A comparison of Neural Network and SVM on the multi-label classification of Quran verses topic in English translation

Fahmi Salman Nurifikri and Adiwijaya
Telkom University, Telekomunikasi Street No. 1, Bandung 40257, Indonesia

Abstract. Indeed the Quran is the main guideline for Muslims. The enticing thing in the Quran is one verse of the Quran can be classified into more than one topic of discussion in the Quran, for example one verse discusses about prayer and faith or arkanul Islam and religions, thus this called multilabel case. In this study, we compared the multi-label Quran verse topic classification techniques using Naïve bayes, ANN and SVM. Based on the experiment conducted, combination Naïve Bayes algorithm with Mutual Information or Information Gain gives the best performance when it applied to construct the model with Hamming Loss of 0.0938.

1. Introduction

In Islam, the source of Islamic teachings and law will always be oriented on two sources, namely the Quran and Sunnah (Hadith). The Quran is the word of God and the last revelation made as a guide for mankind. The enticing thing in the Quran is one verse of the Quran can be classified into more than one topic of discussion in the Quran, for example Surah Al Bayyinah verse 98 which can be classified into verse that discusses about arkanul Islam and religions. Since one verse in the Quran can have more than one topic of discussion, it is called a multilabel case [1, 2]. Contrast to a singlelabel case, which one verse can be mapped to one topic of discussion or class only.

To find out the topic of discussion of each verse in the Quran, we can ask it to someone who certainly understand the division of topics of discussion. Yet availability of either place, time, and resources are limited. Therefore digitalization is required for multilabel classification Quran verses based on the topic of discussion; with machine learning approach, where is the machine or computer used to categorize the verses of Quran. On the other hand, Quran structure is different from any text representation [3, 4].

Generally in text classification, features are terms or words contained in the text. Usually a document or textual data contains considerable amount of words that can cause high computational complexity and decrease accuracy because some attributes may be irrelevant [5, 6].

To overcome these problem, it takes a feature reduction. One way to perform the feature reduction is to use the feature selection [7]. Feature selection is a process to
select the relevant features to be used in the classification process.

2. Related work

2.1. Multilabel Classification

Text classification is part of supervised learning that is used to group data into a particular category. There are two types of classification, such as single label classification and multilabel classification. In single label classification, each data can only be grouped into one category, whereas in a multilabel classification each data can be grouped into more than one category.

There are two approaches that can be used in multilabel classification, which are problem transformation and algorithm adaption. Problem transformation approach will turn multilabel problems into a single label classification problem. While the algorithm adaption approach will classify multilabel data by using algorithms that have been designed to handle multilabel classification problems.

The method that will be used in this paper is problem transformation method, based on the result that is obtained from previous experiment [5, 8, 9]. The algorithm to be used in this research is Naïve Bayes (NB), Artificial Neural Network (ANN), and Support Vector Machine (SVM).

2.2. Selection and Feature Selection

Feature Selection in text classification is used to improve the effectiveness of the model, computation and classification efficiency [8]. Some feature selection include Mutual Information, Information Gain and Document Frequency, where the methods can produce the best results in previous experiments [7, 10].

Mutual Information (MI) select a subset if relevant attributes (words). The selected attributes are considered to be more informative and effective for the classification model. The output of MI is a matrix of size n x n, where n is the number of variables. Each value in the matrix represents a corresponding value between two variables [11, 12]. In this research, the authors used only MI values between word and class variable. Calculation of MI is shown in (1).

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x) \cdot p(y)} \right)$$ (1)

Information Gain (IG) is used to measure the relevance level between attribute A and class C. The higher the value between attribute A and class C, the higher the relevance between attribute A and class C [13]. Calculation of IG is shown in (2).

$$I(C, A) = H(C)H(C|A)$$ (2)

Document Frequency (DF) thresholding is the simplest technique for dimension reduction. DF easy scale for very large corporations with computational complexity is
linear in the number of training documents.

3. System Design

The system built in this research is a system capable of classifying multilabel topics verses of the Quran. Figure 1 visualizes the design of system.

