Document Classification using Naïve Bayes for Indonesian Translation of the Quran

1st Syopiansyah Jaya Putra, 2nd Yuni Sugiarti, 3rd Galuh Dimas, 4th Muhamad Nur Gunawan, 5th Tata Sutabri, 6th Agung Suryatno
[1,2,3]*Information System Department Faculty of Science and Technology*
Syarif Hidayatullah State Islamic University
[4]*Faculty of Information Technology*
University of Respati Indonesia
[5]*Faculty of Computer*
University of MH. Thamrin
Jakarta, Indonesia
syopian@uinjkt.ac.id, yuni.sugiarti@uinjkt.ac.id, galuhfitriani84@gmail.com, nur.gunawan@uinjkt.ac.id, tata.sutabri@gmail.com, agungsr@ieee.org

***Abstract*—Classification for Indonesian language documents was increased. But the application of classification for question and answer system needs is still few. The purpose of this paper is to maximize the classification of Indonesian documents especially the Qur'an translation to support the question and answer system. In the process of creating a question and answer system that is still ongoing, testing the Naïve Bayes algorithm becomes very important besides other algorithms. The Naïve Bayes method is the first choice in this test as it has practicality in calculating. The result of this study is the classification of ITQ documents with 4 categories: morality, faith, knowledge, and Muamalah. The average accuracy rate of 90.5% indicates that the Naïve Bayes method is still relevant for use.**

Keywords—Classification, Text Mining, Naïve Bayes, Indonesian

I. INTRODUCTION

The development of digital documents has been very rapidly, characterized by the increasing number of data volumes on the Internet [1]. The information contained in digital documents can be accessed by search engines such as Google, Bing, and other tools [2-4]. However, not all these documents deserve a proper or valid reference source. Users need the tool again to parse the information according to their wishes. In terms of sorting information from digital document text, a classifying role is very useful [5-7].

Special Q&A (Question and Answer) system that develops today for digital documents that use Bahasa Indonesia with the translation of the Qur'an (ITQ) is still a little [8-15], the role of classification is much needed, especially in determine the topic or theme of the document.

The research on the classification of Quran translations has been conducted [16], the Decision Tree algorithm used to classify translations of Quranic verses in the category of science i.e. Biology, Physics, and Chemistry in Android applications and on applications created can display the paragraph classification information. The results of the success percentage of applications created using the decision tree are 75.73%. From this result, it is known that the decision tree can accelerate the process of classifying with many data.

Method Naïve Bayes Classifier for classification documents was conducted by researchers for example in the [17], the classification of documents can be reviewed from the process take action based on pre-existing data. Therefore, the classification of documents with this method can be personalized, the meaning is the process of document classification can be customized according to the nature and needs of each person.

Other research on Classification for Indonesian Text is partners searching website using Naïve Bayes methods [18]. The method was automatically categorized and classify unstructured data. The success rate of this method relies on the initial knowledge given, for example, the user enters personality data, such as follow: "I am an honest, cheerful, friendly, patient, and humorous person. This experiment managed to get the type of personality and found a partner based on personality type using the text mining method with Naïve Bayes for personality classification.

The other method to classification is using K-Nearest Neighbor (K-NN) [19], in this study which text documents used for the testing process were not based on the text document. However, the text documents used in the training and trial processes are taken from www.suaramerdeka.com, www.kompas.com, and www.detik.com.The Document text is selected manually, and their classes are written and saved in a text file format. This research data is six classes consisting of "disgust" (disgust), "shame" (shame), "angry" (angry), "sad" (sad), "Happy" (happy), and "Afraid" (fear).

The problems that occur today are the search results sometimes give very broad information, so that the information received by the user is less relevant after conducting experiments when classified data is slight. Then the less the accuracy of the resulting data, then also any translations that do not yet have a class label and some have more than one label (multi-label). This study aims to classify the translation of the Quran into four topics include morality, faith, knowledge, and Muamalah.

