Journal of Computer Science 8 (9): 1441-1446, 2012
ISSN 1549-3636

Poetry Classification Using Support Vector Machines

Noraini Jamal, Masnizah Mohd and Shahrul Azman Noah
Knowledge Technology Research Group,
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia

Abstract: Problem statement: Traditional Malay poetry called pantun is a form of art to express ideas, emotions and feelings in the form of rhyming lines. Malay poetry usually has a broad and deep meaning making it difficult to be interpreted. Moreover, few efforts have been done on automatic classification of literary text such as poetry. **Approach:** This research concerns with the classification of Malay pantun using Support Vector Machines (SVM). The capability of SVM through Radial Basic Function (RBF) and linear kernel function are implemented to classify pantun by theme, as well as poetry or non-poetry. A total of 1500 pantun are divided into 10 themes with 214 Malaysian folklore documents used as the training and testing datasets. We used tfidf for both classification experiments and the shape feature for the classification of poetry and non-poetry experiment alone. **Results:** The results of each experiment showed that the linear kernel achieved a better percentage of average accuracy compared to the RBF kernel. **Conclusion:** The results show the potential of SVM technique in classifying poems into various classification of which previous approaches only focused on classifying prose only.

Key words: Text classification, support vector machines, malay poetry, Radial Basic Function (RBF), express ideas, Malaysian folklore, classify pantun

INTRODUCTION

Poetry is a form of literary art in which language is used for its aesthetic and evocative qualities in addition to, or in lieu of, its' apparent meaning. Classical poetry showed works of art with diverse styles from the history development of past literary, tradition and the nature of the data is unique and creative. Poetries are usually mean to deliver expression such as love, kindness and dignity. Thus, there are various categories of poetries. However, efforts in performing automatic classification of poetries are very rare. This is because the forms and features of poetry text are different from normal text as bound by a factor of lines, stanzas, rhyme, elements of style and beauty of sound and rhythm. Furthermore, poetries are usually in the form of short textual paragraphs with little discriminative value word features for automatic classification purposes. Therefore, the classification of poetries proved to be a challenging task. Classification of poetry is important particularly in information retrieval (or poetry retrieval) as its retrieval is not according to a simple keyword matching but involves the context, classes and themes of the poetries. Apart from that, the ability to recognize poetry from prose is potentially important, particularly for search machines with particular emphasis for poetry mining.

This study presents evaluative experiments on the performance of Support Vector Machine (SVM) in Malay poetry classification. Malay poetry is a traditional poem of Malay in the form of oral poetry and branches of the oldest Malay literature. Classical Malay poetry inspired by the earliest Malays is the historical treasures that contains valuable source of knowledge that reflects the character of ancient Malay civilization. Malay poetry usually has a broad and deep meaning making it difficult to be interpreted even by the Malays (Winstedt and Wilkinson, 1957). Classification of poetries presented in this study includes two types: Classification of poetries into themes and classification of texts (or documents) into poetry and non-poetry classes.

Background and related research: Pantun or the Malay poetries are categorized by audience, shape and theme (Bakar, 1983). Audience means the division of the poetry from the perspective of both parties, who recited and heard the poetry according to their age. Therefore, poetries are appropriate to be 'heard by' or 'recited for' children, adults or the elderly. The shape refers to the division of poetry from the point of length lines structure either consisting of two, four, six and up to sixteen lines. While the theme refers to the categorization of poetry based on philosophical concept, experience, emotion, interpretation and human understanding. For example, poetry that expresses love to God is classified under the 'religion' theme.

