First, build a joint corpus J, including K_i and U, and identify the 500 most frequent words in it.

2. Calculate the proportion of the joint corpus made up of the candidate author's tokens (AuShare).

AuShare = len(token K_i)/len(token Jcorpus)

3. Look at the 500 most common words in the candidate author's corpus and compare the number of times they can be observed to what would be expected if the author's file and the disputed file were both random samples from the same distribution.

4. Calculate how often we really see each of the 500 most common words, cw[x] (x = 1, 2, … 500), in K_i and U respectively with:

- Kcw_ob: observed number of cw in K_i
- Kcw_ob: observed number of cw in U

5. Calculate how should we see each cw in K_i and U respectively with:

- Kcw_ex: expected number of cw in K_i
- Ucw_ex: expected number of cw in U

6. Calculate a chi-squared distance of K_i and U:

$$\chi^2 = \chi^2_{K_i} + \chi^2_U$$

Respectively calculate chi-squared of K_i and U:

$$\chi^2_{K_i} = \sum_x \frac{(Kcw\_ob - Kcw\_ex)^2}{Kcw\_ex}$$

$$\chi^2_U = \sum_x \frac{(Ucw\_ob - Ucw\_ex)^2}{Ucw\_ex}$$

Figure 2. Equations for the chi-squared statistic of K_i and U.

The smaller the chi-squared value, the more similar the two corpora. Therefore, we will calculate a chi-squared for the difference between each file of the candidate author dataset Known and disputed file U; the smaller value will indicate which of Known is the most similar to U.

Measure 3: John Burrows' Delta measure

The Delta measure, proposed by John F. Burrows as a tool to solve the problem of copyright, measures the difference between two sets of text.

1. Combine all files in Known into a single corpus and get n frequency distribution words (test in n=20, n=30 respectively)

2. Calculating n[y] (y =1, 2, …, n) presence for each subcorpus K_i.

3. Calculating n[y] means (μy) and standard deviations (σy).

4. Calculating z-scores:

$$\boldsymbol{Z_i = \frac{C_i - \mu_i}{\sigma_i}}$$

C_i: the observed frequency

μ_i: means

σ_i: standard deviation

Figure 3. z-scores calculate the z-score in the test set.

5. Calculating features and z-scores for our test file

6. Calculating Delta

Find Delta point to compare the test set with each author.

$$\Delta_c = \sum_i \frac{|\boldsymbol{Z_{c(i)} - Z_{t(i)}}|}{\boldsymbol{n}}$$

$Z_{c(i)}$: z-score for feature i in subcorpus 'c'

$Z_{t(i)}$: z-score for feature i in the test set

Figure 4. Delta measure

3.3 Results

Measure 1: Mendenhall's characteristic curves

The results are shown in Figure 5. We observe that each author has the following features: Author59's longest word contains 17 characters, while that of Author203 and Author1050 only has 14.

Every author uses words having between 2 and 4 characters the most. The most prevalent word has 3 characters.

Author1035 and Author1262 yield different results from the other authors. Each of them uses 3letter words the most, followed by 4-letter words instead of 2-letter words like the other eight authors. The authors' lexical diversity: When examined with the same 58,088 tokens, the author having the highest lexical diversity (Caroll index R) is Author1035 with 0.146 whereas the one with the lowest diversity is Author83 with 0.092. The results are shown in Table 2.

No.	Authors	Lexemes	Vocabulary richness (R)
1	Author59	5961	0.103
2	Author83	**5327**	**0.092**
3	Author88	6781	0.117
4	Author97	6922	0.119
5	Author203	8206	0.141
6	Author1028	6467	0.111
7	Author1035	**8456**	**0.146**
8	Author1050	6136	0.106
9	Author1262	7498	0.129
10	Author1289	6259	0.108

