USING CONTENT AND NON-DICTIONARY WORDS FOR AUTHOR PROFILING OF VIETNAMESE FORUM POSTS

Duong Tran Duc[1], **Pham Bao Son**[2], **Tan Hanh**[1]
[1] Posts and Telecommunications Institute of Technology
[2] University of Engineering and Technology, Vietnam National University, Hanoi

Abstract: This paper reports the results of author profiling task for Vietnamese forum posts to identify personal traits, such as gender, age, occupation, and location of the author using content and non-dictionary words. Experiments were conducted on different types of features, including stylometric features (such as lexical, syntactic, structural features), content-based features (the most important content words), non-dictionary words (such as slangs, abbreviations) to compare the performance and on datasets we collected from popular forums in Vietnamese. Three learning methods, consisting of Decision Tree, Bayes Network, Support Vector Machine (SVM), were tested and SVM achieved the best results. The results show that these kinds of features work well on such a kind of short and informal messages as forum posts, in which, content words features yielded much better results than stylometric and non-dictionary words features when used individually. However, the combination of stylelometric and non-dictionary words also achieved good results.

Keywords: author profiling, machine learning, content-based features, non-dictionary words.

I. INTRODUCTION

The rapid growth of World Wide Web (WWW) has created a lot of online channels for people to communicate, such as email, blogs, social networks, etc. However, online forums are still among the most popular channels for people to share opinions and discuss about the topics which are interested in common. Forum posts created by users can be considered as informal and personal writings. Authors of these posts can indicate their profiles for other people to view as a function of forum. But not many users reveal their personal information, because of information privacy issues on online systems. Moreover, personal information of users is not mandatory to input when they register as a user of forums. Therefore, most of people do not provide their personal information or input the incorrect/unclear data.

As a result, the task of automatically classifying the author's properties such as gender, age, location, occupation, etc. becomes important and essential. Applications of this task can be in commercial field, in which providers can know which types of users like or do not like their products/services (for targeted marketing and product development). In social research domain, researchers also want to know profiles of people who have a specific opinion about some social issues (when doing a social survey). It can also be used to support the court, in term of identifying if a text was created by a criminal or not.

Profiling the author of forum posts is also a challenging task when compared to doing this on other formal types of text such as article or novel or even the other types of online texts such as blog posts or email. Forum posts are often short and written in free style, which may contain grammar errors or informal sentence structures.

Most of earlier works in author profiling were conducted on other types of text (blog posts, email) and focused on using the stylometric features (or only small part of content-based features). This

Corresponding Author: Duong Tran Duc
Email: duongtranduc@gmail.com
Manuscript received: 23/7/2016, revised: 30/8/2016, accepted:03/9/2016

work presents a study in which we applied the machine learning algorithms to predict profiles of authors of forum posts using both types of features. Motivations for this work are:

- Only few previous works (e.g. [13]) on author profiling were done on forum posts, especially none of them was tested on Vietnamese. The work of Abbasi and Chen [1] was conducted on forum posts, but for author attribution, not author profiling task.
- Only one research in author profiling was done in Vietnamese [6], but was tested on blog posts, and used stylometric features only. Our work is not only conducted on a more informal and noisier type of document, but also explored the use of content and non-dictionary words features.

The organization of the paper is as follows. In the section II, we present the related work on author analysis problem. Section III describes our methods and system. Section IV presents results and discussion. In section V, we draw a conclusion and future work.

II. RELATED WORK

The problem of authorship analysis has been studied for decades, mostly on English and some other languages (Dutch, French, Greek, Arabia etc.). In the early stage, it was often conducted on the long and formal documents such as article or novel. However, since 1990s, when the WWW grew and created a large amount of online text, the task of author analysis has moved the focus to this type of text.

According to Zheng et al. [23], the authorship analysis studies can be classified into three major fields, including authorship attribution, authorship profiling, and similarity detection.

Authorship attribution is the task of determining if a text is likely written by a particular author or not. It also is the technique to identify which one from a set of infinite authors is the real author of a disputed document. Therefore, it is also called authorship identification. The first study in this field dates back to 19th century when Mendenhall (1887) investigated Shakespeare's plays. But the work which was considered the most thorough study in this field was conducted by Mosteller and Wallace (1964) when they analyzed the authorship of Federal List Papers. From that point, a number of works have been conducted by various researchers, including [2], [5], [7], [10], [18], [20], [23].

