IJID International Journal on Informatics for Development, *e-ISSN :2549-7448*
Vol. 8, No. 1, 2019, Pp. 47-51

Comparative Analysis of Text Mining Classification Algorithms for English and Indonesian Qur'an Translation

Rahmat Hidayat[1], Sekar Minati[2]
Informatics Engineering, Faculty of Science and Technology
Sunan Kalijaga State Islamic University
Yogyakarta, Indonesia
rahmat.hidayat@uin-suka.ac.id[1], sekar.minati@gmail.com[2]

***Abstract*—Qur'an, As-Sunnah, and Islamic old book have become the sources for Islam followers as sources of knowledge, wisdom, and law. But in daily life, there are still many Muslims who do not understand the meaning of the sentence in the Qur'an even though they read it every day. It becomes a challenge for Science and Engineering field academicians especially Informatics to explore and represent knowledge through intelligent system computing to answer various questions based on knowledge from the Qur'an. This research is creating an enabling computational environment for text mining the Qur'an, of which purpose is to facilitate people to understand each verse in the Qur'an. The classification experiment uses Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor (kNN), and J48 Decision Tree classifier algorithms with Al-Baqarah verses translated to English and Indonesian as the dataset which was labeled by three most fundamental aspects of Islam: '*Iman*' (faith), '*Ibadah*' (worship), and '*Akhlaq*' (virtues). Indonesian translation was processed by using the *sastrawi* package in Python to do the pre-processing and StringToWord Vector in WEKA with the TF-IDF method to implement the algorithms. The classification experiments are determined to measure accuracy, and f-measure, it tested with a percentage split 66% as the data training and the rest as the data testing. The decision from an experiment that was carried out by the classification results, SVM classifier algorithms have the overall best accuracy performance for the Indonesian translation of 81.443% and the Naïve Bayes classifier has the best accuracy for the English translation, which achieved 78.35%.**

Keywords-SVM; Naïve Bayes; kNN; J48; text classification; Qur'an

I. INTRODUCTION

Qur'an was passed down to Prophet Muhammad when he was at the age of 40 for 23 years. It is the basic guidance for Muslims. It is arranged originally with Arabic and composed of 114 chapters and about 6,236 verses with 77,477 words [1]. After Prophet Muhammad dies, Qur'an, As-Sunnah, and old book Islamic have become the source for the Islam follower to make a source of knowledge, wisdom, and law. Therefore, this makes the Qur'an so important in daily life. However, there are still many Muslims especially in Indonesia do not understand the meaning of the sentence in the Qur'an even though they read it every day. It caused, there are still a few people who can understand Arabic fluently because Arabic is difficult to learn for most people and rarely people who have Arabic education since childhood.

From these problems, it becomes a challenge for Informatician to explore and represent knowledge through intelligent system computing to answer various questions based on knowledge from the Qur'an. This makes the background of the author to research in terms of creating an enabling computational environment for text mining the Qur'an, whose purpose is to facilitate people to understand each verse in the Qur'an. This study also aims to compare the performance of four text mining classification algorithms that are applied to the translation of the verses of the Qur'an.

Classification is one of the processing in text mining aiming to divide things according to classes. Text mining is similar to data mining, data mining is usually used to process data structured like data from a database, while text mining is used to process data unstructured or semi-structured such as e-mails, text documents, etc. [2]. The classification algorithms that are most often used in text mining include Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor (KNN), and Decision Tree.

Hassan et *a*l. [3] implemented a k-Nearest Neighbor (kNN) algorithm to classify the Holy Qur'an Tafseer verses into predefined categories. To achieve this task, a database of 1000 verses of the Qur'an was divided into two document sets with the training set consists of 800 verses and the test set consists of 200 verses. Seven predefined categories of the Tafseer texts were chosen (Marriage, Inheritance, Pray, Zakat, Respecting Parents, Halal, and Jihad) after transforming from Arabic to the Malay language. The results were evaluated based on Precision and Recall metrics with ‘marriage' category has the highest recall value of 0.9 and ‘inheritance' has the lowest recall value of 0.74.

II. RELATED WORK

In research titled Comparative analysis for topic classification in Juz Al-Baqarah, Rahman et al. [4] classified a topic in 286 verses of the Al-Baqarah into three categories or class labels, which are '*Iman*' (faith), '*Ibadah*' (worship), and '*Akhlaq*' (virtues) after transforming from Arabic to the Malay language. The classification algorithms used are Naive Bayes, K-Nearest Neighbor, Decision Tree J48, and Support Vector Machine (SVM). The decision from the experiment that was 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 and KNN had the last accuracy result of 71%.