Corresponding Author: Shahrul Azman Noah, Knowledge Technology Research Group, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi Selangor, Malaysia Tel: +6 03 89216178 Fax: +6 03 89256732

Malay poetry consists of short sentences in which each line of poetry has only four to seven words. The lines of Malay poetry are formed in pairs by alternating between each line of poetry. Typically, the numbers of lines in each verse of poetry are either two, four, six, eight or twelve lines. An example of Malay poetry follows (verses in brackets are the translation). Note the rhyming alternate lines:

Dua tiga kucing berlari	(Two or three cats a running)
Manakan sama	(They are not comparable
sikucing belang	to the cat with stripes)
Dua tiga boleh kucari	(Two or three (people)
	I can find)
Manakan sama	(But there are not
puan seorang	comparable to you)

There are two quite separate parts in a verse of Malay poetry. The message (or meaning) that the poetry carries is contained in the second half, i.e., the last two lines. The first two lines simply act as the lead (or indicator) as to what is coming (Ahmad, 2002). The most important role of the indicator is simply to serve as the "rhymer". The first two lines do make sense in themselves but have no relationship in meaning with the second half of the poetry. As such, the selection of appropriate techniques for predicting Malay poetry classification based on the features found in the documents poetry, is a major problem and challenge.

However, the effort in automatic classification of literary text such as poetry is little or none. The only related work for Malay type of poetry is discussed in (Noah and Ismail, 2008), which classifies Malay proverbs using naïve Bayesian method. It achieved a maximum of 72% accuracy using Multinomial model with background knowledge (i.e., meaning of proverbs and example of sentences) and only 42% of accuracy without any background knowledge. Meanwhile, a similar effort in poetry classification in the study of (Tizhoosh *et al*., 2008) only focused on recognizing poem from prose with the best accuracy recorded at 97%.

Naive Bayes (McCallum and Nigam, 1998; Ko and Seo, 2000), Rocchio (Lewis *et al*., 1996), Nearest Neighbor (Yang *et al*., 2002) and SVM (Joachims, 1998) are among the popular supervised learning methods for classification. However, SVM shown the best accuracy performance amongst classification techniques in text classification problem (Joachims, 1998; Dumais *et al*., 1998; Weiss *et al*., 1999). Therefore, we used SVM in this study. SVM has a feature that maximizes the margin that attempts to separate between the two groups of classes more easily.

MATERIALS AND METHODS

Support Vector Machine (SVM): Vapnik (2000) has developed SVM based on the principle of Structural Risk Minimization from Statistical Learning Theory framework. SVM was introduced in the field of text classification by (Joachims, 1998) and subsequently used by (Dumais *et al*., 1998; Drucker *et al*., 1999; Taira and Haruno, 1999; Yang and Liu, 1999). A SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which are used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin). To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function K(x, y) selected to suit the problem. According to (Hsu *et al*., 2010), there are several types of kernel functions, such as the linear kernel, polynomial kernel, Radius Basic Function (RBF) and sigmoid kernel. SVM is a very established method in text classification and it is beyond the scope of this study to discuss it in detail. Interested reader may refer to (Brucher *et al*., 2002; Han and Kamber, 2006).

Model construction: Model construction is a training process in which the SVM is trained using a set of training documents to map the document to the category by class and revenue model calculations. Three processes are involved, which are: Document representation, feature selection and transformation to SVM data format.

The document representation process transforms poetry into terms-documents vector, where the vector represents the weight of terms in documents. Tokenization and stopword removal are among the text processing processes in this stage. Given that poetries are short textual documents, unstemmed single terms are used as features. Therefore, in feature selection the terms weight is based on the tfidf weighting scheme as shown in the following Eq. 1:

$$w_{ik} = tf_{ik} \times \log \frac{N}{n_k} \quad (1)$$

Whereby w_{ik} is the weight of term k in poetry i, tf_{ik} is the frequency of term k in poetry i, N is the total number of poetries and n_k is the number of poetries with terms k. Words are used to classify poetries by theme. On the other hand, we also perform classification as to whether the text is a poetry or non-poetry. In this case, the shape structures of poetries are used. The shapes and characteristics of Malay poetry were gained through analysis of literatures particularly from (Ahmad, 2002). Features of the shape structure were applied in the calculation of the structural features of textual documents. The resulting values were used as a model in the classification of poetry and non-poetry.