Authorship profiling, also known as authorship characterization, detects the characteristics of an author (e.g. gender, age, educational background, etc.) by analyzing texts created by him/her. This technique is different from the former in that it is often used to examine the anonymous text, which is created by an unknown author, and generates the profile of the author of that text. For this reason, the author profiling task is often conducted on online documents rather than literary texts. Therefore, this field is only more concerned by researchers from the late of 1990s, when more and more online documents are created by Internet's users. The most typical studies in this fields are from [2, 3, 4], [6], [8, 9, 10, 11], [13, 14, 15], [17], [19], [21].

Similarity detection, on the other hand, doesn't focus on determining the author or his/her characteristics, but analyzes two or more documents to find out if they are all created by the same author or not. This technique is also used to verify if a piece of text is written by the author himself/herself or copied from the product of other authors. This task is mostly used for plagiarism detection. Some of the most convincing studies in this field were conducted by [2], [5], [7], [10].

Regarding the process of authorship analysis, there are two main issues that may significantly affect the performance, namely features set and analytical techniques [23].

Features set can be considered as a way to represent a document in term of writing style. With a chosen features set, a document can be represented as a features vector in which entries represent the frequency of each feature in the text

[11]. Although various types of features have been examined, there is no features set that is the best to all the cases. According to Argamon et al. [4], there are two types of features that often can be used for authorship profiling: stylometric features and content-based features.

Stylometric features can be grouped into three types, including lexical, syntactic, and structural features. Lexical features are used to measure the habit of using characters and words in the text. The commonly used features in this kind consist of the number of characters, word, frequency of each kind of characters, frequency of each kind of words, word length, sentence length [7], and also the frequency of individual alphabets, special characters, and vocabulary richness [10]. Syntactic features include the use of punctuations, part-of-speeches, and function words. Function words feature is the interesting kind of features, which is examined in a number of studies and yielded very good results ([10], [19], [23]). The set of function words used is also varying, from 122 to 650 words. Structural features show how the author organizes his/her documents (sentences, paragraphs, etc.) or other special structures such as greetings or signatures ([5], [10]).

Content-based features are often specific words or special content which are used more frequent in that domain than in other domains [22]. These words can be chosen by correlating the meaning of words with the domain ([2], [10], [22]) or selecting from corpus by frequency or by other feature selection methods [4].

Non-dictionary words such as slangs, abbreviations are used commonly by online community recently. They often are the intentionally misspelled or shorten words which may contain meaning or not. Therefore, they may belong to the stylometric or content-based type. As far as we have known, there are only few author profiling works on this kind of feature [8].

Also the investigation of Zheng et al. [22] showed that, in early studies most authorship analytical techniques were statistical methods, in which the probability distribution of word usage in the texts of each author was examined. Although these methods achieved good results in authorship analysis, there are still some limitations, such as the ability to deal with multiple features or the stability over multiple domains. To overcome those limitations, the extensive use of machine learning techniques has been investigated. Fortunately, the advent of powerful computers allows researchers to conduct the experiments on complicated machine learning algorithms, in which Support Vector Machine (SVM) shows the better results in many cases ([1], [2], [5, 6, 7], [10, 11], [15], [17], [19], [23]). Some other machine learning algorithms also have been examined and yielded good results, including Bayesian Network, Neural Networks, Decision Tree ([4], [10], [19], [22]). In general, machine learning methods have advantages over statistical methods because they can handle the large features sets and the experiments also shown that they achieved the better results.

III. SYSTEM DESCRIPTION

A. System overview

In this work, we built a system which can take sample texts from web crawlers, then used text and linguistic processing components to extract features to create the data sets for the purpose of training the classifier. The classifier then can be used to predict the profile of the author of an anonymous forum post.

In the data processing step, data is cleaned and grouped by author profiles. Unlike the gender and location trait, which can be divided into two groups (male/female, north/south), the other traits are grouped by more than 2 classes. For age trait, we categorized our data into 3 subclasses (less than 22/24-27/more than 32). Age is categorized according to the life stages of a person (students or pupils/young working adults/middle-age people) and age periods are not continuous because distinguishing two contiguous ages is almost impossible. With the occupation trait, we tried to identify three occupations which are the most popular (business, sale, and administration/ technical and technology/education and healthcare).

