ISSN 2047-3338

Comparison on Some Machine Learning Methods for Lao Text Categorization

Souksan Vilavong[1] and Khanh Phan Huy[2]
[1]Champasak University, Ban Chat Sanh, Pakse, Laos
[2]Danang University of Technology, The University of Danang, Danang, Vietnam
[1]ssuchedu@yahoo.com, [2]khanhph29@gmail.com

***Abstract*—Text categorization is one of the most important role in many applications in natural language processing (NLP). The task of text classification is assignment of free text document to one or more predefined categories based on their content. Whereas a wide range of methods have been applied to English text classification, relatively very few studies have been done on Lao text. In this paper, we present methodology for Lao document presentation and two of the best machine learning techniques, which have namely Radial Basis Function (RBF) network and support vector machines (SVM), to classify the documents. Experimental results revealed that these approaches could achieve an average about 82% accuracy. Additionally, we also analyze the advantages and disadvantages of each approach to find out the best method in specific circumstances.**

***Index Terms*— Machine Learning, Comparison, Natural Language Processing (NLP) and Lao Text Categorization**

I. INTRODUCTION

TEXT categorization has been one of the most popular problem in natural language processing. It is based module in many applications and most of categorization systems classify documents into one or more given predefined categories. It has been utilized in many application areas such as, spam filtering [25]; language identification [18]; genre classification [27]; customer relationship management [7], web page classification [23], text sentinel classification [31] and astronomy [15]. Depending on the user's requirement, each document can be categorized into multiple, exactly one, or no category at all [2].

Text categorization can be considered and resolve by many methods and machine learning is one of the most performance method. Text categorization based on machine learning techniques is a general inductive process automatically builds

a classifier by learning, from a set of predefined categories of documents, the characteristics of the categories. For instance, Naive Bayes classifiers [32], N-gram [16], k-Nearest Neighbor (k-NN) [13] and Support Vector Machine (SVM) [14],... Many research publications that studies evaluated the performance of these methods based on accuracy of document classification, and most of the studies focused on automatic text categorization for documents, which were written in English. Previous works on Lao text categorization is very limited. This may be due to the nature of Lao language and also to the lack of Lao language resources such as labeled corpus.

The automatic text categorization process foresees a set of tasks universally recognized by the research community [1]. These tasks include features extract, feature selection and feature weighting processes are performed. Moreover, these tasks also include training task in a machine learning classifier is trained using a set of labeled documents. Finally, the last task is the Testing task the accuracy of the classifier is evaluated by using a set of pre-labeled documents (i.e. test-set) that are not used in the training phase. In this paper, we have used two supervised classification models for Lao text categorization. It presents an empirical comparison of these supervised machine learning classifiers (SVN, RBF). In order to evaluate those classification models, we collected Lao predefined categorized from online Lao newspapers archives, namely; Lao News Agency, and Vientianetimes News. This corpus was collected from five categories: Politics, Economics, Crime, Education, Tourism and Sport[1].

The rest of this paper is organized as follows: Section II presents an overview about some related works in the area of Lao text categorization. Some basic concepts for text categorization and our approaches for text categorization are described in Section III. Next in Section IV, we present the experiment setup and discuss on the experimental results. Finally, we describe about conclusion and indicate future works in Section V.

II. RELATED WORKS

Text categorization stands at the cross junction to modern information retrieval and machine learning. In last decade, there are many previous works to categorize english documents automatically [8]. Applications of machine learning techniques help to reduce the manual effort required in analysis and the accuracy of the systems also improved through the use of these techniques. For instance, N-gram [4,

[1] http://kpl.gov.la.

11]; k-NN [4, 30]; Decision Tree [3] and SVM [14].

