Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 11 (2013) 748 – 754

Procedia Technology

The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013)

Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization

Hamood Alshalabi*, Sabrina Tiun, Nazlia Omar, Mohammed Albared

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Abstract

Due to the rapid growth of documents in digital form, research in automatic text categorization into predefined categories has witnessed a booming interest. Although, there is a wide range of supervised machine learning methods have been applied to categorize English, relatively, only a few studies have been done on Malay text categorization. This paper reports our comparative evaluation of three machine learning methods on Malay text categorization. Two feature selection methods (Information gain (IG) and Chi-square) and three machine learning methods (K-Nearest Neighbor (k-NN), Naive Bayes (NB) and N-gram) were investigated. The three supervised machine learning models were evaluated on categorized Malay corpus, and experimental results showed that the k-NN with the Chi-square feature selection gave the best performance (Macro-F1 = 96.14).

Selection and peer-review under responsibility of the Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.

Keywords: Feature selection;Machine Learning Methods; N-gram; Naïve Bayesian; K-Nearest Neighbour.

1. Introduction

Due to the rapid growth in the compilation of documents, Information Retrieval (IR) system faces much harder task in accessing and retrieving relevant documents. In order to improve the IR performance system, text categorization plays key roles for handling and organizing the text data. Most of categorization systems categorized documents into a number of pre-defined categories. Each document can be categorized into multiple, exactly one, or no category at

* Corresponding author.
E-mail address: hmoud.shalabi@gmail.com

Selection and peer-review under responsibility of the Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.
doi:10.1016/j.protcy.2013.12.254

all [1].

There are many machine learning methods which have been proposed for text categorization, such as N-gram, Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Decision Tree, and Support Vector Machine (SVM). Many studies evaluated the performance of these methods based on general Newswire articles, and most of the studies focused on automatic text categorization for documents written in English. Previous works on Malay text categorization is very limited. This may be due to the nature of Malay language and also to the lack of Malay language resources such as labelled corpus.

The automatic text categorization process foresees a set of tasks universally recognized by the research community [2]. These tasks include features design in which the corpus processing, extraction of relevant information, feature selection and feature weighting processes are performed. In addition, these tasks also include training task in which a machine learning classifier is trained using a set of labelled documents. Finally, the last task is the Testing task the accuracy of the classifier is evaluated by using a set of pre-labelled documents (i.e. test-set) that are not used in the training phase

In this paper, we have designed several supervised classification models for Malay text categorization. It presents an empirical comparison of two feature selection methods (Information Gain (IG) and Chi-Square), and three supervised machine learning classifiers (NB, N-Gram and k-NN). In order to evaluate those classification models, we collected Malay pre-defined categorized documents from online Malay newspapers archives, namely; Bernama, and Utusan on-line. This corpus has eight categories: *Arts, Business, Crime, History, Healthy, Medicine, Politics, Religion and Sport*.

The rest of this paper is organized as follows: Section 2 reviews related work in the area of text categorization. Methodology and the different key techniques and approaches are described in Section 3 and Section 4. In Section 5 and 6, we present the experiment setup and discuss on the experimental results. Finally, we conclude our work and indicate its future directions in Section 7.

2. Related work

Text categorization methods were first proposed in the 1950s where the word frequency was used to categorize documents automatically [2]. 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. In addition, many machine learning methods have been proposed for text categorization in the past studies, such as N-gram [3-6], k-NN [3, 7-9], Decision Tree [10, 11] and SVM [12, 13].

Many researchers attempted to obtain better classification algorithms performance for automatic text categorization. K-NN is considered as the common method to text categorization[16]. Therefore, k-NN has been treated as the base method for categorizing text. Thus, in this paper, we include k-NN, N-gram and NB, and the other two feature selections, IG and Chi-square, in our experiment to find out the most suitable method in categorizing Malay text.

3. Feature selection

Feature selection method (FSS) is one of the most crucial tasks that will make the performance of text categorization improved, as they will select the most predictive features. FSS improves the performance of text categorization tasks in terms of learning speed and effectiveness. FSS also reduces the number of data dimensions, additionally, it removes irrelevant, redundant, and noisy data[1]. In this section, we give a brief introduction of the used feature selection methods:

Information Gain (IG) : IG measures the number of information bits that is obtained to predict categorization by recognizing whether a term in a document is present or absent [14], [15] and [16]. As result, IG is recruited in this context to select features that disclose the most information about the classes [17] . The values of IG were calculated as the following:

$$\mathrm{IG(t)} = -\sum_{i=1}^{|c|} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{|c|} p(c_i|t)\log p(c_i|t) + p(\bar{t})\sum_{i=1}^{|c|} p(c_i|\bar{t})\log p(c_i|\bar{t}) \quad (1)$$

In which $p(c_i)$ denotes the probability that class c_i occurs; $p(t)$ denotes the probability that word t occurs; $p(\bar{t})$ denotes the probability that word $\bar{t}$ does not occurs.