Linguistic processing is the task of tokenizing the text into sentences or word and the tagging for part-of-speeches. These tasks are important for extracting the word and syntactic features in the next step. In this work, we used existing tools from [16].

In the next sections, we describe the features and techniques which were used for classification in detail.

B. Features

As mentioned earlier, various features can be used to identify the characteristics of an author. In this work, we used both stylometric and content-based features.

Stylometric features include character-based, word-based, structural, and syntactic features. Character-based features include the number of characters in total and the ratio of each type of characters (number, letter, special, etc.) or each individual character (letters from a to z, and the special characters such as @, #, etc.) to the total number of characters. Some other features related to character are the average number of characters per word, per sentence, the number of upper case letters or how the author uses upper case letters in a word, etc. Word-based features group consists of the total number of words of a post, the average number of words per sentence, and the ratio of some types of word to the total number of words, such as words with a specific length, special words, the vocabulary richness (hapax legomena, hapax dis legomena etc.). Syntactic features indicate the use of punctuations such as "!","?", function words, and part-of-speech tags. Function words chosen are words which have little lexical meaning and express the grammatical relationship with other words in a sentence (212 Vietnamese function words). Part-of-speech tags include 18 word types, such as noun, verb, preposition, etc. Structural features present the structure of a post, such as the number of paragraphs, number of lines, etc.

Content-based features used in our work were chosen from the corpus, which are content words that can discriminate best between classes of each trait. Firstly, these words were selected based on the frequency of them in the corpus (separately by classes of each trait). Then the Information Gain method was applied to select the best features. Information Gain is one of the most popular feature selection methods, which attempts to measure the significance of each feature in distinguishing between classes. This method was tested on various previous works and yielded the good result.

For gender trait, we selected 3000 words which were used most frequently by male/female separately. After eliminating the identical words and applied the Information Gain method, we chose 1000 words which have highest significance.

Using the similar process, we chose about 1000 most significant words to use as content-based features for discriminating the age, occupation, and location traits.

We also extracted 170 non-dictionary words from the corpus and used as features for our author profiling work. They are slangs or abbreviations which are commonly used by forum users to express the emotion or save time of typing. Some of them express the meaning but others only serve as function words.

All of these features are extracted from the text and store in a numeric vector. For features which need some kinds of linguistic processing activities, such as the word segmentation or the part-of-speech tagging, we used existing tools available for Vietnamese. Extracted features are stored in the features containers, then are sent to classifiers for training purposes and prediction models are built for classifying the new data.

We also conducted experiments on subsets of features, including stylometric features, content words, non-dictionary words, and combinations for analysis of performance of each type.

IV. EXPERIMENTS

A. Data

There are a number of Vietnamese forums, of which we can collect the data. However, each of them often serves for a specific type of user only

(e.g. for ladies or gentlemen) or for a specific subject of interest such as technology, automobile etc. Therefore, we selected three forums to collect data to ensure that the data collected will cover a wide range of users and subjects.

- Webtretho forum (www.webtretho.com/forum): A forum for girls and ladies to discuss about the variety of subjects in life and work.
- Otofun forum (www.otofun.net/forum): A forum for mostly the men to exchange about issues of automobile and related subjects.
- Tinhte forum (www.tinte.vn/forum): A forum for young people to exchange the topics about technological devices and interests.

Users of these forums can indicate the personal information such as name, age, gender, interest, job etc. in their profiles. However, none of them is the explicit field in the user's profile. As a result, we must use both of methods, automatic and manual, to collect and annotate the data.

After the last step, we obtained a collection of 6831 forum posts from 104 users (736.252 words in total), for which we also received at least one of the information about age, gender, location, occupation of the author of each post. The length of each post is also restricted in the range from 250 to 1500 characters to eliminate the too long or too short posts (too long post may contain the text copied from other sources).