Table 1: Distribution set of Malay poetry for experiment 1

Poetry themes	Number of poetry documents	Set of training documents	Set of testing documents
Customs and Traditions of Human	119	84	35
Religion	98	69	29
Travel and Foreign Place	249	174	75
Heroism	139	98	41
Riddle	17	12	5
Love	257	179	78
Kindness	132	93	39
Advice and Education	166	115	51
Proverbs	103	72	31
Metaphor	220	154	66
Total	1500		

Table 2: Distribution set of training and testing documents for experiment 2

Set of training documents -70%			Set of testing documents -30%		
Total	Malay Poetry	Malaysian folklore	Total	Malay poetry	Malaysian folklore
50	25	25	22	11	11
100	50	50	42	21	21
150	75	75	64	32	32
200	100	100	84	42	42
250	125	125	108	54	54
300	150	150	128	64	64

The shape structure features used are: the number of lines in the document(N_{line}); the number of words contained in the document (N_{word}); the average number of words per line ($A_{word} = N_{word}/N_{line}$); the number of syllables in the document, ($N_{syllable}$); the average number of syllables per line($A_{syllable}= N_{syllable}/N_{line}$); the number of rhyme at the end of line (Nrhyme); the number of paragraphs($N_{paragraph}$) and the average number of words per paragraph ($A_{paragraph}= N_{word}/N_{paragraph}$).

Then, the weights values from feature selection process are transformed to SVM data format using the SVM classifier. The model built by the classifier is used to predict the classes of poetry and non-poetry.

Experiment setup: As mentioned earlier, we conducted two types of experiments: i.e. classification of poetry based on theme; and classification of either poetry or non-poetry based on the shape structure and tfidf features. The dataset for classical Malay poetries were mainly derived from (Bakar, 1983), which is actually a compilation of poetries from various sources arranged according to themes. The non-poetry dataset is represented by the Malaysian folklore collections obtained from (Puteh and Said, 1995). Both of these datasets are necessary during the training phase as SVM requires both positive and negative documents. Two kernel functions of SVM such as RBF and Linear kernel were tested using LIBSVM software (Fan *et al*., 2005). The 5-fold and 10-fold cross validation technique were used to validate the estimation on percentage accuracy of classification. The following sections describe the two experiments.

Experiment 1: Classification of poetry according to themes: In this experiment, we evaluated the performance of SVM in classifying poetries according to themes. We divided the poetries into two groups i.e., poetries that contain the meaning only and poetries that contain both the indicator and the meaning. This intention of this division is to observe the best features for performing the classification. Table 1 shows the ten themes used in this experiment together with their breakdown of numbers. We used the tfidf term weighting scheme to represent the features. Two types of test were carried out for each group; i.e. with stopwords and without stopwords.

Experiment 2: Classification as poetry and non-poetry: In this experiment, the intention is to distinguish and extract appropriate poetic features with which Malay poetry can be accurately differentiate from other type of texts. The features used during the experiment are the shape structure of the poetry and the tfidfterm weighting. Removal of stopwords is not implemented in the former feature in order to maintain the shape structure of the original document. In this experiment, we used 428 documents comprising of 214 classical Malay poetry documents and 214 Malaysian folklore documents. The Malaysian folklore documents are in the form of short paragraphs and not bound by number of lines, number of words per line and number of syllables per line. Six repetitions of training and testing on the SVM classification were carried out with an increase of 50 training documents for each test as shown in Table 2. The shape features of Malay poetry were based on the eight types of shape features previously described.

RESULTS

Evaluation is based on the classification accuracy and effectiveness. We compared two types of SVM i.e., RBF kernel and Linear kernel. Follows are the result for both experiments.

