Indonesian Journal of Electrical Engineering and Computer Science
Vol. 12, No. 1, October 2018, pp. 406~411
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v12.i1.pp406-411

Comparative Analysis for Topic Classification in Juz Al-Baqarah

Mohamad Izzuddin Rahman, Noor Azah Samsudin, Aida Mustapha, Adeleke Abdullahi
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia

Article Info

Article history:

Received May 1, 2018
Revised Jul 10, 2018
Accepted Jul 25, 2018

Keywords:

Quran Verse Classification
Text Mining

ABSTRACT

In Islam, Quran is the holy book that was revealed to the Prophet Muhammad. It functions as complete code of life for the Muslims. Remarks from Allah which contains more than 77,000 words that was passed down through Prophet Muhammad to the mankind for 23 years started in 610 ce. The Quran was divided into 114 chapters. Arabic language is the original text. The need for the Muslims across the world to find the meaning to understand the content in the Quran is necessary. Nevertheless, understanding the Quran is an interest for the Muslims as well as the attention of millions of people from the faiths. Following the generation, lots of content that related to the Quran has been broadcast by Muslims scholars in the way of the tafsirs, translation and the book of hadiths. Problem has happened at current is most Muslim in Malaysia do not understand sentences in the Quran due to language barrier. The purpose of this research is classified topic in each verses of the Quran sentence based on its specific theme. It involves the objective of text mining which are based on linguistic information and domain. The usage of corpus helps to perform various data mining tasks including information extraction, text categorization, the relationship of concepts, association discovery, the evaluation of pattern and assessed. This research project is aiming to create computing environment that enable us use to text mining the Quran. The classification experiment is using the Support Vector Machine to find themes in Juz' Baqarah. The SVM performance is then compared against other classification algorithms such as Naive Bayes, J48 Decision Tree and K-Nearest Neighbours. This research project aims at creating an enabling computational environment for text mining the Qur'an and to facilitate users to understand every verse in Juz' Baqarah.

Corresponding Author:

Noor Azah Samsudin,
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
Email: azah@uthm.edu.my

1. INTRODUCTION

Al-Quran [1] was passed down for 23 years in 610 ce, when Prophet Muhammad was at the age of 40. First sentence was passed down in Mecca city. Prophet Muhammad begin to spread these holy verses secretly and then openly. Al-Quran theme of holy verses is emphasis in doctrine of monotheism which only worship to Allah and fight against polytheistic belief in Mecca people. Therefore, many among Mecca people at that time opposing the new religion and gradually his follower increase. Prophet Muhammad mission continued further for 13 years in Mecca.

After Prophet Muhammad dies, Al-Quran, Al-Sunnah and old book Islamic has become the source for the Islam follower to make source of knowledge, wisdom and law. The challenge for Computer Science is way to extract and represent knowledge, wisdom and law in computer system which are the smart system that can answer the various questions based on knowledge from Al Quran and also Al Sunnah. This application will help the society either the Muslim or not, to understand and grow the love for the religion of Islam. Technically, this project is to understand how the new technique in mining text can extract Islam knowledge from the source and represent this knowledge. Can help society, both Muslim and non-Muslim, to understand and appreciate the Islamic religion.

We believed that Al-Quran be the only holy document in history to preserve its content, structure, words, points, and accents conservatively. Muslims across the globe, regardless of their native language, should recite the Quran in Arabic, including during prayers. This aspect is extremely important when analyzing the Quran, as the set of words in Quran is fixed and widely memorized. Each verse in the chapters of the Quran has a significant meaning. Hence, the massive number of Quran interpretation books and studies. All these features make Al-Quran as tempting target to computes the text addition which is to aim to look for new information from Al-Quran in terms of relationship, pattern, coincidence, association and hidden trend.

Compared to other textual data, there are a few research studies in Quranic verse classification based on the translation of English [2], [3]. Several studies focused on Al-Quran verse classifying in Arabic [4]-[6]. Jamil et al. [7] proposed the use of term frequency in subject identification in order to classify text groups into certain subject. Their dataset for the experiment consisted of 286 Al-Quran verse and the verse contains 16 keywords. Label that are prepared for classify verse chosen is *Iman*, *Ibadah* and *Akhlak*. The performance metrics used included the ranking score and classification accuracy of the verses identified as subject in the Al-Baqarah chapter.

This research focuses on Al-Quran web portal checking and forecast by using data mining approach. The dataset used is sourced from the Web Alexa's company of Amazon.com, which provides analysis to all websites. The objective for the study is to see access pattern of some website that use data mining tool based on classification.

