VNU Journal of Science: Comp. Science & Com. Eng., Vol. 33, No. 1 (2017) 37-46

Author Profiling of Vietnamese Forum Posts - An Investigation on Content-based Features

Duong Tran Duc[1,*], Pham Bao Son[2], Tan Hanh[1]

[1]*Posts and Telecommunications Institute of Technology, Hanoi, Vietnam*
[2]*VNU University of Engineering and Technology*

Abstract

In this paper, we investigate the author profiling task for Vietnamese forum posts to predict demographic attributes, such as gender, age, occupation, and location of the author. Although we conducted the experiments on different types of features, including style-based and content-based features, we focused more on analyzing the effects of content-based features. We used machine learning approaches to perform classification tasks on datasets we collected from popular forums in Vietnamese. The results show that these kinds of features work well on such a kind of short and free style messages as forum posts, in which, content-based features achieved much better results than style-based features.

Received 28 June 2016; Revised 10 December 2016 & 08 February 2017; Accepted 18 February 2017

Keywords: Author profiling, machine learning, content-based features.

1. Introduction

The rapid growth of World Wide Web has created a lot of online channels for people to communicate, such as email, blogs, social networks, etc. However, online forum is still one of the most popular channels for people to share the 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 the 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 target marketing and product development). For the social research domain, researchers also want to know the profile 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 [1].

Profiling the author of forum posts is also a challenging task in comparison to doing this on other formal types of text such as article, novel, or even the other types of online texts such as blog posts or emails. Forum posts are often

* Corresponding author. E-mail.: ducdt@ptit.edu.vn
https://doi.org/10.25073/2588-1086/vnucsce.136

short and written in free style, which may contain grammar errors or informal sentence structures.

Although most of previous works in author profiling were conducted on online texts (blog posts, emails), there are a litter works on more informal style of texts such as forum posts. These works also focused on the popular languages such as English, Dutch, Chinese, Greek, etc. [1, 4, 16, 23, 26]. As far as we have known, there is only one work on author profiling conducted in Vietnamese, but on blogs and used style-based features only [6]. In this work, we investigate the use of both style-based and content-based features for author profiling of Vietnamese forum posts, in which we report a deeper analysis on content-based features. This work is also an extension version of our paper on author profiling which presented at ACIIDS'16 [8]. In this paper, we investigated further about the content-based features, such as the best number of content-based features for each trait (which yields the highest result), the list of the most important features for each trait with their weights and provide some analysis about them. In addition, we also improve the prediction results on some traits by applying the Grid Search algorithm to select the best parameters for SVM algorithm.

The organization of the paper is as follows. In section 2, we present the related work on the author analysis problem. Section 3 describes the methods and the system. Section 4 presents the result and discussion. In section 5, we draw a conclusion and future work.

2. 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, such as email, blog posts, forum posts [1, 7, 24].

According to Zheng et al. [26], 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) [14] investigated the Shakespeare's plays. But the work which was considered the most thorough study in this field was conducted by Mosteller and Wallace (1964) [15] when they analyzed the authorship of FederalList Papers. From that point, a number of works have been conducted by various researchers, including [2, 5, 7, 11, 21, 23, 26].

Authorship profiling, also known as authorship characterization, detects the characteristics of an author (e.g. gender, age, educational background, etc.) by analyzing the 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 the 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, 9, 10, 11, 12, 16, 17, 18, 20, 22, 24].

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] and [11].

Regarding the process of authorship analysis, there are two main issues that may significantly affect the performance, namely features set and analytical techniques [26].

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 [12]. 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: Style-based features and content-based features.

Style-based 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 [11]. 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 ([11, 22, 26]). 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, 11]).

Content-based features are often specific words or special content which are used more frequent in that domain than in other domains [25. These words can be chosen by correlating the meaning of words with the domain ([2], [11]) or selecting from corpus by frequency or by other feature selection methods [4].

Also the investigation of Zheng et al. [25] 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, 11, 12, 18, 20, 22, 26]). Some other machine learning algorithms also have been examined and achieved good results, including Bayesian Network, Neural Networks, Decision Tree ([4, 11, 22, 25]). 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.

This paper addresses the problem of author profiling for forum posts, which are in type of online text and written in free-style with short length. For this kind texts, it may be difficult to capture the pure style of authors and using content words as discriminating features could improve the author profiling results.