Table 3: Result for Experiment 1 (with stopword)

	Meaning				Indicator + Meaning			
	5-Fold		10-Fold		5-Fold		10-Fold	
Poetry Themes	RBF	Linear	RBF	Linear	RBF	Linear	RBF	Linear
Customs and Traditions of Human	50	54.29	47.14	54.29	58.57	48.57	61.43	52.86
Religion	50.00	56.90	46.55	56.90	44.83	51.72	67.24	53.45
Travel and Foreign Place	53.33	52.00	53.33	51.33	57.33	51.33	54.67	50.00
Heroism	53.66	57.32	54.88	57.32	52.44	57.32	53.66	58.54
Riddle	40.00	80.00	40.00	80.00	50.00	80.00	50.00	80.00
Love	57.69	58.33	57.05	58.97	57.69	55.77	60.26	57.69
Kindness	65.38	73.08	64.10	71.79	62.82	60.26	61.54	60.26
Advice and Education	54.90	57.84	57.84	57.84	51.96	55.88	51.96	51.96
Proverbs	54.84	48.39	43.55	48.39	53.23	50.00	50.00	50.00
Metaphor	43.94	46.21	40.91	46.21	53.03	47.73	53.79	47.73
Average	52.37	58.44	50.54	58.30	54.19	55.86	56.45	56.25

Table 4: Result for Experiment 2 (without stopword)

	Meaning				Indicator + Meaning			
	5-Fold		10-Fold		5-Fold		10-Fold	
Poetry Themes	RBF	Linear	RBF	Linear	RBF	Linear	RBF	Linear
Customs and Traditions of Human	47.72	57.14	5	50.00	61.43	54.29	68.57	52.86
Religion	48.28	46.55	48.28	46.55	63.79	62.07	70.69	63.79
Travel and Foreign Place	52.67	58.00	50.00	58.00	54.67	48.00	52.00	47.33
Heroism	58.54	67.07	63.41	65.85	58.54	47.56	58.54	56.10
Riddle	50.00	70.00	50.00	70.00	50.00	60.00	50.00	60.00
Love	51.92	54.49	52.92	53.21	53.21	50.00	48.72	48.08
Kindness	69.23	52.56	69.23	52.56	53.85	65.38	53.85	60.26
Advice and Education	50.90	50.98	48.04	50.00	52.94	50.98	56.86	50.98
Proverbs	58.06	54.84	59.68	54.84	56.45	46.77	56.45	50.00
Metaphor	54.55	51.52	53.79	59.24	49.24	49.24	50.00	49.24
Average	53.99	56.32	54.43	55.03	55.41	53.43	56.57	53.86

Experiment 1: The result of the experiment is as shown in Table 3 and 4 for the dataset with and without stopwords, respectively. The test results showed that the Linear kernel obtained the best percentage of classification accuracy for the riddle theme, followed by the kindness theme for the dataset with stopwords. Deeper analysis indicated that there are common terms used in the poetry for the riddle and kindness themes that subsequently boost the percentage of classification accuracy for both. This finding conforms to (Tizhoosh *et al*., 2008) who found that the accuracy can be measured easily when there is a repetition of phrases that are common features of a lyric poems. Overall, the Linear kernel produced the better accuracy (58.44%) for the dataset with stopwords. The findings showed that the class feature meaning of poetry can be well separable by the Linear kernel method. Furthermore, the result showed that removal of stopwords is not necessary as been normally done for typical information retrieval systems. This is due to the fact that poetry data consists of short verses and small vocabulary stopwords do have discriminating values.

Table 5: Result for Experiment 2 (structural features)

	Content			
	5-Fold		10-Fold	
Repetitions of Test	RBF	Linear	RBF	Linear
Test 1	90.91	100	90.91	100
Test 2	95.24	100	90.48	100
Test 3	96.88	100	92.19	100
Test 4	97.62	100	97.62	100
Test 5	100	100	100	100
Test 6	100	100	100	100
Average	96.77	100	95.2	100

Another interesting finding in this experiment is the comparison between poetry that contains the meaning alone and poetry with the meaning as well as the indicator. The result illustrated that the poetry with meanings alone produced the best average accuracy as indicated by the linear method for the 5-Fold validation dataset. In this case, meanings of poetry give better features for the SVM classifier.

