A Multi-Label Classification of Al-Quran Verses Using Ensemble Method and Naïve Bayes

Muhammad Rizqi Choirulfikri, Kemas Muslim Lhaksamana*, Said Al Faraby**

School of Computing, Telkom University, Bandung, Indonesia
Email: [1]rizqichoirulfikri@student.telkomuniversity.ac.id, [2,*]kemasmuslim@telkomuniversity.ac.id,
[3]saidalfaraby@telkomuniversity.ac.id
Email Penulis Korespondensi: kemasmuslim@telkomuniversity.ac.id

Abstract−Al-Quran is the holy book as a guide and also a source of law for muslims. Thus, understanding and studying Al-Quran is very important for muslims. To make it easier for muslims to understand and study the Qur'an, it is necessary to classify the verses of the Al-Qur'an. This study built a system that can perform multi-label classification of Al-Quran verses. Multi-label means that the classification will divide each verse of the Al-Quran into more than 1 topic. The model is built using the ensemble method by combining several Naïve Bayes algorithms. The ensemble method was chosen because research with different datasets can obtain good performance. The naïve Bayes algorithm was also chosen because it has a simple calculation so it requires a fairly short computation time. The preprocessing step is also carried out to see the comparison of performance results. To measure the performance of the system that has been built, the calculation of hamming loss is used. Based on the experimental results with several testing scenarios, the best performance results are obtained by combining Multinomial NB and Bernoulli NB with a hamming loss value of 0.1167. Thus, the use of the ensemble method can improve performance compared to without the ensemble method. This research can also of course build a multi-label classification model for the verses of Al-Quran with the ensemble method.

Keywords: Al-Quran; Multi-label; Classification; Ensemble Method; Naïve Bayes

1. INTRODUCTION

In 2015, muslims worldwide reached 1.8 billion people. Al-Quran is a holy book and a way of life for muslims [1]. Studying and understanding the verses of the Al-Quran is an obligation for muslims. Thus, a muslim is obliged to study the holy book Al-Quran kaffah or thoroughly. Al-Quran consists of more than 6000 verses and each verse has a different topic, even one verse can have more than one topic in each verse [2]. The topics contained in the Al-Quran are very diverse, ranging from Islamic history, charity, morals, and others. In the interpretation of the Al-Quran Cordova published by Syaamil Quran, Bandung, there are 15 different topics [3]. One way that can be done to make it easier to learn the Al-Quran is to classify the existing topics. Therefore, it is necessary to carry out a classification process on the verses of the Qur'an. The classification of verses in the Qur'an can also be categorized in multi-label text classification [4]. Multi-label means that the classification will divide each verse of the Quran into more than 1 topic. With this Al-Quran classification system, muslims in the world are expected to be able to easily distinguish and study the category of one verse from another.

Text classification is a process to group text into certain classes [5]. Text classification is used for several cases of document text in many fields with different purposes [6]. One application of text classification on the topic of Al-Quran verses has been carried out by Abdullah Adeleke [7]. In this study, it is explained about the comparison of algorithms on the topic of Al-Quran verses. The research was conducted by Ananda Pane [2]. In this study, the researchers divided the verses of the Al-Quran into 15 different classes. Multi-label classification is a case of text classification. In the case of multi-label text classification, each text or document can be grouped into more than one class [8]. The multi-label classification also illustrates the problems that exist in the world. In this study, the multi-label classification used is to classify the verses of the Al-Quran into 15 different classes.

In several previous kinds of related works to text classification that has been carried out, the classification model using the naïve Bayes algorithm produces a fairly high performance [2] and [5]. Ananda Pane et al. [2], have researched the case of multi-label text classification of Al-Quran verses in English translation. In this study, the multinomial naïve Bayes algorithm was used and focused on the use of stemming on preprocessing. The use of stemming can also accelerate the computing speed up to 29.44%. By using multinomial naïve Bayes, the resulting hamming loss of 0.1247 is the best performance. However, these figures are obtained without using stemming. To overcome this, they suggest using other selection features to get different performances. In Shou Xu's research [5], a classification text study was conducted using naïve Bayes. There are 3 naïve Bayes algorithms used in this study. Multinomial naïve Bayes, Bernoulli naïve Bayes and Gaussian naïve Bayes. Multinomial can classify text with an f1 score of 82%, followed by Bernoulli with an f1 score of 77%, and finally Gaussian at 70%.