Lao is similar to other South East Asian languages, such as: Chinese, Thai, Vietnamese, etc. A text is a string of symbols with no explicit word boundary. Spaces between syllables are rarely used for separating is main difficulty in many tasks of natural language processing. May text categorization publications on these language has played important role for our purpose on Lao language such as Association rule [5], N-Gram [28], Naive Bayes [17] for Chinese; SVM [19]; Bag Of Words(BOW) [12]; N-Gram [12], L-KNN [20],... Especially in terms of spoken and writing system, Lao has closest relationships with Thai language, so many researches about text categorization on Thai have an direct influence on Lao language. For instance, SVM [6] Naive Bayesian, Decision Tree, k-Nearest Neighbor and RBF network [21]. These results have several implications for Lao text classification problem. Many researchers attempted to obtain better classification algorithms performance for automatic text categorization. SVM and RBF network are considered as the common methods to text categorization [24, 29]. Therefore, they have been treated as the base method for categorizing text. Thus, in this paper, we include SVM and RBF network, and our feature selection, in our experiment to find out the most suitable method in categorizing Lao text.

III. OUR APPROACHES

In our approach, we proposed model for Lao text categorization as Figure 1. This model can be summarized in two sub-components that are pre-processing and classification component. Pre-processing component will transform internet documents to text documents (*plain text*), after reading the input text document by the proposed system which divides that text document into features which are also called (*tokens, words, terms or attributes*); indexing corpus and weights computation on these texts for building model. It represents that text document in a vector space as a vector whose components are that features and their weights which are computed by the frequency of each feature in that text document.

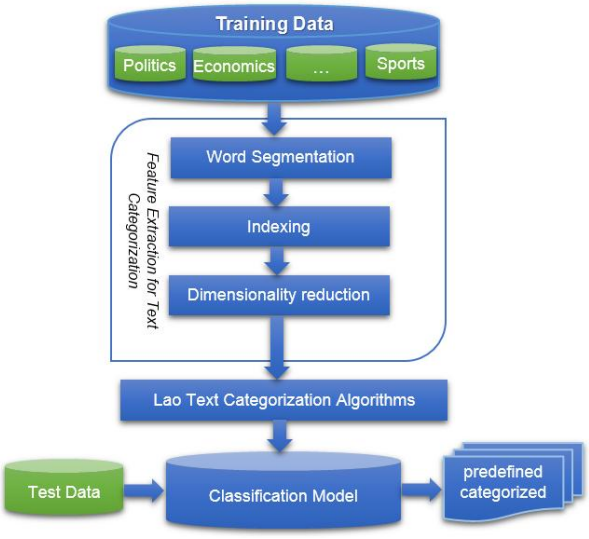

Figure 1. The process of Lao text categorization system

A. Word segmentation

We used Htttrack[2] tool for get html code of all news from website the they is divided into two sets: Training set and Test set. The first step in text categorisation is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. Because the corpus were gotten from website so that the text transformation usually involves of the following processes: removing HTML tags, removing stopwords as Figure 2, and performing word stemming. The stopwords are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions etc.). After text transformation, we use word segmentation method in research of Srithirath et al [26]. The results of this step will be used for features statistics processing.

#ID	Word/ Words	#ID	Word/ Words	#ID	Word/ Words	#ID	Word/ Words
1	ສ່ວນ	4	ຂອງ	7	ໃນເວລາທີ	10	ໃນຖານະເປັນ
2	ໃນ	5	ແລະ	8	ເນື່ອງຈາກວ່າ	11	ຕື່ມ
3	ໂດຍ	6	ແຕ່	9	ນັບຕັ້ງແຕ່	12	ເພາະສະນັ້ນ,

Figure 2. The list some stopwords in Lao

B. Indexing

Vector space model is one of the most commonly document representation methods for text categorization. In the vector space model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by a word by document matrix ***M***, where each entry represents the occurrences of a word in a document.

$$M = (m_{ik}) \qquad (1)$$

where m_{ik} is weight of word *i* in document *k*. The number of rows, N of the matrix M corresponds to the number of words in dictionary and can be very large. So that, the high dimensionality of the feature space is a major characteristic, or difficulty of text categorization problems. We will present methods to reduce dimensionality of the matrix in next section. There are several ways of determining the weight m_{ik} of word *i* in document *k*. In this work, we use one of the most effective weighting scheme, which is Entropy weighting [9]. In this scheme, m_{ik} is given by:

$$m_{ik} = \log(f_{ik} + 1.0)\left(1 + \frac{1}{\log(ND)}\sum_{j=1}^{ND}\left[\frac{f_{ij}}{n_i}\log\left(\frac{f_{ij}}{n_i}\right)\right]\right) \quad (2)$$

where:

- f_{ik} be the frequency of word *i* in document *k*;
- ND be the number of documents in the collection;

[2] https://www.httrack.com

- n_i the total number of times word *i* occurs in the whole collection.

$$\frac{1}{\log(ND)}\sum_{j=1}^{ND}\left[\frac{f_{ij}}{n_i}\log\left(\frac{f_{ij}}{n_i}\right)\right] \quad (3)$$

is the average uncertainty or entropy of word *i*. This quantity is -1 if the word is equally distributed over all documents and 0 if the word occurs in only one document.

