Automating quranic verses labeling using machine learning approach

A. Adeleke, N. Samsudin, A. Mustapha, S. Ahmad Khalid
Software Engineering Department, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received Jan 9, 2019
Revised Mar 2, 2019
Accepted Mar18, 2019

Keywords:

Classifiers
Feature selection
Holy Quran
Machine learning
Text classification

ABSTRACT

Classification of Quranic verses into predefined categories is an essential task in Quranic studies. However, in recent times, with the advancement in information technology and machine learning, several classification algorithms have been developed for the purpose of text classification tasks. Automated text classification (ATC) is a well-known technique in machine learning. It is the task of developing models that could be trained to automatically assign to each text instances a known label from a predefined state. In this paper, four conventional ML classifiers: support vector machine (SVM), naïve bayes (NB), decision trees (J48), nearest neighbor (*k*-NN), are used in classifying selected Quranic verses into three predefined class labels: faith (*iman*), worship (*ibadah*), etiquettes (*akhlak*). The Quranic data comprises of verses in chapter two (*al-Baqara*) of the holy scripture. In the results, the classifiers achieved above 80% accuracy score with naïve bayes (NB) algorithm recording the overall highest scores of 93.9% accuracy and 0.964 AUC.

Corresponding Author:

Abdullahi Adeleke,
Software Engineering Department,
Universiti Tun Hussein Onn Malaysia,
86400, Parit Raja, BatuPahat, Malaysia.
Email: hi10046@siswa.uthm.edu.my

1. INTRODUCTION

Machine learning is an important and recognized field in information technology as well as artificial intelligence. As the world advances over the years with massive technological growth, it becomes more demanding and necessary for AI systems to be able to make decisions automatically and independently within a time frame. In order to achieve this, computers need to learn without being explicitly programmed [1].

The field of machine learning (ML) focuses on the study that gives AI system the capability to improve its performance (decision making) over a time period through acquiring new knowledge and skills (learning/training), as well as its ability to reorganize the existing knowledge based on the newly acquired knowledge [2]. Thus, what clearly differentiates an intelligent AI system from other computing systems is its ability to learn and make decisions.

The basic concept of ML is typically the goal of modeling machines for critical decision making purposes. One of the most important and widely studied techniques in machine learning is classification [3], which is the problem of identifying to which set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known [4]. A frequently applied area of data classification is text (also referred to as text categorization). It is the task of automatically sorting a set of documents into categories from a predefined set [3, 5-8].

The Quranic text is an important holy book of Muslims' faithful [1, 9-10]. It is one of the most widely read and referenced resource. There are interesting features in the Quran which make automating the textual data an attractable task in ML. These features include the arrangement of words in the Quran, the grouping of words into verses, verses into chapters, chapters into *juz.*

For many years, Quranic scholars have devoted much attention and efforts in producing classical works among which are: Quran commentaries, science of *hadith* (prophetic sayings), and grammar (*Nahw*). However, in recent times, with the advancements in information technology and machine learning, automating the Quranic verses (and other related works) for the purpose of knowledge discovery becomes a necessity.

Research in automating the Quranic text has gained attention in recent times. Some of the existing works as found in literatures include: text classification applications on the Holy Quran [1, 3, 11-13]; ontology-based applications [14-17]; digitized Holy Quran applications [18-22]. Furthermore, conventional among machine learning algorithms often implemented in ML tasks include: naïve bayes (NB) [4], decision trees (J48) [23], neural networks [24], support vector machines (SVM) [25], and *k*-nearest neighbour (*k*-NN) [26].

An exhaustive review of the existing works in Quranic text classification showed the Quranic data experimented were from individual Quranic sources. However, this study opined that combining multiple related data sources such as the Quranic translation and commentary (*tafsir*) could provide more relevant information. Furthermore, most of the existing works are based on the Arabic which is the primary language of the Quran. However, study has shown that only about 15% of the world Muslims' population [27] are Arabs or Arab speaking. Thus, there is a need to extend the Quranic text classification tasks to other languages most importantly the English language which arguably is one of the most spoken languages in the world.