Table 2. Vocabulary Richness

STT	Authors / Authors	Author59	Author83	Author88	Author97	Author203	Author1028	Author1035	Author1050	Author1262	Author1289
1	Author59	**2,549.649**	6,775.924	7,995.894	7,241.717	4,740.681	5,447.395	8,319.614	9,016.777	23,040.670	14,850.658
2	Author83	9,678.276	**2,359.664**	8,415.828	6,905.708	4,433.277	4,812.310	9,282.713	9,058.835	23,650.604	13,255.750
3	Author88	8,473.703	6,471.436	**2,154.291**	7,169.004	3,825.872	5,402.131	7,798.770	7,443.396	14,637.402	8,640.629
4	Author97	9,013.078	7,526.163	10,346.465	**1,874.816**	4,372.070	4,435.556	7,266.010	10,843.907	25,638.256	17,720.348
5	Author203	6,527.751	5,967.048	4,146.237	5,678.605	**2,267.813**	3,991.429	5,421.921	5,989.636	16,631.327	9,929.698
6	Author1028	10,234.820	8,515.822	11,195.689	7,291.693	6,462.341	**3,348.786**	8,587.273	12,500.521	23,682.286	17,239.991
7	Author1035	7,622.525	8,220.965	9,796.556	6,694.397	4,728.346	4,963.033	**3,366.667**	9,957.941	21,031.091	16,496.578
8	Author1050	9,416.742	6,611.444	6,650.636	8,935.540	4,336.624	6,179.540	9,129.678	**1,787.953**	17,136.267	10,778.297
9	Author1262	16,449.314	10,470.761	9,555.598	13,669.770	7,137.432	9,292.752	12,779.942	13,189.267	**2,850.582**	14,699.759
10	Author1289	10,584.509	6,665.983	5,901.024	8,819.146	5,094.171	6,564.798	10,761.940	8,324.811	19,483.718	**3,339.960**

Table 3. Chi-squared results

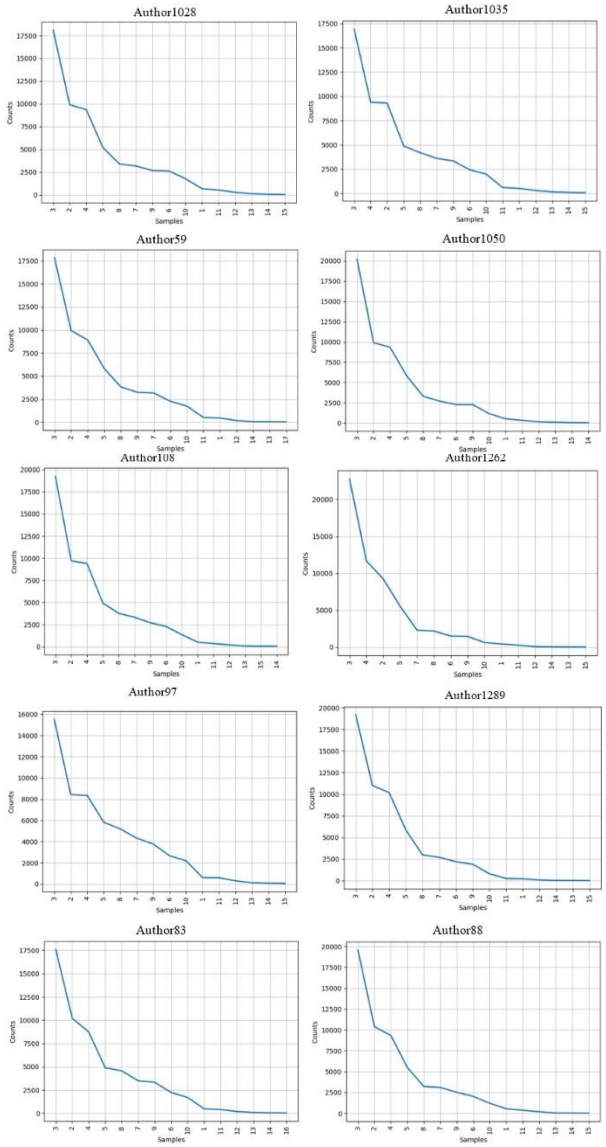

Figure 5. The Mendenhall's characteristic curves

Measure 2: Kilgariff's Chi-squared

Grieve (2007) assumed that the lower the chisquared measure between two texts, the more likely that they were written by the same author. Therefore, the known text giving the smallest Chi-squared value would be written by the author most likely to have written the unknown text. Table 3 below shows the Chi-squared results when we tested the text sample for each of the authors.

Measure 3: John Burrows' Delta

Table 4 and Table 5 display the results of Delta measure when we tested on each of the authors' text (according to the headings). Examined by rows, the smaller the Delta value is, the closer to the author's style the test work is.

After we tested on 30 signatures, the result yields 40%, matching the prediction on 4 out of 10 authors: Author59, Author97, Author1262, Author1289. The results are shown in Table 4.

