Authorship Attribution Analysis of Thai Online Messages

Rangsipan Marukatat[1], Robroo Somkiadcharoen, Ratthanan Nalintasnai, Tappasarn Aramboonpong
Department of Computer Engineering, Faculty of Engineering
Mahidol University
Nakhon Pathom, Thailand
[1] rangsipan.mar@mahidol.ac.th

Abstract—**This paper presents a framework to identify the authors of Thai online messages. The identification is based on 53 writing attributes and the selected algorithms are support vector machine (SVM) and C4.5 decision tree. Experimental results indicate that the overall accuracies achieved by the SVM and the C4.5 were 79% and 75%, respectively. This difference was not statistically significant (at 95% confidence interval). As for the performance of identifying individual authors, in some cases the SVM was clearly better than the C4.5. But there were also other cases where both of them could not distinguish one author from another.**

Keywords— authorship attribution; identification

I. Introduction

With the growing popularity of online media such as blogs, webboards, Facebook, and Twitter, there have been concerns about illegal and unethical activities. ThaiCERT reported that internet fraud accounted for almost 40% of computer crimes reported in Thailand, in year 2013 [1]. The country's ongoing political conflicts also lead to hoax and abusive messages on many webboards, and thus the increase in webboard policing. Unfortunately, people who post such messages often use fake accounts or even impersonate innocent users. These situations highlight the need for authorship attribution.

Authorship attribution is a process of identifying who wrote an offending message. It compares the offending message with sample messages written by suspects. The comparison is based on writing attributes such as the use of capital letters, special symbols, emoticons, slangs, profanities, etc. In some domains, the focus may be on certain profiles of the author (e.g. gender or personality traits) rather than the identity [2][3].

Although there have been many studies on English texts [2][4], and recently on Arabic [3] and Chinese [5] texts, their methods may not be applicable to Thai. Unlike English, Thai text does not have word boundary. First-person pronouns can be gender-specific or gender-neutral (males tend to use the former, females tend to use the latter). Some spoken sentences end with gender suffixes. Because there are as many as 44 consonants, 18 vowel symbols that make up many compound vowels, 4 tone marks, and a few special symbols, typing words can be frustrating especially with mobile devices. Therefore, misspellings, shorten words, and a mix of English spellings of Thai words are common.

This paper presents a framework for authorship attribution analysis of Thai online messages. It evaluates classification techniques and writing attributes selected for this task. The rest of the paper is structured as follows. Section II reviews some related background. Sections III and IV describe our research methodology and experimental results, respectively. Section V concludes the paper.

II. Related Background

A. Authorship Attribution Analysis

Identifying the author of an offending message is treated as a classification problem, where each class is a suspect. Sample messages are collected to train the classifiers. From literature survey, including [2]-[6], a message should be at least 100-200 words long to cover some writing attributes. The sample size should be at least 20-50 messages per author to capture his or her consistency in writing style.

The number of writing attributes required for the analysis ranges from a few tens to hundreds. Cheng et al. [2] found that character-based attributes and function words most contributed to gender classification; many others seemed to be redundant. As they reduced the number of attributes from 545 to 157, the accuracy decreased by merely 3%. Similarly, de Vel et al. [4] reported that character-based attributes were essential for identifying authors while content-related attributes were not, particularly when sample messages contained a variety of topics. Pearl and Steyvers [6] reported that 81 stylometric attributes (mostly character and grammatical measures) could already distinguish authors. Adding content-related attributes did not improve accuracy much because their sample messages had only 400 words on average.

B. Classification Techniques

Among several classification techniques, ones suitable for authorship attribution task should be able to handle a large set of features and support multiple classes. For binary classifiers, the task is transformed into determining whether the author is each suspect.

Support vector machine (SVM) is naturally robust to high dimensional data. In SVM training, its parameters are adjusted

This research is sponsored by the National Research Council of Thailand (NRCT).