II. METHOD

A. Dataset

The Indonesian Translation of Al-Qur'an was used in this study [13]. The ITQ consist of 114 surah which divides into the training set and test set. The Training Set consists of 224 documents which have label manually into 4 topics; Morality, Faith, Knowledge, and *Muamalah* as seen in Table I.

TABLE I. TOPICS OF A TRAINING SET

No	Topics	Count
1	Morality	64

No	Topics	Count
2	Faith	90
3	Knowledge	22
4	Muamalah	48

B. Process

The classification process begins with extracting the translated documents, text preprocessing, Train Dataset, Test Dataset, and Classification, then processed using the Naïve Bayes algorithm as follows [17, 18, 20-25]:

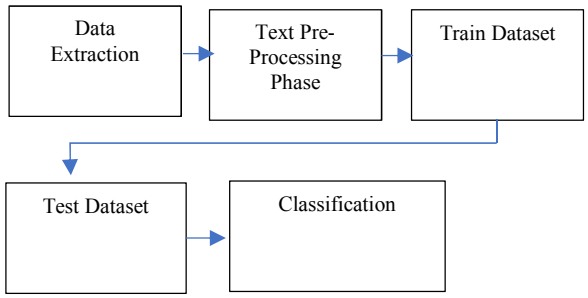

Fig.1 The Classification Process

C. Data Extraction

In this stage, the dataset starts from a whole file of the Qur'an translation obtained from a valid source that is the Ministry of religious affairs as an open dataset.

While the system in Python needs data already categorized in advance to retrained. Therefore, the data in the extract again becomes a part according to the predefined topic.

The extraction phase begins by separating the unnecessary label in the dataset, after that the data is converted to text file as an input to next phase which is text processing phase.

D. Text Preprocessing

This phase is the initial process for preparing the text into the data to be processed further. A set of connecting characters (text) should be broken by a more meaningful element. It can be done on several different levels. The data that has been translated is tokenized based on space and letters generate tokens and then the token result is in filtering/transformation and recorded the Term Frequencies-Inverse Document Frequencies (TF-IDF) value.

E. Train Dataset

In this process, the train data of 224 documents are labeled manually and divided into 4 topics. The number of train datasets is more than the test dataset, this is done so that the machine learns more about the data that will be tested, resulting in a high level of accuracy.

F. Test Dataset

At this stage data test as much as 216 documents from ITQ in test by machine using Naïve Bayes classifier. The Target of this data test is to produce 216 documents into 4 classifications according to the setting at the beginning of training.

G. Classification

At this stage classifications by Model Naïve Bayes are performed by calculating the number of possibilities or probability of each word – the word in the dataset compared to the existing document.

In this study using Multinomial Naïve Bayes as the basis of classification calculations, in this algorithm of distributed data in existing classes, data is represented as word vector.

At [20], Naïve Bayes classifier has a feature in each class independently, based on the formula:

$$P(X|C) = \prod_{i=1}^{n} P(X_i|C) \qquad (1)$$

Where,

X = (X_1...,X_n) is a feature vector, and C is a class.

$P(X|C)$ is a probability of X in class C.

On testing using Naïve Bayes Classifier, each test data will be counted and at once compared to train data at a time, resulting in faster testing.

H. Performance Measures

The algorithm used is in Python programming language with the settings of the for using Intel VGA, Windows operating system, and 8 GB memory. Training datasets use Python-customized libraries for Indonesian datasets.

In this study, performance measurements used precision, recall, F1 score, and accuracy. Based on the confusion matrix, there are four variables that determine in performance, namely:

- True Positive (TP): the documents which is correct and include in the result of classification.
- True Negatives (TN): the documents which is not correct but include in the result of classification.
- False Positives (FP): the documents which is correct, but not include in the result of classification
- False Negatives (FN): the documents which is not correct and not include in the result of classification

Performance formula based on 4 variables above is as follows [26]:

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

$$F1\ score = \frac{2TP}{2TP+FP+FN} \qquad (5)$$

Precision to measure the exact number of documents according to a category, recall measuring the number of appropriate documents, Accuracy to measure the accuracy of all the test dataset and F1 score to detect overall results. A

good result for the performance of classification according to [20, 21, 24, 25] is the average accuracy of more than 80%.