Chi Square: It measures the absence of independence between t (term) and c (category) [15, 16] . Chi-square can be calculated as follows:

$$\chi^2(c,t) = \frac{N \times (\text{AD-BC})}{(A+C)(B+C)(A+B)(C+D)} \tag{2}$$

$$\chi^2_{max}(t) = max_i(\chi^2(t,c_i)) \tag{3}$$

where A is the number of documents that contain the term, t, and also belong category, c. B is the number of documents that contain the term, t, but do not belong to category, c. C is the number of documents that do not contain the term, t, but belong to category, c. D is the number of documents that do not contain the term, t, and do not belong to category, c. N is the number of training documents [18].

4. Classification methods

As we mentioned before, we choose three classifier methods that are used in Malay text classification; the k-NN, NB and N-gram methods due to their simplicity, effectiveness and accurateness. Brief descriptions of these methods are given, as follows:

4.1. k-Nearest Neighbor (k-NN)

The k-NN is a well-known example-based classifier. It is one of the most popular classification techniques due to its simplicity and accuracy. The k-NN is also known as lazy learners, since it delays the decision on how to generalize beyond the training data until each new query instance is encountered. In order, to categorize a document, the k-NN classifier ranks scores of the document's neighbors among the training documents. Then, the k-NN uses the class labels of the k most similar neighbors.

Given a test document d, the system finds the K nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used. The weighted sum in k-NN classification is written as follows:

$$Score(d_i,d) = \sum_{d_j=KNN(\mathrm{d})} sim(d,d_j) \cdot \partial(d_j,c_j) \tag{4}$$

where KNN(d) indicates the set of K nearest neighbors of document d. If d_j belongs to c_i, $\delta(d_j, c_i)$ equals 1, or otherwise 0. For test documentd, it should belong to the class that has the highest resulting weighted sum. In order to compute $sim(d, d_j)$, we use the Euclidean distance, which represents the usual manner in which humans think of distance in the real world [19]:

$$D_{\text{Euclidean}}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{5}$$

4.2. The Naive Bayes (NB)

The NB algorithm is a widely used as an algorithm for document classification. It is a probability based classifier, based on the features independent probability value is calculated for each and every model. NB is often used in text category tasks based on Bayes' formula:

$$P(C_i|d) = \frac{P(C_i)P(d|C_i)}{P(d)} \tag{6}$$

Where $P(C_i|d)$ is the posterior probability of class C_i given a new document d, $P(C_i)$ is the probability of class C_i which can be calculated by:

$$P(C_i) = \frac{N_i}{N} \tag{7}$$

Where N_i is the number of documents assigned to class C_i, and N is the number of classes, $P(d|C_i)$ is the probability of a document d given a class C_i, and P(d) is the probability of document d, and because the independence assumption of NB, the probability of document d can be calculated by:

$$P(C_i|d) = P(C_i)\prod_{k=1}^{n} p(t_k|C_i) \tag{8}$$

where t_k is a feature that occurs with class C_i, and also we can calculate $p(t_k|C_i)$ by:

$$P(t_k|C_i) = \frac{1+n_{ki}}{1+\sum_{h=1}^{l} n_{hk}} \tag{9}$$

where n_{ki} is the total number of documents that contain feature t_k and belong to class C_i, l is the total number of distinct features in all training documents that belong to class Ci.

NB calculates posterior probability for each class, and then assigns document d to highest posterior probability's class, i.e.

$$C(d) = \text{argmax}_{i=}|C|\ (P(C_i|d)) \tag{10}$$

4.3. N-gram Classifier

An N-gram is a continuous sequence of n characters or n words of a longer portion of a text. In this work, character level N-grams classifier has been used. In the N-gram training process, the N-gram profile needs to be generated. The generated N-gram profile consisted of the text which is spilt into tokens consisting letters only. Only the most frequent N-grams are kept. This gives us the N-gram profile for the document. In order for each document to be classified, each document need to go through the text preprocessing phase, then, the N-gram profile will be generated as described above. The N-gram profile of each document will then be compared against the profiles of all documents in the training classes (class profile) in terms of similarity. Specifically, cosine similarity measurement is used. It measures the similarity between two documents training document Di and test document D_j:

$$Sim_{cosine}(D_i, D_j) = \frac{\sum_{k=1}^{m}(W_{ik} \times W_{jk})}{\sqrt{\sum_{k=1}^{m} W_{ik}^2 \times \sum_{k=1}^{m} W_{jk}^2}} \tag{11}$$