978-1-4799-4441-5/14/$31.00 ©2014 IEEE

in order to construct a hyperplane that separates classes with maximum margins. So, the training itself is an optimization problem. It can be effectively solved by sequential minimal optimization (SMO) method [7]. SVM uses a kernel to define a feature space in which the data will be classified, allowing transformation from a non-linear separable space into a linear separable one. As a result, it tends to perform well even with hard-to-classify data and is a popular technique in authorship attribution studies [2]-[5]. But despite high accuracy, it lacks expressive decision model, which is possibly requested when implicating a person of a crime.

Decision tree classifier grows a tree by picking attributes, one by one, to be its nodes. The selection is based on how well each of them separates the data; irrelevant ones are naturally left out. One of the most popular algorithms to build a tree is C4.5 [8]. Because the attributes are considered individually – rather than together, as in SVM – their interactions, if existing, are not modeled in the tree. Although its performance may not be comparable to that of SVM in some cases, the decision tree presentation is much more expressive.

III. RESEARCH METHODOLOGY

Shown in Fig. 1, our framework is quite straightforward. It consists of 2 main parts: attribute extraction and authorship attribution analysis. The attribute extraction module uses Thai Lexeme Tokenizer (LexTo) [9] to segment a message into words. The attribute values are kept in attribute-relation file format (ARFF), which can be further processed by Weka [10].

The total of 53 writing attributes are extracted. They are categorized into 6 groups. Numeric attributes are normalized by either the number of characters (C1) or the number of words (W1) in each message.

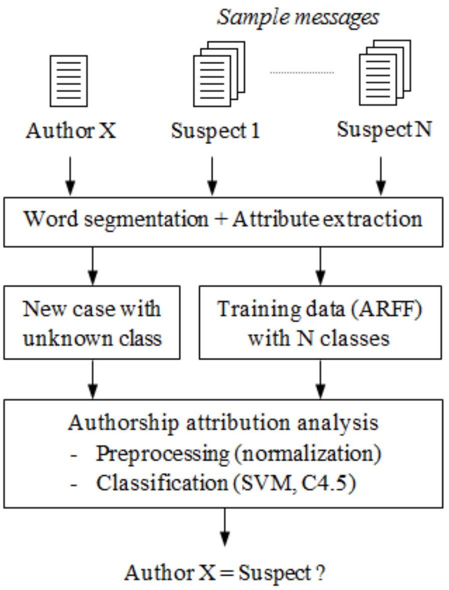
Sample messages
Author X
Suspect 1
Suspect N
Word segmentation + Attribute extraction
New case with unknown class
Training data (ARFF) with N classes
Authorship attribution analysis
- Preprocessing (normalization)
- Classification (SVM, C4.5)
Author X = Suspect ?

Fig. 1. An overview of authorship attribution framework

1. Attributes for character analysis (C1-C9). For example, the number of Thai letters, upper-case and lower-case English letters, digits, and special symbols.
2. Attributes for word analysis (W1-W10). For example, average length of words; the number of unique words, words longer than 6 letters, words shorter than 3 letters; and word variation measures adapted from [2] (e.g. Yule's characteristic K, Simpson's diversity index, and Honore's vocabulary richness). These word variation measures are not content-specific.
3. Attributes for punctuation analysis (P1-P10). For example, the number of periods, commas, colons, semi-colons, question marks, exclamation marks, etc.
4. Attributes for emotion analysis (E1-E10). For example, the number of consecutive 5s (555 is read "ha ha ha"), tilted emoticons with and without noses such as :-) and :), and vertical emoticons with and without cheeks such as (^_^) and ^_^.
5. Attributes for structure analysis (S1-S5). For example, average length of sentences and paragraphs; the number of sentences, lines, empty lines, etc.
6. Attributes for content analysis (T1-T9). For example, the number of shorten words, profanities, exclamations; male, female, and neutral first-person pronouns; and male and female sentence ending words.

Given advantages and limitations of each technique, SVM and C4.5 decision tree are selected for authorship attribution analysis.

Messages posted on pantip.com webboard and 2 fanpages were collected, as summarized in Table I. The ones from the webboard were about actors and politics, whereas those from the fanpages were mostly about politics in satirical tone. The webboard imposed some censorship. The fanpages did not.