The ensemble method is a combined method of several models that can be used for classification [9]. In the ensemble concept, several models that have been built will do majority vote to find the best classification results [10]. This method has been proven in studies [10] and [11] by producing better accuracy than without using the ensemble method. With some research and completion by existing methods, researchers will focus on research using the ensemble method to differentiate from previous research. The ensemble method combining the Gaussian naïve Bayes, Multinomial naïve Bayes, Complement naïve Bayes, and Bernoulli naïve Bayes algorithm.

This research was conducted to build a classification model system using the ensemble method and several naïve Bayes algorithms. The model aims to classify the English translation of the Al-Quran into certain topics. This research was conducted to analyze the effect of using several combinations of preprocessing steps on the results of hamming loss performance. This study also aims to analyze the effect of the naïve Bayes algorithm with the ensemble method and without the ensemble method.

2. RESEARCH METHODOLOGY

2.1 Research Flow

In this research, a multi-label text classification system was built. The classified text is an English translation of the Al-Quran verse. In building the text classification system, there are several steps carried out. First, the researcher prepared a labeled Al-Quran dataset. Then, the dataset will be processed at the data preprocessing step which aims to make the data higher quality. Furthermore, feature extraction will be carried out using TF-IDF. At the classification step, there are 2 processes, first is building a model with 4 naïve Bayes algorithms and the second is combining these models into an ensemble method using majority voting. An ensemble method is a new innovation method for Al-Quran dataset. An overview of the system to be built can be seen in Figure 1.
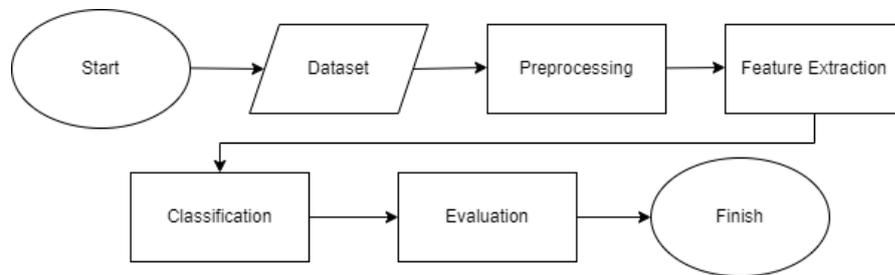
Figure 1. Research Process

2.1.1 Dataset

The dataset used in this research is an English translation of the Al-Quran verses that have been labeled in an excel file. The label consists of 15 topics (classes) according to the Tafsir Al-Quran Cordova published by Syaamil Quran, Bandung. The Al-Quran dataset used is more than 6000 verses of the Al-Quran which can be accessed on the Dataverse [13]. Each verse has at least 1 label (class).

2.1.2 Preprocessing

At this preprocessing step, the data train and data test were treated the same preprocessing. There are several preprocessing steps including case folding, punctual removal, stemming, stopword and tokenization. The preprocessing steps can be seen in Figure 2.
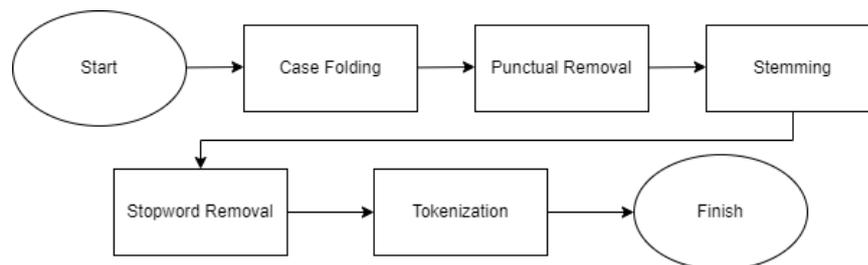
Figure 2. Preprocessing Steps

In Figure 2 it can be seen that the first step is case folding which is used to convert each word into lowercase letters. Then there is punctual removal to remove punctuation marks such as semicolons and others. Furthermore, there is stemming which is used to remove the initial or final affix to change it into a basic word. Then there is stopword removal which will remove common words or conjunctions that have no meaning. The last step is tokenization to cuts the words in each sentence. The examples of input and output in preprocessing can be seen in Table 2.