C. Dimensionality reduction

In text categorization field, features are extracted from documents are often expressed as vector pattern (keyword, weight). As more number of documents, more number of words are extracted and the feature space could contain more than several hundreds to thousands words. One of the most important module in dimensionality reduction is feature selection method. The weight can be computed by different methods, such as information gain (IG) and gain ratio [10]. Information Gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a word in at document. Let $c_1, c_2, ..., c_K$ denote the set of possible categories. The information gain of a word *w* is defined as follow:

$$\begin{aligned} IG(w) = &-\sum_{j=1}^{K}\log P(c_j) + P(w)\sum_{j=1}^{k}P(c_j|w)\log P(c_j|w) \\ &+ P(\bar{w})\sum_{j=1}^{k}P(c_j|\bar{w})\log P(c_j|\bar{w}) \end{aligned} \quad (4)$$

where $P(c_j)$ can be estimated from the fraction of documents in the total collection that belongs to category c_j and $P(w)$ from the fraction of documents in which the word *w* occurs. Moreover $P(c_j|w)$ can be computed as the fraction of documents from category c_j that have at least one occurrence of word *w* and $P(c_j|\bar{w})$ as the fraction of documents from category c_j that does not contain word *w*. The information gain is computed for each word of the training set and the words whose information gain is less than some predetermined threshold are removed. Gain ratio is an extension of information gain which selects words that have maximized the ratio of its gain divided by its entropy [9]. The gain ratio of word *w* is defined as:

$$G(w) = \frac{H(cate) - H(cate|w)}{H(w)} \quad (5)$$

where *H* is the entropy.

D. Classification methods

As we mentioned before, we choose two of the best efficient classifier methods for some languages which has very close relationship with Lao and used them for Lao text classification; SVM and RBF neural network methods due to their simplicity, effectiveness and accurateness. Brief descriptions of these methods are given, as follows:

1) Support Vector Machine Learning

SVM is a robust machine learning methodology which shows high performance on text classification [14]. Depending on the purpose of classification, the SVM can be constructed as a linear or nonlinear model. Let us consider a for instance, given that the training dataset *X* contains *n* labeled sample vectors $(x_1, y_1), ..., (x_n, y_n)$, where each x_i is a feature vector of the document *i* and each y_i is the class label of the document *i*. The linear SVM uses a weight vector *w* and a bias term *b* to classify *a* new example *x*, by creating a predicted class label *f(x)* as given in below:

$$f(x) = sign(\langle w, x\rangle + b) \quad (6)$$

For the non-separable case, the training errors are allowed so that the linear SVM finds the vector *w* by minimizing the objective function over all *n* training samples as shown:

$$T(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{n}^{i=1}\xi_i \quad (7)$$

under the constraints that

$$\forall i = \{1..n\} : y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \xi_i \geq 0$$

However, for the problem of text classification using SVM method, the selection of properties for each classification are extremely important issues, it decided to classify the efficiency of algorithm. In this work, we employed the Platt's SMO algorithm [22] with default parameter for building the support vector machine classification model.

2) Radial Basis Function Network

RBF network has been widely used for pattern classification, function approximation and text classification. The RBF network is a three-layer feed-forward neural network, between the input and the output layers there is a "hidden layer". In the training phase, vectors are input to the first layer and fanned out to the hidden layer. In the latter, a cluster of RBF functions turn the input to output, adjusting the weight of the input to the hidden layer. Then, under the target vector's supervising, the weigh of the output vector of the hidden layer is adjusted. When clustering texts, the Euclidean Distance between the input vectors and the weight vectors, which have been adjusted by training process, is calculated. Each input sample is sorted to a class. Then the output layer collects samples belonging to same classes and organizes an output vector, the final clustering.