Table I. Corpus Statistic

Trait	Total posts	Class	Percent in corpus
Gender	4.474	Male	54%
		Female	46%
Age	3.017	Less than 22	21%
		From 24 to 27	27%
		More than 32	52%
Location	3.960	North	57%
		South	43%
Occupation	3.453	Business, Sale, Admin	36%
		Technique, Technology	31%
		Education, Healthcare	33%

B. Results and discussion

We conducted experiments on 4 traits of authors as mentioned earlier (gender, age, location, occupation) using the Weka[2] toolkit. The results were verified through a ten-fold cross validation process.

Table II shows the results of author profiling experiments of 4 traits.

Table II. The results of author profiling experiments

Trait	Feature	J48	SVM	Bayes-Net
Gender	Stylometric	73.31	82.94	77.17
	Content Words	83.36	89.97	87.58
	NonDict Words	69.75	73.18	69.89
	Style-Content	83.35	90.47	87.35
	Style-NonDict	75.03	86.23	80.11
Age	Stylometric	52.03	62.14	56.17
	Content Words	55.24	61.74	62.55
	NonDict Words	54.56	61.21	59.84
	Style-Content	55.76	63.96	63.92
	Style-NonDict	55.30	64.91	60.89
Location	Stylometric	65.73	70.39	66.99
	Content Words	69.23	79.39	75.01
	NonDict Words	67.28	67.62	68.07
	Style-Content	69.32	80.06	74.54
	Style-NonDict	70.05	75.72	70.31
Occupation	Stylometric	43.97	51.77	46.44
	Content Words	43.32	55.38	51.34
	NonDict Words	38.71	44.70	40.60
	Style-Content	43.41	56.98	50.65
	Style-NonDict	42.81	51.95	45.29

As the results shown in Table II, we can observe that content words outperformed stylometric features and non-dictionary words when used individually. Although content words are often considered

2. http://www.cs.waikato.ac.nz/ml/weka/

domain-specific and may be less accurate when moving the other domains, the results in this task are still promising. Firstly, the data in corpus was collected from various source, therefore it is not so domain-specific. Secondly, even the results are domain-specific to some extent, it is still useful when we conduct the research or apply the results in that domain. However, the results of stylometric features are also good, especially for gender and location. The non-dictionary words have not achieved good results when used solely, but yielded better results when combined with stylometric features (better than stylometric features only).

Regarding the learning methods, the SVM outperformed the other two methods, in which Bayesian Network gave better results than Decision Tree. This is a reasonable result and again proves that SVM is a good algorithm for classifying the author characteristics.

In comparison to the results of previous works, although forum posts are shorter and noisier than other types of online messages such as blog posts or emails, but the results can be considered as promising, especially for gender and location traits. The accuracy of 90.47% when predicting the gender is even better than the results of most of previous works which were conducted on blogs or emails (which had base-line about 80%). The percentage of age prediction (63.96%) is not as good as the results conducted on blog posts or emails (which had the base-line around 77% for blog posts), but much better compared to the result of a research on forum posts conducted by [13], which is only 53%. The same evaluation can be used when saying about the location trait, but the occupation prediction is not so good. The main reason is that occupation information is very noisy and subtle. For example, a person who studied about technical but then works as a sale person is not an easy case when predict his/her job. This needs to be investigated further in later researches.

When comparing with the only previous work on author profiling in Vietnamese by [6], for the gender trait, we achieved the better result (90.47% and 83.3%) when using content words features or style-based/non-dictionary words combination (86.23% and 83.3%), and the same result (82.94% and 83.3%) with stylometric features. It showed that our approach when adding the content and non-dictionary words features has improved the results significantly. The same evaluation can be said when comparing the results of location trait. But for other traits, our results are less accurate, but it is understandable and still promising, because our experiments were conducted on a shorter and more informal type of text than blog posts.

V. CONCLUSION AND FUTURE WORK

In this study, we showed that it is feasible to classify authorial characteristics of the informal online messages as forum posts based on linguistic features, in which using content and non-dictionary words features improved the results significantly. Experiments conducted show the promising results, although some aspects still need to be improved such as the solutions for noisy information in occupation trait or the result for age prediction should be better and so on. This also showed that the SVM algorithm outperformed the other classifiers, while Decision Tree gave the poor results.