Table 1. Input and Output Preprocessing Example

Process	Input	Output
Case folding	And the moon becomes dark,	and the moon becomes dark,
Punctual removal	and the moon becomes dark,	and the moon becomes dark
Stemming	and the moon becomes dark	and the moon become dark

Process	Input	Output
Stopword removal	and the moon become dark	moon become dark
Tokenization	moon become dark	['moon', 'become', 'dark']

2.1.3 Feature Extraction (TF-IDF)

Feature extraction is used to give weight to a word. In this research, the weights are calculated using the TF-IDF calculation. TF describes the number of occurrences of words in the document and IDF describes the value of how important a word is to the document. To calculate the value of TF-IDF can be seen in equation (1).

$$w_{ij} = tf_{ij} * log\frac{D}{df_i} \tag{1}$$

Where w_{ij} is word weight t_j for document d_i, tf_{ij} is number of words t_j in document d_i, D is number of document, and df_i is a number of the word in document D.

2.1.4 Classification

The multi-label text classification steps in this study have several steps carried out, including the distribution of data train and data test, model making, model training, prediction of each label on each model built, combining the prediction results of the model built using the majority voting. The Ensemble Method (Majority Voting) is a newness method that used for Al-Quran dataset. The more detailed classification steps of each fold in ensemble method are presented in Figure 3.
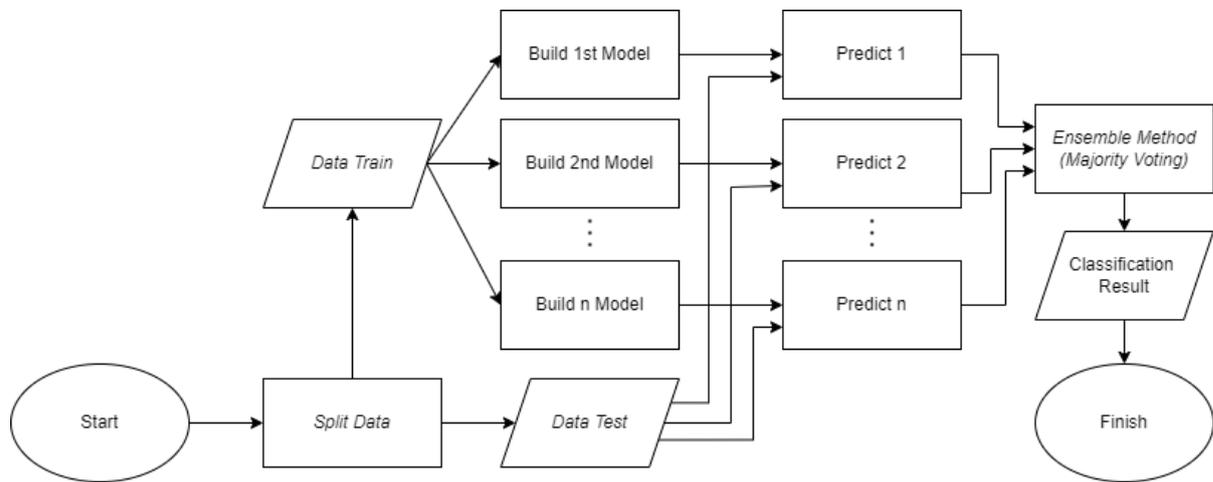
Figure 3. Classification Step for Each Fold in Ensemble Method (Majority Voting)

At the data split step to dividing data train and data test, the researcher used K-fold cross-validation with K=5. Each fold is built several classification models using the Naïve Bayes algorithm as many as 4 models. The naïve Bayes algorithms used includes Gaussian naïve Bayes (GNB), Multinomial naïve Bayes (MNB), Complement naïve Bayes (CNB), and Bernoulli naïve Bayes (BNB). Basically, every model with a naïve Bayes algorithm will produce a single label classification. In this case, each model will classify 15 times for each label.