Al-Kabi et al. [5] worked on the classification of different Qur'anic verses (ayat) according to their topics using four different classification algorithms (Decision tree, kNN, SVM, NB) to evaluate their effectiveness. To achieve this, they identified three distinct predefined categories (ignorant of religion; oneness of God; the penalty of Apostates). A total number of 1,227 Ayat (verses) were used out of the entire 6,236 Ayat of the Holy Qur'an to train and evaluate the selected classifiers. The results show Naive Bayes (NB) accuracy score is 99.9099%, Support Vector Machine (SVM) accuracy score is 99.8649%, k-Nearest Neighbor (kNN) accuracy score is 99.8198%, and J48 Decision has the lowest accuracy score, it is 99.5946%.

Besides the research above, there are a few research studies in Qur'anic verse classification with various algorithms based on the translation of English [6], [7] and several studies focused on Al-Qur'an verse classifying in Arabic [8]–[13]. In this research, the authors will research text mining with the Qur'an being the object. The dataset for the experiment consisted of 286 Qur'an verses of Al-Baqarah that have transformed from Arabic to Indonesian. The label that is prepared for classify verses were chosen is 'Iman' (faith), 'Ibadah' (worship), and 'Akhlaq' (virtues). The classification experiment is using Support Vector Machine (SVM), Naive Bayes, k-Nearest Neighbor (kNN), and J48 Decision Tree.

III. METHODOLOGY

A. Dataset

The dataset for the experiment consisted of 286 Qur'an verses of Al-Baqarah that have transformed from Arabic to Indonesian. Dataset taken from this is a translation from the Qur'an classical by the Indonesian Ministry of Religious Affairs. It achieved from www.Qur'andatabase.org website.

B. Text Preprocessing

Text pre-processing is the process of converting unstructured data into structured data according to needs, which is carried out for further mining processes. The steps in text pre-processing are tokenizing, case-folding, stop-words filtering.

1) Tokenizing

Tokenizing is the stage of cutting input text into words, terms, symbols, punctuation marks, or other elements that have a meaning called tokens [14]. In the process, tokens which are punctuation that is deemed unnecessary such as periods (.), Commas (,), exclamation marks (!), etc. will be deleted.

2) Case Folding

Case-folding is the process of matching the case in an article, this is because not all text articles are consistent in the use of capital letters. Therefore, case-folding is used to convert all text into a standard form.

3) Stop-words Filtering

The process carried out at this stage is to remove stop-word. Stop-word is a word that is not unique in an article or general words that are usually always in an article. [15].

C. Classification

The comparative experiments in this research used four algorithms, which are Support Vector Machine (SVM), Naïve Bayes, k-Nearest Neighbor (kNN), and J48 Decision Tree.

1) Support Vector Machine

Support Vector Machine (SVM) [16] is supervised learning models with associated learning algorithms used for classification and regression analysis. The SVM algorithm will plot each data item or instance as a point in dimensional space where there are several features with the value of each feature being the value of a particular coordinate. Next, the algorithm performs classification by finding the hyperplane that separates the two classes as shown in Figure 1.

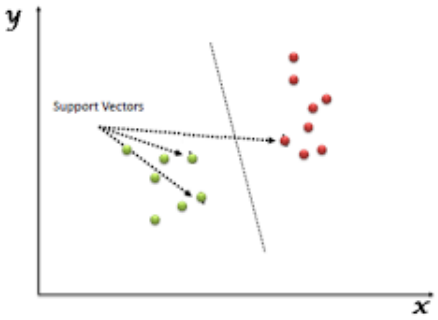

Figure 1. SVM Hyperplane

2) Naïve Bayes

Naive Bayes classification is a collection of simple opportunities based on the application of the Bayes theorem with the assumption that the explanatory variables are 'naïve' independence. Given by class variable and vector feature depend on, Bayes theorem shows the relationship as shown in Equation 1,

$$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots,x_n \vee y)}{P(x_1,\dots,x_n)} \quad (1)$$

using this naïve independence assumption in Equation 2 for all *i*.

$$P(x_1 \vee y, x_1,\dots,x_{i=1},x_{i+1},\dots,x_n) = P(x_i|y) \quad (2)$$

The simplified relationship is shown in Equation 3.

$$P(y|x_1,\dots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_1|y)}{P(x_1,\dots,x_n)} \quad (3)$$

Since the probability of through is constant given the input, we can use the following classification rule in Equation 4.

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

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. In turn to helps alleviate problems stemming from the curse of dimensionality [4].

3) K-Nearest Neighbor

K-Nearest Neighbor [17] is the simplest algorithm in data mining. There are three key elements of this approach: a set of labeled objects, e.g., a set of stored records, a distance or similarity metric to compute the distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object. The issue that affects the performance of kNN is the choice. It is too small, then the result can be sensitive to noise points. And if is too large, then the neighborhood may include too many points from other classes.

