Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 2, November 2019, pp. 730~736
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i2.pp730-736

A two-step feature selection method for quranic text classification

A. Adeleke[1], N. A. Samsudin[2], Z. A. Othman[3], S. K. Ahmad Khalid[4]
[1,2,4]Software Engineering Department, Universiti Tun Hussein Onn Malaysia, Malaysia
[3]Faculty of Information Science and Technology, National University of Malaysia, Malaysia

Article Info

Article history:

Received Jan 25, 2019
Revised Apr 17, 2019
Accepted May 18, 2019

Keywords:

Classifier
Feature selection
Holy Quran
Text classification

ABSTRACT

Feature selection is an integral phase in text classification problems. It is primarily applied in preprocessing text data prior to labeling. However, there exist some limitations with the FS techniques. The filter-based FS techniques have the drawback of lower accuracy performance while the wrapper-based techniques are highly computationally expensive to process. In this paper, a two-step FS method is presented. In the first step, chisquare (CH) filter-based technique is used to reduce the dimensionality of the feature set and then wrapper correlation-based (CFS) technique is employed in the second step to further select most relevant features from the reduced feature set. Specifically, the ultimate aim is to reduce the computational runtime while achieving high classification accuracy. Subsequently, the proposed method was applied in labeling instances of the input data (Quranic verses) using standard classifiers: naïve bayes (NB), support vector machine (SVM), decision trees (J48). The results report the proposed method achieved accuracy result of 93.6% at 4.17secs.

Corresponding Author:

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

1. INTRODUCTION

With the advancements in information technology, the amount of documents being processed over the years have continually increased. This has made the fields of artificial intelligence (AI) and machine learning (ML) attract attention and unceasing developments. The increasingly demand for processing large documents within a short time necessitate the automating of documents processing. Automated text classification (ATC) [1] is the steps and techniques involve in automatically classifying textual data to predefined class/label. To do this, there is a need for machines to learn [2]. In the field of machine learning, the goal is to develop models that give computing machines the capability of learning [2]. Thereafter translates the acquired knowledge into decision making.

An essential phase in text classification task is feature selection; a dimensionality reduction method that helps to reduce the complexity of dimensionality usually associated with text [2]. In text, there is presence of large feature space which often results in high level of dimensionality. This often occurs when the feature set consists of both relevant, irrelevant, as well as redundant features (or attributes) which could lead to misclassification by the classifiers (overfitting) [2]. Thus, to ensure the optimization of the classification algorithms' performance, feature selection methods and techniques can be applied in preprocessing the textual data prior to the actual labeling tasks. Quran is a unique text and a good source of information with about 78,000 words. These words are very rich in text and systematically arranged by experts into various sections, groups [1]. With such great importance and characteristics, features

(or keywords) could be extracted from the divine text in order to help improve the literacy level of its readers and researchers. In addition, there are multiple sources of the Quranic text available from the extensive academic works of the religious scholars. From such sources are the Quran translations (translated from Arabic to almost all languages), as well as the Quran commentaries.

However, experimenting these sources independently has its setbacks. For example, the Holy Quran translation as a source may not be sufficient for the purpose of analyzing Quranic verses for the labeling task. Thus, there is a need to combine the sources (otherwise termed group-based) while extracting features for the classification task. The field of Holy Quran study has witnessed a quite number of research works. These include: text classification applications of the Holy Quran [1-6]; ontology-based applications [7-10]; digitized Holy Quran applications [11-14]. Furthermore, from among the techniques that have been widely applied to text classification problems, include the Bayes probabilistic approach [15], decision trees [16], neural networks [17], support vector machines [18], and *k*-nearest neighbor [19]. In addition, the research work in [20] is based on classifying sonar targets using information gain FS algorithm for attribute evaluation. The experimental work focused on training networks for the purpose of discriminating between the sonar signals.

The experimental results showed IG attribute evaluation significantly improved the classification task. A Genetic algorithm wrapper-based feature selection method was proposed in [21] for classifying hyperspectral images using SVM classifier. The feature selection process involves three steps: creating the training and testing sets using ENVI software; setting up required parameters; running the model. The FS algorithm was used to optimize the kernel parameters and feature subsets. [22] introduced in their work a feature selection method using support vector machine to find dependency between the attributes of high dimensional data extracted from UCI data repository and then decide the appropriate class attributes values. The results showed that the FS method had promising results. In addition, from other research works in feature selection include *but not limited* to [23-24]. Due to the limitations found with the available FS techniques such as high computational cost (as associated with wrapper-based FS techniques) and lower accuracy performance (as associated with filter-based techniques), the study proposes a two-step FS method. The study aimed at reducing the computational complexity while achieving high classification accuracy results.

