International Journal on
Advanced Science
Engineering
Information Technology

Vol.10 (2020) No. 1
ISSN: 2088-5334

Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature

Mahendra Dwifebri Purbolaksono[#1], Feddy Dea Reskyadita[#2], Adiwijaya[#3], Arie Ardiyanti Suryani[#4], Arief Fatchul Huda[*]

[#] *School of Computing, Telkom University, School of Computing, Telkom University, Bandung, 40257, Indonesia*
E-mail: [1]mahendradp@telkomuniversity.ac.id; [2]feddydr@telkomuniversity.ac.id; [3]adiwijaya@telkomuniversity.ac.id; [4]ardiyanti@telkomuniversity.ac.id

[*] *Sunan Gunung Djati State Islamic University, Jl. A.H. Nasution No.105, Cipadung, Cibiru, Kota Bandung, 40614, Indonesia*
E-mail: afhuda@uinsgd.ac.id

***Abstract*— The second fundamental source of law for Moslems is the Hadith. The Hadith can be used to explain Quranic texts. However, Hadith still needs to be translated according to each national language to easily understand its meaning [1]. In Indonesia Hadith more usually refers to a special class of relevance to more particular religious concern [1]. Base on that, this research will Classify the translation Hadith Text into three classes: Obligation, Prohibition, and Information. From previous research, the Back Propagation Neural Network (BPNN) showed good performance in classifying hadith text. Therefore, BPNN was used to solve the problem of hadith text classification in this study. However, the dataset has a huge number of varied bag-of-words, which are features that will be used in the classification process. Hence, Information Gain (IG) was utilized to select influential features, and as the sequential process before the classification process. To measure the performance of this system, the Macro F1-Score was used. The F1-Score enables one to observe exactness from precision and completeness from recall. The Macro F1-score is also needed for the performance evaluation of more than two classes. Based on the experiment conducted, the system was able to classify hadith text using BPNN, IG, and without stemming, yielding the highest F1-score of 84.63%. However, the system performance that included the stemming process yielded an F1-score of 80.92%. This shows that the stemming process could decrease classification performance. This decreasing performance is due to some influential words merging with more noninfluential words.**

***Keywords*— feature selection; information gain; text mining; neural network; classification.**

I. Introduction

Hadith is the secondary holy text for Moslems after the Quran. Hadith contains guidance from the Prophet Muhammad (PBUH). Hadith is narrated by *Ulamas/Muftis*. Not only do these scholars have a great knowledge of Islam, they must also have good morals and *sanad* (inheritance). This research used the Hadith narrated by Al-Bukhari. Hadith originally wrote in Arabic text. There are still many Moslems that do not know about Hadith. However, Moslems must perform every obligation, and must not commit any prohibited acts. Not only that, by just knowing about Hadith, Moslems can reap many benefits. Hadith could also serve as an explanation for the Quran, which contains the words from *Allah Subhanahu wa ta'ala*. Therefore, all Moslems use Hadith to understand the law of Islam as well as the many instructions from the Prophet Muhammad (PBUH) [1]. Due to the many topics of the Hadith, it can be classified according to three main classifications, namely Obligation, Prohibition, and Information.

Moslems are spread out all over the world. All authentic Moslem holy books are written in the Arabic language. For better understanding and to reach a wider international audience, the books have been translated into many languages including Bahasa Indonesia [1]. Indonesia is the country with the biggest Moslem population in the world. To address this gap, this research focuses on classifying hadith text in the Indonesian Language. Many previous works have been conducted in Arabic and English, but in contrast, only a few works have been conducted in the Indonesian Language.

