http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 7 (7): 1016-1022, 2008
ISSN 1812-5638

Automatic Classifications of Malay Proverbs Using Naïve Bayesian Algorithm

S.A. Noah and F. Ismail
Department of Information Science, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi Selangor, Malaysia

Abstract: This research presented an experimental study on automatic classification of Malay proverbs using Naïve Bayesian algorithm. The automatic classification tasks were implemented using two Bayesian models, multinomial and multivariate Bernoulli model. Both models were calibrated using one thousand training and testing dataset which were classified into five categories: family, life, destiny, social and knowledge. Two types of testing have been conducted; testing on dataset with stop words and dataset with no stop words by using three cases of Malay proverbs, i.e., proverb alone, proverb with meaning and proverb with the meaning and example sentences. The intuition was that, since proverbs were commonly short statement, the inclusion of its meaning and associated used in sentences could improve the accuracy of classification. The results showed that a maximum of 72.2 and 68.2% of accuracy have been achieved respectively by the Multinomial model and the Multivariate Bernoulli for the dataset with no stop words using proverb with the meaning and example sentences. This experiment has indicated the capability of the Naïve Bayesian algorithm in performing proverbs classification particularly with the inclusion of meaning and example usage of such proverbs.

Key words: Document classification, information retrieval, Naïve Bayesian algorithm

INTRODUCTION

Proverbs are generally known as short sentence of the folk which contains wisdom, truth, morals and traditional views in a metaphorical, fixed and memorizable form and which is handed down from generation to generation (Mieder, 2004). In Malay culture, proverbs play an important role and appear in political speeches, literature, popular culture and everyday life. A proverb has its own customary meaning (Norrick, 1984). Therefore, a proverb must be used within the correct context or meaning of a speech or a literature. For example the sentence or speech such as: "Perselisihan antara Azmi dan sepupunya Fazli tidak lama, akhirnya mereka berbaik semula" can be best described by the proverb "biduk lalu kiambang bertaut"; which literally means "The disagreement between Azmi and his cousin Fazli didn't stand for long, finally they reconciled together". This Malay proverb roughly means "when a small yacht passes through water plants, the water plants will get back together". Similarly for other proverbs in different languages such as English as illustrated in the following excerpt from the popular literature of Jules Vern, Journey to the Center of the Earth.

"The whole floor, composed of sand and malachite, was strewn with bones, freshly gnawed bones of reptiles and fish, with a mixture of mammals. My very soul grew sick as my body shuddered with horror. I had truly, according to the old proverb, fallen out of the frying pan into the fire. Some beast larger and more ferocious even than the shark-crocodile inhabited this den."

The aforementioned examples clearly indicate that proverbs must be used within the correct context of the sentences and speeches. As a result, the task for the automatic classification of proverbs which can assist users in selecting a suitable proverb according to a given context proved to be important and interesting but challenging. This is due to the nature of proverbs which contain short statements and pose limitation to existing text classification techniques to derive desirable features. Furthermore, to date there is no such commonly known classification scheme for proverbs (Takkinen, 1995).

Text classification is a problem in information science (Makkonen *et al*., 2004). The task is to assign text or document into categories based on their content (Sebastini, 2002). Mathematically, such task can be viewed as assigning a Boolean value to each pair $<t_j, c_i> \in T \times C$; where, T denotes the set of text or documents and C a set of predefined categories. The classification objective is to reduce the detail and diversity of data: eliminate information overload, to simplify access to and processing of information. In the classification research

Corresponding Author: Shahrul Azman Noah, Department of Information Science,
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
43600 UKM Bangi Selangor, Malaysia

community, automatic classification becomes more important due to the growing amount of the document in digital form. The machine learning paradigm has become one of the main classification approach in this area. It generates a classifier from the training set based on the characteristics of the documents already classified. Then it uses the classifier to classify the new documents. In the machine learning fields, text classification tasks can be divided into two types: supervised text classification and unsupervised text classification. Supervised text classification assumed that a preclassified text indicating the categories such text belongs to is made available; whereas unsupervised assumed no such data exist. Techniques form machine learning such as Naïve Bayesian (Hoare, 2008; Peng *et al*., 2004), Neural Network (Ruiz and Srinivasan, 2002), Rocchio (Xu *et al*., 2005) and Genetic Algorithm (Hirsch *et al*., 2005; Peng *et al*., 2006) has been a popular approach for supervised classification.

