Text Categorization on Hadith Sahih Al-Bukhari using Random Forest

IOP Conf. Series: Journal of Physics: Conf. Series **971** (2018) 012037 doi:10.1088/1742-6596/971/1/012037

Text Categorization on Hadith Sahih Al-Bukhari using Random Forest

Muhammad Fauzan Afianto[1], Adiwijaya[2], Said Al-Faraby[3]
Telkom University

E-mail: fauzanafianto@student.telkomuniversity.ac.id[1], adiwijaya@telkomuniversity.ac.id[2], saidalfaraby@telkomuniversity.ac.id[3]

Abstract. Al-Hadith is a collection of words, deeds, provisions, and approvals of Rasulullah Shallallahu Alaihi wa Salam that becomes the second fundamental laws of Islam after Al-Qur'an. As a fundamental of Islam, Muslims must learn, memorize, and practice Al-Qur'an and Al-Hadith. One of venerable Imam which was also the narrator of Al-Hadith is Imam Bukhari. He spent over 16 years to compile about 2602 Hadith (without repetition) and over 7000 Hadith with repetition. Automatic text categorization is a task of developing software tools that able to classify text of hypertext document under pre-defined categories or subject code[1]. The algorithm that would be used is Random Forest, which is a development from Decision Tree. In this final project research, the author decided to make a system that able to categorize text document that contains Hadith that narrated by Imam Bukhari under several categories such as suggestion, prohibition, and information. As for the evaluation method, K-fold cross validation with F1-Score will be used and the result is 90%.

1. Introduction

Islam based on 2 fundamental laws : Al-Qur'ana as the set of words of Allah and Al-Hadith that documenting words, deeds, provisions, and approvals of Rasulullah as the prophet of Allah. Hadith was compiled and categorized by several Imam such as Imam Bukhari, Imam Muslim, and Imam Tirmidzi, etc. All of them based on one source Rasulullah Muhammad PBUH. Imam Bukhari is one of the known Imam that according to Ulama, Hadith that he narrated proven hold onto the most sahih among the other. Imam Bukhari spent 16 years on compiling about 2602 hadith (without repetition) and more than 7000 hadith (with repetition)[1]. Referring to[2], Figure 1 is the component of Hadith.

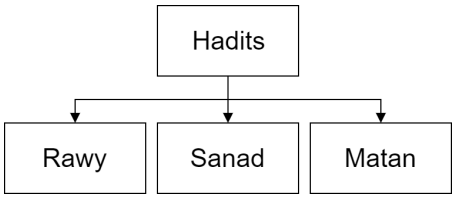

Figure 1. Hadith Components.

Sanad is the chain of conveyor of each Hadith, this component present on the beginning of

Hadith. Matan is the content of Hadith, present after the Sanad, and finally Rawy, this is the person or Imam that compile Hadith such as Imam Bukhari.

By the exponential growth of digitalized document, emerge the need of a system that able to extract high quality information, that is why automatic text categorization become popular.

Automatic text categorization is a task of developing software tools that able to classify text of hypertext document under pre-defined categories or subject code[1].

2. Literature Study

In this reseach, we only used 1650 Hadith from 7008 documented Hadith from "Hadits 9 Imam" from Lembaga Ilmu dan Dakwah serta Publikasi Sarana Keagamaan (LIDWA) in Bahasa Indonesia. Then we divide that into 3 category suggestion, prohibition, and information with 550 record each. As for the label we used hand labelling.

as for the related study, there are several paper that used hadith as their documents such as:

Table 1. Related Works

No	*Author*	*Method*	*Accuracy*
1	Harrag et al. (2011b)[4]	Dictionary-lookup Stemming, and ANN	94%
2	Harrag et al. (2009)[3]	ID3 Decission Tree	81%
3	Najeeb (2014)[5]	Associative Classication	72%

Each of them using different dataset in English Language. In the research by Harrag et al. (2011b)[4]. They use various stamming method in preprocessing. As for the best method they use dictionary lookup-stemming, it is proven to reduce 4055 variance of words into 739. as for the class, they divide the hadith based on 14 class and use ANN classifier, the result is 94% Accuracy.