Each prediction result of each model is entered in the ensemble method. The ensemble method used in this study is majority voting. The majority voting is taken from the predictions of each n model. In determining the majority voting it can be illustrated that if 3 models are entered in the ensemble method, then 2 or more than 3 models that produce 1 prediction result will produce a vote of 1. An example of the classification results using 3 models can be seen in Table 4.

2.1.4 Evaluation (Hamming Loss)

The evaluation step that will be used in this research is hamming loss. Hamming loss is used because it is suitable for multi-label classification cases. The smaller the hamming loss, the better. To calculate the hamming loss [5] follow a equation (2).

$$Hamming\ Loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} I(\hat{y}_j^{(i)} \neq y_j^{(i)}) \tag{2}$$

Where N is number of data, L is column of label, $\hat{y}_j^{(i)}$ is multi-label classification target and $y_j^{(i)}$ is multi-label classification output.

3. RESULT AND DISCUSSION

The evaluation step of this research was carried out on the dataset of the English translation of the Al-Quran. The dataset that used for testing is above 6000 data(text). The testing scenario of this research is carried out on preprocessing the data, and testing the classification method. The first scenario carried out in the preprocessing will show the effect of the performance results using all preprocessing steps, then without using stopwords and without stopwords as well as stemming. The second testing scenario was carried out on the classification using the ensemble method and the naïve Bayes algorithm.

Hamming loss evaluation used will produce a value with 4 digits behind the comma. This is done because the calculation of each fold, the hamming loss will be calculated for 1247 lots of data and 15 labels. And every wrong prediction will give a hamming loss value of 0.00005. The performance results of each fold will be divided by 5 according to the number of K-folds.

3.1 Result

The first testing scenario was carried out on the preprocessed data. In this scenario, each model using the Multinomial naïve Bayes algorithm, Bernoulli naïve Bayes and Complement naïve Bayes algorithms are tested with full preprocessing, and without using stopwords and without stopwords as well as stemming on preprocessed data. The experimental results in the first testing scenario can be seen in Table 2.

Table 2. Result of Scenario in Hamming Loss with Full Preprocessing, Preprocessing without Stopword and Preprocesing Without Stopword and Stemming

Algorithm Model	Full Preprocessing	Without Stopword	Without *Stopword* and *Stemming*
Gaussian naïve bayes (GNB)	0.3243	0.3204	0.2690
Multinomial naïve bayes (MNB)	0.1600	0.1539	0.1497
Complement naïve bayes (CNB)	0.1894	0.1785	0.1701
Bernoulli naïve bayes (BNB)	0.1175	0.1266	0.1221

The second testing scenario was carried out on the classification using the ensemble method and the naïve Bayes algorithm. In this scenario, 4 models with naïve Bayes algorithm are included in the ensemble method with different combinations. The first combination in finding the most votes is to enter 2 models in the ensemble method. The results of the first combination can be seen in Table 3.

Table 3. Testing Result of Scenario in Ensemble Method for 2 Model in Hamming Loss Without Stopword and Stemming

Majority Voting 2 Models	Without Stopword and Stemming
Gaussian NB and Multinomial NB	0.1491
Gaussian NB and Complement NB	0.1594
Gaussian NB and Bernoulli NB	0.1349
Multinomial NB and Complement NB	0.1498
Multinomial NB and Bernoulli NB	0.1167
Complement NB and Bernoulli NB	0.1168

Another combination in finding the most votes is to include 3 models and 4 models in the ensemble method. The results of the first combination can be seen in Table 4.