To carry out this task, 1000 Quran verse from database were divided into two sets of training and testing dataset, respectively. The training set consists of 800 sentences and the testing set consists of 286 sentences. Three classes were chosen (Iman, Ibadah, Akhlak) after transforming from Arabic to Malay language. Results are evaluated based on Precision and Recall metrics with Support vector machine the highest recall value of 86% and Naïve Bayes has the lowest recall value of 71%.

2. RESEARCH METHOD

CRISP-DM refers to cross process industry for data mining. CRISP-DM methodology [17] provides structure approach to plan a project especially data mining. It is a methodology that right and suitable. It has the strength that is distinctive, flexibility and used to analyses to resolve issues that complicated. The Following is CRISP-DM model sample shown in Figure 1.

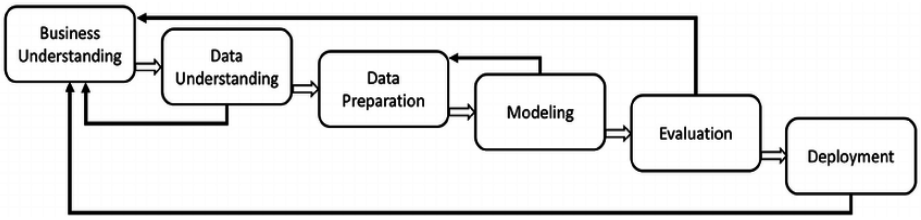

Figure 1. CRISP-DM methodology

Figure 2 shows the order of incident that is ideal. In most tasks it can be done in various arrays that are different and needed to go to tasks predecessor and repeat the action. This model will not block any route will through process data mining.

The classification methodology in Figure 2 will begin with a collection of Quranic verses from the surah Al-Baqarah in the Quran. Next, the translation for the verses are prepared for text processing including tokenization, stemming, and part-of-speech tagging. Next, the verses are classfied into three categories or

class labels, which are 'Iman' (faith), 'Ibadah' (worship), and 'Akhlak' (virtues). The classification algorithms used are Naive Bayes, K-Nearest Neighbor, Decision Tree J48, and Support Vector Machine (SVM).

2.1 Dataset

The dataset consists of 286 verses from chapter two Surah al-Baqarah of the Holy Quran. Dataset taken from this is translation from Al-Quran classical by Abdullah Yusuf Ali. It achieved from www.qurandatabse.org website. Currently, there is no dataset Al-Quran for machine learning.

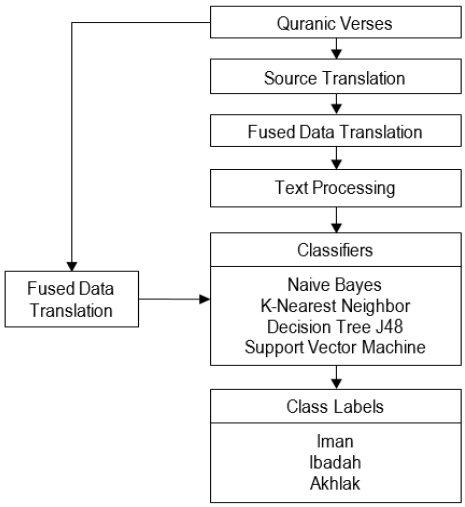

Figure 2. The proposed group-based classification algorithm.

2.2 Evaluation Metric

The objective of the study is to distinguish influence algorithm classification that where significant in classification process. Classification experiment that was carried out is to measure accuracy and area under recipient operation characteristic curve (AUC) which exceeded algorithm in classification [8].

2.1.1 Classification Accuracy

Classification accuracy shown is percentage or proportion from forecast that have been classified. The accuracy can be calculated with the formula as shown in Equation 1.

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

when the TP is positive true, it could be categorized as positive, while if the TN is negative True, it could be categorized correct as negative. FP is positive false, it cannot be categorized as positive, for FN is negative false, it categorized false will be separated as negative.

2.1.2 Area under (ROC) curve (AUC)

ROC has best performance for the classification problem. Roc value shows the overall result for classification performance. This performance show that metric evaluating classification performance. This value could be seen between 0 and 1, where ROC =1 equivalent to classification that right, while ROC=0.5 considered in between of accurate and not accurate and ROC=0 suitable classify by overturned classification. If ROC value inclined towards 1, this mean the algorithm classification is right and correct.