Performing classification of Indonesian-translated hadith text poses a different challenge compared to classification using English or Arabic. The Indonesian language differs in structure when compared to English and Arabic. Hence, the entire process will need a different treatment. In this research, a novel method was proposed to solve this problem. This research identified two problems regarding this dataset. First, the dataset contained some features in which not all had an impact on each class. Hence, more influential features should be selected. This was done using the Information Gain method. Second, this research focused on the classification problem using the Back Propagation Neural Network Classifier. One previous work [2] used this method and observed great system performance. However, the work focused on data mining like paper [3], while this research focuses on text mining. Nevertheless, both problems share some similarities. The similarity between the two is the huge dimension of the dataset. The other similarity between this research and the previous work [2] is that both execute the same task, which is data classification. Another research, Harrag [4], concluded that the Back Propagation Neural Network is better than SVM. Back Propagation is an extension of the Artificial Neural Network that changes weight in every epoch with the aim of improving system performance. Other than that, Back Propagation (BP) is predicted to have greater performance than the basic Neural Network. Classification method is not only method what run in this research, but also there are selection feature methods for text classification. Feature selection method can improve performance

Considering the facts above, Information Gain was used as the feature selection and Back Propagation as the classifier in this study. It is hoped that the proposed method will be able to solve the problem, which categorizes the pattern of the text into three classes, as mentioned previously.

II. Materials and Method

There are many researches on text learning that have utilized machine learning, both in English, Arabic, or Bahasa. One of the researchers that have utilized machine learning in text classification is Al-Kabi [5], who classified Bukhari traditions into 8 classes i.e. "Pray", "Eclipse", "Faith", "Call to Prayer", "Good Manners", "Knowledge", "Fasting", and "Almsgiving". In his research, great results were obtained and an average accuracy of 83.2% was achieved, influenced by stop word and stemming preprocessing and TF-IDF techniques.

Other research such as that of Afianto [6], conducted text classification on the Indonesian-translated Bukhari Hadith text using three classes of hadith. The classes are recommended hadith, prohibited traditions, and information hadith. In the research, the classification of hadith texts was based on Decision Tree with the Random Forest method. TF-IDF was also used as a technique to obtain the weight of values of each word. From the test, 90% performance was achieved. On the other hand, a similar study was conducted by Abu Bakar [7] in 2018 using Backpropagation Neural Network, in which the data was split into training data and testing data using the K-Fold technique. The average result of the performance obtained with Hamming Lost in the test was 0.1158 or 88.42% correctly classified.

Another research on text classification was done by Harrag [8], which used three stemming methods i.e. Root-Based stemming, Light-Stemming, and Dictionary-Lookup stemming, to improve the effectiveness and performance of the results of text classification. In his research, Artificial Neural Network (ANN) was used as a classifier because it had the advantage of being able to handle linear and non-linear problems for text categorization, where linear and non-linear classification has been proven to achieve good results [8]. Besides ANN, Support Vector Machine (SVM) was also used in the study because SVM is a statistical model that has powerful capabilities in handling very large feature sets. As a result, the stemming process was used as a feature reduction method in selecting the features to be used and the Macro-Average F1 Measure score for measuring performance was 94%.

In the paper [9], Jovic et. al. reviewing four different domains for Feature Selection. Those domains are Text Mining, Image Processing and Computer Vision, Industrial Application and juga Bio-informatics]. For Text Mining, there are two cases. That are Text Classification conducted by Forman et. al. [10] and Text Clustering conducted by Liu et. al. [11].

In research [10] conducted by Forman et. al., they compare six Feature Selection methods for classifying text data. Those methods are Probability Ratio (PR), Document Frequency (DF), Odds Ratio (Odds), Chi-Squared, Information Gain (IG) and Bi-Nomial Separation (BNS). Classification method that used is Support Vector Mechine (SVM) BNS got the highest accuracy and recall of that research, but not on precision. For precision and balance in all segment for all experiments, IG is the best from the others. IG also got best accuracy when feature or dimension below 200. Classification for text mining is similar with Classification on Data Mining. In research [12] which is research for Classification Microarray data, Mutual Information is used for Selection Feature method. That research used 5 datasets. The resulting average got 91.06% with F1-Measure. MI has a similar process with IG, but MI only focused on positive class and IG focused on both.

Liu et. al. did Text Clustering research [11] with compare method of Feature Selection too. That research is using 5 Feature Selection methods: Information Gain (IG); Chi-Square (Chi); Document Frequency (DF); Term Strength (TS); Term Contribution (TC). For clustering method, it used K-means with Entropy and Precision for evaluation methods. Chi and IG are the methods with Supervised approaching method. Although clustering is an unsupervised process, it can be processed with both of feature selection methods. The classes for selection feature got from gold standard include on the dataset. Result for experiment with the supervised approach of feature selection is IG better than Chi on both precision and entropy evaluation.