Table 4. Testing Result of Scenario in Ensemble Method for 3 and 4 Model in Hamming Loss Without Stopword and Stemming

Majority Voting 3 and 4 Models	Without Stopword and Stemming
Gaussian NB, Multinomial NB and Complement NB	0.1612
Gaussian NB, Complement NB and Bernoulli NB	0.1472
Multinomial NB, Complement NB and Bernoulli NB	0.1498
Gaussian NB, Multinomial NB and Bernoulli NB	0.1363
Gaussian NB, Multinomial NB, Complement NB and Bernoulli NB	0.1363

3.2 Analysis

Analysis of the testing results was carried out on 2 existing testing scenarios. The results of the first testing scenario can be seen in Figure 4. In Figure 4 it can be seen that the use of all preprocessing steps does not provide a better hamming loss value. The testing also shows that the use of the Bernoulli naïve Bayes algorithm produces the best hamming loss value.
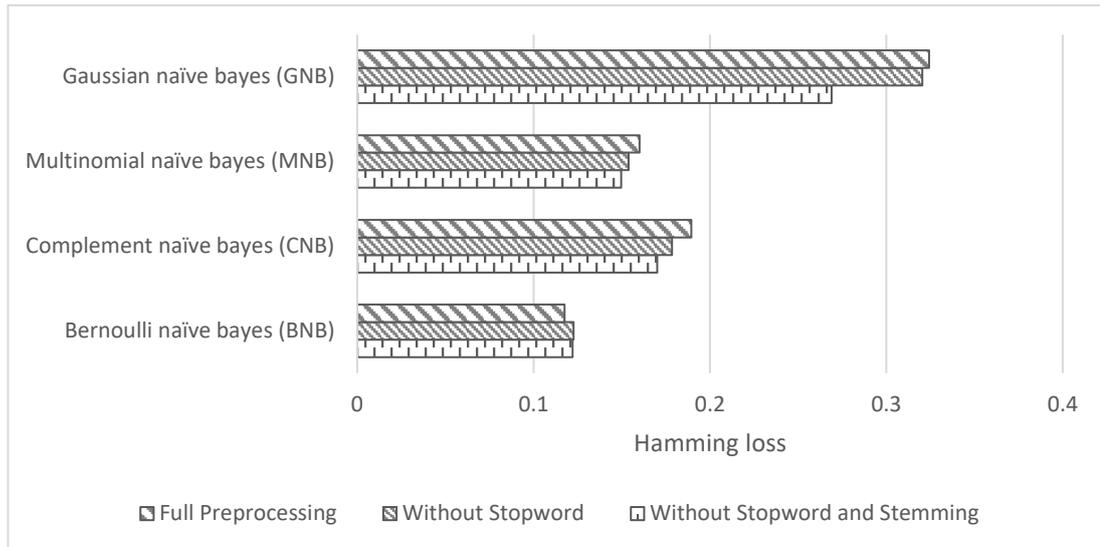
Figure 4. Analysis of Testing Result for Each Model on Preprocessing Testing Scenario.

In the Gaussian NB, Multinomial NB and Complement NB algorithms, the use of preprocessing without stopwords and stemming results in a better hamming loss value than full preprocessing and preprocessing without stopwords. The use of stopwords will provide a filter or reduction of words in the sentence. This will also change the structure of the sentence from initially using a conjunction to not. Therefore, the Al-Quran dataset has a sentence structure that cannot be separated or reduced because every word in the verse can give a certain meaning.

The use of stemming will certainly eliminate the affixes contained in each word. This can give a different meaning to each word so the Al-Quran dataset is not suitable for removing affixes. The Bernoulli naïve Bayes algorithm produces a better hamming loss value than the other 3 algorithms. Bernoulli NB algorithm represents features or words in binary. The same word in every verse of the Quran will be counted as 1 if it is present and 0 if it is not present. So that each word in the verse only has a value of 1 or 0 which is not the frequency of occurrence of the word.

