Author Profiling for Vietnamese Blogs

Dang Duc Pham, Giang Binh Tran, Son Bao Pham
Human Machine Interaction Laboratory
Faculty of Information Technology
College of Technology
Vietnam National University, Hanoi
{dangpd, giangtb, sonpb}@vnu.edu.vn

***Abstract*—This paper presents the first work in the task of author profiling for Vietnamese blogs. This task is important in threat identification and marketing intelligence. We have developed a Vietnamese Blog Profiling framework to automatically predict age, gender, geographic origin and occupation of weblogs' authors purely based on language use. The experiments on the blogs corpus we collected show very promising results with accuracy of around 80% across all traits.**

I. INTRODUCTION

The Internet has created a new way to share information across time and space. Since computer networks enrich human-being life in many aspects, they have also opened a new venue for criminal activities. Especially, these activities spread out quickly on the computer-mediated communication and most of them can be conducted through global electronic networks such as the Internet. One of the predominant activities is the illegal distribution of material in the form of text using popular media such as weblogs, emails, websites, newsgroups or chat rooms. Being able to automatically identify authors of given texts is therefore important in addressing criminal activities in the Internet era.

Automatically identifying authors or analyzing characteristics of authors are also useful for marketing intelligence where specific information about current and potential customers is of high importance. This can help the business to have suitable marketing strategy and develops products to meet the demands of customers.

There have been many tools tackling this task for various languages such as English [3][11], Arabic [1]. In this paper we propose the first work on author profiling for Vietnamese blogs. Specifically, we aim to predict demographic characteristics of a text blog's author namely: gender, age, geographic origin and occupation.

In Section 2, we present related works including hypotheses of relationship between author's profile and language use as well as the studies of author profiling. Section 3 presents our corpus and its collection method. In section 4, we describe our Vietnamese Blog Profiling (VBP) framework and its architecture. Experiments will be described in section 5 while conclusion and future work are presented in section 6.

II. LITERATURE REVIEW

A. Author attribution and author profiling

There are two main tasks of author identification namely the author attribution and author profiling. Authorship attribution is the task of deciding for a given text which author has written it [3]. Authorship attribution has contributed in the fight against cyber crime and in a more general search for reliable identification techniques [1][11][12]. Traditionally, the task of authorship attribution has carried out on data from small sets of authors. This task will be much more difficult when working with a larger set of authors [3]. In such cases, authors' characteristics, or traits, can be a good alternative and open up clues and personal information as to the author's identity.

Author profiling is the task of determining one or more such traits, and an author profile consists of the resulting set of predicted traits [3][11]. Importantly, and contrary to author attribution, the author profiling task is possible even when documents by the author are not in the training data [3]. The more data we have, the higher accuracy of traits determination we get. Most of author profiling work focuses on the prediction of demographic and psychometric traits, e.g. gender, age, native language, neuroticism, agreeableness, extraversion and conscientiousness [7][2][9][3].

Studying with the Weblogs, investigation is carried out on the relationship between language and personality with the five-factor model [8]. In this work, the task of personality profiling is done using both top-down approach and bottom-up approach. In the top-down approach, Nowson analyzed the stylistic factors between authors and linguistic inquiry and word count. In the bottom-up approach, he paid attention on contextually resolvable parts-of-speech. Similarly with studies in personality profiling task, when studying with e-mails, some authors performed a study determining the relationship between personality of a person and language use [4][5].

B. Traits and Language

The link between language use and personal information has been extensively studied. In the view of gender differences, all men behave in a similar manner, and women are equivalently consistent [8]. Additionally, men often use swear words as well as tattoo words. On the other hand, female language use is much more personal and emotional. Moreover, they pay their attentions on more

frequent use of pronouns and references to other people, uncertainty verbs and hedges [8][11].

Age-related changes also affect language use of people. There are four main areas on age-related changes namely emotional experience and expression, identity and social relationship, time orientation and cognitive abilities [10]. The older individuals have a variety of stereotypes with a set of negative characteristics like loneliness and selfishness. Aging comes with a higher level of conscientiousness, agreeableness and adherence to norms. There are some studies showing that the change of age takes the change of language use from parts of speech, function words and so on [3][8][11].