This paper presents the automation of Quranic verses using machine learning approach and technique. In this work, standard machine learning algorithms are applied for the labeling task. The study employed four ML classification algorithms (or classifiers). These classifiers include SVM, NB, J48, and *k*-NN algorithms.Section 2 documents the methodology employed in executing the classification task.

2. METHODS AND MATERIALS

The experimental work comprises of five phases as shown in Figure 1; data gathering, preprocessing (feature generation and selection), classification, and output/result. The input data are Quranic verses extracted from the combined sources of Holy Quran translation and *tafsir*. The study identified the significance of combining multiple related Quranic sources [1, 11] for better understanding of the input verses.

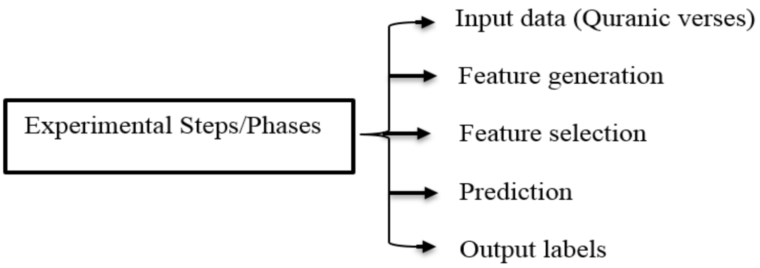

Figure 1. Experimental steps

2.1. Data Gathering

The experimental datasets (*QTrans*, *QTaf*, *QTrans+Taf*) comprise of 286 instances of Quranic data. The Quranic text are the words of Allah in surah Baqara (The Cow) of the holy book. The scripture comprises of 114 chapters (varying in size and order of revelation) in its entirety. Al-Baqara is the longest chapter (also called *surah*) in the Quran revealed in madinah, with a sum total of 286 verses. The input verses are grouped into one of three predefined labels: *faith*, *worship*, and *etiquettes*. These class labels are from the most fundamental aspects of Islam [1, 3].

2.2. Text Preprocessing

Preprocessing is an important step employed when classifying textual data [1, 4]. The step includes feature generation, transformation, and data cleansing. Firstly, features are extracted from the Quranic sources using standard String to Word Vector filter tool [1]. *TF-IDF* weighting method is further applied to access and measure the degree of relevance of the extracted features. Term frequency $Tf(t,d)$ as shown in equation 1 is a method used in knowing the frequency of words in a document$d$ [11].

$$Tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{Maximum\ Occurrences\ of\ words} \tag{1}$$

In addition, inverse-document frequency (IDF) is a method that helps to evaluate how relevant a word to the document. It is given as:

$$idf(t,D) = log\frac{N}{|\{d \in D: t \in d\}|} \tag{2}$$

The combination of *TF-IDF* is given as:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3}$$

2.2.1 Feature Selection

The generated features often come with the problem of high dimensionality. The curse of dimensionality is a known problem usually associated with textual data [11]. High dimensional data usually influence negatively the classifiers' decisions resulting in lower classification accuracy [1]. Dimensionality reduction methods such as feature selection are mostly employed to reduce curse of dimensionality. There are two possible ways to feature selection: the ranking features approach and subset selection approach [1].

The ranking features approach ranks features according to a certain criterion of the feature selection algorithms and the top k features are selected while on the other hand, the subset selection approach selects a minimum subset of features without learning performance deterioration [1].

In this work, information gain (IG) and chisquare (CH) FS algorithms are employed for the dimensionality reduction purpose. InfoGain is one of the most widely applied feature selection algorithm [1]. The filter-based algorithm measures the inter-dependency between features and labels [1]. Mathematically, IG is given as:

$$I(X:Y) = H(X) - H(X|Y) \tag{4}$$

Chisquare filter FS algorithm is used as a test of independence to access the independence of the class label of a particular feature [1]. Given a feature with r different values and c classes, chisquare feature score can be defined as:

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}} \tag{5}$$

2.3. Data Classification

The ultimate aim of this study is to automate the labeling of Quranic verses using machine learning method. To achieve this, four ML classifiers: SVM, NB, J48, and k-NN, will be implemented using the conventional 10-fold cross validation method. The classifiers are trained to predict/classify Quranic text instances into predefined labels.