In the results of the second testing scenario, the diagram can be seen in Figure 5. The previous testing scenario can be seen that the Bernoulli NB is the superior model. The results of this testing indicate that the addition of the Bernoulli NB model can produce a better hamming loss value. These results are because the Bernoulli NB model can provide good predictions so the majority of votes can give the same or even better voting results.
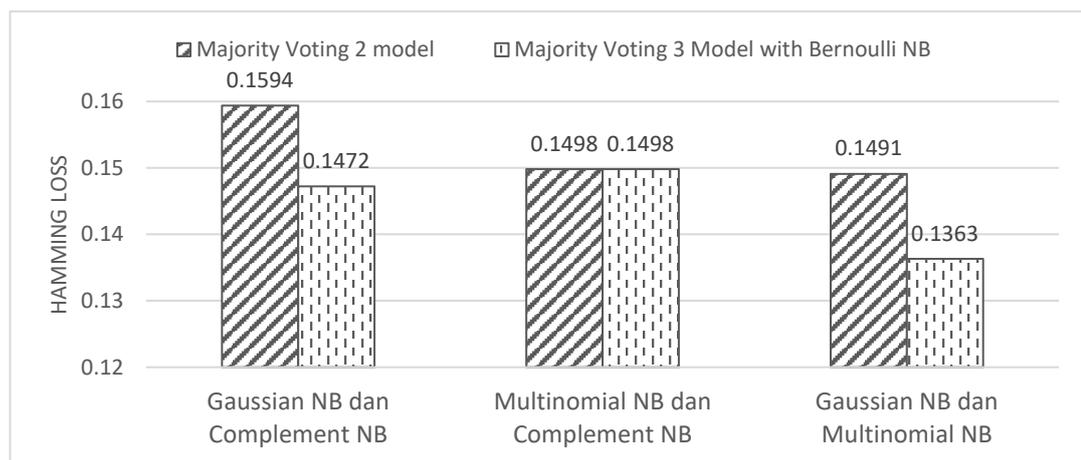
Figure 5. Comparison of Testing Results for Ensemble Method with 2 and 3 Models.

In the same testing scenario, the results show that the use of the ensemble method can indeed provide a better hamming loss value as shown in Figure 6. The more models that provide predictions for the ensemble method, the better the resulting hamming loss value will be. In certain cases, the ensemble method may not provide a better hamming loss value. This is because the models used are very far apart in providing the hamming loss value so that the hamming loss value may not be better, but it will not be too far from the model without the ensemble method.
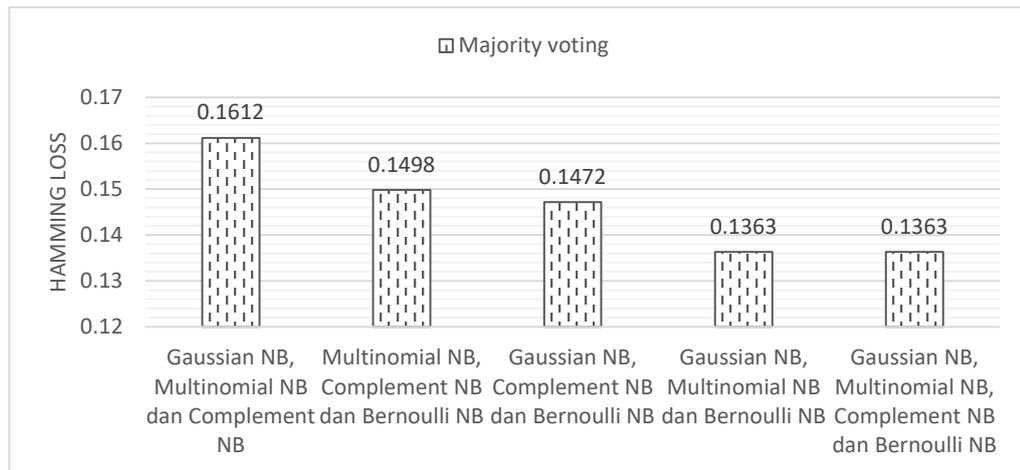
Figure 6. Comparison of Testing Results for Ensemble Method with 3 and 4 Models.

4. CONCLUSION

Based on the results of testing and analysis, this research has several conclusions. Referring to the testing scenario carried out, preprocessing the data without using stop words and stemming can affect the evaluation results of the multi-label text classification in the Al-Quran dataset for the better. The use of the ensemble method can also provide a better hamming loss value. The best hamming loss was obtained using the ensemble method (majority voting) between the multinomial NB and Bernoulli NB algorithm models with a hamming loss value of 0.1167. More models that used for predict the ensemble method (majority voting), will get the better performance. However, based on the analysis of the testing results, the ensemble method will be better if good model combined with good model also. Based on the research problem, a multi-label classification using ensemble method was built. This conclusion is certainly useful to make it easier muslim to learn Al-Quran. This conclusion This research also can give better performance from similar research before. For further research, the use of more diverse and more algorithm models is one way to find different performances. The use of algorithm models with similar quality can also provide different performance.