As for the research by Harrag et al. (2009)[3]. ID3 classifier used with the highest accuracy of 81% while Najeeb (2014)[5] decided to classify Hadith by their authentic, so the class is binary either the hadith is not authentic (daif) or authentic (sahih). They used Association Rule Mining (ARM) and yield 72% as their highest accuracy.

The machine learning method that used in this research is Random Forest. This method developed from Decision Tree that known able to overcome the weakness of Decision Tree, Overfitting. The Decision Tree that used is Classification and Regression Tree (CART) that build By using Gini Index[6]. Basically, this method build lots of CART then doing voting to decide the which category that has the most amount from them. The creation of CART itself use bootstrapping method, in this research we call the bootstrap tree.

to train the Random Forest, we have to select which feature that will be extracted from the documents hence, Term Frequency-Inverse Document Frequency (TF-IDF) used[7].. By using this Feature Extraction, a sparse matrix will be built. the topmost of that matrix will be filled by bag of words and the leftmost will be filled with list of document name, the rest will be filled by the value of TF-IDF that gained by Equation 1

$$Word_{xy} = TermFrequency_{xy} * log \frac{SumDocument}{TermFrequencyDocument_{xy}} \quad (1)$$

Before we extract the feature form Hadith, we have to reduce the variance of each word there, by using the following preprocessing, stemming, case folding, punctuation removal, and stopwords removal.

To evaluate, we decided to use K-Fold Cross Validation to observe the overall F1-Score. This mechanism divide the data training to by the amount of K that depicted in Figure 2

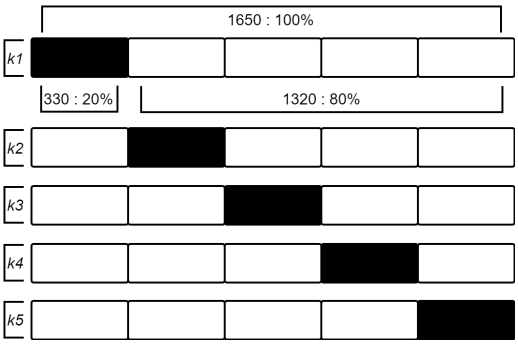

Figure 2. K-Fold Cross Validation.

So, each selected K part would be data testing while the rest become data training. Then the average of F!-Score would be calculated to observe the final scoring.

3. Development and Analyze

In this research, we build the system that able to categorize hadith into 3 category suggestion, prohibition, and information. Here is the figure t represent every step that would be done by the system.

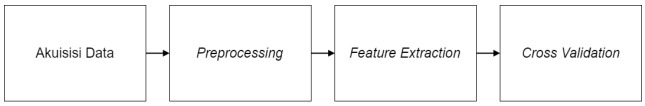

Figure 3. Block Diagram.

According to the Figure 3 The data would go to the preprocessing mechanism to reduce the variance of words, in preprocessing there will be several sub mechanism such as case folding, stemming, punctuation removal, and stopwords removal that proven good enough to reducing the variance of data in Bahasa Indonesia.

after that, the preprocessed data needs to represented by the feature that able to used as the input for classifier. In this research we represent the preprocessed data as sparse matrix with the topmost row filled with bag of word, and the leftmost column filled with the document number, the rest would be filled with the value of words that calculated using TF-IDF formula.

Then, the extracted features ready to be learned by the Random Forest classifier. the amount bootstrap tree would be generated. The evaluation using K-Fold Cross Validation also happened in conjunction with the creation of Random Forest classifier.

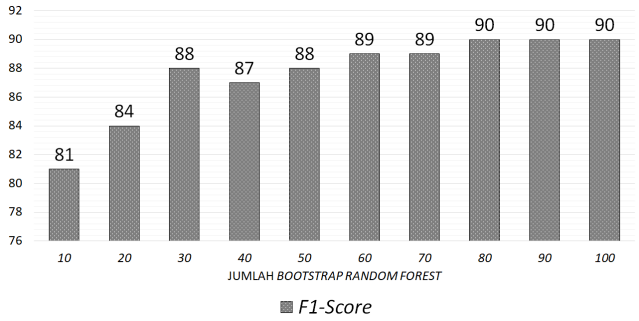

Figure 4. Amount of Bootstrap Tree

Based on Figure 4 The purpose of this scenario is to observe the influence of bootstrap tree towards the score of F1-Score. we divide each test with the addition of 10 bootstrap tree. by the time the amount of bootstrap tree reach 80 there are no improvement of F1-Score, from this we can see that the amount of bootstrap tree able to yield significant effect.