III. RESULTS AND DISCUSSION

Classification result of Indonesian documents with ITQ dataset is divided into two parts using classification 2 and 4 categories. The data test division can be seen in Table II.

The first phase classification uses two categories of morality and faith. While the second stage classification using four categories are morality, faith, knowledge, and Muamalah.

TABLE II. NUMBER OF TEST DATA

No	Topics	Classification 1	Classification 2
1	Morality	50	50
2	Faith	80	80
3	Knowledge	-	15
4	Muamalah	-	10

The division of two phases in this classification is to know the accuracy of the classification on the topic slightly compared to many topics and test the reliability of the method Naïve Bayes.

Based on the classification, the results of each stage are a table confusion matrix and accuracy table.

TABLE III. CONFUSION MATRIX – CLASSIFICATION 1

No	Topics	Precision	Recall	F1 Score
1	Morality	1.00	0.86	0.92
2	Faith	0.92	1.00	0.96

In Table III. The value of F1 Score shows that classifications with two topics produce F1 Score which is quite high above 0.9 and with a difference of 0.01.

TABLE IV. CONFUSION MATRIX – CLASSIFICATION 2

No	Topics	Precision	Recall	F1 Score
1	Morality	1.00	0.84	0.91
2	Faith	0.80	1.00	0.89
3	Knowledge	1.00	0.13	0.24
4	Muamalah	0.82	0.90	0.86

In Table IV. Results F1 Score varied and quite high in the classification with topics morality, faith, and Muamalah, but not for knowledge topics that have F1 Score 0.24.

TABLE V. ACCURACY

No	Measure	Classification 1	Classification 2
1	Accuracy	0.95	0.86
2	Macro Avg	0.94	0.72
3	Weighted Avg	0.95	0.83

In Table V. for the first Classification has Accuracy, macro AVG, and weighted avg of 0.95, 0.94, and 0.95, while the second Classification is 0.87, 0.73, and 0.85. This indicates that the classification of documents with fewer topics has higher accuracy tendencies. However, it also relates to the amount of test data and the data train used. In the second classification, the test data is more than the first classification, it is no surprise that its accuracy value is smaller.

In some references [17, 20, 24], the number of train data is drawn more than the test data, the more train data, the better the accuracy result. This means that machine is better trained to identify or predict test data on ITQ documents when the model uses more train data.

IV. CONCLUSION AND FUTURE WORKS

This study resulted in a document classification for Indonesian documents with a high accuracy rate of average 0.95. Classification with Naïve Bayes Model indicates that this model is still relevant to be used, although some shortcomings such as the smaller its accuracy value when the topic division more and more.

For further study, the optimization of classifications with Naïve Bayes should be done mainly in the selection of datasets for training and datasets for tests, as well as the text processing process.

REFERENCES

[1] Han, , and Chang, K.-C.: 'Data mining for web intelligence', Computer, 2002, 35, (11), pp. 64-70

[2] Graepel, T., Candela, J.Q., Borchert, T., and Herbrich, R.: 'Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine', (Omnipress, 2010)

[3] Killoran, J.B.: 'How to use search engine optimization techniques to increase website visibility', IEEE Transactions on professional communication, 2013, 56, (1), pp. 50-66

[4] Usmani, T.A., Pant, D., and Bhatt, A.K.: 'A comparative study of google and bing search engines in context of precision and relative recall parameter', International Journal on Computer Science and Engineering, 2012, 4, (1), pp. 21

[5] Nigam, K., McCallum, A.K., Thrun, S., and Mitchell, T.: 'Text classification from labeled and unlabeled documents using EM', Machine learning, 2000, 39, (2-3), pp. 103-134