Hybrid approach to feature selection has been successfully experimented in classification problems [25-30]. The proposed method is a combination of chisquare (*CH*) filter-based and wrapper-based *CFS* algorithms. The two-step *CH-CFS* method will be applied in labeling the verses of the Quranic datasets using naïve bayes (NB), support vector machine (SVM), and decision trees (J48) classification algorithms. The input verses are classified into three predefined labels: '*iman*, *ibadah*, and *akhlak*'. These class labels are from the most fundamental aspects of Islam [1, 3].

2. METHODS AND MATERIALS

The experimental design as shown in Figure 1 consists of five steps. The input data are Quranic verses gathered from the combined sources of Holy Quran translation and tafsir. The resulting combined text data is otherwise termed 'grouped-data'. The experimental phases include: data gathering, feature generation, feature selection, classification, and output results.

2.1. Data Gathering

The experimental datasets as tabulated in Table 1 comprises of 451 instances (quranic verses) made up of 286 verses from chapter two (Surah al-Baqarah) and 165 verses from chapter six (Surah al-Anaam) of the Holy Quran. As could be seen from the class weight distribution, the '*iman*' class has the most class members (input verses).

Table 1. Percentage Composition of Class Labels

Datasets	No of Instances	Class Weight		
		Iman	Ibadah	Akhlak
QTrans	451	343.0	44.0	64.0
QTaf	451	345.0	42.0	64.0
QTrans+Taf	451	345.0	42.0	64.0

2.2. Feature Generation

Features are first extracted from the Quranic texts. To do this, the study employed standard StringToWordVector filter tool [1]. Furthermore, TF-IDF weighting method is applied to access and measure the degree of relevance of the extracted features. Term frequency $Tf(t,d)$ as given in (1) is an important

preprocessing method used in measuring the frequencies of words in textual data d [1-2].

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

where $Maximum\ Occurrences\ of\ words$ is denoted with $Max\ \{ft^{|}, d: t^{|} \in d\}$.

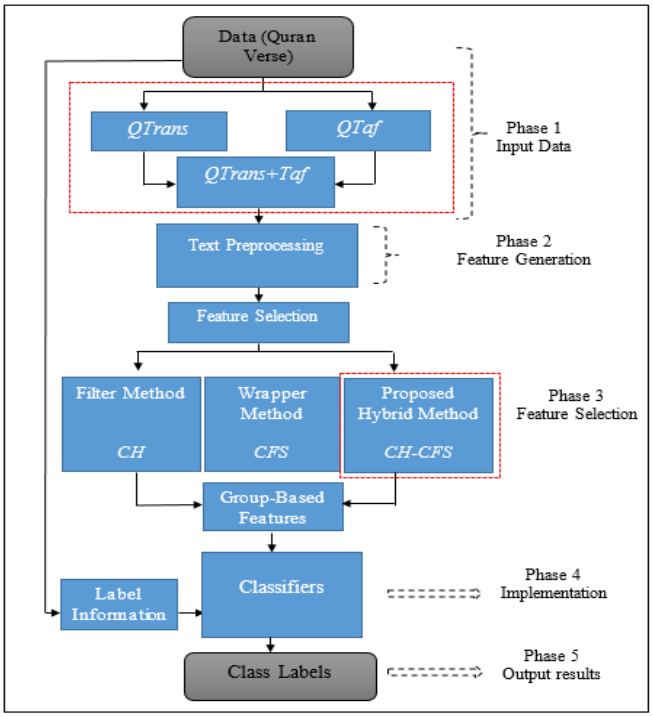

Figure 1. Proposed framework

Also, inverse-document frequency (IDF) helps to measure the relevancy of a given word. The method is given as:

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

Generally, *TF-IDF* is given as:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \quad (3)$$

2.3. Feature Selection

Dimensionality reduction methods such as feature selection are mostly employed to reduce the curse of dimensionality. This problem is often associated with textual data. The presence of high dimensionality may influence negatively classifier's decision making. Feature selection method can be applied either by ranking the features/attributes or through subset selection approach [2]. The ranking features approach ranks features according to a certain criterion of the feature selection algorithms with the top *k* features selected. On the other hand, the subset selection approach selects a minimum subset of features without learning performance deterioration [2].