In the hidden nodes, the activation function is usually chosen as Gaussian Function, the input of node i is the product of threshold b_i and the Euclidean Distance between weight vector W and input vector X:

$$k_i^q = \sqrt{\sum_j \left(x_j^q - w_{ji}\right)^2} * b_i \qquad (8)$$

Where, x_i^q is the j^{th} component of the q^{th} input vector, w_{ji} is the weight between the j^{th} node in the input layer and the i^{th} node in the hidden layer, b_i is a threshold to control the accuracy of the Gaussian Function. The output of the same node is as follows:

$$r_i^q = exp\left(-\left(k_i^q\right)^2\right) = exp\left(\sqrt{\sum_j \left(x_j^q - w_{ji}\right)^2} * b_i\right) \qquad (9)$$

Instead of adjusting b_i, we can use the parameter of spread in Neural Network Toolbox of Matlab 7.0 to control the performance of the network. The larger spread is, the smoother the function approximation will be. The input of the output layer is weight sum of the output of the nodes in hidden layer. The activation function is linear, so the output of the whole network, in response to the qth component of the input, is shown as:

$$y^q = \sum_n^{i=1} r_i * V_i \qquad (10)$$

where v_i is i^{th} component of weight vector from the hidden layer to the output layer.

As is discussed above, RBF has a strong capability of approximation to the kernel vector in a limited part of the whole net. The training of the RBF network should be divided into two processes. The first is unsupervised learning, which adjusts the weight vector between the input and hidden layer. The other is supervised learning, which adjusts the weight vector between the hidden and output layer. Three parameters should be given before training: input vector, target vector and the threshold value, in Matlab, the spread.

IV. Experiments and Results

A. Experimental setup

In order to evaluate the used classification algorithms, several experiments have been conducted. We have measured the performance of these classification algorithms on manually classified Lao corpus collected from online Lao newspapers archives from Vientianetimes News, and Lao News Agency. This corpus contains 4000 documents that are different in length and divided into eight categories: Politics, Economics, Crime, Education, Tourism, and Sport. The training corpus contains 90% (3600) documents in six different categories, and the rest 10% (400) documents are used as testing samples, with average about 65 documents in each category. To measure the performance of these classification methods, we use the results of calculating Precision and Recall:

1. *Accuracy (A)*: Is the ratio between the number of text documents which were correctly categorized and the total number of documents.

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \qquad (11)$$

where TP_i (true positives) is the number of text documents correctly classified in category c_i , TN_i (true negatives) is the number of text documents correctly classified as not belonging to category c_i, FP_i (false positives) is the number of text documents incorrectly classified in category c_i, and FN_i (false negatives) is the number of text documents incorrectly classified as not belonging to category c_i.

2. *Error rate (E)*: Is the ratio between the number of text documents which were not correctly categorized and the total number of text documents.

$$E_i = 1 - Ac_i = \frac{FP_i + FN_i}{TP_i + TN_i + FP_i + FN_i} \qquad (11)$$

3. ***Precision (P)***: Is the percentage of correctly categorized text documents among all text documents that were assigned to the category by the classifier.

$$P_i = \frac{TP_i}{TP_i + FP_i} \qquad (12)$$

4. ***Recall (R)***: Is the percentage of correctly categorized text documents among all text documents belonging to that category.

$$R_i = \frac{TP_i}{TP_i + FN_i} \qquad (13)$$

B. Experimental results

Some our experimental results are listed in the following:

Categorized	Based	SVM Classification							
		TP	TN	FP	FN	Accuracy	Errorrate	Precision	Recall
Politics	63	51	225	12	0	95.83%	4.17%	80.95%	100.00%
Economics	68	59	217	9	12	92.93%	7.07%	86.76%	83.10%
Crime	68	54	222	14	11	91.69%	8.31%	79.41%	83.08%
Education	61	48	228	13	17	90.20%	9.80%	78.69%	73.85%
Tourism	68	53	223	15	18	89.32%	10.68%	77.94%	74.65%
Sport	72	62	214	10	15	91.69%	8.31%	86.11%	80.52%