Table 6: Result of Experiment 2 (*tfidf* features)

	Content							
	With stopwords				With stopwords			
	5-Fold		10-Fold		5-Fold		10-Fold	
Repetitions of Test	RBF	Linear	RBF	Linear	RBF	Linear	RBF	Linear
Test 1	90.91	95.45	90.91	95.45	81.82	90.91	90.91	90.91
Test 2	95.24	100	95.24	100	92.86	95.24	92.86	95.24
Test 3	96.88	100	96.88	100	93.75	79.69	93.75	79.69
Test 4	97.62	98.81	97.62	98.81	96.43	97.62	96.43	97.62
Test 5	98.15	100	98.15	100	97.22	98.15	97.22	98.15
Test 6	98.44	100	98.44	100	96.09	96.88	96.09	96.88
Average	96.2	99.04	96.2	99.04	93.03	93.08	94.54	93.08

Experiment 2: The result of the experiment is as shown in Table 5. The result demonstrated that by using the structural features and the removal of punctuation marks, the classification produced by the Linear kernel is better than RBF kernel as evidenced by the 100% classification accuracy for all the six trial tests. These findings are consistent with (Tizhoosh *et al*., 2008) that mentioned features such as length of line as an ideal feature for text classification of poems as opposed to prose. Interestingly, the number of training documents and vocabulary used did not have any significant impact on the performance of the Linear kernel classifier. Thus, it can be concluded that the classification model produced by the Linear kernel was fully optimized and the data sets are linearly well separated.

DISCUSSION

The highest average percentage of 58.44% accuracy was found somewhat less satisfactory for the classification of poetry by theme as shown in Experiment 1. This is due to the nature of poetry that contain beautiful words and, thus, has a rather limited specific words to distinguish among themes. This finding is in-line with (Tizhoosh *et al*., 2008), which stated that a poem written using any word that is desired by a poet and not limited to the keywords specified. On the other hand, the findings seem better than the Naïve Bayesian classifier applied on proverb as illustrated by the study of (Noah and Ismail, 2008) with 42% accuracy for the dataset without any background knowledge or information.

The result for classifying poetry and non-poetry based on the tfidf feature is comparable with the previous experiment as shown in Table 6. Similarly, the Linear kernel produced the best result. However, overall the structural feature is seen as a more appropriate feature for classifying poetry or non-poetry.

CONCLUSION

Malay poetry classification is a challenging and interesting research. Unfortunately, it has been largely overlooked in the past research. This study presents an experimental study on automatic classification of Malay poetry using the SVM technique. The RBF and Linear kernel functions are used for the classification task. Previous research in poetry classifications are mainly concerned with recognizing poetry from prose. This study however extends such a recognition task into a more challenging area, which is classification of poetry according to themes. In addition, applying methods to semantically identify poetry according to themes can improve the quality of poetry retrieval or even finding suitable usage of poetry according to certain contextual situations. In the case of recognizing poetry from prose, this study proposes poetic features and compares subsequently compare the effectiveness of these features.

A collection of 1500 and 428 documents were gathered from various resources for experimenting the aforementioned classification respectively. The tfidf term weighting features were used for both tests and the poetry shape structure was used for the second test. Overall, the Linear kernel function performs the best for both test with 58.44 and 100% accuracies have been achieved for thematic-classifying of poetry and recognizing of poetry from prose respectively.

The feature used for the thematic-based poetry classification in this study was simply based on the conventional statistical feature. As the content of Malay poetry was built on poetic word elements and characteristic, further research work is required to identify the suitable poetic features. Various features can be explored, integrated and even semantically linked with external resources in order to produce the best features. Exploring or combining with other classification techniques are also some of the potential research works.