In this paper, the ranking features approach was experimented using chisquare (CH) filter-based algorithm while for the *CFS* wrapper-based algorithm was applied for the subset selection approach. Filter method is less computationally expensive in comparison with the wrapper method. The filter method selects features independent of the classifiers. This makes the method simple, fast, and less expensive to run. On the other hand, wrapper method utilizes the performance of the classification algorithms to evaluate and select the feature subsets. This makes wrapper FS method perform better but with high computational cost.

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

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (4)$$

Where n_ij is the number of samples with the i^th feature value. CFS wrapper-based method finds the subsets of features that are individually highly correlated with the class but have low inter-correlation [2]. The method can be calculated using:

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k+k(k-1)\overline{r_{ii}}}} \quad (5)$$

The experimental workflow of the proposed methodology follows the following steps:

Step 1: Input data (quranic verse)
Step 2: Generate features from input data using StringToWordVector and TF-IDF
Step 3: Feature preprocessing using *CH* feature selection algorithm
Step 4: Implement features from (step 3) with the classifiers
Step 5: Select features from (step 2) using *CFS* algorithm
Step 6: Implement selected features from (step 5) with the classifiers
Step 7: Load the generated features from (step 2)
Step 8: Iterate (step 3)
Step 9: Apply *CFS* algorithm on selected features from (step 8)
Step 10: Implement resulting features from *CH-CFS* with the classifiers
Step 11: Evaluate results

2.4. Classification (Labeling)

For the labeling task, the experimental work implemented three of the conventional classification algorithms: NB, SVM, and J48 classifiers. These algorithms are widely applied to several classification problems [2]. For data partitioning, we employed the standard 10-fold cross validation method. NB classifier is a simple probabilistic model based on the Bayes rule [2]. Given a class C, the probabilty of a particular document d to belong to C is given as:

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

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 [2]. The equation for hyperplane is:

$$w^T x + b = 0 \quad (7)$$

In decision tree classification algorithm, each node specifies a test to be performed on a single attribute [1]. The goal is to create a model that predicts the value of a target variable based on several input variables. The data generally takes the form:

$$(x, Y) = (x_1, x_2, x_3, \cdots, x_k, Y) \quad (8)$$

Finally, the study validated the experimental results using accuracy performance. Classification accuracy is one of the widely used performance metrics in text classification problems [2]. Given a confusion matrix, the accuracy metric is calculated as:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

The FS algorithms experimented on the Quranic datasets produced mixed results as shown in Tables 2-4. The datasets include: *QTrans, QTaf, and QTrans+Taf.* The classifiers were implemented using

the entire generated features from the Quranic datasets as well with the selected features using *CH*, *CFS*, and the proposed *CH-CFS* algorithms. The classification results obtained established the significant influence of feature selection process in text classification tasks.

Table 2. Classification Accuracy using NB Classifier

FS Algorithm	QTrans		QTaf		QTrans+Taf	
	Time	ACC (%)	Time	ACC (%)	Time	ACC (%)
All Features	-	83.7	-	87.4	-	77.5
CH	1.43s	91	1.63s	89.1	1.53s	92.9
CFS	135s	90.5	97.6s	88.3	119.6s	92.9
CH-CFS	2.8s	90.5	3.75s	88.3	4.17s	90.2

Table 3. Classification Accuracy using SVM Classifier

FS Algorithm	QTrans		QTaf		QTrans+Taf	
	Time	ACC (%)	Time	ACC (%)	Time	ACC (%)
All Features	-	87.4	-	90.4	-	76.1
CH	1.43s	89.8	1.63s	92.3	1.53s	92.8
CFS	135s	90.5	97.6s	92.8	119.6s	93.6
CH-CFS	2.8s	90.5	3.75s	92.8	4.17s	93.6

Table 4. Classification Accuracy using J48 Classifier

FS Algorithm	QTrans		QTaf		QTrans+Taf	
	Time	ACC (%)	Time	ACC (%)	Time	ACC (%)
All Features	-	82.3	-	85.7	-	64.3
CH	1.43s	85.7	1.63s	86.7	1.53s	89.5
CFS	135s	85.2	97.6s	86.7	119.6s	87.3
CH-CFS	2.8s	85.1	3.75s	86.7	4.17s	87.1

Working with the entire generated features (without feature selection) could affect the classifiers' performance as shown in the experimental results. Applying all the features directly produced the least accuracy result of 64.3% with J48 algorithm on the group-based *QTrans+Taf* dataset. The result was obtained as a result of high dimensionality in the text data. To optimize the classification results, feature selection techniques were applied. The curse of dimensionality could be solved by applying FS algorithms on the text data prior to classification. Consistently, the feature selection algorithms obtained above 80% accuracy results.