2.1.3 Classification Algorithm

The comparative experiments in this research used four algorithms, which are Naive Bayes, J48 and K-Nearest Neighbor apart from the proposed Support Vector Machine.

a) Naïve Bayes

Naïve Bayes method is one of the sets algorithm learning taken based on usage from theory that given by Bayes' theorem name by implying ‘naïve’ independence to every feature. Given by Y class variable and vector feature depend on x1, bayes theorem shows the relationship as shown in Equation 2,

$$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)} \quad (2)$$

using this naïve independence assumption in Equation 3 for all *i.*

$$P(x_i|y,x_1,\dots,x_{i=1},x_{i+1},\dots,x_n = P(x_i|y), \quad (3)$$

The simplified relationship is shown as Equation 4.

$$P(y|x_1,\dots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\dots,x_n)} \quad (4)$$

Since the probability of x_1 through x_n is constant given the input, we can use the following classification rule in Equation 5.

$$P(y|x_1,\dots,x_n)\alpha\, P(y)\prod_{i=1}^{n} P(x_i|y) \qquad \hat{y} = \arg\max P(y)\prod_{i=1}^{n} P(x_i|y) \quad (5)$$

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

b) Decision Tree J48

A decision tree (C4.5 algorithm) is an algorithm that uses a tree-like structure to represent the features/attributes as well as the class labels. In Weka, the C4.5 algorithm is called J48 [12]. C4.5 has several parameters, by the default visualization (when you invoke the classifier) only shows C. C is Confidence value (default 25%): lower values incur heavier pruning. Then M is Minimum number of instances in the two most popular branches (default 2). The experiment uses all the default parameter settings for this classifier.

c) Support Vector Machine

Support Vector Machine (SVM) [13] is a supervised machine learning algorithm that is well-known for both classification or regression problems. The SVM algorithm will plot each data item or instance as a point in an *n*-dimensional space where *n* is number of features with the value of each feature being the value of a particular coordinate. Next, the algorithm performs classification by finding the hyper-plane that separates the two classes as shown in Figure 3.

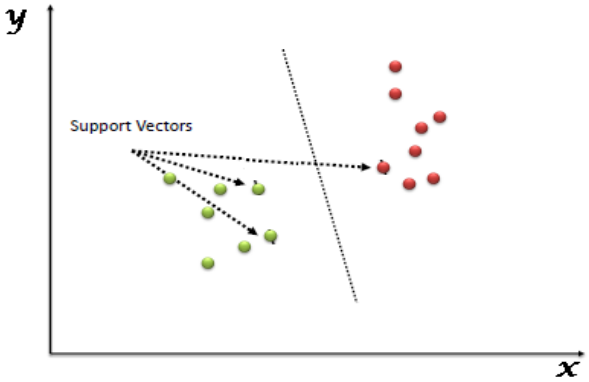

Figure 3. SVM Hyperplane

d) K-nearest neighbors classifier

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. When KNN is used for classification, the output can be calculated as the class with the highest frequency. Each instance in essence votes for their class and the class with the most votes and it taken as the prediction.

Class probabilities can be calculated as the normalized frequency of samples that belong to each class in the set of K most similar instances for a new data instance. For example in Equation 6, in a binary classification problem (class is 0 or 1).

$$p(class = 0) = \frac{count(class=0)}{count(class=0)+count(class=1)} \quad (6)$$

3. RESULTS AND ANALYSIS

This part presents results of the research work with four classifications algorithms, which are Naïve Bayes, J48, SVM, and k-NN. The results are presented in terms of classification accuracy and AUC. Tables 1 to 5 show the comparative results in terms of accuracy.

Table 1. Results for Naïve Bayes Classifier

Class	Precision	Recall	F-Measure	ROC Area
Iman	0.839	0.897	0.867	0.837
Ibadah	0.417	0.375	0.395	0.756
Akhlak	0.531	0.405	0.459	0.828
Weighted Avg	0.735	0.752	0.741	0.824

Table 2. Results for J48 Classifier

Class	Precision	Recall	F-Measure	ROC Area
Iman	0.780	0.902	0.836	0.621
Ibadah	0.261	0.150	0.190	0.483
Akhlak	0.407	0.262	0.319	0.664
Weighted Avg	0.652	0.703	0.670	0.608

Table 3. Results for SVM Classifier

Class	Precision	Recall	F-Measure	ROC Area
Iman	0.827	0.936	0.878	0.730
Ibadah	0.708	0.425	0.532	0.708
Akhlak	0.613	0.452	0.521	0.785
Weighted Avg	0.779	0.794	0.777	0.735

Table 4. Results for K-NN Classifier

Class	Precision	Recall	F-Measure	ROC Area
Iman	0.800	0.824	0.812	0.655
Ibadah	0.356	0.400	0.376	0.630
Akhlak	0.387	0.286	0.329	0.592
Weighted Avg	0.677	0.685	0.680	0.642

Table 5. Comparative results for all classification algorithms

Algorithm	Precision	Recall	F-Measure	ROC Area
Naïve Bayes	0.735	0.897	0.741	0.824
J48	0.653	0.703	0.670	0.608
SVM	0.779	0.794	0.777	0.735
K-NN	0.677	0.685	0.680	0.642

Based on the experimental results, the proposed algorithm achieved classification accuracy above 85% using the two classification algorithms: SVM and J48 classifiers. Nonetheless, a relatively low accuracy result was achieved when using the Naïve Bayes and the K-NN classifiers.