In this research, we explore the possibility of automating the task of proverb classification. The experiment was centered around three dimensions, i.e. proverbs alone; proverbs and meanings; and proverbs, the meanings of proverbs and the associated sentences that use such proverbs. Our intuition was that, since proverbs were commonly short statement, the inclusion of its meaning and associated used in sentences could improve the accuracy of classification. We decided to use the Naïve Bayesian algorithm technique as it was shown to be effective in many document classification tasks (McCallum and Nigam, 1998; Peng *et al*., 2004). It is a powerful, simple and language independent method.

MATERIALS AND METHODS

The Naïve Bayesian model: Bayesian classifiers have become widely accepted and increasingly used in the information science fields and have been found to perform surprisingly well (Friedman *et al*., 1997). These probabilistic approaches make strong assumptions about how the data is generated and posit a probabilistic model that embodies these assumptions; then they use a collection of labeled training data to estimate the parameters of the generative model. The Naïve Bayesian classifier is the simplest Bayesian models in that it assumes that all attributes of the examples are independent of each other given the context of the class (McCallum and Kamal, 1998).

Despite of simplest model and unrealistic assumption, the resulting classifier known as Naïve Bayesian is remarkably successful in practice and often compete well with more sophisticated technique (Friedman *et al*., 1997; Hilden, 1984; Domingos and Pazzani, 1997). It has proven effective in many practical applications including text classification, medical diagnosis and systems performance management. Naïve Bayesian has been successfully applied to document classification in many research efforts and has a good reputation for working in this area. It perform as well as newer classifiers, more sophisticated method and also shows a good runtime-behavior during the classification of new documents. Hovold (2005) in his research for spam filtering using a variant of the Naïve Bayesian classifier showed that it is possible to achieve very good classification performance using a word-position-based variant of Naïve Bayes. The simplicity and the low time complexity of the algorithm thus makes Naïve Bayes a good choice for end-user application.

Although Naïve Bayesian classifier is a simple technique used in document classification area, it has been implemented by different researchers with two different models. The two Naïve Bayesian models which commonly used in text classification are the multivariate Bernoulli model and the multinomial model (McCallum and Nigam, 1998). With the multivariate Bernoulli model, we make the Naïve Bayesian assumption that the probability of each word occurring in a proverb (and its varieties) is independent of the occurrences of other words in a document. Therefore, the probability of a proverb given its category is simply the product of the probability of the attribute values over all word attributes as follows:

$$P(d_i/c_j)=\prod_{t=1}^{V}\left[B_{it}P(w_t/c_j)+(1-B_{it})(1-P(w_t/c_j))\right] \quad (1)$$

where, V is the number of features in the vocabulary, $B_{it}\in(0,1)$ indicates whether feature t appears in proverb i and $P(w_t|c_j)$ indicates the probability that feature w_t appears in a document whose category is c_j. For the multivariate Bernoulli model, $P(w_t|c_j)$ is the probability of feature w_t appearing at least once in a proverb whose category is c_j. It is calculated from the training sample as:

$$P(w_t|c_j)=\frac{1+\sum_{t=1}^{D_j}B_{it}}{J+D_j} \quad (2)$$

where, D_j is the number of proverbs in the training group scored c_j and J is the number of category. The value 1 in the numerator and J in the denominator are Laplacian values meant to adjust for the fact that this is a sample probability and to prevent $P(w_t|c_j)$ from equaling to zero or unity (Rudner and Liang, 2002).