In the future, this study can be expanded to other domains, such as social networks or user comments/ product reviews. The data in these domains is even shorter and noisier than forum posts, so it is more challenging task. But the results of such kind of works have promising applications in commercial fields, such as analyzing market trends or user behaviors prediction etc.

We also have planned to investigate more about the use of content and non-dictionary words features in this kind of task. We have conducted experiments and found that these features work very well on the author profiling task for Vietnamese text. However, more insightful analytics should be investigated to show why they are better than stylometric features and which kinds of content are more significant.

REFERENCES

[1] A. Abbasi, H. Chen, Applying authorship analysis to extremist-group Web forum messages, IEEE Intelligent Systems (2005)

[2] A. Abbasi, H. Chen, Writeprints: A Style-based approach to identity-level identification and similarity detection in cyberspace, ACM Transactions on Information Systems, 26 (2), pp: 1-29 (2008)

[3] S. Argamon, M. Koppel, J. Fine, and A. Shimoni, Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August (2003)

[4] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler, Automatically Profiling the Author of an Anonymous Text, Communications of the ACM , in press (2008)

[5] M. Corney, O. DeVel, A. Anderson, and 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)

[6] P. Dang, T. Giang, and P. Son, Author profiling for Vietnamese blogs, International Conference on Asian Language Processing (2009)

[7] O. De Vel, A. Anderson, M. Corney, and G. Mohay, Mining e-mail content for author identification forensics, SIGMOD Record 30(4), pp. 55-64 (2001)

[8] S. Goswami, S. Sarkar, and M. Rustagi, Style-based analysis of bloggers' age and gender, In Proceedings of the Third International ICWSM Conference. The AAAI Press (2009)

[9] G. Gressel, P. Hrudya, K. Surendran, S. Thara, A. Aravind, and P. Prabaharan, Ensemble learning approach for author profiling, Notebook for PAN at CLEF (2014)

[10] F. Iqbal, Messaging Forensic Framework for Cybercrime Investigation. A Thesis in the Department of Computer Science and Software Engineering - Concordia University Montréal, Canada (2010)

[11] M. Koppel, S. Argamon, and A. R. Shimoni, Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), pp : 401-412 (2002)

[12] T. Kucukyilmaz, C. Aykanat, B. B. Cambazoglu, and F. Can, Chat mining: predicting user and message attributes in computer-mediated communication. Information Processing and Management, 44(4), pp - 1448-1466 (2008)

[13] D. Nguyen, Noah A. Smith, and Carolyn P. Rosé, Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '11, pages 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics (2011)

[14] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "How old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (2013)

[15] C. Peersman, W. Daelemans, and 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. ACM (2007)

[16] L. H. Phuong, N. T. Huyen, M. Rossignol, and A. Roussanaly, An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. In Proceedings of Traitement Automatique des Langues Naturelles (TALN-2010), Montreal, Canada (2010)

[17] F. Rangel, and P. Rosso, Use of language and author profiling: Identification of gender and age. In Natural Language Processing and Cognitive Science, p. 177 (2013)

[18] J. Savoy, Authorship attribution based on specific vocabulary. ACM Trans. Inf. Syst. 30, 2 (2012)

[19] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, Effects of Age and Gender on Blogging. In 43 proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (2006)

[20] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Automatic text categorization in terms of genre and author, Computational Linguistics 26(4), pp. 471-495 (2000)

[21] C. Zhang, and P. Zhang, Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA (2010)

[22] R. Zheng, H. Chen, Z. Huang, and Y. Qin, Authorship Analysis in Cybercrime Investigation (Eds.): ISI 2003, LNCS 2665, pp : 59-73 (2003)

[23] R. Zheng, J. Li, H. Chen, and 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)

Duong Tran Duc received the Master degree from University of Leeds, UK in 2004. Currently, he works at tology as a lecturer of Faculty of Information Technology. His research interests are big data, machine learning, and data mining.

Pham Bao Son received the PhD degree from University of New South Wales in 2007. Currently, he is Vice rector of University of Engineering and Technology, Vietnam National University, Hanoi. His research interests are natural languge processing, machine learning, and data mining.

Tan Hanh received the PhD degree from Grenoble Institute of Technology, France. Currently, he is Vice president of Posts and Telecommunications Institute of Technology. His research interests are machine learning, image processing, and data mining.