The 30 signatures include: 'là', 'không', 'và', 'của', 'có', 'một', 'người', 'tôi', 'những', 'cho', 'được', 'các', 'thì', 'trong', 'với', 'đó', 'đã', 'cũng', 'để', 'phải', 'mà', 'ở', 'như', 'khi', 'này', 'mình', 'đến', 'về', 'sẽ', 'đi'.

Authors / Authors	Author 59	Author 83	Author 88	Author 97	Author 203	Author 1028	Author 1035	Author 1050	Author 1262	Author 1289
Author59	***0.003***	***0.005***	0.059	0.017	0.013	***0.005***	***0.005***	0.048	0.095	0.039
Author83	0.014	0.014	0.050	0.027	0.004	0.007	0.018	0.036	0.086	0.027
Author88	0.055	0.057	0.007	0.070	0.040	0.050	0.059	***0.007***	0.043	0.016
Author97	0.019	0.018	0.081	***0.004***	0.035	0.024	0.017	0.066	0.116	0.058
Author203	0.019	0.021	0.043	0.034	0.004	0.014	0.023	0.031	0.078	0.022
Author1028	0.018	0.017	0.049	0.031	***0.003***	0.011	0.022	0.035	0.085	0.026
Author1035	0.042	0.042	0.103	0.029	0.057	0.049	0.039	0.091	0.139	0.082
Author1050	0.061	0.061	0.003	0.074	0.046	0.054	0.065	0.011	0.039	0.020
Author1262	0.062	0.064	***0.002***	0.076	0.046	0.057	0.063	0.014	***0.037***	0.023
Author1289	0.027	0.027	0.037	0.040	0.012	0.020	0.030	0.023	0.072	***0.014***

Table 4. Experimental results of 30 most frequent lexemes

After we tested on 20 signatures, the result yields 50%, matching the prediction on 5 out of 10 authors: Author83, Author203, Author1035, Author1262, Author1289. The results are shown in Table 5. The 20 signatures include: 'là', 'không', 'và', 'của', 'có', 'một', 'người', 'tôi', 'những', 'cho', 'được', 'các', 'thì', 'trong', 'với', 'đó', 'đã', 'cũng', 'để', 'phải'.

Authors / Authors	Author 59	Author 83	Author 88	Author 97	Author 203	Author 1028	Author 1035	Author 1050	Author 1262	Author 1289
Author59	0.039	0.053	0.042	0.033	0.064	***0.004***	0.081	0.042	0.096	0.068
Author83	***0.012***	***0.002***	0.087	0.081	0.115	0.047	0.132	0.091	0.147	0.020
Author88	0.061	0.067	0.022	0.016	0.050	0.029	0.067	0.026	0.082	0.088
Author97	0.076	0.091	0.004	0.007	0.026	0.042	0.044	0.004	0.059	0.106
Author203	0.098	0.112	0.023	0.029	***0.007***	0.063	0.022	0.019	0.038	0.127
Author1028	0.078	0.093	***0.004***	0.009	0.024	0.044	0.041	***0.003***	0.057	0.108
Author1035	0.134	0.149	0.059	0.065	0.032	0.100	***0.015***	0.056	0.002	0.164
Author1050	0.076	0.090	0.005	***0.007***	0.027	0.041	0.044	0.004	0.060	0.105
Author1262	0.137	0.151	0.061	0.068	0.034	0.102	0.017	0.058	***0.002***	0.166
Author1289	0.013	0.007	0.090	0.081	0.115	0.047	0.132	0.091	0.147	***0.017***

Table 5. Experimental results of 20 most frequent lexemes

4 Discussion

Among the three measures mentioned above, Delta measure does not yield good results as we expected.

In Chi-square statistic, we convert everything to lowercase so that we won't count word tokens that begin with a capital letter because they appear at the beginning of a sentence and lowercased tokens of the same word as two different words. Sometimes this may cause a few errors, for example when a proper noun and a common noun are written the same way except for capitalization, but usually it increases accuracy. In addition, Chi-squared is a coarse method. For one thing, words that appear very frequently tend to carry a disproportionate amount of weight in the final calculation. Sometimes this is fine; other times, subtle differences in style represented by the ways in which authors use more unusual words will go unnoticed. (Laramée, 2018).

The algorithm based on taking the number of the most common words (words with highest frequency) in the corpus as a feature. In the VnExpress corpus, we get texts from the "Perspective" section, which offers a wide variety of topics, such as finance, society, lifestyle, health, etc. Not all authors write about the same topics, and relativity among topics leads to an inconsistency of topics in the corpus.