Support vector machines (SVM) algorithm is typically used for learning classification, regression, or ranking function. The algorithm works by searching a seperating hyperplane to seperate between samples with a maximal margin [1]. The equation for hyperplane is:

$$w^T x + b = 0 \tag{6}$$

Naïve bayes (NB) classifier is a simple probabilistic model based on the bayes rule [1]. Given a class C, the probabilty of a particular document d to belong to C is given as:

$$P(C_i|\ d) = \frac{P(d|\ C_i) * P(C_i)}{P(d)} \tag{7}$$

The decision tree (J48) classifer is a simple representation for classifying data samples as shown in equation 8. Structurally, the algorithm functions like a tree where each internal node [28-29] is labeled with an input feature x.

$$(x, Y) = (x_1, x_2, x_3, \dots, Y) \qquad (8)$$

where vector x composed of input features while variable Y represent the target variable to be classified.

The nearest neighbor algorithm is one of the most widely applied classifiers in pattern recognition [30]. The algorithm (also known as lazy learning) predict instances by measuring the distances between sample points using the famous Euclidean distance formula as shown in equation 9.

$$d(x, x_i) = \sqrt{\sum_{i=1}^{n}(x_j - x_{ij})^{\wedge}{}_{2}} \qquad (9)$$

2.4. Evaluation Metrics

Three of the most conventional metrics [1, 3, 11] are used in evaluating the performance of the ML classifiers. These include: accuracy, AUC, and ROC curve. Combining these metrics provide a more accurate and balance performance evaluation [1]. Given a confusion matrix, accuracy is obtained using:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (10)$$

where TP is True Positive (instances correctly classified as Positive), TN is True Negative (instances correctly classified as Negative), FP is False Positive (instances incorrectly classified as Positive), and FN is False Negative (instances incorrectly classified as Negative).

3. EXPERIMENTAL RESULTS AND ANALYSIS

Implementation was carried out using four conventional machine learning classification algorithms together with information gain and chisquare feature selection algorithms. The experimental results obtained were evaluated and compared in terms of classification accuracy (ACC) and AUC. Furthermore, ROC curve metric was used in visualizing the classifiers' performance. Tables 1 to 3 respectively shows the classification results of the classifiers. The classifiers were implemented in WEKA using all generated features (without feature selection) as well as with feature selection.

Table 1. Classification Performance in Terms of Accuracy (ACC) and AUC (without Feature Selection)

ML	QTrans		QTaf		QTrans+Taf	
Classifiers	ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC
SVM	86.2	0.76	88.1	0.748	88.6	0.754
NB	87.4	0.884	89.7	0.904	90.7	0.925
J48	79.3	0.601	83.2	0.7	82.5	0.679
k-NN	81.1	0.679	81.4	0.499	83.3	0.519

Table 2. Classification Performance in Terms of Accuracy (ACC) and AUC (with Infogain FS Algorithm)

ML	QTrans		QTaf		QTrans+Taf	
Classifiers	ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC
SVM	88.4	0.793	91.4	0.859	90.2	0.832
NB	90.7	0.936	91.8	0.96	93.9	0.964
J48	84.1	0.757	85.1	0.75	83.7	0.677
k-NN	83.2	0.72	88.3	0.751	86.9	0.774

Table 3. Classification Performance in Terms of Accuracy (ACC) and AUC (with Chisquare FS Algorithm)

ML	QTrans		QTaf		QTrans+Taf	
Classifiers	ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC
SVM	88.6	0.765	91.4	0.859	90.2	0.832
NB	90.4	0.935	91.8	0.96	93.9	0.964
J48	84.4	0.769	84.6	0.75	83.7	0.689
k-NN	83.2	0.738	88.3	0.751	86.9	0.774

From the classification results, the machine learning algorithms consistently achieved above 80% accuracy performance across all experimental datasets. However, an exemption to this is the decision tree (J48) algorithm which achieve the least accuracy score of 79.3% with the *QTrans* dataset. This could be as a result of high dimensionality of the features set. As previously noted, the curse of dimensionality associated with text data influence the decisions of the classifiers. The ML classifiers had promising results with the feature selection algorithms. This again established the significance of feature selection process in data classification.