Figure 3. Experimental Result of the SVM algorithm

Categorized	Based	RBF Classification							
		TP	TN	FP	FN	Accuracy	Errorrate	Precision	Recall
Politics	63	55	275	8	4	96.49%	3.51%	87.30%	93.22%
Economics	68	54	276	14	10	93.22%	6.78%	79.41%	84.38%
Crime	68	52	278	16	22	89.67%	10.33%	76.47%	70.27%
Education	61	51	279	10	17	92.44%	7.56%	83.61%	75.00%
Tourism	68	55	275	13	9	93.75%	6.25%	80.88%	85.94%
Sport	72	63	267	9	8	95.10%	4.90%	87.50%	88.73%

Figure 4. Experimental Result of the RBF algorithm

Categorized	Based	SMV Classification		RBF Classification	
		Correct	Percent	Correct	Percent
Politics	63	51	80.95%	55	87.30%
Economics	68	59	86.76%	54	79.41%
Crime	68	54	79.41%	52	76.47%
Education	61	48	78.69%	51	83.61%
Tourism	68	53	77.94%	55	80.88%
Sport	72	62	86.11%	63	87.50%
			81.64%		**82.53%**

Figure 5. Experimental Result for comparison of the RBF network and SVM

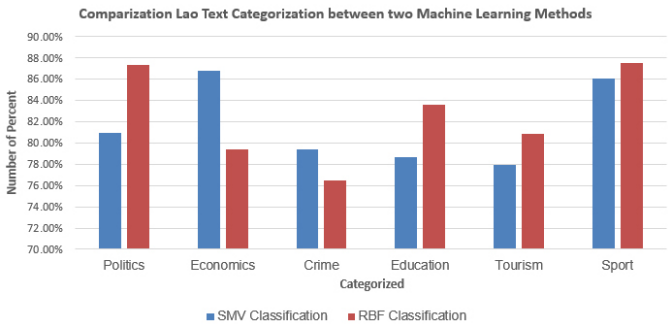

Figure 6. Diagram for comparison of the RBF network and SVM

V. Conclusion and Future Work

This paper presents some research on Lao text categorization using machine learning techniques. SMV algorithm is more simple than RBF but not easy to find set of parameters for many language. In general, the RBF algorithm is selected for constructing the classification model since it gives better results than other. However, with both algorithm, there are many documents were misclassified into other groups. We investigated this problem and observed that the training dataset contains a small number research's corpus and the results of word segmentation has clearly affect for actual text categorization. Future work will investigate the impact of skewed class distribution of the training dataset to the accuracy of the classification model. Moreover, we will improve the accuracy of the classification model by considering improving word segmentation task and other features such as part of speech of words, semantic of phrase, etc.

References

[1] KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2007. ACM. 618070.

[2] Hamood Alshalabi, Sabrina Tiun, Nazlia Omar, and Mohammed Albared. Experiments on the use of feature selection and machine learning methods in automatic malay text categorization. Procedia Technology, 11(0):748 _ 754, 2013. 4th International Conference on Electrical Engineering and Informatics, {ICEEI} 2013.

[3] Chidanand Apté, Fred Damerau, and Sholom M. Weiss. Automated learning of decision rules for text categorization. ACM Trans. Inf. Syst.,12(3):233_251, July 1994.

[4] William B. Cavnar and John M. Trenkle. N-grambased text categorization. In In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161-175, 1994.

[5] Ding-An Chiang, Huan-Chao Keh, Hui-Hua Huang, and Derming Chyr. The chinese text categorization system with association rule and category priority. Expert Systems with Applications, 35(1_2):102 -110, 2008.

[6] N. Chirawichitchai, P. Sa-nguansat, and P. Meesad. Developing an e_ective thai document categorization framework base on term relevance frequency weighting. In Knowledge Engineering, 2010 8th International Conference on ICT and, pages 19-23, Nov 2010.

[7] Kristof Coussement and Dirk Van den Poel. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Information Management, 45(3):164-174, 2008.

[8] Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski, and Michael W. Mahoney. Feature selection methods for text classi_cation. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, pages 230-239, New York, NY, USA, 2007. ACM.

[9] Susan T. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, 23(2):229_236, June 1991.

[10] Irene Diaz Jose Ranilla Elias F. Combarro, Elena Montanes and Ricardo Mones. Introducing a family of linear measures for feature selection in text categorization. IEEE Transactions on Knowledge and Data Engineering, pages 1223_1232.