Even though we have processed on the sets with the same token number of each author, the disparity in topics may be the reason why the chosen features are biased towards certain authors, rather than representing the whole corpus.

4.1 Conclusion

Research in authorship identification in Vietnamese text is uncommon despite its high applicability in many fields. In fact, researchers face difficulties in finding a corpus with sufficient size and information about authors.

In this paper, we have presented three different measures of authorship identification; these are the basic methods of determining an author's style such as lexical diversity, number of characters in a word, and word frequency (to find the most frequent words). The Chi-squared measure yields 100% accuracy; whereas Burrows' Delta measure yields 40% accuracy with 30 features, and 50% accuracy with 20 features.

In future research, we will the examining on a corpus with a wide variety of topics to increase lexical variety. At the same time, we will prepare a richer annotated corpus so as to work on authorship identification using machine learning.

References

Alex I. Valencia Valencia, Helena Gomez Adorno, Christopher Stephens Rhodes & Gibran Fuentes Pineda. 2019. *Bots and Gender Identification Based on Stylometry of Tweet Minimal Structure and n-grams Model*. Notebook for PAN at CLEF.

Andrea Bacciu, Massimo La Morgia, Eugenio Nerio Nemmi & Valerio Neri. 2019. *Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features*. Notebook for PAN at CLEF.

Antonio Pascucci, Vincenzo Masucci & Johanna Monti. 2019. *Computational Stylometry and Machine Learning for Gender and Age Detection in Cyberbullying Texts*. IEEE.

Carmen Klaussner & Carl Vogel. 2015. *Stylochronometry: Timeline Prediction in Stylometric Analysis*. Springer International Publishing, Switzerland.

Divjak, D. 2019. *Frequency in Language Memory, Attention and Learning*. Cambridge University Press.

Eder, M. 2015. *Rolling stylometry*. Oxford University Press on behalf of EADH.

Hoover, D.L. 2004. *Testing Burrows's Delta. Literary and Linguistic Computing*. 19(4):453-475.

Imene Bensalem, Paolo Rosso & Salim Chikhi. 2014. *Intrinsic Plagiarism Detection using N-gram Classes.* Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Cornell University, arXiv: 1810.04805v2.

Kyung-Ah Sohn, Alemu Molla Kebede & Kaleab Getaneh Tefrie. 2014. *Anonymous Author Similarity Identification*. IEEE Symposium on Security and Privacy.

K. Surendran, O. P. Harilal, P. Hrudya, Prabaharan Poornachandran & N. K. Suchetha. 2017. *Stylometry Detection Using Deep Learning*. Springer Nature Singapore Pte Ltd.

Love, H. 2002. *Attributing authorship: An introduction.* Cambridge University Press, pp. 133.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming & F. J. Smith. 2003. *Extension of Zipf's Law to Word and Character N-grams for English and Chinese*. The Association for Computational Linguistics and Chinese Language Processing, 77-102.

Le Thanh Nguyen & Dinh Dien. 2019. *English-Vietnamese Cross-Language Paraphrase Identification Method.* Springer.

Le Thanh Nguyen, Nguyen Xuan Toan & Dinh Dien. 2016. *Vietnamese plagiarism detection method*. University of Florida, ACM, 44–51. https://doi.org/10.1145/3011077.3011109.

Mahmoud Khonji & Youssef Iraqi. 2017. *De-anonymizing Authors of Electronic Texts: A Survey on Electronic Text Stylometry*. Preprints.

Mendenhall, T. C. 1887. *The Characteristic Curves of Composition. Science*, 9(214): 237-249.

Sadia Afroz, Aylin Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt & Damon McCoy. 2014. *Doppelgänger Finder: Taking Stylometry To The Underground.* IEEE Symposium on Security and Privacy.

Shaina Ashraf, Hafiz Rizwan Iqbal & Rao Muhammad Adeel Nawab. 2016. *Cross-Genre Author Profile Prediction Using Stylometry-Based Approach*. Notebook for PAN at CLEF.

Shuiyuan Yu, Chunshan Xu & Haitao Liu. 2018. *Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation*. Cornell University, arXiv:1807.01855.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang & Damon Woodard. 2017. *Surveying stylometry techniques and applications.* University of Florida, ACM Comput, Surv. 50, 6, Article 86. https://doi.org/10.1145/3132039.

Zipf, G.K. 1968. *The psycho-biology of language: an introduction to dynamic philology*. M.I.T. Press.