Exceptional among the classification algorithms is the naïve bayes (NB) classifier which consistently achieved the best classification results with the feature selection algorithms. The classifier achieved the overall highest classification result of 93.9% and AUC value of 0.964 with the *QTaf* and *QTrans+Taf* datasets. Nearest neighbour (*k*-NN) classifier achieved the least AUC score of 0.499. Again, this maybe as result of the high dimensionality of the features set.In addition, classifiers are sensitive to the nature of the experimental data. This probably could be the reason why varying classification results were obtained in the experimental work.

Furthermore, the classifiers' performance was plotted for better visualization using the receiver operating characteristics (ROC) curve evaluation metric. The ROC curves of the classification results with *QTrans+Taf* dataset are shown in Figures 2 and 3.

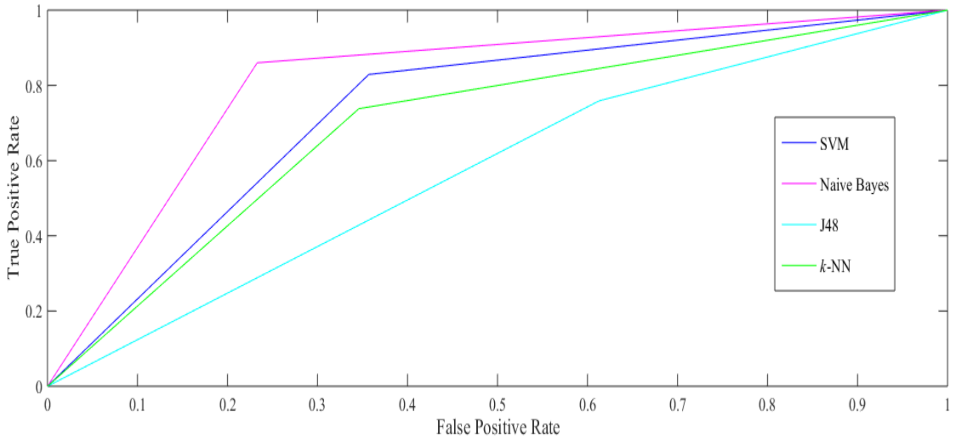

Figure 2. ROC curve of the ML classifiers using all features (without feature selection)

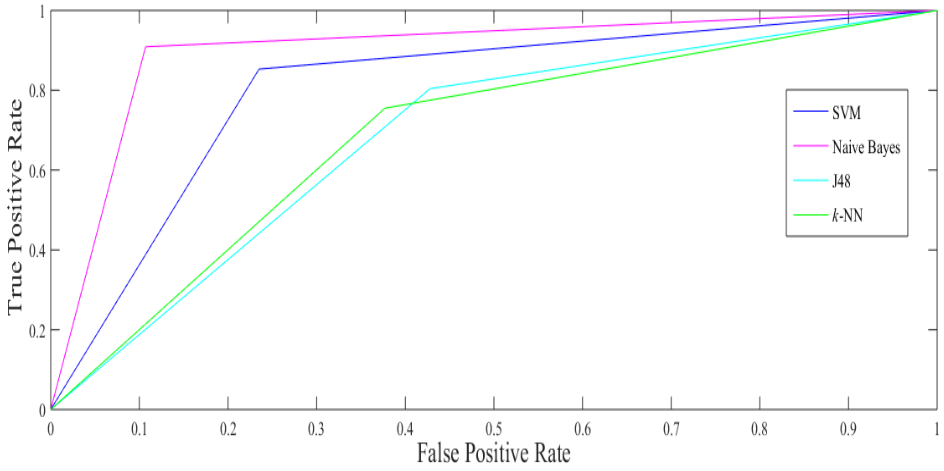

Figure 3. ROC curve of the ML classifiers with InfoGain FS

4. CONCLUSION

The classification of Quranic verses into predefined categories is an essential task in Quranic studies. In this paper, we presented an automated machine learning approach for classifying the input Quranic verses. To achieve this purpose, we employed four conventional machine learning algorithms: SVM, NB, J48, and *k*-NN.

Features were generated from the Quranic textual data using standard machine learning techniques. Furthermore, InfoGain and chisquare FS methods were used to preprocess the input data in order to reduce the curse of dimensionality. The preprocessed textual data along with the label information were used in training the classifiers for the purpose of the labeling task. Constantly, throughout the experimentation, the conventional 10-fold cross validation method was employed.