TABLE I. SUMMARY OF EXPERIMENTAL DATA

Author	No. of Messages	Message Length		Topics
		Characters	***Words***	
1	25	426-701 mean = 535	123-225 mean = 165	Actors (pantip.com)
2	25	401-882 mean = 575	98-189 mean = 131	Actors (pantip.com)
3	25	360-841 mean = 588	92-196 mean = 142	Actors (pantip.com)
4	25	380-697 mean = 562	90-174 mean = 146	Politics (pantip.com)
5	25	387-781 mean = 587	107-208 mean = 151	Politics (fanpage 1)
6	25	415-726 mean = 532	105-188 mean = 133	Politics (fanpage 2)
Overall		mean = 563	mean = 143	

Table II summarizes main parameters for Weka's SVM and C4.5 decision tree. The SVM was based on SMO training and polynomial kernel of degree 2. Due to small training data set, there were a few considerations for C4.5 parameters. To handle likely small tree nodes, only binary split was applied and frequencies in leaf nodes were Laplace smoothed. The pruning was based on subtree raising without reduced error pruning, that is, all training data were used only for growing tree.

The experiment was repeated 10 times. In each time, 20 messages of each author were randomed to be training data (120 training messages in total), leaving the rest to be testing data (30 testing messages in total).

TABLE II. PARAMETER SETUP

Method	Parameters

SVM
(Weka's SMO)
	Epsilon = 1.0E-12 (default)
	Filter type = normalize training data
	Kernel = polynomial kernel, degree 2
	Tolerance parameter = 0.001 (default)

C4.5 Decision tree (Weka's J48)	Binary split = true Confidence factor = 0.25 (default) Use Laplace = true Pruning strategy = subtree raising (unprune = false, subtree raising = true, reduced error pruning = false)

IV. RESULTS AND DISCUSSION

Tables III reports the overall accuracies achieved by SVM and C4.5 decision tree. Although the SVM performed better than the C4.5, the difference between their average accuracies was not statistically significant (at 95% confidence interval, or $\alpha = 0.05$).

Furthermore, in each experimental run with each classifier, the true positive rate of each author was computed. The true positive rate, or recall, of author X is the proportion of testing messages written by X that were correctly identified. The true positive rates achieved by both classifiers were compared by using paired samples t-test. Fig. 2 shows the average measures of each author. The SVM was significantly better ($\alpha = 0.05$) at identifying Author 4, and was about as good as the C4.5 at identifying the other authors.

Likewise, in each experimental run, the F-measure of each author was computed. It is a harmonic mean of precision and recall. The precision of author X is the proportion of testing messages identified as X's that were actually written by X. The F-measures achieved by both classifiers were compared by using paired samples t-test. From Fig. 3, the SVM yielded significantly higher F-measures ($\alpha = 0.05$) than the C4.5 when identifying Authors 2 and 4.

Relating the results with writing styles seen in the sample messages, we note the following observations. Firstly, both classifiers could identify authors with clearly different styles, such as Authors 1 and 5. Secondly, the SVM could identify Authors 2 and 4, despite their less characterizable styles (e.g. neutral words and few special symbols). The C4.5 gave many wrong predictions in this case. In addition, both classifiers struggled to distinguish between Authors 3 and 6 even though their styles were very different in human eyes. The messages written by Author 3 were polite and fair, whereas those written by Author 6 were polite but sarcastic. The framework could not detect sarcasm or other feelings conveyed in plain messages, unless explicit emotion-related symbols were present.