5. Evaluation and result

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 Malay corpus collected from online Malay newspapers archives, namely; Bernama, and Utusan on-line. This corpus contains 3040 documents that are different in length and divided into eight categories: *Arts, Business, Crime, History, Healthy, Medicine, Politics, Religion and Sport.*

All algorithms are evaluated using 5-fold cross-validation. To measure the performance of these classification methods, we use the Macro-averaged (Macro-F1) measure. This measure combines Recall and Precision in the following way metrics:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

$$\text{F}_1 = \frac{2 * \text{Recall} * \text{Precision}}{(\text{Recall} + \text{Precision})}$$

$$F_1^{\text{macro}} = \frac{1}{m}\sum_{i=1}^{m} F_1(i)$$

In which, True Positive (TP) is the set of document that is correctly assigned to the given category, False Positive (FP) is the set of documents that incorrectly assigned to the category, False Negative (FN) is the set of documents that is incorrectly not assigned to the category and True Negative (TN) is the set of the set of documents correctly not assigned to the category.

5.1. Experimental results

In order to test the efficiency of three classifiers k-NN, NB and N-gram, and in combination with the two feature reduction methods on Malay text categorization, we evaluate these methods individually, in which, features are selected from feature space at different size: 100, 200, 300, 400, 500 and 600. All of those classifiers have been evaluated using the 5-fold cross-validation. The results presented are in terms of macro-averaged F-measure where averaged values calculated across all 5-fold cross-validation experiments. We have examined the overall performance of the NB, N-gram and k-NN classifiers with the two feature selection methods, Chi-square and IG, applied to reduce the dimension of feature spaces. In the phase, the effects of the individual feature selection method on classifiers performances have been examined. Result on the performance (see Table 1) is displayed with features ranked in decreasing order and feature space at different size: 100, 200, 300, 400, 500 and 600. In Table 1, the best performance of 96.14 is the k-NN classifier when 400 of the features selected using by Chi-square feature selection. In addition, the best accuracy of 95.60 with NB classifier is achieved when 600 of the features selected by IG method are used, and the highest performance with N-gram classifier has been obtained when 600 of the features by IG method are used. When the classifier performances are compared, the k-NN algorithm achieves a higher performance than the NB and N-gram algorithms. We can see that the highest performance is obtained when the feature selection operations made by Chi-square. This observation indicates that the k-NN and NB classifiers are both suitable for Malay text categorization.

In order to examine the overall performance based on document categories, all of parameters for the three classifiers, k-NN, NB and N-gram, are fixed according to their best results in Table 1. The experimental results with the k-NN, NB and N-gram for Malay text categorization are shown in Fig 1. As seen in Fig. 1, the KNN achieves the best result in the *Religious, Business, Crime, Politic, Art* and *Health* domain, meanwhile the NB achieves its best result in *History* and *Sport* domains.

Table 1: The performance (Macro-F1) of NB, N-gram, and KNN classifiers (feature selection methods vs. features sizes).

	k-NN		NB		N-gram	
	Chi	IG	Chi	IG	Chi	IG
100	94.97	94.16	84.84	84.26	75.62	78.36
200	95.95	95.66	90.20	91.37	81.40	83.32
300	95.34	**95.75**	93.62	94.19	84.17	84.64
400	**96.14**	95.43	93.89	94.11	86.13	87.67
500	94.29	94.92	**95.11**	94.83	87.34	88.10
600	92.78	94.91	94.13	**95.60**	**87.80**	**90.05**