The filter-based *CH* and the wrapper-based *CFS* algorithms obtained with SVM classifier 92.8% at 1.53secs and 93.6% at 119.6secs accuracy results respectively. Analysis of these results showed the limitations earlier identified with the existing FS methods. techniques are less efficient and relatively achieve lower accuracy results. However, the wrapper-based techniques are computationally expensive to work on as seen in the results. It took *CFS* algorithm a high computational runtime of 119.6secs to select features from the feature set. Consequently, the proposed *CH-CFS* algorithm achieved the overall highest accuracy performance of 93.6% at a very less computational runtime of 4.17secs with SVM classifier.

4. CONCLUSION

Feature selection process as experimented in this study has proven to be an integral phase in text classification tasks. The study identified some limitations with the existing FS techniques. To address these setbacks, the study proposed a hybridized FS method. The proposed method is a two-step combination of filter-based chisquare and wrapper-based CFS algorithms.

The specific goal of the study is to apply FS algorithms in automating the labeling of Quranic verses. The set target is to achieve with the proposed *CH-CFS* higher classification accuracy performance at lower computational runtime. The proposed technique achieved the overall accuracy result of 93.6% at 4.17secs in comparison with the wrapper-based CFS algorithm which achieved the same accuracy result but at a high computational runtime of 119.6secs. In future work, the study will focus on extending the proposed hybrid *CH-CFS* algorithm to other classification problems.

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*., "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.
[2] 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.
[3] N. S. Jamil, *et al*., "A subject identification method based on term frequency technique," *J. of Advanced Computer Research*, vol. 7, pp. 103-110, 2017.
[4] 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.
[5] 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.
[6] E. A. 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] A. H. Ahmed and S. M. Abdo, "Verification System of Quran Recitation Recordings," *Int. J. of Computer* Applications, vol. 163, pp. 6-11, 2017.
[14] 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.
[15] J. Tang, *et al*., "Feature Selection for Classification: A Review," in *Data Classification: Algorithms and Applications. CRC Press*, 2014.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] J. Novakovic, "Using Information Gain Attribute Evaluation to classify Sonar Targets," *17th Telecommunications Forum*, pp. 1351-1354, 2009.
[21] L. Zhuo, *et al*., "A Genetic Algorithm based Wrapper Feature Selection method for Classification of Hyperspectral Images using Support Vector Machine," *The Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVII, pp. 397-402, 2008.
[22] A. Veeraswamy and S. A. Balamurugan, "*An Effective Performance of Feature selection with Classification of Data mining using SVM Algorithm*," In Proceedings of the National Conf. on Recent Trends in Mathematical Computing, pp. 427-431, 2013.
[23] T. K. Mansoori, *et al*., "Feature selection by Genetic Algorithm and SVM Classification for Cancer Detection," *Int. J of Advanced Research in Computer Sci and Software Eng*, vol. 4, pp. 357-365, 2014.
[24] V. Molano, *et al*., "Feature selection based on sampling and C4.5 Algorithm to improve the Quality of Text Classification using Naïve Bayes," *Springer*, 2011.
[25] M. Aladeemy, *et al*., "A new hybrid approach for feature selection and support vector machine model selection based on self-adaptive cohort intelligence," *Expert Systems with Applications*, vol. 88, pp. 118-131, 2017.
[26] H. Wang and S. Liu, "An Effective Feature Selection Approach Using the Hybrid Filter Wrapper," *Int. J. of Hybrid Information Technology*, vol. 9, pp. 119-128, 2016.
[27] A. K. Uysal, "An improved global feature selection scheme for text classification," *Expert Systems with Applications*, vol. 43, pp. 82-92, 2016.

[28] A. S. Ghareb, *et al*., "Hybrid feature selection based on enhanced genetic algorithm for text categorization," *Expert Systems with Applications*, vol. 49, pp. 31-47, 2016.
[29] 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.
[30] H. Hui-Huang, *et al*., "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications*, vol. 38, pp. 8144-8150, 2011.

BIOGRAPHIES OF AUTHORS

Abdullahi Adeleke is a Ph.D. student at Faculty of Computer Science and Information Technology, Universiti Tun 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, Universiti Tun 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.

Zulaiha Ali Othman is currently an Associate Professor at National University of Malaysia. Her research interest in on artificial intelligence, Big Data, and optimisation algorithms in various problem domain including network intrusion, human talent, climate change and pollution.

Shamsul Kamal Ahmad Khalid is a senior lecturer at Faculty of Computer Science and Information Technology, Universiti Tun 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 Universiti Tun Hussein Onn Malaysia. His research interest includes information security, watermarking, steganography, and network security.