TABLE III. OVERALL ACCURACY (OVER 10 RUNS)

Measures	SVM	C4.5
Minimum accuracy	70%	60%
Maximum accuracy	90%	86.67%
Mean accuracy Std. error of mean	79% 2.00	75% 2.29
Std. deviation	6.30	7.24
Paired samples t-test	t = 1.616, df = 9 Sig. (2-tailed) = 0.140	

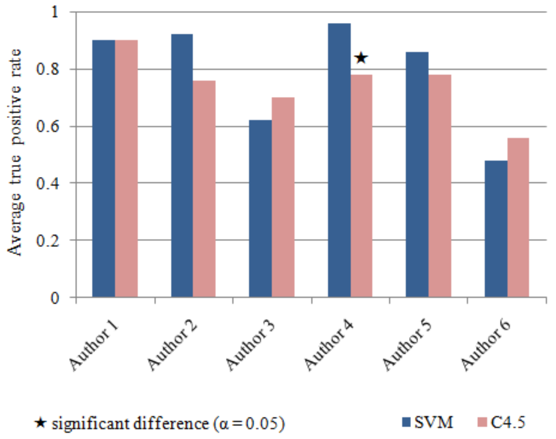

Fig. 2. Average true positive rates (over 10 runs) of each author

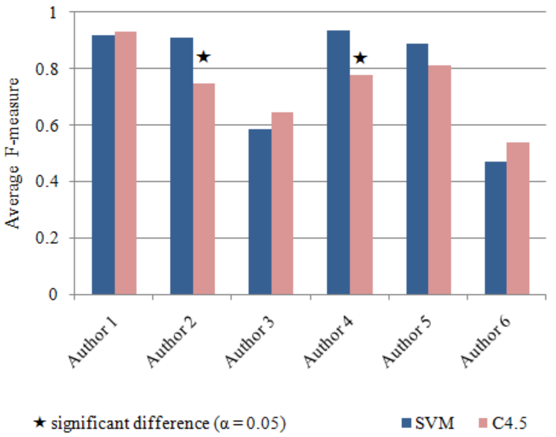

Fig. 3. Average F-measures (over 10 runs) of each author

V. CONCLUSION

A framework for authorship attribution analysis of Thai online messages has been presented. It uses support vector machine (SVM) and C4.5 decision tree to identify the authors, based on 53 writing attributes. From the experiment, the SVM yielded better results than the C4.5, especially when authors' writing styles were not clearly different. However, the overall accuracy of the C4.5 was still not much worse. Currently, this research uses fewer writing attributes than others mentioned in Section II. Adding more attributes may allow finer analysis of the writing style and thus improve the classifier performance.

Ensembled classifiers and other classifiers that produce expressive decision models will be examined in details. Instead of letting the classifier pick just one author from a group of suspects, the framework may report the matching scores (or probabilities) of all suspects. This will be more flexible in real practice, as the authorities can use their own judgement to further investigate a few probable authors.

REFERENCES

[1] Thailand Computer Emergency Response Team (ThaiCERT), Statistics for the year 2013. http://www.thaicert.or.th/statistics/statistics-en.html.

[2] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," Digital Investigation, 8, pp. 78-88, 2011.

[3] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson, "TAT: an author profiling tool with application to Arabic emails," Proceedings of the Australasian Language Technology Workshop, Melbourne, Australia, 2007, pp. 21-30.

[4] O. de Vel, A. Anderson, M. Corney, and G. Mohey, "Mining e-mail content for author identification forensics," SIGMOD Record, 30(4): 55-64, 2001.

[5] J. Ma, Y. Li, G. Teng, F. Wang, and Y. Zhao, "Sequential pattern mining for Chinese e-mail authorship identification," Proceedings of the 3rd International Conference on Innovative Computing Information and Control (ICICIC 08), Washington, DC, USA, 2008, pp. 73-76.

[6] L. Pearl and M. Steyvers, "Detecting authorship deception: a supervised machine learning approach using author writeprints," Literary and Linguistic Computing, 27(2): 183-196, 2012.

[7] J. Platt, "Fast training of support vector machine using sequential minimal optimization," In B. Schoelkopf, C. Burges, and A. Smola (eds), Advances in Kernel Methods – Support Vector Learning, 1998, pp. 185-208.

[8] R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers, 1993.

[9] National Electronic and Computer Technology Center (NECTEC) of Thailand, LexTo: Thai Lexeme Tokenizer [opensource]. http://www.sansarn.com/lexto/.

[10] University of Waikato, Weka 3.7.10 [opensource]. http://www.cs.waikato.ac.nz/ml/weka/downloading.html.