3. System description

3.1. 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. Fig.1. shows the overall structure of the system.

In the data processing step, data is selected, cleaned and grouped by author profiles. Only posts with length from 50 to 300 words (250 to 1500 characters) were used. We also applied both automatic and manual text processing activities such as eliminating the spam texts, abnormalities, updating training labels, etc. 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, administration /technical, technology/education, 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 [19].

Lastly, the value of each feature is calculated to form a feature vector and saved to training datasets.

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

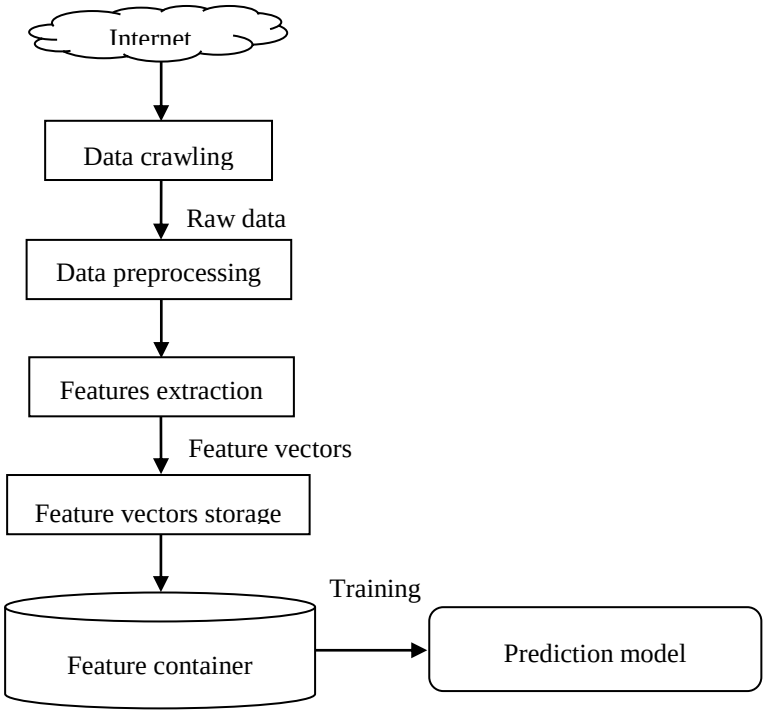

Fig 1. Overall architecture of the system.

3.2. Features

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

Style-based features include character-based, word-based, structural, and syntactic features. In this work, we used common style-based features which were used from the previous work in other languages, such as the number of characters/words, ratio of each type, etc. We

also chose 212 Vietnamese function words, which have little lexical meaning. Part-of-speech tags include 18 word types, such as noun, verb, preposition, etc.

Content-based features used in our work were chosen from the corpus, which are the 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, a feature selection method, was applied to select the best features. For gender trait, we selected 2000 words which were used most frequently by male/female separately. After eliminating the identical words and applied the Information Gain method, we chose the words which have highest significance. Using the similar process, we chose the most significant words to use as content-based features for discriminating the age, occupation, and location traits.

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 Style-based features, content-based features, and all features for analysis of performance of each type.

3.3. Learning methods

In this work, we used Support Vector Machine (SVM) as the learning method to build the classifiers for input messages. Support Vector Machine is a learning method having an advantage that it does not require a reduction in the number of features to avoid the problem of over-fitting. This property is very useful when dealing with large dimensions as encountered in the area of text categorization [5]. SVM has been used in many previous works in author analysis and in most case achieved the better result than other classifiers. Although SVM is a binary classifier, it can handle the multiclass problem (as in case of age and occupation) by building the classifiers which distinguish between every pairs of classes and then using the voting strategy to determine the instance classification.

In addition to SVM, we also used Information Gain method for feature selection and Grid Search for parameter tuning to select the best features and parameters. Information Gain is one of the most popular feature selection methods, which attempts to measure the significance of each feature in distinguishing between classes [24]. This method was tested on various previous works and yielded the good result.

We also experimented on the subsets of features (Style-based, Content-based, All) to investigate the performance of each type.

4. Experiments

4.1. Data

There are a number of Vietnamese forums 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): A forum for girls and ladies to discuss about the variety of subjects in life and work.
- Otofun forum (www.otofun.net): A forum for mostly the men to exchange about issues of automobile and related subjects.
- Tinhte forum (www.tinte.vn): 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. Therefore, we collect only the data which contain information about at least one author trait.