Based on Figure 1, the flow of system in this study can be described as follows.

(i) Dataset

The topic of classification dataset used is the topic of the classification of the verses of Al Quran Cordova in the English translation. The number of Quran verses used in this study is 6236 verses which are divided into 15 classes. Table 1 presents lists of topics in the classification of the verses of Al Quran Cordova.

(ii) Preprocessing, we prepared the dataset to be used in the classification process, removal of attributes that are not needed, thus the data becomes more structured. Figure 2 visualizes the design of system.
Table 1: List of topics

| Chapter | Topic                        |
|---------|------------------------------|
| I       | Arkanul Islam                |
| II      | Faith                        |
| III     | Al Quran                     |
| IV      | Science and Its Branches     |
| V       | Charity                      |
| VI      | Dakwah                       |
| VII     | Jihad                        |
| VIII    | Humans and Relations         |
| IX      | Morals                       |
| X       | Regulations Related to Property |
| XI      | Matters Related to Law       |
| XII     | Country and Society          |
| XIII    | Agriculture and Trade        |
| XIV     | History and Stories          |
| XV      | Religions                    |

- **Case Folding**
  Case Folding is converted all letters in document into lowercase. At this stage, character other than letters, such as numbers and punctuation marks is removed.

- **Tokenization**
  Tokenization turn the sentence in document into tokens. In this study, we cut sentence into pieces of word.

- **Stopword Removal**
  It removes common words considered as less relevant word in order to reduce the number of words to be processed. In this study, the authors used a stopwords list from Wordnet.

- **Lemmatization**
  Lemmatization is the process of returning word in the token list to a common base form of the word. Lemmatization grouping together the inflected forms of a word so they can be analysed as a single item.

(iii) Feature extraction, we used the representation of bag of words model. It count each word in vocabulary list obtained from training set for each document.

(iv) Feature selection, we used MI, IG and DF to select features contained in data. The result of feature selection will be used to create a classification model.

(v) Training Classifier, we create a classification model using NB, SVM, and ANN by using words that pass the feature selection process.

(vi) Classification, we classify the data using NB, SVM, and ANN with problem
transformation approaches.

(vii) In final stages, the performance of system is done by using Hamming Loss method.

The scenarios that were used in this experiment is to build the model by combining the feature selection method and the classification algorithm. The first scenario is to build a classification model using all attributes (without feature selection). The second scenario is to apply of feature selection method before construct of the classification model.

4. Results and Discussion

The first scenario was conducted to determine which classification algorithm that can produce classification models that have the best results. The algorithm used is Multinomial Naive Bayes, SVM with RBF kernels and $C = 1.0$, and ANN using stochastic gradient descent or LBFGS optimizer.

Table 2 shows experiment results of the model built without implementing feature selection method.

| Method   | Hamming Loss |
|----------|--------------|
| Naive Bayes | 0.1141       |
| SVM       | 0.1102       |
| ANN       | 0.1271       |

From Table 2 it can be seen that SVM has the smallest Hamming Loss value. This is because, SVM can indeed suffer in high dimensional spaces where many features are irrelevant [14].

The second scenario was conducted to determine which feature selection algorithm that can produce classification models that have the best results. In this experiment, several number were applied as threshold for the feature selection, which were 0.1 to 0.9.

Figure 3, 4, and 5 show the performance of Naive Bayes, SVM and ANN after feature selection using DF, IG, and MI. It can be seen in Figure 3, and 5 that MI and IG gives the best result on average, while in Figure 4 DF gives the best result on average.

Figure 3, resulted comparison between MI, IG and DF using Naive Bayes algorithm, shows influence of number of input variables to classification performance. it can be seen that the classification model that using IG and MI, the more number of input variables tends to decrease classification rates, but the number of input variables does not apply to DF. It can be seen from Figure 3 that when the number of input variables is 20% it will produce the smallest hamming loss or the best result.
In Figure 4, it can be seen that the difference in the number of inputs variables has no significant effect, this is caused by SVM can overcome high dimensional problems, so svm can work well even if it doesn’t use feature selection.