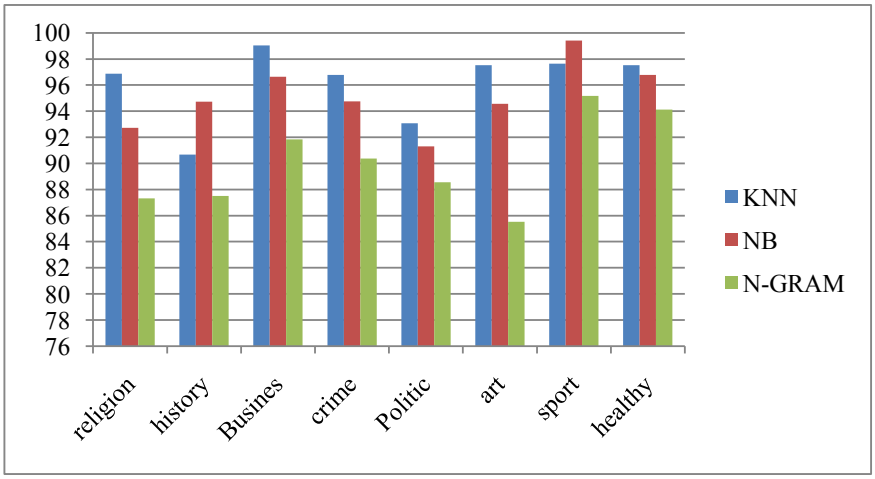

Fig. 1. The performance (F-measure) on each class of k-NN, NB and N-gram classifiers

6. Conclusion and future works

This paper presents our results on categorizing Malay documents. We have evaluated six classification models which are formed by combining two traditional feature selection methods, information gain (IG) and Chi-square, with three learning methods, k-NN, NB and N-gram

In order to conduct the experiment, we have collected categorized Malay documents from online Malay newspapers archives, namely; Bernama, and Utusan on-line, and treated it as our Malay corpus. The corpus contains 3040 documents that are different in length and divided into eight categories: *Arts, Business, Crime, History, Healthy, Medicine, Politics, Religion* and *Sport.*

Based on the carried out experiments, the obtained results showed that Chi-square feature selection method performed the best for terms selection, and both k-NN and NB exhibit the best classifiers for Malay text categorization.

In the future, we plan to will expand the size of the corpus and to add more categories for evaluation. We also plan to add other advanced classification models to be tested on the expanded Malay corpus.

References

[1] F. Sebastiani. Machine learning in automated text categorization. ACM computing surveys (CSUR) 2002; 34: 1-47.

[2] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney. Feature selection methods for text classification. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining; 2007. p. 230-239.

[3] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. Ann Arbor MI 1994; 48113: 161-175.

[4] J. Fürnkranz. A study using n-gram features for text categorization. Austrian Research Institute for Artifical Intelligence 1998; 3:1-10.

[5] F. Mohammed, L. Zakaria, N. Omar, and M. Albared. Automatic Kurdish Sorani text categorization using N-gram based model. International Conference on,Computer & Information Science. 2012. p. 392-395.

[6] M. Farhoodi, A. Yari, and A. Sayah. N-gram based text classification for Persian newspaper corpus. International Conference on Digital Content, Multimedia Technology and its Applications (IDCTA); 2011. p. 55-59.

[7] P. Soucy and G. W. Mineau. A simple KNN algorithm for text categorization. International Conference on Data Mining; 2001.p. 647-648.

[8] S. Tan. An effective refinement strategy for KNN text classifier. Expert Systems with Applications 2006; 30: 290-298.

[9] S. Manne, S. Kotha, and S. Sameen Fatima. Text Categorization with K-Nearest Neighbor Approach. Proceedings of the International Conference on Information Systems Design and Intelligent Applications; 2012 .p. 413-420.

[10] C. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems (TOIS) 1994;12: 233-251.

[11] D. E. Johnson, F. J. Oles, T. Zhang, and T. Goetz. A decision-tree-based symbolic rule induction system for text categorization. IBM Systems Journal 2002; 41: 428-437.

[12]T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98;1998. p. 137-142.

[13]T. Joachims. Learning to classify text using support vector machines: Methods, theory and algorithms. USA: Kluwer Academic Publishers Norwell; 2002.

[14]S. Li, R. Xia, C. Zong, and C.-R. Huang. A framework of feature selection methods for text categorization. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNL 2009; 2: 692-700.

[15]Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE;1997.p. 412-420.

[16]M. Rogati and Y. Yang. High-performing feature selection for text classification. Proceedings of the eleventh international conference on Information and knowledge management; 2002. p. 659-661.

[17]R. Mukras, N. Wiratunga, R. Lothian, S. Chakraborti, and D. Harper. Information gain feature selection for ordinal text classification using probability re-distribution. Proceedings of the Textlink workshop at IJCAI; 2007.

[18]F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi. Naïve Bayesian based on Chi Square to Categorize Arabic Data. Proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt,; 2009. p. 930-935.

[19]J. He, A.-H. Tan, and C.-L. Tan. A comparative study on Chinese text categorization methods. Proceedings of PRICAI'2000 International Workshop on Text and Web Mining; 2000. p. p24-35.