In contrast with the multivariate Bernoulli model, the multinomial model captures word frequency information in proverbs. In the multinomial model, a document is an

ordered sequence of word events from the same vocabulary V. We assumed that the lengths of proverbs are independent of categories. As in the Multivariate Bernoulli model, we again assumed that the probability of each word event in a proverb is independent of the word's context and position in the proverb. The probability of each category for a given proverb is computed as the product of the probabilities of the features contained in the proverb as follows:

$$P(d_i \mid c_j)=\prod_{i=1}^{V}\frac{P(w_t \mid c_j)^{N_{it}}}{N_{it}} \quad (3)$$

where, N_{it} is the number of times feature w_t appears in proverb i. For the multinomial model $P(w_t|c_j)$ is the probability of feature w_t being used in a proverb whose category is c_j. It is calculated from the training sample as:

$$P(w_t \mid c_j) = \frac{1+\sum_{i=1}^{D_j} N_{it}}{V+\sum_{i=1}^{N_d} N_{it}} \quad (4)$$

where, N_d is the total number of proverbs.

The key difference in both models is the computation of $P(w_t|c_j)$ (Rudner and Liang, 2002). The Bernoulli checks for the presence or absence of the feature in each proverb. The multinomial on the other hand accounts for multiple uses of the feature within a proverb. Generally, the multinomial model works much faster as only features in the proverb need to be examined. In contrast, the multivariate Bernoulli requires all the features in the vocabulary to be examined. McCallum and Nigam (1998) suggest that with a large vocabulary the multinomial model is more accurate than the Bernoulli model for many classification tasks.

Stopwords: Similar to English and other languages, Malay language contains large number of stopwords, i.e. words that have little content-bearings. Some researches on text classification have reported improved accuracy with the removal of stopwords (Mitchell, 1997; Lazarinis, 2007) because the goal is to remove words that do not carry discriminative meaning of the document contents, while others have reported otherwise (Silva and Ribeiro, 2003). As a result, in this experiment we will consider data sets with and without stopwords.

Stemming: Stemming is the process of reducing inflected (or sometimes derived) words to their word stem. For example, the words compute, computer, computation and computational will all be stemmed to the word comput if based on the classic Porter's algorithm (Porter, 1980). Stemming is usually language dependent, except for certain algorithm such as the N-gram technique (Mayfield and McNamee, 2003). For Malay language, only a few stemming algorithm have been proposed by Ahmad *et al*. (1962). The original intention of stemming in the study of information retrieval is to reduce the number of indexes thus increases the speed of retrieval. However, in the case of short documents such as the proverbs, stemming is not really necessary. This is also valid as most of the words in proverbs are classical words which may be overstemmed by the current Malay stemming algorithm. Therefore, stemming is not incorporated in this study.

Feature selection: The objective of features selection is to eliminate those features that provide on few or less important information (Lee and Lee, 2006). Feature selection is done by selecting words that have the highest average mutual information with the category (Bratko and Filipic, 2006). We used the information gain measure which is based on the entropy in information theory as proposed by Shannon and Weaver (1998) and further discussed by MacKay (2003). Information gain measure is claimed to be one of the best method in selection of features as compared to the simple term frequency and its varieties approaches. Entropy is defined as:

$$H(S)=\sum_{j=1}^{J}-p_j \log_2 p_j \quad (5)$$

where, p_j is the probability of belonging to category J.

Entropy can be viewed as a measure of uniformness of a distribution and has a maximum value when $p_j = 1/J$ for all J. The goal is to have a peaked distribution of p_j. The potential information gain is the reduction in entropy:

$$H(S_0)-H(S_i) \quad (6)$$

where, $H(S_0)$ is the initial entropy based on the prior probabilities and $H(S_i)$ is the expected after scoring of feature t.

Experimental approach: Two Bayesian models for proverb classification were examined, a multivariate Bernoulli model and a multinomial model, using words as features. We are at the moment not considering phrases. In our testing, the data set consists of proverbs only; proverbs with meanings; and proverb with meanings as well as example of sentences (involving the use of such

proverbs). Overall there are 1000 data, of which 50% of them are used for training and the remaining for testing. The proverbs under the training data have been classified either as family, life, destiny, social and knowledge. Each classification contains 100 data. The training phase shows between 96-99% of accuracy for all the three different cases (i.e., the data set consists of proverbs only; proverbs with meanings and proverb with meanings as well as example of sentences).