[6] Khan, A., Baharudin, B., Lee, L.H., and Khan, K.: 'A review of machine learning algorithms for text-documents classification', Journal of advances in information technology, 2010, 1, (1), pp. 4-20

[7] Manning, C., Raghavan, P., and Schütze, H.: 'Introduction to information retrieval', Natural Language Engineering, 2010, 16, (1), pp. 100-103

[8] Naf'An, M.Z., Mahmudah, D.E., Putra, S.J., and Firmansyah, A.F.: 'Eliminating Unanswered Questions from Question Answering System for Khulafaa Al-Rashidin History', (IEEE, 2016, .), pp. 140-143

[9] Putra, S.J., and Gunawan, M.N.: 'Search by Concept for Indonesian Translation of Al-Qur'an'(UIN Syarif Hidayatullah Jakarta, 2017)

[10] Putra, S.J., Gusmita, R.H., Hulliyah, K., and Sukmana, H.T.: 'A semantic-based question answering system for indonesian translation of Quran', (ACM, 2016, .), pp. 504-507

[11] Putra, S.J., Hulliyah, K., Hakiem, N., Iswara, R.P., and Firmansyah, A.F.: 'Generating weighted vector for concepts in indonesian translation of Quran', (ACM, 2016, .), pp. 293-297

[12] Putra, S.J., and Khalil, I.: 'Context for the Intelligent Search of Information'2017 pp. Pages

[13] Putra, S.J., Mantoro, T., and Gunawan, M.N.: 'Text mining for Indonesian translation of the Quran: A systematic review', (IEEE, 2017 .), pp. 1-5
[14] Putra, S.J., Naf'an, M.Z., and Gunawan, M.N.: 'Improving the Scoring Process of Question Answering System in Indonesian Language Using Fuzzy Logic', (IEEE, 2018.), pp. 239-242
[15] Larasati, S.D., and Manurung, R.: 'Towards a semantic analysis of bahasa Indonesia for question answering', (2007, .), pp. 273-280
[16] Setiawati, D., Taufik, I., Jumadi, J., and Zulfikar, W.B.: 'Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile', Jurnal Online Informatika, 2016, 1, (1), pp. 24-27
[17] Samuel, N.: 'Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen', http://informatika. stei. itb. ac. id/~ rinaldi. munir, 2010, .), pp.
[18] Lestari, N.M.A., Putra, I., and Cahyawan, A.: 'Personality types classification for indonesian text in partners searching website using naïve bayes methods', IJCSI International Journal of Computer Science Issues, 2013, 10, pp. 1-8
[19] Purnama, K.E.: 'Classification of emotions in indonesian texts using K-NN method', International Journal of Information and Electronics Engineering, 2012, 2, (6), pp. 899-903
[20] Rish, I.: 'An empirical study of the naive Bayes classifier', (2001), pp. 41-46
[21] Ting, S., Ip, W., and Tsang, A.H.: 'Is Naive Bayes a good classifier for document classification', International Journal of Software Engineering and Its Applications, 2011, 5, (3), pp. 37-46
[22] Sutabri, T., Putra, S.J., Effendi, M.R., Gunawan, M.N., and Napitupulu, D.: 'Sentiment Analysis for Popular e-traveling Sites in Indonesia using Naive Bayes', (IEEE, 2018, .), pp. 1-4
[23] Zhang, H.: 'The optimality of Naïve Bayes. American Association for Artificial Intelligence' (2004.), pp.
[24] Zulfikar, W.B., Irfan, M., Alam, C.N., and Indra, M.: 'The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter', (IEEE, 2017,.), pp. 1-5
[25] Rennie, J.D., Shih, L., Teevan, J., and Karger, D.R.: 'Tackling the poor assumptions of naive bayes text classifiers'(2003, .), pp. 616-623
[26] Allehaibi, K. H. S., Nugroho, L. E., Lazuardi, L., Prabuwono, A. S., & Mantoro, T. (2019). Segmentation and Classification of Cervical Cells Using Deep Learning. IEEE Access, 7, 116925-116941.