4) J48 Decision Tree

In the WEKA data mining tool, J48 [18] is an open-source Java implementation of the C4.5 algorithms. C4.5 builds decision trees from a set of training data using the concept of information entropy. The training data is a set of already classified samples. Each sample consists of a p-dimensional vector, where they represent attribute values or features of the sample, as well as the class in which falls. C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other at each node of the tree. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithms then recurses on the partitioned sub-lists.

D. Evaluation Metrics

The objective of this research is to compare the performance of four classification algorithms in classifying English-language texts and Indonesian-language texts, in this case, the translation of the text of the Quran. Classification experiments are determined to measure accuracy, precision, recall, f-measure, the area under the receiver operating characteristics curves (AUC), and algorithms on classifiers. The AUC value closer to 1 indicates a better classification result [19].

E. Tools

This research used two tools to do data processing and application of algorithms, such as python and WEKA.

1) Python

Python [20] is created by Guido van Rossum and first released in 1991. It is an interpreted, high-level, general-purpose programming language that has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Python is an interpretative programming language that is considered easy to learn and focuses on code limitations. In other words, python is claimed to be a programming language that has a programming code that is very clear, complete, and easy to understand [21].

2) WEKA

The Waikato Environment for Knowledge Analysis (WEKA) [22] is computer software developed at the University of Waikato in New Zealand. WEKA is a free Java software available under the GNU General Public License. It is a collection of machine learning algorithms to solve data mining problems. WEKA has become one of the most widely used data mining systems while it offers many powerful features. WEKA supports many different data mining tasks such as data pre-processing and visualization, attribute selection, classification, prediction, model evaluation, clustering, and association rule mining. It is written in Java and can be run with almost any computing platforms. WEKA can be used to pre-process without writing any program code, and it comes with a graphical user interface to provide easily used tools for beginner users to identify hidden information from database and file systems in a simple way by using options and visual interfaces. There is a specific default. ARFF data file format that WEKA accepts. The data should be as a single flat file or relation; the data can be imported from a Comma Separated Value (.CSV) file, a database, a URL, etc. where each data point is described by a fixed number of attributes. WEKA supports numeric, nominal, date and string attributes types.

IV. Result and Discussion

The experiment used python to pre-processing the Indonesian dataset with the sastrawi package because WEKA cannot do the pre-processing Indonesian text. Pre-processing that have been done are case-folding to change cases to lowercase, and stop words filtering.

Next, the processed text was tokenizing by using StringToWordVector in WEKA with the IDF-TF method. The results of the research work with four classifications algorithms, which are SVM, Naive Bayes, k-NN, and J48 are presented in terms of classification accuracy and AUC. Tables 1 and 2 show the comparative results in terms of accuracy and f-measure with a percentage split of 66% as the data training and the rest as the data testing.

TABLE I. Comparative results for Indonesian translation using all classification algorithms

Algorithm	Accuracy (%)	F-Measure
SVM	81.4433	0.825
Naive Bayes	76.2887	0.834
kNN	72.1649	0.756
J48	74.2268	0.690

TABLE II. Comparative results for English translation using all classification algorithms

Algorithm	Accuracy	F-Measure
SVM	73.1959	0.619
Naive Bayes	78.3505	0.788
kNN	75.2577	0.704
J48	75.2577	0.739

Based on the experimental results, the highest classification accuracy achieved by the SVM algorithm for the Indonesian Translation dataset, achieved of 81.4% and 78.35% for the English Translation dataset using the Naïve Bayes classifier. Nonetheless, a relatively low accuracy result was achieved when using the kNN classifier for Indonesian Translation text, had classification accuracy 72.165% and the SVM classifiers for English translation text, had classification accuracy 73.2%. In term of f-measure, the Naïve Bayes classifier gets the largest f measure score, in Indonesian Translation and English Translation which are 0.834 and 0.788 respectively. It means that the Naïve Bayes classifier has the best combination of precision and recall.

V. Conclusion

Classifying the Qur'anic verses into pre-defined categories is an essential task in Qur'anic studies. The research presented a classification Quranic verses into the most fundamental aspects of Islam: 'Iman' (faith), 'Ibadah' (worship), and 'Akhlaq' (virtues). Dataset consisting of 286 sentences from Al-Baqarah surah, had processed by using the sastrawi package in Python and StringToWordVector in WEKA with IDF-TF method. Four different algorithms that were chosen for classification that usual and subset selection approach has been used to aim experiment in a classification task.