All the proverbs together with its classification were derived from classic collections (Ahmad, 1962) as well as recent collections (Salleh, 1988). The results of the testing are shown as a precision breakeven point, a standard information retrieval measure for binary classification, which is defined as:

$$\text{Precision} = \frac{\text{No. of correct positive classifications}}{\text{No. of positive classifications}} \quad (7)$$

RESULTS AND DISCUSSION

The result of the experiment is as shown in Table 1 and 2 for dataset without stopword and with stopword, respectively. Detail comparisons of these results are shown in Fig. 2-7.

As can be show from Table 1 and 2 the multinomial model shows a slightly better performance as compared to the Multivariate Bernoulli. As expected providing more information to the proverbs give better performance as illustrated by the result of the proverb+meaning dataset and proverb+meaning+example dataset with an increase of accuracy by 32.8 and 34.0%, respectively for the multinomial model and 33.2 and 34.6%, respectively for the multivariate Bernoulli model.

The slightly better performance of the multinomial model is might due to the nature of the model that captures word frequencies information. Instead, the multivariate Bernoulli model only considers the presence or absence of the feature (words) in each proverb (and its varieties). Present testing results are consistent with that of McCallum and Nigam (1998) findings who found that, with large vocabulary sizes, multinomial model is said to be more accurate over multivariate Bernoulli to many classification tasks.

In terms of stopwords removal, although some (Silva and Ribeiro, 2003) have indicated the insignificance of it, present results have shown that dataset without stopwords have better accuracy. This therefore is coincided with Mitchell (1997) statements that classification accuracy can be increased with the removal

Table 1: Testing by removing stopword

	Proverb		Proverb+meaning		Proverb+meaning +example	
# Doc.	MNM	MVB	MNM	MVB	MNM	MVB
10	20.0	20.0	30.0	32.0	36.0	44.0
20	27.0	22.0	47.0	47.0	49.0	43.0
30	30.7	30.7	54.0	53.3	52.0	55.3
40	30.5	27.5	59.0	54.5	54.5	55.0
50	30.8	28.8	62.8	62.4	56.8	56.8
60	31.7	31.3	62.0	61.0	61.3	59.7
70	34.6	34.3	67.7	66.3	67.1	63.7
80	34.0	35.2	68.8	68.8	68.8	67.2
90	35.8	36.2	69.3	68.7	69.6	68.4
100	38.2	36.8	71.0	70.0	72.2	71.4

*MNM = Multinomial Model, MVB = Multivariate Bernoulli

Table 2: Testing with stopword

	Proverb		Proverb+meaning		Proverb+meaning +example	
# Doc.	MNM	MVB	MNM	MVB	MNM	MVB
10	28.0	18.0	24.0	30.0	30.0	38.0
20	30.0	31.0	40.0	47.0	43.0	37.0
30	31.3	32.7	48.7	50.7	52.0	48.0
40	35.5	31.0	55.5	55.0	56.5	51.0
50	35.2	30.0	61.2	59.6	55.6	54.8
60	35.0	32.0	61.3	60.3	57.0	55.0
70	38.3	34.6	64.6	63.7	63.4	59.4
80	39.0	37.0	65.8	64.8	66.2	62.5
90	39.6	38.4	67.8	65.8	68.2	63.8
100	41.4	39.0	69.2	67.6	71.2	68.2

*MNM = Multinomial Model, MVB = Multivariate Bernoulli

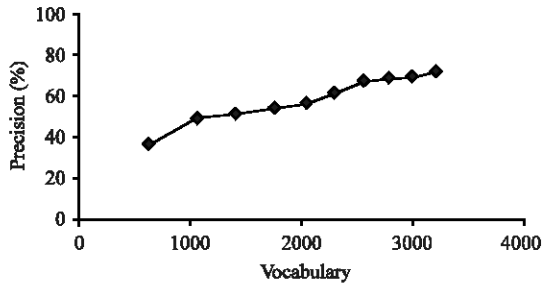

Fig. 1: The effect of vocabulary terms with precision

of stopwords. Apart from that, the removal of stopwords has another important benefit which is reducing the size of features. As a result, better features will be generated because stopwords is considered has little content bearing meaning (which is not suitable to be used as features). As mentioned earlier, present testing has showed that the more information given on the proverbs will give better classification results. This is also supported by the increased of classification accuracy with the increased of vocabulary terms as shown in Fig. 1. As can be seen, the classification precision of accuracy increase from 36% at 632 words size of vocabulary to 72.2% at 3203 words size of vocabulary.