The related studies above show that the ANN classifier and Information Gain is suitable for our dataset.

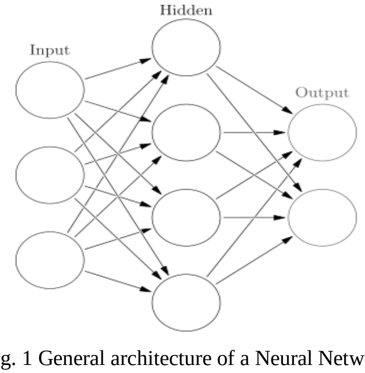

Fig. 1 General architecture of a Neural Network

Figure 1 shows the Neural Network architecture implemented in this research. The network architecture on

Neural Networks in this study used 1 hidden layer. To implement the entire process in this research, the process from the beginning of raw data processing to the testing of the data was conducted, as per Figure 2. There are three main parts of the process i.e. preprocessing, learning and testing, and accuracy calculation. All these processes are explained in the following flowchart of Figure 2.

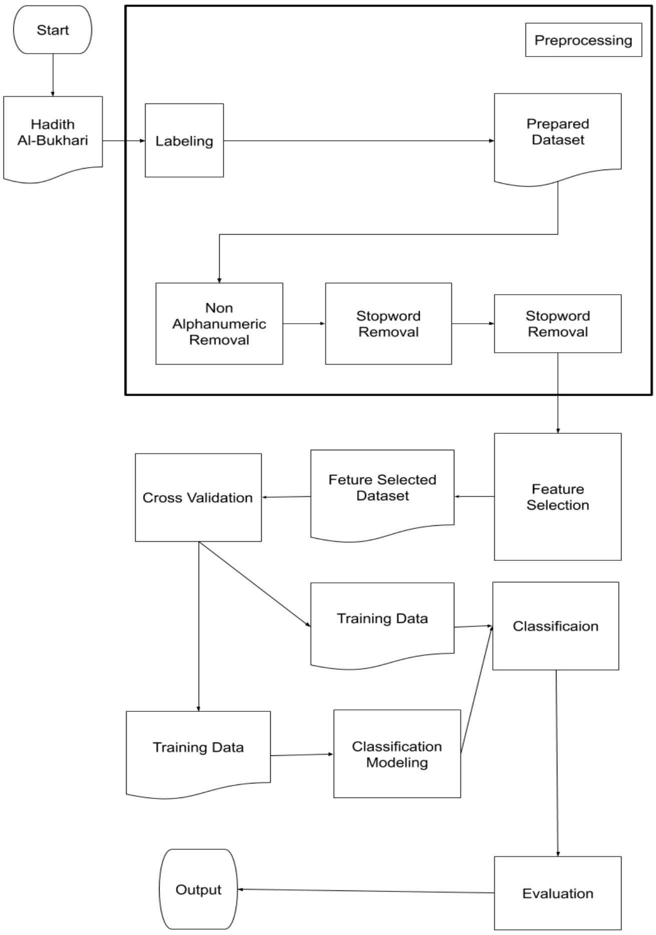

Fig. 2: Flowchart

A more detailed explanation of the processes in Figure 2:

1) *Preparation of raw data i.e. the Bukhari Hadith*: In this research, we use labeled into three categories: Obligation, Prohibition, and Information in Al-Bukhari Hadiths (1651 hadiths).

2) *Preprocessing the dataset:*

- Manual annotations of the dataset from the hadith expert judgement
- Removal of non-alphabet characters
- Stop word removal to eliminate meaningless words
- Stemming to eliminate additives so that only the basic words remain using *Sastrawi's* stemming model.