[11] Johannes Fürnkranz. A study using n-gram features for text categorization, 1998.

[12] Vu Cong Duy Hoang, Dien Dinh, Nguyen Le Nguyen, and Hung Quoc Ngo. A comparative study on vietnamese text classi_cation methods. In Research, Innovation and Vision for the Future, 2007 IEEE International Conference on, pages 267_273, March 2007.

[13] Shengyi Jiang, Guansong Pang, Meiling Wu, and Limin Kuang. An improved k-nearest-neighbor algorithm for text categorization. Expert Syst. Appl., 39(1):1503_1509, January 2012.

[14] Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, ECML '98, pages 137_142, London, UK, UK, 1998. Springer-Verlag.

[15] Huaizhong Kou, Amedeo Napoli, and Yannick Toussaint. Application of text categorization to astronomyeld. In Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005, Proceedings, pages 32_43, 2005.

[16] Jingyang Li and Maosong Sun. Non-independent term selection for chinese text categorization. Tsinghua Science Technology, 14(1):113 _ 120, 2009.

[17] Lei LI, Yu guang HUANG, and Zhong wan LIU. Chinese text classi_cation for small sample set. The Journal of China

Universities of Posts and Telecommunications, 18, Supplement 1(0):83 _ 89, 2011.

[18] Kavi Narayana Murthy and Guntur Bharadwaja Kumar. Language identi_cation from small text samples. Journal of Quantitative Linguistics, 13(1):57_80, 2006.

[19] Giang-Son Nguyen, Xiaoying Gao, and Peter Andreae. Vietnamese document representation and classi_cation. In Ann Nicholson and Xiaodong Li, editors, AI 2009: Advances in Arti_cial Intelligence, volume 5866 of Lecture Notes in Computer Science, pages 577_586. Springer Berlin Heidelberg, 2009.

[20] Tu-Anh Nguyen-Hoang and Kiem Hoang. Frequent subgraph-based approach for classifying vietnamese text documents. In Joaquim Filipe and José Cordeiro, editors, ICEIS, volume 24 of Lecture Notes in Business Information Processing, pages 299_308. Springer, 2009.

[21] Nattira Muangmala Phimphaka Taninpong. Classification of thai independent study in statistics using data mining techniques, 2013.

[22] John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

[23] Xiaoguang Qi and Brian D. Davison. Web page classi_cation: Features and algorithms. ACM Comput. Surv., 41(2):12:1_12:31, February 2009.

[24] Monica Rogati and Yiming Yang. Highperforming feature selection for text classi_cation. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM '02, pages 659_661, New York, NY, USA, 2002. ACM.

[25] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, and P. Stamatopoulos. A memory-based approach to antispam filtering for mailing lists. Information Retrieval, 6(1):49_73, 2003. cited By 117.

[26] A. Srithirath and P. Seresangtakul. A hybrid approach to lao word segmentation using longest syllable level matching with named entities recognition. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference on, pages 1_5, May 2013.

[27] Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. Automatic text categorization in terms of genre and author. Comput. Linguist., 26(4):471_495, December 2000.

[28] M. Suzuki, N. Yamagishi, and Yi-Ching Tsai. Chinese text categorization using the character ngram. In Information Theory and its Applications (ISITA), 2012 International Symposium on, pages 722_726, Oct 2012.

[29] Ah-Hwee Tan and et al. Text categorization, supervised learning, and domain knowledge integration. In IN PROCEEDINGS OF KDD-2000: WORKSHOP ON TEXT MINING, pages 113_114. ACM, 2000.

[30] Songbo Tan. An e_ective re_nement strategy for knn text classi_er. Expert Syst. Appl., 30(2):290_298, February 2006.

[31] Suge Wang, Deyu Li, Xiaolei Song, Yingjie Wei, and Hongxia Li. A feature selection method based on improved _sher's discriminant ratio for text sentiment classi_cation. Expert Systems with Applications, 38(7):8696 _ 8702, 2011.

[32] Wei Zhang and Feng Gao. An improvement to naive bayes for text classi_cation. Procedia Engineering, 15(0):2160 _ 2164, 2011. {CEIS} 2011.