III. CORPUS DEVELOPMENT

The corpus of Vietnamese weblogs is collected from various sources conforming to the following criteria:

- Author of the Weblog pages must be native in Vietnamese language and the main language in their blog pages must be Vietnamese.
- Only the blog pages written in the last 4 years are collected because the period of 4 years affects occupation and age traits.
- Each author must have more than 10 entries.
- The number of words of each entry must be greater than 150 (as ten lines).
- The weblog pages, or blog entries, must be written by the weblog author. Copied entries or multiple authored entries are omitted.

We attract subjects, or weblogs authors, to the experiment by distributing advertisement in forums, newsgroups, instant messages and through direct contact. For subjects who agree to participate in our experiment, we sent them an email or write a blog post directly in their weblog pages explaining which data is collected and why the study is performed. The content of emails contains questions to get the traits of author profiles namely name, gender, age, occupation and geographic origin.

Finally, we chose 73 subjects with 29 males and 44 females from people agreed to participate in the experiment and provided us with their personal information. Our subject selection is to get the balance as much as possible for traits in the. All subjects are native Vietnamese writers with age ranging from 16 to 40. The occupation spreads out from high school student to postgraduate student, model, and singer. The location spreads out from the North to the South of Vietnam, and others locations outside Vietnam. The summary of the corpus is shown in table 1.

TABLE I. CORPUS SUMMARY

Bloggers	Pages Total	Words Total by blogger	Average words by blogger
73	3524	74196	1016

IV. VIETNAMESE BLOG PROFILING (VBP) FRAMEWORK

The VBP Framework has 4 processing components and 3 data containers corresponding with each intermediate processing component. Each processing component is a processing module that permits us working with objects like documents of Vietnamese weblog pages. Figure 1 shows the high-level diagram of VBP Framework's architecture. This architecture ís language independent, which allows us to apply the framework to tasks in different languages using corresponding linguistic processing modules.

A. Preprocessing Component

The task of preprocessing is to standardize input data since weblog pages are created in various formats. For each weblog page or entry, we extract the main text content ignoring none-content blocks such as menus, friend list etc.

B. Linguistic Processing Component

The Linguistic Processing component is a pipelined collection of taggers aimed at linguistically analyzing the preprocessed input documents (i.e. Weblog pages). Results of these taggers are annotations that are a medium for inter-communication between the taggers and will be used for feature calculation at a later stage. These taggers analyze writing styles at different levels namely lexical, syntactic and structural. Furthermore, they detect topic words belonging to specific domains such as computer, science, politics, education etc.

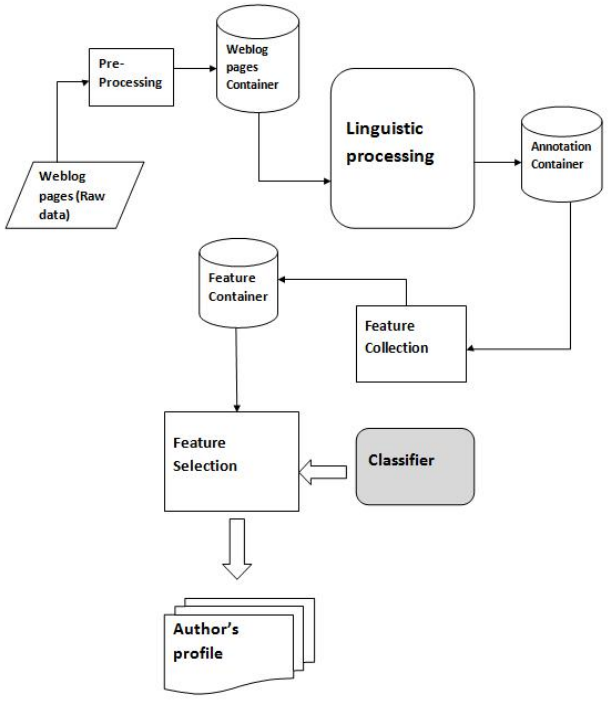

Figure 1. High-level diagram of the VBP Framework

This module performs following analysis:

1. Tokenization: the input document is split into paragraphs, sentence and tokens.

2. Word segmentation: the sentences in the document are segmented into Vietnamese words. This process is important because word boundaries in Vietnamese are not simply spaces. A word can contains multiple tokens, or syllables. We use the word segmentation tool developed by [14].