After the last step, we obtained a collection of 6.831 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). The average lenth of posts in words is 107 (the short test post contains 50 words, the longest post contains 300 words).

Table 1. Corpus Statistic

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

4.2. Results and discussion

We conducted experiments on 4 traits of authors as mentioned earlier using the Weka[1] toolkit. The results were verified through a 10-fold cross validation process, in which the training set is randomly partitioned into 10 equal size subsets and 9 subsets were used as training data and the remaining subset is retained for testing. This process is then repeated 10 times with each of 10 subsets is used exactly once as the validation data. Using Grid Search for SVM on PolyKernel with two parameters *c* and *exponent*, together with some modifications in the feature extraction step, the results improved noticeably compared with results in [8], specially on age, location, and occupation traits (e.g. the best parameters for gender trait are *c*=3.0 and *exponent*=1.0). Table 2 shows the results of author profiling experiments of 4 traits.

General evaluation. As the results shown in Table 2, we can observe that content-based features outperformed Style-based features. Although content-based features are often considered 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. Besides, the results of Style-based features are also good, especially for gender and location. Generally, using content-based features increases the accuracy from 7% to 8%, but the improvement is more than 11% for the location trait. Therefore, we may infer that prediction of location is more sensitive on content-based features than other traits. It is reasonable because people from north and south of Vietnam often use different local words in casual communication.

Table 2. The results of author profiling experiments

Feature	Gender	Age	Location	Occup-ation
All Features	90.55	70.70	83.13	61.04
Style-based	83.47	62.76	71.22	52.46
Content-based	90.01	70.05	82.98	60.99

Number of content-based features. As mentioned earlier, to reduce the complexity and improve the accuracy of the model, we applied a feature selection method to eliminate the irrelevant features. We experimented the classification with different number of content words which were chosen by Information Gain

[1] http://www.cs.waikato.ac.nz/ml/weka/

method, ranging from 100 to 1000. Fig 2. shows the best number of features for each trait.

The figure shows that the highest score of gender prediction is achieved when using 600 content words. The best number of words for age and location traits is 400 and the occupation trait is 200. The reason for this is probably the noise in occupation data and therefore, not many words can be used to discriminate between the classes of occupation. Table 3 shows some of the most important content words with their weights for each trait (the bigger absolute value of weight is, the more important the feature is).

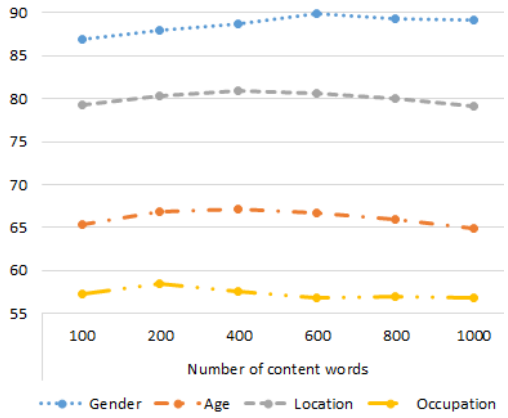

Fig 2. Prediction accuracy for different numbers of content words.

Table 3. The top important content words for each trait

(a) Important words for gender prediction

Male				**Female**			
feature	*weight*	*feature*	*weight*	*feature*	*weight*	*feature*	*weight*
mục tiêu	-1.35	quy định	-1.18	cảm ơn	1.91	hồng	1.46
dữ liệu	-1.34	máy ảnh	-1.09	khách sạn	1.79	bếp	1.43
doanh nghiệp	-1.32	điện tử	-1.07	cưới	1.76	sữa	1.31
kỹ thuật	-1.31	triển khai	-1.03	bác sĩ	1.56	chia sẻ	1.27
xử lý	-1.26	kiểm tra	-1.02	vải	1.51	áp lực	1.18

(b) Important words for age prediction

Younger		Middle		Older	
feature	*weight*	*feature*	*weight*	*feature*	*weight*
học hỏi	-1.50	nhu cầu	-1.29	xài	1.24
lịch sử	-1.32	triệu	-1.20	luật	1.11
nguyên do	-1.25	khắp nơi	-0.90	quy định	0.66
hành động	-1.05	lang thang	-0.74	chi phí	0.62
thể thao	-0.80	bỏ qua	-1.03	hỗ trợ	0.58