3) *Feature selection using the dataset to produce a classification*. The steps involved in computing the Information from the histogram of the training data are given below. The data set is arranged in the ascending order based on the output. The output class label (Y) is divide into two groups and the initial entropy H(Y) is calculated using:

$$H(Y) = -\sum_{j=1}^{Ny} P(Y\text{–}i).\log(P(Y\text{–}i)) \qquad (1)$$

The input genes (X) are divided into ten levels and their conditional entropies H(Y|X) are evaluated using:

$$H(Y|X) = -\sum_{i=1}^{Nx} P(X\text{–}i).$$

$$\sum_{j=1}^{Ny} P(Y\text{–}j|X\text{–}i).\log(P(Y\text{–}j|X\text{–}i)) \qquad (2)$$

$$\log(P(Y\text{–}j|X\text{–}i))$$

Next, the mutual information of each gene with respect to the output is computed using:

$$I\ (Y;\ X) = H(Y)-H(Y|X) \qquad (3)$$

Where is:
X-i = Input number-i {X$_1$: 2.5, X$_2$: 1.5 Xn: 2.25}
Y-i = Class number-i {Y$_1$: ALL, Y$_2$: AML}

The result of Mutual Information will arrange in the ascending. First attribute is the highest value and will be selected as the most informative gene.

4) *Cross Validation is the process for divide dataset into training and testing data by proposition that decided before.* Figure 3 show the illustrate of Cross Validation works.

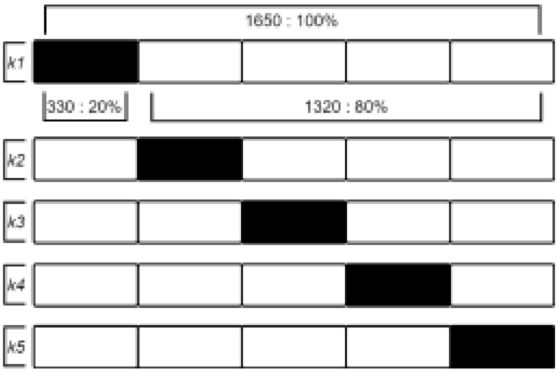

Fig. 3: Cross Validation works.

5) *Classification for classifying prepared data using the Backpropagation Neural Network*. In this research, we used 1 hidden layer with 10 nodes. Size of input layer's node is depending on output of feature selection term. For output layer, we used only a node which is a single class for every data.

6) *Testing Data and classification model output an evaluation.* The evaluation aims to assess the performance value of the system using F1-Measure. Here is formula for F1-Measure:

$$Precision = \frac{True\ Positive}{Total\ Predicted} \qquad (4)$$

$$Recall = \frac{True\ Posotive}{Total\ Target} \tag{5}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

The dataset based on the Bukhari Hadith text was annotated first with three classes. The classes are *"Larangan"* (Prohibition), "*Anjuran*" (Commend), and "Informasi" (Information). To determine the topic of the Hadith, the following text is given as an example:

> "*Seorang muslim adalah orang yang Kaum Muslimin selamat dari lisan dan tangannya, dan seorang Muhajir adalah orang yang meninggalkan apa yang dilarang oleh Allah. Abu Abdullah berkata; dan Abu Mu'awiyyah berkata; Telah menceritakan kepada kami Daud, dia adalah anak Ibnu Hind, dari 'Amir berkata; aku mendengar Abdullah, maksudnya ibnu 'Amru, dari Nabi Dan berkata Abdul A'laa dari Daud dari 'Amir dari Abdullah dari Nabi*"

After preprocessing, the sentences below are obtained:

> *"orang muslim adalah orang yang kaum muslimin selamat dari lisan dan tangan dan orang muhajir adalah orang yang tinggal apa yang larang oleh allah abu abdullah kata dan abu mu awiyyah kata telah cerita kepada kami daud dia adalah anak ibnu hind dari amir kata aku dengar abdullah maksud ibnu amru dari nabi dan kata abdul a laa dari daud dari amir dari abdullah dari nabi"*

The problem with the dataset is that there are many repetitive words for explaining the origin of the hadith. For example, *"...dari Nabi Dan berkata Abdul A'laa dari Daud dari 'Amir dari Abdullah dari Nabi"*. At the time of preprocessing, these words are deemed unimportant. Then, many names as Hadith carriers are also considered as new words, so that they are entered in the dictionary of words made.