3. Part-of-speech tagging: there are about 40 classes of part of speech such as conjunctions, prepositions, pronouns, nouns, verb, etc. We use an existing part of speech tagger [15].

4. Topic recognition: words or word phrases are categorized into some topics such as computer, education, emotion, politics, money, etc.

5. Character case expression: following cases of tokens properties are identified as in [3]:

- Upper case: all characters of the Token are in upper case.
- Lower case: all characters of the Token are in lower case.
- CamelCase: words combined together like "WeAreTheWorld"
- First UpperCase: the first character is in upper case; the rest is in lower case.
- SlowShiftRelease: two or more upper case characters, the rest is in lower case.
- SingletonUpperCase: a single character in upper case.

C. Features Collection Component

This component generates a feature vector for every input document as its representation. A feature vector is a set of features and their corresponding values. A feature element of a feature vector, or attribute, is a relationship among annotations. It expresses a property of the input document.

A feature is calculated based on the annotations generated by the linguistic processing component. For example, with some annotations like "alphabetic A" for the *'A'* and "space', 'tab' for *space character* and *tab character* respectively, character based features will be calculated using the Character annotations: Count (alphabetic A), Ratio (space) and (tab), Mean Length (char) in (Line). In general, there are 3 ways to generate a feature using annotations arrived from the previous component:

- Count (X): is number of elements that have annotation X appearing in the document.
- Mean Length (X) in (Y): is the mean length of element with annotation X in the bigger set of element with annotation Y.
- Ratio (X) and (Y): is the ratio between the number of the element X and the number of the element Y.

D. Classifier and Feature Selection

A classifier is used to match an input document with a trait value. In this framework, we use 10 machine learning algorithms from the Weka toolkit [13] namely ZeroR, Decssion Tree J4.8, Random Forest, Bagging, IBk (IB1), Support Vector Machine (SMO), NaiveBayes, BayesNetwork, Neuron Network (Multilayer Perceptron) and RandomTree. For each author trait, one best classifier will be chosen through a cross validation process. The machine learning algorithms are used together with feature selection methods namely Chi Square, Information Gain and Consistency Subset Evaluator in the Weka toolkit [13].

TABLE II. LIST OF CLASSES AND THEIR DESCRIPTION FOR CLASSIFICATION

Trait Name	Class	Description	Percent in corpus
Gender	Male	People have male gender	40 %
	Female	People have female gender	60 %
Age	Age Level 1	People with age <= 22 year olds	45.8 %
	Age Level 2	People with age in 23-26 year olds	28.7 %
	Age Level 3	People with age >= 27 year olds	26.5 %
Location	The North	People who live in The North Vietnam	57.2 %
	The South	People who live in the South Vietnam	32.8 %
	Other	People who don't live neither the North nor the South	10 %
Occupation	Student	People are students	42.4 %
	Singer	People are singers	43.8 %
	Model	People are models	14.8 %

V. EXPERIMENT

We carry out the experiment on the corpus of 3524 Vietnamese Weblog pages described in section III. The corpus filters for balance as much as possible. For each Weblog page, a feature vector is generated by the VBP framework. In total, we have 298 features including document-based, Word-based, Character-based, Function words, Structural, Line-based, Paragraph-based, Lexicon, Content-Specific, POS-based features. Features can be classified into three categories:

- CharFeat: Character based features (70 features)
- WordFeat: Word based features (200 features)
- Other: Other features (28 features)

For example, properties of a Line can be expressed via Characters and Words such as *the number of Characters in Line*, *number of Words in Line*, *Ratio of Upper Characters and Lower Characters in Line*, etc.

We experimented with 4 traits of author profile namely age, gender, location (geographic origin) and occupation. Table 2 summarizes the data distribution for each trait. For traits with numerical values such as age, we divide them into three classes using the first and third quartiles.

For each trait, we find the best classifier among the 10 algorithms using five fold cross-validation on the collected corpus. The results of our experiments for each trait are shown in Table 3, 4, 5, 6.