In conclusion Support Vector Machine (SVM), Naïve Bayes, k-Nearest Neighbor (kNN), and J48 Decision Tree were implemented according to algorithm classification to determine class membership to every sentence and count decision in terms of accuracy and f-measure. The decision from an experiment that was carried out by the classification results, SVM classifier algorithms have the overall best accuracy performance for the Indonesian translation of 81.443% and the Naïve Bayes classifier has the best accuracy for an English translation, which achieved 78.35%.

With this research expected it can help the next research or another researcher to expand the dataset with a complete set of Qur'an and the label as was explained by the Qur'an.

Besides that, this method is can be used for another text document like Hadith or Tafseer. Next researchers can also compare the accuracy performance with several ensemble method algorithms.

REFERENCES

[1] [1] M. Osman, A. Hilal, and M. Alhawarat, "Fine-Grained Quran Dataset," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 12, 2016.

[2] [2] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," J. Emerg. Technol. Web Intell., vol. 1, no. 1, pp. 60–76, 2009.

[3] [3] G. S. Hassan, S. K. Mohammad, and F. M. Alwan, "Categorization of 'Holy Quran-Tafseer' using K-Nearest Neighbor Algorithm," Int. J. Comput. Appl., vol. 129, no. 12, pp. 1–6, 2015.

[4] [4] M. I. Rahman, N. A. Samsudin, A. Mustapha, and A. Abdullahi, "Comparative analysis for topic classification in Juz Al-Baqarah," Indones. J. Electr. Eng. Comput. Sci., vol. 12, no. 1, pp. 406–411, 2018.

[5] [5] Mohammed N. Al-Kabi, Belal M. Abu Ata, Heider A. Wahsheh, and Izzat M. Alsmadi, "A Topical Classification of Quranic Arabic Text," Proc. 2013 Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Its Sci., no. December, pp. 272–277, 2013.

[6] [6] S. K. Hamed and M. J. A. Aziz, "A question answering system on Holy Quran translation based on question expansion technique and Neural Network classification," J. Comput. Sci., vol. 12, no. 3, pp. 169–177, 2016.

[7] [7] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Dharmalaksana, "Clustering the verses of the Holy Qur'an using K-means algorithm," Asian J. Inf. Technol., vol. 15, no. 24, pp. 5159–5162, 2016.

[8] [8] M. K. Siddiqui, S. Naahid, and M. N. I. Khan, "A REVIEW of QURANIC WEB PORTALS THROUGH DATA MINING," VAWKUM Trans. Comput. Sci., vol. 5, no. 2, pp. 1–7, 2015.

[9] [9] A. Hilal and N. Srinivas, "Analytical of the Initial Holy Quran Letters Based on Data Mining Study," Am. Int. J. Res. Formal, Appl. Nat. Sci., vol. 10, no. 1, pp. 1–8, 2015.

[10] [10] M. Akour, I. Alsmadi, and I. Alazzam, "MQVC: Measuring quranic verses similarity and sura classification using N-gram," WSEAS Trans. Comput., vol. 13, pp. 485–491, 2014.

[11] [11] N. S. Jamil et al., "A subject identification method based on term frequency technique," Int. J. Adv. Comput. Res., vol. 7, no. 30, pp. 103–110, 2017.

[12] [12] M. Alhawarat, "Extracting Topics from the Holy Quran Using Generative Models," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 12, pp. 288–294, 2016.

[13] [13] M. N. Al-Kabi, H. A. Wahsheh, I. M. Alsmadi, and A. Moh'd Ali Al-Akhras, "Extended Topical Classification of Hadith Arabic Text," Int. J. Islam. Appl. Comput. Sci. Technol., vol. 3, no. 3, pp. 13–23, 2015.

[14] [14] S. Vijayarani, J. Ilamathi, and Nithya, "Preprocessing Techniques for Text Mining - An Overview," Int. J. Comput. Sci. Commun. Networks, vol. 5, no. 1, pp. 7–16, 2018.

[15] [15] F. Z. Tala, "A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia," 2003.

[16] [16] S. Amarappa and S. V Sathyanarayana, "Data Classification Using Support Vector Machine (SVM), a simplified approach," Int. J. Electron. Comput. Sci. Eng., vol. 3, no. 4, pp. 435–445, 2019.

[17] [17] H. Motoda et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2007.

[18] [18] Wikipedia Contributor, "C4.5 algorithm," Wikipedia, The Free Encyclopedia, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=C4.5_algorithm&oldid=883549387.

[19] [19] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. Nawi, "Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 4, p. 1419, 2017.

[20] [20] D. Kuhlman, A Python Book: Beginning Python, Advanced Python, and Python Exercises. Platypus Global Media, 2012.

[21] [21] Jubilee Digital, Pemrograman Python Untuk Pemula. Yogyakarta: Jubilee Solusi Enterprise, 2016.

[22] [22] I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Cerra, Diane, 2011.