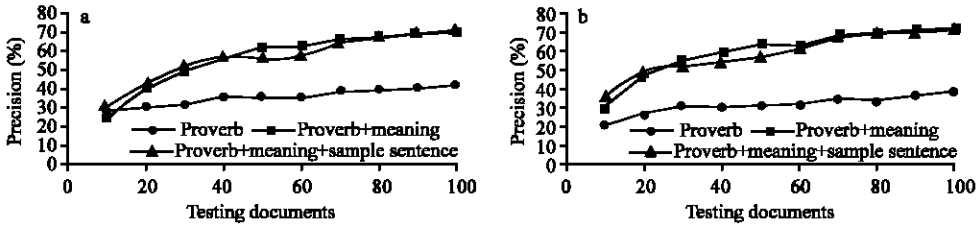

Fig. 2: Comparison of multinomial data set (a) with and (b) without stopwords

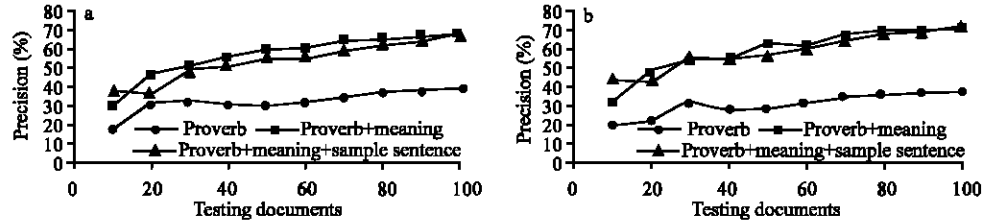

Fig. 3: Comparison of multivariate Bernouli data set (a) with and (b) without stopwords

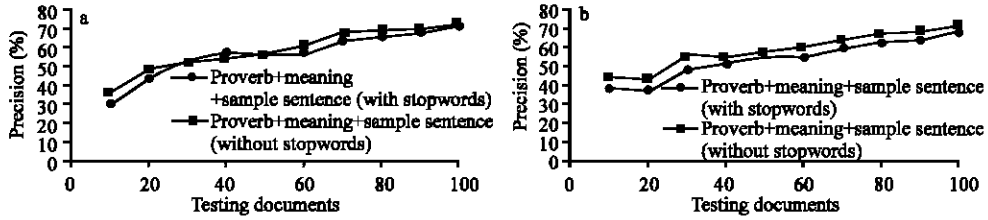

Fig. 4: Comparison of (a) multinomial and (b) multivariate Bernouli (with stopwords vs without stopwords)

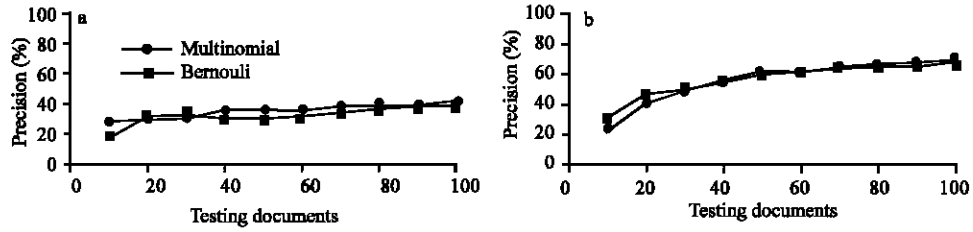

Fig. 5: Comparison of multinomial vs Bernouli (a) proverb with stopwords and (b) proverb+meaning with stopwords