TABLE III. RESULTS OF RUNNING AUTHORS' PROFILING FOR AGE TRAIT IN ACCURACY (%)

	Feature Sel.	CharFeat+Other	WordFeat + Other	All
Baseline (ZeroR)	InfoGain	45.8002	45.8002	45.8002
J 4.8	InfoGain	49.6595	71.9921	71.4813
Random Forest	None	71.0556	76.5323	76.8445
Random Tree	CfsSubset	68.7287	70.5732	71.1975
IBk (IB1)	**None**	**71.1975**	**77.0999**	**77.2701**
Bagging	InfoGain	67.1112	74.2906	75.2838
BayesNet	None	54.9943	56.2429	55.4200
Naïve Bayes	None	51.6913	49.4892	49.3473
MultilayerPerceptron	ChiSquare	54.5687	57.2325	61.4926
SMO	None	51.7026	58.3144	58.4279

TABLE IV. RESULTS OF RUNNING AUTHORS' PROFILING FOR GENDER TRAIT IN ACCURACY (%)

	Feature Sel.	CharFeat+Other	WordFeat + Other	All
Baseline (ZeroR)	InfoGain	59.9035	59.9035	59.9035
J 4.8	InfoGain		80.1078	80.3916
Random Forest	None	76.4756	83.2577	82.378
Random Tree	CfsSubset	76.1635	77.5539	78.8593
IBk (IB1)	**None**	**76.6459**	**83.0874**	**83.3428**
Bagging	InfoGain	76.5891	81.4983	82.2077
BayesNet	None	59.8751	64.2452	64.1033
Naïve Bayes	None	53.6039	45.4881	45.3462
MultilayerPerceptron	ChiSquare	59.1373	69.7934	74.4608
SMO	None	59.9035	65.5789	65.2951

TABLE V. RESULTS OF RUNNING AUTHORS' PROFILING FOR LOCATION TRAIT IN ACCURACY (%)

	Feature Sel.	CharFeat+Other	WordFeat + Other	All
Baseline (ZeroR)	InfoGain	44.1544	44.1544	44.1544
J 4.8	InfoGain	62.8263	72.5596	71.9353
Random Forest	None	71.4813	77.9512	77.2701
Random Tree	CfsSubset	69.126	72.9285	72.0204
IBk (IB1)	**None**	**70.2611**	**77.6674**	**78.0079**
Bagging	InfoGain	66.941	75.454	76.1635
BayesNet	None	51.1067	57.2361	57.2361
Naïve Bayes	None	32.9739	35.244	35.244
MultilayerPerceptron	ChiSquare	48.5528	59.8653	60.2724
SMO	None	47.1056	59.1941	59.2225

TABLE VI. RESULTS OF RUNNING AUTHORS' PROFILING FOR OCCUPATION TRAIT IN ACCURACY (%)

	Feature Sel.	CharFeat+Other	WordFeat + Other	All
Baseline (ZeroR)	InfoGain	57.2361	57.2361	57.2361
J 4.8	InfoGain	69.5233	76.958	76.8161
Random Forest	**None**	**77.639**	**82.2077**	**82.1226**
Random Tree	CfsSubset	74.0352	75.5675	78.3276
IBk (IB1)	None		82.0942	82.0375
Bagging	InfoGain	73.9501	79.6538	79.9376
BayesNet	None	62.4007	56.9523	57.0658
Naïve Bayes	InfoGain	59.8751	55.7321	58.598
MultilayerPerceptron	ChiSquare	61.521	70.0057	69.0409
SMO	None	57.2361	65.0681	65.1249

TABLE VII. BEST RESULTS OF RUNNING AUTHORS' PROFILING FOR FOUR TRAITS IN ACCURACY (%)

Trait	ML Algorithm	Features	Feature Sel.	Baseline	Result	Improvement
Age:	IBk (IB1)	all	None	45.80	77.27	+21.47 (47,1 %)
Location	IBk (IB1)	all	None	44.15	78.01	+33.86 (76.7%)
Gender:	IBk (IB1)	all	None	59.90	83.34	+23.44 (39.1 %)
Occupation	Rand.Forest	all	None	57.23	82.12	+24.89 (43.5%)

As can be seen from table 7, which summarizes the best classifier for each trait, the classification accuracy for all four traits exceeds 77% and significantly outperforms the baseline by at least 39%. This demonstrates that our approach is effective across all author traits.