REFERENCES

Ahmad, Z.A., 2002. Ilmu Mengarang Melayu. Dewan Bahasa dan Pustaka. Kuala Lumpur.

Bakar, Z. A., 1983. Kumpulan Pantun Melayu. Dewan Bahasa dan Pustaka, Kementerian Pelajaran Malaysia.

Brucher, H., G. Knolmayer and M. Mittermayer, 2002. Document classification methods for organizing explicit knowledge. Proceedings of the 3rd European Conference on Organizational Knowledge, Learning and Capabilities Athens, (OKLCA' 02), Bern, Switzerland, pp: 1-25.

Drucker, H., W. Donghui and V.N. Vapnik, 1999. Support vector machines for spam categorization. IEEE Trans. Neural Netw., 10: 1048-1054. DOI: 10.1109/72.788645

Dumais, S., J. Platt, D. Heckerman and M. Sahami, 1998. Inductive learning algorithms and representations for text categorization. Proceedings of the 7th International Conference on Information and Knowledge Management. Nov. 2-7, ACM Press, Washington, DC, USA., pp: 148-155. DOI: 10.1145/288627.288651

Fan, R.E., P.H. Chen and C.J. Lin, 2005. Working set selection using second order information for training support vector machines training. J. Mach. Learn. Res. 6: 1889-1918.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann. 2nd Edn., Morgan Kaufmann, Amsterdam, ISBN-10: 1558609016, pp: 770.

Hsu, C.W., C.C. Chang and C.J. Lin, 2010. A practical guide to support vector classification. Bioinformatics.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Mach. Learn., 1398: 137-142. DOI: 10.1007/BFb0026683

Ko, Y. and J. Seo, 2000. Automatic text categorization by unsupervised learning. Proceedings of the 18th International Conference on Computational Linguistics, (CL' 00), ACM Press, USA., pp: 453-459. DOI: 10.3115/990820.990886

Lewis, D.D., R.E. Schapire, J.P. Callan and R. Papka, 1996. Training algorithms for linear text classifiers. Proceedings of the 19th International Conference on Research and Development in Information Retrieval, Aug. 18-22, ACM Press, Zurich, Switzerland, pp. 289-297. DOI: 10.1145/243199.243277

McCallum, A. and K. Nigam, 1998. A comparison of event models for Naive Bayes text classification. Dimension Contemporary German Arts Lett.,752: 41-48.

Noah, S.A. and F. Ismail, 2008. Automatic classifications of Malay proverbs using Naïve Bayesian Algorithm. J. Inform. Technol., 7: 1016-1022.

Puteh, O. and A. Said, 1995. Himpunan 366 Cerita Rakyat Malaysia. 1st Edn., Universiti Utara, Malaysia, ISBN-10: 9676105058, pp: 286.

Taira, H. and M. Haruno, 1999. Feature selection in SVM text categorization. Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence Conference Innovative Applications of Artificial Intelligence, (AAAI '99), ACM Press, CA, USA., pp: 480-486.

Tizhoosh, H.R., F. Sahba and R.A. Dara, 2008. Poetic features for poem recognition: A comparative study. J. Patt. Recogn. Res.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. 2nd Edn., Springer-Verlag, New York, ISBN-10: 0387987800, pp: 314.

Weiss, S.M., C. Apte, F.J. Damerau, D.E. Johnson and F.J. Oles, 1999. Maximizing text-mining performance. IEEE Intell. Syst. Appl., 14: 63-69. DOI: 10.1109/5254.784086

Winstedt, R.O. and R.J. Wilkinson, 1957. Pantun Mělayu. 3rd Edn., Malaya Publishing House, Singapura, ISBN-10: 0404168809, pp: 209.

Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM Press, Berkeley, CA, USA, pp: 42-49. DOI: 10.1145/312624.312647

Yang, Y., S. Slattery and R. Ghani, 2002. A study of approaches to hypertext categorization. J. Intell. Inform. Syst., 18: 219-241. DOI: 10.1023/A:1013685612819