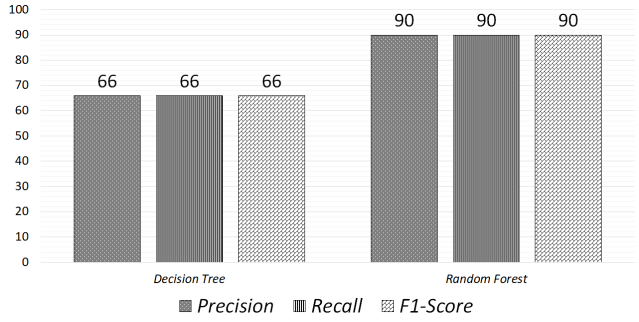

Figure 5. Decision Tree and Random Forest Comparison

The purpose of this scenario is to observe the potency of Random Forest towards Decision Tree as depicted on Figure 5 . As we can see the difference of F1-score while using K-Fold Cross Validation with 5 as the Amount of K in respect of 2:8 of testing and training ratio, yield 66% : 90% of the F1-Score. This is prove that in categorizing Hadith Sahih Al-Bukhari Random Forest able to overcome the weakness of Decision Tree.

4. Conclusion

In conclusion automatic text categorization on Hadith Sahih Al-Bukhari using Random Forest with sample data of 1650 hadith with 550 hadith of each category using TF-IDF as Feature Extraction and 100 as the amount of bootstrap tree evaluated by K-Fold Cross Validation with 5 as the amount of Fold, yield 90% of F1-Score.

The most influential mechanism is the amount of bootstrap tree on the Random Forest, while the rest does not really significant. Also in this case Random Forest also proven able to overcome the weakness of Decision tree by the ratio of 90% : 66% of F1-Score using K-Fold Cross Validation.

As for the further research it would be good if all of the hadith used, other than that combining hadith from the other Imam such as Imam Muslim, Imam Tirmidzi, etc. would be a good idea since the variance would be more.

5. References

[1] Naji Al-Kabi, M., Kanaan, G., Al-Shalabi, R., Al-Sinjilawi, S.I. and Al-Mustafa, R.S., 2005. Al-Hadith text classifier. Journal of Applied Sciences, 5, pp.584-587.

[2] Saloot, M.A., Idris, N., Mahmud, R., Jaafar, S., Thorleuchter, D. and Gani, A., 2016. Hadith data mining and classification: a comparative analysis. Artificial Intelligence Review, 46(1), pp.113-128.

[3] Harrag, F., El-Qawasmeh, E. and Pichappan, P., 2009, July. Improving Arabic text categorization using decision trees. In Networked Digital Technologies, 2009. NDT'09. First International Conference on (pp. 110-115). IEEE.

[4] Harrag, F., El-Qawasmah, E. and Al-Salman, A.M.S., 2011, April. Stemming as a feature reduction technique for Arabic text categorization. In Programming and Systems (ISPS), 2011 10th International Symposium on (pp. 128-133). IEEE.

[5] Najeeb, M.M., 2014. Towards innovative system for Hadith Isnad processing. Int J Comput Trends Technol, 18(6), pp.257-259.

[6] Arifin, A.H.R.Z., Mubarok, M.S. and Adiwijaya, A., 2016, September. Learning Struktur Bayesian Networks menggunakan Novel Modified Binary Differential Evolution pada Klasifikasi Data. In Indonesia Symposium on Computing (IndoSC) 2016.

[7] Aziz, R.A., Mubarok, M.S. and Adiwijaya, A., 2016, September. Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. In Indonesia Symposium on Computing (IndoSC) 2016.

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