The most effective machine learning algorithms are IBk (IB1) and Random Forest. These two algorithms consistently appear in the top two classifiers for all traits. It is surprising to note that support vector machine does not perform well in our experiment. This needs to be investigated further but our conjecture is that the number of features we use is still small for support vector machine to work at its best.

The results on running machine learning algorithms using "CharFeat+Other" and "WordFeat+Other" features reveals that Word-based features gives better results than Character-based features. While character-based features are mostly language independent, word-based features includes Vietnamese word segmentation and parts-of-speech information. This is indicative that Vietnamese specific features are important in getting high performance for the task of author profiling for Vietnamese texts.

It also confirms that age, gender, location and occupation can be predicted with promising results. Moreover, it provides a conclusion that there are certain relationship among language use in blogs and personal information of author.

VI. CONCLUSION

We have presented the first work to tackle the task of author profiling for Vietnamese blogs. We have also developed a Vietnamese Blog Profiling framework to predict author traits using his/her weblogs. Experimental results on our collected corpus of Vietnamese weblogs show promising results with accuracy exceeding 77% across all traits.

This demonstrates that age, gender, location and occupation can be reliably predicted from language use in text. This is significant in the area threat identification on the Internet or marketing intelligence.

In the future we plan to collect more data by inviting more subjects to participate the experiment. Carrying out error analysis to identify what features work best for what traits would give us more insight into how to improve the system. Furthermore, we would also like to apply the framework to predict more author traits including psychometric traits.

The corpus we have collected for this study will be made available for the research community.

Acknowledgement

This work is partly supported by the research fund from College of Technology, Vietnam National University, Hanoi.

References

[1] Abbasi, A., Chen, H. "Applying authorship authorship to extremist group web forum messages". *Homeland security*. IEEE Intelligence System, 2005.

[2] Argamon, S., Koppel, M., Fine, J., and Shimoni, A. "Gender, genre, and writing style in formal written texts". Text, 2003, 23 (3).

[3] Estival D., Gaustad T., Pham S. B., Radford W., and Hutchinson B. "Author Profiling for English Emails". *10th Conference of the Pacific Association for Computational Linguistics (PACLING, 2007),* 2007.

[4] Gill, A., Harrison, A., and Oberlander, J. "Interpersonality: Individual differences and interpersonal priming". *In Proceedings of the 26th Annual Conference of the Cognitive Science Society,* Hillsdale, NJ: Lawrence Erlbaum Associates, 2005, pp. 464–469.

[5] Gill, A.J. "Personality and Language: The projection and perception of personality in computer-mediated communication". *Doctoral Thesis,* University of Edinburgh*,* 2004.

[6] Groom, C.J., and Pennebaker, J.W. "The language of love: sex, sexual orientation, and language use in online personal", 2005.

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

[8] Nowson, S. "The Language of Weblogs: A study of genre and individual differences". *Doctoral thesis,* University of Edinburgh, 2006.

[9] Oberlander, J., and Gill, A. "Individual difference and implicit language: personality, parts-of-speech and pervasiveness". *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: LEA, 2004, (pp. 1035–1040).

[10] Pennebaker, J.W., Mehl, M.R., and Niederhoffer, K.G. "Psychological Aspects of Natural Language Use: Our Words, Our Selves". *Annual Review of Psychology*, 2003, 54, 547-577.

[11] Schler, J., Koppel, M., Argamon, S., and Penebaker, J. "Effects of Age and Gender on Blogging". *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03, 2006.

[12] Zheng, R, & Qin, Y, Huang, Z, and Chen, H. "Authorship ananlysis in Cybercrime Investigation. Intelligence and Security Informatics", *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, IEEE, 2003, 59-73

[13] Witten, I. H., and Frank, E. Data mining: Practical machine learning tools and techniques, Morgan Kaufmann, San Francisco, second edition, 2005.

[14] Pham D. D., Tran B. G and Pham S. B. "A Hybrid Approach to Vietnamese Word Segmentation using Part of Speech tags". *IEEE International Conference on Knowledge System Engineering*, Vietnam, 2009.

[15] Nguyen T. M. H., Vu X. L. and Le. H. P. "Using QTAG POS tagging for Vietnamese documents". *ICT.rda'03*, Vietnam, 2003.