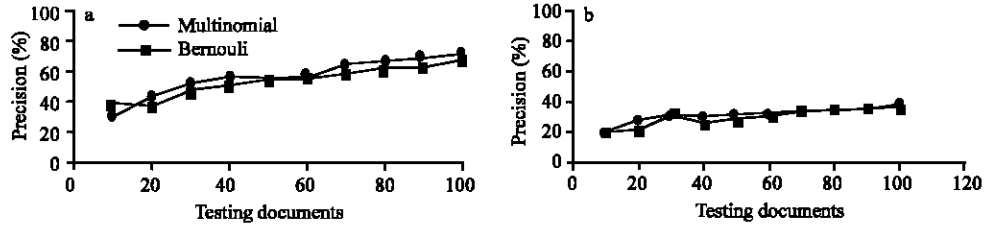

Fig. 6: Comparison of multinomial vs Bernouli (a) proverb+meaning+sample sentence with stopwords and (b) proverb without stopwords

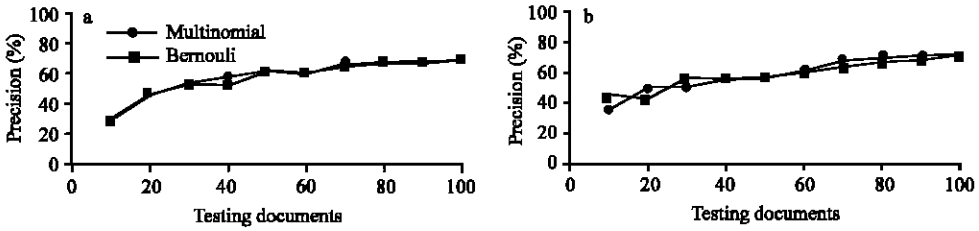

Fig. 7: Comparison of multinomial vs Bernouli (a) proverb+meaning without stopwords and (b) proverb+meaning+sample sentence without stopwords

CONCLUSION AND FUTURE WORK

This study has presented research on automatic classification of proverbs using Naïve Bayesian algorithm. Two models which have frequently used in text classifications; multinomial and multivariate Bernoulli have been used during the experiment. We have divided the test data into three categories: proverb; proverb+meaning and proverb+meaning+example. The results are very encouraging with the multinomial model shows a better accuracy. The results also suggesting that more information provided in the proverbs will increase the classification accuracy as exhibited by the results of the proverb+meaning+example data set. The result shows the potential of the Naïve Bayesian in performing the proverbs classification. We believe that better results can be obtained if more test and training data being used.

Our future research is looking at the possibility of providing a kind of assistance tool for writers who wish to use proverb in their writings by looking at the context of sentences they are constructing. We may combine the Naïve Bayes approach and an approach from semantic sentence similarity as proposed by Noah *et al.* (2007). Further testing is also required for other classification algorithm such as the Support Vector Machine (SVM) (Simon and Koller, 2001), Bayesian Network (Denoyer and Gallinari, 2004) and k-Nearest Neighbor (Kwon and Lee, 2003). These algorithms were known to be successful in text classification for web documents; but their application to literary documents such as proverbs has yet to be tested.

REFERENCES

Ahmad, F., M. Yusoff and T.M.T. Sembok, 1996. Experiments with a stemming algorithm for Malay words. J. Am. Soc. Inform. Sci., 47: 896-908.

Ahmad, Z.A., 1962. Knowledge of Malay Writing. 6th Edn., Dewan Bahasa and Pustaka, Kuala Lumpur, ISBN: 9836275975.

Bratko, A. and B. Filipic, 2006. Exploiting structural information for semi-structured document categorization. Inform. Process. Manage., 42: 679-694.

Denoyer, L. and P. Gallinari, 2004. Bayesian network model for semi-structured document classification. Inform. Process. Manage., 40: 807-827.

Domingos, P. and M. Pazzani, 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn., 29: 103-130.

Friedman, N., D. Geiger and M. Goldszmidt, 1997. Bayesian network classifiers. Mach. Learn., 29: 131-163.

Hilden, J., 1984. Statistical diagnosis based on conditional independence does not require it. Comput. Biol. Med., 14: 429-435.

Hirsch, L., M. Saeedi and R. Hirsch, 2005. Evolving text classification rules with genetic programming. Applied Artificial Intell., 19: 659-676.

Hoare, Z., 2008. Landscapes of naïve bayes classifiers. Pattern Anal. Appl., 11: 59-72.