Finally, the classifiers' performances were evaluated and compared. Consistently, the classifiers achieved above 80% accuracy score except for J48 algorithm which obtained the least accuracy score of 79.3% with the *QTrans* dataset. Naïve bayes (NB) classification algorithm achieved the overall highest accuracy result of 93.9% and AUC value of 0.964 while *k*-NN classifier obtained the least AUC value of 0.499. The research study further hopes to explore several other classification application domains.

ACKNOWLEDGEMENTS

This research study was supported by a grant from Universiti Tun Hussein Onn Malaysia (UTHM) Vot U611.

REFERENCES

[1] A. O. Adeleke, *et al*., "A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses," in *R. Ghazali et al. (eds.), Recent Advances on Soft Computing and Data Mining, Advances in Intelligent Systems and Computing 700*, pp. 282-297, 2018.

[2] A. Talwar and Y. Kumar, "Machine Learning: An Artificial Intelligence Methodology," *J. of Engineering and Computer Science*,vol. 2, pp. 3400-3404, 2013.

[3] A. O. Adeleke, *et al*., "Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses," *Int. J. on Advance Science, Engineering and Info. Tech*,vol. 7, pp. 1419-1427, 2017.

[4] J. Tang, *et al*., "Feature Selection for Classification: A Review," in *Data Classification: Algorithms and Applications. CRC Press*, 2014.

[5] A. Faraz, "An Elaboration of Text Categorization and Automated Text Classification Through Mathematical and Graphical Modelling," *Computer Science & Engineering: An International J*, vol.5, pp. 1-11, 2015.

[6] S. Bhumika, *et al*., "A Review Paper on Algorithms used for Text Classification," *International Journal of Application or Innovation in Engineering & Management*,vol. 2, pp. 90-99, 2013.

[7] T. H. Nguyen and K. Shirai, "*Text Classification of Technical Papers Based on Text Segmentation*," 18th International Conference on Applications of Natural Language to Information Systems, 2013.

[8] M. K. Dalal and M. A. Zaveri, "Automatic Text Classification: A Technical Review," *International Journal of Computer Applications*,vol. 28, pp. 37-40, 2011.

[9] A. Hilal and N. Srinivas, "Analytical of the Initial Holy Quran Letters Based on Data Mining study," *American International Journal of Research in Formal, Applied & Natural Sciences*,vol. 10, pp. 1-8, 2015.

[10] M. Alhawarat M, "Extracting Topics from the Holy Quran using Generative Models,"*International Journal of Advanced Computer Science and Applications*,vol. 6, pp. 288-294, 2015.

[11] A. Adeleke and N. Samsudin, "A Hybrid Feature Selection Technique for Classification of Group-based Holy Quran Verses," *International J of Engineering & Technology*, vol. 7, pp. 228-233, 2018.

[12] M. Goudjil, *et al*., "*Using Active Learning in Text Classification of Quranic Sciences*," Int. Conf. on Advances in Information Technology for the Holy Quran and Its Sciences, pp. 209-213, 2015.

[13] G.S. Hassan, *et al*., "Categorization of Holy Quran Tafseer' using k-Nearest Neighbour Algorithm," *Int. J. of Computer Applications*vol. 129, pp. 1-6, 2015.

[14] E. A. Ibrahim, *et al*., "Provisions of Quran Tajweed Ontology (Articulations Points of Letters, UN Vowel Noon and Tanween)," *Int. J. of Science and Research* vol. 6, pp. 756-761, 2017.

[15] M. Alqahtani and E. Atwell, "*Arabic Quranic Search Tool Based on Ontology*," 21st Int. Conf. on Applications of Natural Language to Information Systems,pp. 478-485, 2016.

[16] S. M. Alrehaili and E. Atwell, "*Computational Ontologies for Semantic tagging of the Quran: A survey of past approaches*," Ninth Int. Conf. on Language Resources and Evaluation, 2014.

[17] Y. Abdelhamid, *et al*., "*Using Ontology for Associating Web Multimedia Resources with the Holy Quran*,"Taibah University Int. Conf. on Advances in Information Technology for the Holy Quran and its Sciences, pp. 266-271, 2013.