Then, there are special treatments for several words to prevent changes in meaning that occur due to the stemming process. For example, with the word *"berikan"* (give), if stemming is done, it could change the meaning of the word, via cutting the additions that are not appropriate such that it becomes the basic word *"ikan"* (fish). This will affect all the meanings of a sentence and even the entire Hadith. Therefore, a few exceptions can be made to the stemming process, especially when working with the Indonesian language.

There are 1651 Hadith, which was divided into 1340 hadiths as learning data. Then, 310 hadiths were used as testing data. The classifications were then classified into three classes, namely command, prohibition, and information.

III. Result and Discussion

Testing scenario was done using BPNN; 2 scenarios were run. The first scenario involves observing the effect of Information Gain (Dimensional Reduction). The second scenario aims to observe the effect of Stemming. All scenarios used K-Fold Cross-Validation with K = 5.

A. First Scenario: Dimensional Reduction

In this scenario, feature selection was performed using Information Gain (IG). Information Gain is a useful method for reducing the term as a feature. Therefore, all the attributes can be observed by assessing the importance of each attribute. This scenario is important because it will solve the problem of text classification, which involves huge amounts of terms and thus measure the performance of IG itself. To ascertain the importance of IG, the system performance can be measured via several experiments with threshold variations. The results obtained are presented in Table 1

TABLE I
THE RESULT OF CLASSIFICATION WITH THE INFORMATION GAIN DIFFERENTIATOR

IG Threshold	K-Fold	F1 Macro Precision	F1 Macro Recall	F1 Macro Score	Max F1 Macro Score (%)
IG = 0	1	0.8217	0.8102	0.8159	81.99
	2	0.8142	0.8231	0.8187	
	3	0.8074	0.8114	0.8094	
	4	0.8192	0.8207	0.8199	
	5	0.8056	0.8122	0.8088	
IG = 0.5	1	0.8154	0.8037	0.8095	82.94
	2	0.8154	0.8282	0.8217	
	3	0.7880	0.7845	0.7862	
	4	0.8255	0.8333	0.8294	
	5	0.7996	0.7940	0.7968	
IG = 0.6	1	0.8269	0.8218	0.8244	84.64
	2	0.8339	0.8592	0.8464	
	3	0.7935	0.7899	0.7917	
	4	0.8290	0.8493	0.8391	
	5	0.8034	0.8002	0.8018	
IG = 0.7	1	0.8358	0.8298	0.8328	84.29
	2	0.8284	0.8579	0.8429	
	3	0.7898	0.7857	0.7878	
	4	0.8133	0.8272	0.8202	
	5	0.8225	0.8189	0.8207	
IG = 0.8	1	0.7948	0.7853	0.7901	83.98
	2	0.8328	0.8469	0.8398	
	3	0.7593	0.7557	0.7575	
	4	0.7525	0.7650	0.7588	
	5	0.7835	0.7790	0.7813	

The overall analysis obtained from analyzing each of the thresholds is outlined below:

- Increasing the IG threshold did not necessarily improve the performance result. This is shown by the performance results, where the performance results were not affected by the increment in value range. On the other hand, without IG (threshold IG = 0), performance decreased. Therefore, dimension reduction is still needed to obtain the important features (term). This is proven by the increased performance of up to 3%.
- There is a max point for the IG threshold = 0.6. At first, the performance increased until it reached this max point after which it decreased.

The analysis mentioned above is caused by different amounts of terms, which became a unique term for each K, and resulted in diverse data distribution.

B. Second Scenario: Stemming

In the second scenario, preprocessing was done with three schemas, which are classification with stemming, without stemming, and with stemming modification. The stemming method was based on *Sastrawi's* model (Indonesia's Stemming Model). This scenario was run to measure the performance of Stemming itself. Stemming modification eliminated some rules that exist in the stemming model itself. This was done to prove the extent of the effect of the affixes on the word and system performance. The results obtained are presented in Table 2.