REFERENCES

[1]   N. S. Huda, M. S. Mubarok, and Adiwijaya, "A multi-label classification on topics of quranic verses (english translation) using backpropagation neural network with stochastic gradient descent and adam optimizer," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, hal. 1–5, 2019, doi: 10.1109/ICoICT.2019.8835362.

[2]   R. A. Pane, M. S. Mubarok, N. S. Huda, and Adiwijaya, "A multi-lable classification on topics of Quranic verses in English translation using multinomial naive bayes," *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, no. April 2019, hal. 481–484, 2018, doi: 10.1109/ICoICT.2018.8528777.

[3]   A. M. K. Izzaty, M. S. Mubarok, N. S. Huda, and Adiwijaya, "A multi-label classification on topics of quranic verses in English translation using Tree Augmented Naïve Bayes," *2018 6th Int. Conf. Inf. Commun. Technol. ICoICT 2018*, vol. 0, no. c, hal. 103–106, 2018, doi: 10.1109/ICoICT.2018.8528802.

[4]   G. I. Ulumudin, A. Adiwijaya, and M. S. Mubarok, "A multilabel classification on topics of qur'anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012026.

[5]   S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, hal. 48–59, 2018, doi: 10.1177/0165551516677946.

[6]   F. Gurcan, "Multi-Class Classification of Turkish Texts with Machine Learning Algotirthms," *ISMSIT 2018 - 2nd Int. Symp. Multidiscip. Stud. Innov. Technol. Proc.*, hal. 18–22, 2018, doi: 10.1109/ISMSIT.2018.8567307.

[7]   A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "Comparative analysis of text classification algorithms for automated labelling of Quranic verses," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 4, hal. 1419–1427, 2017, doi: 10.18517/ijaseit.7.4.2198.

[8]   M. Y. Abu Bakar, Adiwijaya, and S. Al Faraby, "Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation)Using Information Gain and Backpropagation Neural Network," *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, hal. 344–350, 2019, doi: 10.1109/IALP.2018.8629263.

[9]   H. Yin, Z. R. Yang, and R. Everson, "Preface," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3177, no. May, 2004, doi: 10.1007/978-3-540-28651-6.

[10] M. H. D. M. Ribeiro and L. dos Santos Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Appl. Soft Comput. J.*, vol. 86, hal. 105837, 2020, doi: 10.1016/j.asoc.2019.105837.

[11] A. R. Arrahimi, M. K. Ihsan, D. Kartini, M. R. Faisal, and F. Indriani, "Teknik Bagging And Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa," *J. Sains dan Inform.*, vol. 5, no. 1, hal. 21–30, 2019, doi: 10.34128/jsi.v5i1.171.

[12] A. W. Romadon, K. M. Lhaksmana, I. Kurniawan, and D. Richasdy, "Analyzing TF-IDF and Word Embedding for Implementing Automation in Job Interview Grading," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, hal. 1–4,

2020, doi: 10.1109/ICoICT49345.2020.9166364.

[13] Adiwijaya, Said Al Faraby, and Mohamad Syahrul Mubarok. 2021. Indonesian Translation of the Holy Quran (Multi-label), Dataverse Telkom University, DOI : https://doi.org/10.34820/FK2/XQCNPN

[14] T. B. Chandra, K. Verma, B. K. Singh, D. Jain, and S. S. Netam, "Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble," *Expert Syst. Appl.*, vol. 165, no. May 2020, hal. 113909, 2021, doi: 10.1016/j.eswa.2020.113909.

[15] M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, and J. L. Fernández Alemán, "Reviewing ensemble classification methods in breast cancer," *Comput. Methods Programs Biomed.*, vol. 177, hal. 89–112, 2019, doi: 10.1016/j.cmpb.2019.05.019.