[18] S. K. Hamed and M. J. Ab Aziz, "A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification," *J. of Computer Sciences*, vol. 12, pp. 169-177, 2016.

[19] H. Abdelnasser, *et al*.,"*Al-Bayan: An Arabic Question Answering System for the Holy*," Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing,pp. 57-64, 2014.

[20] A. N. Akkila and S. S. Abu Naser, "Teaching the right letter pronunciation in reciting the holy Quran using intelligent tutoring system,"*Int. J. of Advanced Research and Development*, vol. 2, pp. 64-68, 2017.

[21] A. H. Ahmed and S. M. Abdo, "Verification System of Quran Recitation Recordings,"*Int. J. of Computer Applications*, vol. 163, pp. 6-11, 2017.

[22] H. O. Aljaloud, *et al*., "Stemmer Impact on Quranic Mobile Information Retrieval Performance,"*Int. J. of Advanced Computer Science and* Applications, vol. 7, pp. 135-139, 2016.
[23] A. S. Zharmagambetov and A. A. Pak, "*Sentiment analysis of document using deep learning and decision trees*,"Twelve IEEE Int. Conf. on Electronics Computer and Computation, pp. 1-4, 2015.
[24] J. H. Wang and H. Y. Wang, "Incremental Neural Network Construction for Text Classification,"*IEEE Int. Symposium on Computer Consumer and Control*, pp. 970-973, 2014.
[25] T. Sabbah and A. Selamat, "*Support Vector Machine based approach for Quranic words detection in online textual content*,"8th IEEE Malaysian Software Engineering Conference, Malaysia, pp. 325-330, 2014.
[26] K. R. Townsend, *et al*., "*k-NN text classification using an FPGA-based sparse matrix vector multiplication accelerator*,"IEEE Int. Conf. on Electro/Information Technology, pp. 257-263, 2015.
[27] K. Houssain, "The World Muslim Population, History & Prospect," Research Publishing, 2014.
[28] P. Sewaiwar and K. K. Verma, "Comparative Study of various Decision Tree Classification Algorithm using WEKA,"International J of Emerging Research in Management &Technology, vol. 4, pp. 87-91, 2015.
[29] S. Teli and P. Kanikar, "A Survey on Decision Tree Based Approaches in Data Mining,"International J of Advanced Research in Computer Science and Software Engineering, vol. 5, pp. 613-617, 2015.
[30] F. S. Gharehchopogh, et al., "A New Approach in Bloggers Classification with Hybrid of k-Nearest Neighbor and Artificial Neural Network Algorithms,"Indian J of Science and Technology, vol.8, pp. 237-246, 2015.

BIOGRAPHIES OF AUTHORS

AbdullahiAdeleke is a Ph.D. student at Faculty of Computer Science and Information Technology, UniversitiTun Hussein Onn Malaysia (UTHM) since March 2018. He had his master degree in Information Technology (MIT) at UTHM. His research includes machine learning, data mining, document (text) classification, and feature selection.

Noor A. Samsudin is a senior lecturer at Faculty of Computer Science and Information Technology, UniversitiTun Hussein Onn Malaysia (UTHM) since March 2004. She received her bachelor degree in Computer Science from University Missouri-Columbia in 1996. Then, she accomplished her master degree at National University of Malaysia. She received his Phd from The University of Queensland, Australia. Her research interest includes machine learning, data mining and ICT applications in education.

Aida Mustapha received the B.Sc. degree in Computer Science from Michigan Technological University and the M.IT degree in Computer Science from UKM, Malaysia in 1998 and 2004, respectively. She received her Ph.D. in Artificial Intelligence focusing on dialogue systems. She is currently an active researcher in the area of Computational Linguistics, Soft Computing, Data Mining, and Agent-based Systems

Shamsul Kamal Ahmad Khalid is a senior lecturer at Faculty of Computer Science and Information Technology, UniversitiTun Hussein Onn Malaysia (UTHM) since March 2004. He received his bachelor degree in Computer Science from New York University in 1995. Then, he accomplished his master degree at National University of Malaysia. He received his Phd from UniversitiTun Hussein Onn Malaysia. His research interest includes information security, watermarking, steganography, and network security.