TABLE II
THE RESULT OF CLASSIFICATION WITH STEMMING DIFFERENTIATOR

Stemming	K-Fold	F1 Macro Precision	F1 Macro Recall	F1 Macro Score	Max F1 Macro Score (%)
With Stemming	1	0.7956	0.7935	0.7946	80.92
	2	0.7998	0.8058	0.8027	
	3	0.7546	0.7550	0.7550	
	4	0.8094	0.8089	0.8092	
	5	0.7607	0.7681	0.7643	
No Stemming	1	0.8269	0.8218	0.8243	84.64
	2	0.8339	0.8592	0.8464	
	3	0.7935	0.7899	0.7917	
	4	0.8290	0.8493	0.8391	
	5	0.8034	0.8002	0.8018	
Stemming Modification	1	0.8191	0.8192	0.8192	81.92
	2	0.7998	0.8057	0.8027	
	3	0.7741	0.7667	0.7703	
	4	0.7967	0.7965	0.7966	
	5	0.7574	0.7640	0.7607	

Rules of stemming and could therefore improve performance. Based on this result, most of the words cannot be the same as its basic word. It is important to therefore conduct the feature classification process carefully.

IV. CONCLUSION

Based on the results of this research, the following conclusions are derived: The Back Propagation Neural Network classification method has been proven to work well for hadith text classification, yielding a maximum F1-Score of 84.63%. The stemming process was carried out based on the literature but the results were not as good as the process without stemming. This was because there were several words that if transformed into other words would change the meaning of the particular *hadith* and would eliminate some information from the words. This is proven from the process of Stemming Modification. Stemming modification obtained better performance than when using Full Stemming.

The evaluation process used K-Fold with K = 5. Each fold experienced changes that fluctuated quite a bit. Therefore, the use of K-Fold greatly affected the maximum results of the implementation process. All paragraphs must be justified alignment. With justified alignment, both sides of the paragraph are straight.

It is quite possible to further develop the system used today. Other feature selection methods should be explored, so that the terms obtained are more varied. SVM, Bayesian, and Decision Tree methods could affect the results obtained. The dataset that was also, and is still in use, is 1600 of the total Bukhari Hadith that numbers 7000 Hadith. Big data could affect the quality of the model produced during the training process.

REFERENCES

[1] K. A. Aldhlan, A. M. Zeki, A. M. Zeki, and H. A. Alreshidi, "Novel mechanism to improve hadith classifier performance," in *Proceedings - 2012 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2012*, 2013.

[2] H. Aydadenta and Adiwijaya, "On the classification techniques in data mining for microarray data classification," in *Journal of Physics: Conference Series*, 2018.

[3] S. Nurcahyo, F. Nhita, and Adiwijaya, "Rainfall prediction in kemayoran Jakarta using hybrid genetic algorithm (GA) and partially connected feedforward neural network (PCFNN)," in *2014 2nd International Conference on Information and Communication Technology, ICoICT 2014*, 2014.

[4] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Stemming as a feature reduction technique for Arabic text categorization," in *Proceedings of the 10th International Symposium on Programming and Systems, ISPS' 2011*, 2011.

[5] M. N. A.-K., G. K., R. A.-S., S. I. A.-S., and. R. S. A.-M., "Al-Hadith Text Classifier," *J. Appl. Sci.*, 2009.

[6] M. F. Afianto, Adiwijaya, and S. Al-Faraby, "Text Categorization on Hadith Sahih Al-Bukhari using Random Forest," in *Journal of Physics: Conference Series*, 2018.

[7] 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," in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 2019.

[8] F. Harrag and E. El-Qawasmah, "Neural network for Arabic text classification," in *2nd International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2009*, 2009.

[9] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*, 2015.

[10] G. Forman, I. Guyon, and A. Elisseeff, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, 2003.

[11] T. Liu, S. Liu, Z. Chen, and W. Ma, "An evaluation on feature selection for text clustering," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[12] M. D. Purbolaksono, K. C. Widiastuti, M. S. Mubarok, Adiwijaya, and F. A. Ma'ruf, "Implementation of mutual information and bayes theorem for classification microarray data," in *Journal of Physics: Conference Series*, 2018.