(c) Important words for location prediction

North				South			
feature	*weight*	*feature*	*weight*	*feature*	*weight*	*feature*	*weight*
buổi	-1.22	rẽ	-0.78	máy lạnh	1.52	gởi	1.09
đỗ	-1.18	quay	-0.73	coi	1.51	đậu	1.04
mạch	-1.05	sinh	-0.70	gạt	1.48	xài	1.00
liệu	-1.00	ảnh	-0.65	nhơn	1.46	uổng	1.00
nộp	-1.00	chịu khó	-0.53	quẹo	1.35	dơ	0.91

(d) Important words for occupation prediction

Business/Sale/Admin		Technology/Technique		Education/Healthcare	
feature	*weight*	*feature*	*weight*	*feature*	*weight*
lịch	-1.64	phát triển	1.68	tâm lý	1.61
cuộc	-1.62	cấu hình	1.60	hình ảnh	1.58
lang thang	-1.21	kết hợp	1.53	xã hội	1.43
đến nơi	-0.88	kỹ thuật	1.30	học	1.13
cung cấp	-0.77	tài liệu	1.20	từ thiện	1.09

The words in tables suggest that the men tend to discuss about work, technology, regulation etc. while the women often talk about life, health, pressure, and so on. Young people like to discuss about learning, action, etc. The middle age people talk about the needs, travel, and the older people often exchange the views on expenses, law, etc. There many local words that the northern and southern people often used differently from each other, but in our corpus, we found some of them as in the Table 3 (c). Table 3 (d) shows that the people working in business, sale field often used words related to schedule, appointments, travel, while the people working in technology field like to talk about development, machine, etc., and the people which have jobs in education/healthcare fields often discuss about the social, learning, charity issues.

Comparison with previous works. 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 90.55% 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 (70.70%) 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 [16], 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 (90.55% and 83.3%) when using content-based features, and the same result (83.47% and 83.3%) without content-based features. It showed that our approach when adding the content-based 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.

5. Conclusion

In this study, we investigate the author profiling task on a different language (Vietnamese) and different type of text (forum posts) than previous works. The results show that it is feasible to classify authorial characteristics of the informal online messages as forum posts based on linguistic features, in which using content-based features improved the results significantly. We also have a

thorough analysis on content-based features, such as the best number of content words and the list of important words for each trait. 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.

In 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 about the use of more grammar-based features in this kind of task. Vietnamese has many interesting linguistic features such as tones, spells, and we can exploit these features to improve the author profiling results.

Acknowledgements

This work has been supported by Vietnam National University, Hanoi (VNU), under Project No. QG.16.91

References

[1] Abbasi, A., Chen, H. Applying authorship analysis to extremist-group Web forum messages, IEEE Intelligent Systems, 20(5), pp.67-75 (2005).

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

[4] Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. Automatically Profiling the Author of an Anonymous Text, Communications of the ACM , 52(2), pp.119-123 (2008).

[5] Corney, M., DeVel, O., Anderson, A., Mohay, G. 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] Dang, P., Giang, T., Son, P. Author profiling for Vietnamese blogs. International Conference on Asian Language Processing (2009).

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

[8] Duc, D.T., Son, P.B., Hanh, T. Using Content-based Features for Author Profiling of Vietnamese Forum Posts. In: Recent Developments in Intelligent Information and Database Systems, pp. 287–296. Springer International Publishing, Berlin (2016).

[9] Goswami, S., Sarkar, S., and Rustagi.M. Style-based analysis of bloggers' age and gender. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, ICWSM. The AAAI Press (2009).

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

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

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

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

[14] Mendenhall, T.C. The characteristic curves of composition. Science, 11(11), 237–249 (1887).

[15] Mosteller, F., Wallace, D.L. Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley (1964).

[16] Nguyen, D., 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).

[17] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. "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).

[18] Peersman, C., Daelemans, W., and Vaerenbergh. L.V. 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).

[19] Phuong, L., H.,. In Proceedings of Traitement Automatique des Langues Naturelles (TALN-2010), Montreal, Canada (2010).

[20] Rangel, F., Rosso, P. Use of language and author profiling: Identification of gender and age. In Natural Language Processing and Cognitive Science, p. 177 (2013). Huyen, N., T., M., Rossignol, M., Roussanaly, A. An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts.

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

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

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

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

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

[26] Zheng, R., Li, J., Chen, H. and Huang, Z. "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).