Hovold, J., 2005. Naïve Bayes spam filtering using word-position-based attributes. Proceeding of the 2nd Conference on Email and Anti-Spam, July 21-22, Stanford University, California, USA., pp: 1-8.

Kwon, O.W. and J.H. Lee, 2003. Text categorization based on k-nearest neighbor approach for web site classification. Inform. Process. Manage., 39: 25-44.

Lazarinis, F., 2007. Engineering and utilizing a stopword list in Greek web retrieval. J. Am. Soc. Inform. Sci. Technol., 58: 1645-1652.

Lee, C. and G.G. Lee, 2006. Information gain and divergence-based feature selection for machine learning-based text categorization. Inform. Process. Manage., 42: 155-165.

MacKay, D.J.C., 2003. Information Theory, Inference and Learning Algorithms. 1st Edn., Cambridge University Press, Cambridge, ISBN: 978-0521642989.

Makkonen, J., H. Ahonen-Myka and M. Salmenkivi, 2004. Simple semantics in topic detection and tracking. Inform. Retrieval, 7: 347-368.

Mayfield, J. and P. McNamee, 2003. Single n-gram stemming. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-August 1, ACM New York, USA., pp: 415-416.

McCallum, A. and N. Kamal, 1998. A comparison of event models for naive bayes text classification. Proceedings of AAAI-98 Workshop on Learning for Text Categorization. July 26-27, AAAI Press, Madison, Wisconsin, USA, pp: 41-48.

Mieder, W., 2004. Proverbs: A Handbook. 1st Edn., Greenwood Press, Westport, ISBN: 978-0-313-32698-3.

Mitchell, T.M., 1997. Machine Learning. 1st Edn., McGraw-Hill, New York, ISBN: 978-0070428072.

Noah, S.A., A.Y. Amruddin and N. Omar, 2007. Semantic similarity measures for malay sentences. Lecture Notes Comput. Sci., 4822: 117-126.

Norrick, N.R., 1984. How Proverb Means: Semantic Studies in English Proverbs. 1st Edn., Mouton, Berlin, ISBN: 3-11-0101963.

Peng, F., D. Schuurmans and S. Wang, 2004. Augmenting naive bayes classifiers with statistical language models. Inform. Retrieval, 7: 317-345.

Peng, T., F. He and W. Zuo, 2006. Text classification from positive and unlabeled documents based on GA. VECPAR'06, 7th International Meeting, July 10-12, Rio de Janeiro, Brazil, pp: 32-38.

Porter, M.F., 1980. An algorithm for suffix stripping. Read. Inform. Retrieval, 14: 313-316.

Rudner, L.M. and T. Liang, 2002. Automated essay scoring using bayes theorem. The J. Technol. Learn. Assess., 2: 3-21.

Ruiz, M. and P. Srinivasan, 2002. Hierarchical text categorization using neural networks. Inform. Retrieval, 5: 87-118.

Salleh, Z., 1988. Practical Malay Proverbs. 1st Edn., Berita Publishing, Kuala Lumpur, ISBN: 9679691837.

Sebastini, F., 2002. Machine learning in automated text categorization. ACM Comput. Surveys, 34: 1-47.

Shannon, C.E. and W. Weaver, 1998. The Mathematical Theory of Communication. 1st Edn., University of Illinois Press, Illinois, ISBN: 978-0252725463 .

Silva, C. and B. Ribeiro, 2003. The importance of stop word removal on recall values in text categorization. Proceedings of the International Joint Conference on Neural Networks, July 20-24, IEEE, pp: 1661-1666.

Simon, T. and D. Koller, 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2: 45-66.

Takkinen, J., 1995. An adaptive approach to text categorization and understanding a preliminary study. Fifth IDA Graduate Conference on Computer and Information Science, November 22, Department of Computer and Information Science, Linköping University, pp: 39-42.

Xu, X., B. Zhang and Q. Zhong, 2005. Text categorization using SVMs with Rocchio ensemble for Internet information classification. Networking and Mobile Computing, 3rd International Conference, August 2-4, ICCNMC, pp: 1022-1031.