4. CONCLUSIONS

Classifyng Al-Quran sentence into category that had been given is important task in al-Quran study. In this project, we display algorithm classification to label Quran verses by automatic to connect sentences with three fundamental aspects in Islam namely ‘Iman’, 'religious worship (religious worship)', and ‘Akhlak’. Dataset consisting of 286 sentences from Al-Baqarah's surah, had denormalization by using StringToWordVector in WEKA with IDF-TF method. Four different algorithm that was chosen for classification that usual and subset selection approach has been used to aim experiment in classification task.

In conclusion, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, and J48 decision tree was implemented according to algorithm classification to determine class membership to every sentence and count decision in terms of accuracy and area under AUC curva operation feature. Decision from experiment that were carried out by the classification results, support vector machine (SVM) and J48 classifier algorithm have the overall best accuracy performance of 85% while Naïve Bayes had the last accuracy result of 71%. This shows that there is no specific best classification algorithm. Last but not least, we hope that can expand dataset's complete set Koran and consequently widen al-Quran holy verses into label as was explained by al-Quran. Besides that, this method is can be used for Hadith.

ACKNOWLEDGEMENT

This project is supported by the Fundamental Research Grant Scheme, vot 1609 from Ministry of Higher Education, Malaysia.

REFERENCES

[1] Abdul Baquee Muhammad, 2012, "Annotation of conceptual Co-reference and Text Mining the Quran."
[2] Hamed, S.K., Ab Aziz, M.J.: A question answering system on holy Quran translation based on question expansion technique and neural network classification. J. Comput. Sci. 12, 169–177 (2016)
[3] Hamoud, B., Atwell, E.: Quran question and answer corpus for data mining with WEKA, pp. 211–216. IEEE Conference of Basic Sciences and Engineering Studies, Leeds (2016)
[4] Siddiqui, M.K., Naahid, S., Khan, M.N.I.: A review of Quranic web portals through data mining. VAWKUM Trans. Comput. Sci. 5, 1–7 (2014)
[5] Hilal, A., Srinivas, N.: Analytical of the initial holy Quran letters based on data mining study. Am. Int. J. Res. Formal Appl. Nat. Sci. 10, 1–8 (2015)
[6] Akour, M., Alsmadi, I., Alazzam, I.: MQVC: measuring Quranic verses similarity and Surah classification using N-Gram. WSEAS Trans. Comput. 13, 485–491 (2014)
[7] Jamil, N.S., Ku-mahamud, K.R., Din, A.M., Ahmad, F., Chepa, N., Ishak, W.H.W., Din, R., Ahmad, F.K.: A subject identification method based on term frequency technique. J. Advanc. Comput. Res. 7,
[8] Santra, A.K., Christy, C.J.: Genetic algorithm and confusion matrix for document clustering. Int. J. Comput. Sci. Iss. 9, 322–328 (2012)
[9] Yang, J., Qu, Z., Liu, Z.: Improved Feature Selection Method Considering the Imbalance Problem in Text Categorization. Scientific World J. 1–17 (2014)
[10] Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Mining Knowledge Manag. Process. 5, 1–11 (2015)
[11] scikit-learn http://scikit-learn.org/stable/modules/naïve _bayes.html
[12] Weka Decision Tree http://scikit-learn.org/stable/ modules/naive_bayes.html
[13] Amarappa, S., Sathyanarayana, S.V.: Data classification using support vector machine (SVM), a simplified approach. J. Electron. Comput. Sci. Engineering. 3, 435–445 (2014)
[14] Gharehchopogh, F.S., Khaze, S.R., Maleki, I.: A new approach in bloggers classification with hybrid of k-nearest neighbor and artificial neural network algorithms. Indian J. Sci.Technol. 8, 237–246 (2015)
[15] Dey, L., Chakraborty, S., Biswas, A., Bose, B., Tiwari, S.: Sentiment analysis of review datasets using Naïve Bayes' and k- NN classifiers. J. Informat. Eng. Electron. Business. 4, 54–62 (2016)
[16] Jason Brownlee, 'K-Nearest Neighbors for Machine Learning', https://machinelearning mastery. com/k-nearest-neighbors-for-machine-learning/
[17] Smart Vision-Europe, https://www.sv-europe.com/