In Figure 5, the use of feature selection is quite influential on classification rate. Based on the Figure 5 that the best results are obtained when using 70% number of input variables. In the Figure 5, it can be seen that the performance changes when using 70% number of input variables, this is due to the presence of attributes that are
not very influential in training data but have an effect on testing data.

Table 3 shows the best experiment results of each model.

| Feature Selection          | NB   | SVM  | ANN  |
|----------------------------|------|------|------|
| No Feature Selection       | 0.1141 | 0.1102 | 0.1271 |
| Mutual Information         | 0.0938 | 0.1158 | 0.1023 |
| Information Gain           | 0.0938 | 0.1158 | 0.1023 |
| Document Frequency         | 0.1116 | 0.1102 | 0.1161 |

Based on the table 3, that experiment using MI or IG gave the best result, because its can select features that have high relevance to the output class. So the result, generated feature matrix has fewer zero values. The less number of inputs used, the fewer zero values are found in the matrix.

While the experiment using DF showing results that are not too significant compared without using DF, because DF only selects features based on counts the most common words and does not pay attention to the relevance of the class. So that a feature can often to appear but has no relevance to the output class.

On the other hand, Naïve Bayes model would consider each words probability independent of any other word. This model obviously makes several rough, information-discardling assumption, but its work very well for text classification.

5. Conclusion

Multilabel classification that is conducted in this experiment gave the best performance with Hamming Loss of 0.0938. Those best result was achieved by implementing a model using Naïve Bayes algorithm, with the implementation of Mutual Information or Information Gain, where Mutual Information and Information Gain produce similar performance of the classifier. This is because both Mutual Information and Information Gain can choose the most relevant attributes to be included in the classification model.

References

[1] Mena D, Montas E, Quevedo J, JR and del Coz J An overview of inference methods in probabilistic classifier chains for multilabel classification 2016 Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(6), pp.215-230

[2] Tsoumakas G and Katakis I Multi-label classification: An overview 2007 International Journal of Data Warehousing and Mining (IJDWM), 3(3), pp.1-13
[3] Pane R A, Mubarok M S, Huda N S and Adiwijaya A multi-label classification on topics of quranic verses in english translation using multinomial naive bayes 2018 6th International Conference on Information and Communication Technology (ICoICT) IEEE

[4] Izzaty A M K, Mubarok M S, Huda N S and Adiwijaya A multi-label classification on topics of quranic verses in english translation using tree augmented nave bayes 2018 6th International Conference on Information and Communication Technology (ICoICT) IEEE

[5] Harrag F, El-Qawasmah E and Al-Salman A M S Stemming as a feature reduction technique for arabic text categorization 2011 10th International Symposium on Programming and Systems 128 – 133

[6] Nurifikri F S, Mubarok M S and Adiwijaya News topic classification using mutual information and bayesian network 2018 6th International Conference on Information and Communication Technology (ICoICT) IEEE

[7] Xu Y, Jones G, Li J, Wang B and Sun C M A study on mutual information-based feature selection for text categorization 2007 Journal of Computational Information Systems 3:3 1007–1012

[8] Irsan I C and Khodra M L Hierarchical multilabel classification for indonesian news articles 2016 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)

[9] Mubarok M S, Adiwijaya and Aldhi M D Aspect-based sentiment analysis to review products using nave bayes 2017 AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060) AIP Publishing

[10] Yang Y and Pedersen J O A comparative study on feature selection in text categorization 1997 Proceedings of the Fourteenth International Conference on Machine Learning 412–420

[11] Adiwijaya 2014 Aplikasi matriks dan Ruang Vektor (Graha Ilmu: Yogyakarta)

[12] Adiwijaya 2016 Matematika Diskrit dan Aplikasinya (Alfabeta: Bandung)

[13] Pratiwi A I and Adiwijaya On the feature selection and classification based on information gain for document sentiment analysis 2018 Applied Computational Intelligence and Soft Computing

[14] Weston J, Mukherje S, Chapelle O, Pontil M, Poggio T and Vapnik V Feature selection for svms 2001 Proceedings of the 13th International Conference on Neural Information Processing Systems 647–653