Comparison of Decision Tree C4.5 Algorithm with K-Nearest Neighbor (KNN) Algorithm in Hadith Classification

Glen Nur Awaludin[1,7], Yana Aditia Gerhana[1,2,8], Dian Sa'adillah Maylawati[1,5,9], Wahyudin Darmalaksana[3,10], Nunik Destria Arianti[4,5,11], Ali Rahman[1,6,12], Muhamad Musli[4,13]
[1]Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
[2]Department of Information and Communication Technology, Asia e-University, Malaysia
[3]Department of Ilmu Hadits, UIN Sunan Gunung Djati Bandung, Indonesia
[4]Department of Informatics, Universitas Nusaputra, Indonesia
[5]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[6]School of Informatic Management and Computer, LIKMI Bandung, Indonesia

e-mail: [7]1157050062@gmail.com , [8]yanagerhana@uinsgd.ac.id, [9]diansm@uinsgd.ac.id, [10]yudi_darma@uinsgd.ac.id, [11]nunik@nusaputra.ac.id, [12]ali@uinsgd.ac.id, [13]muhamad.muslih@nusaputra.ac.id

2020 6th International Conference on Computing Engineering and Design (ICCED) | 978-1-7281-9189-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCED51276.2020.9415796

Abstract—**Previous scholars always made an effort to make various formulations that were used to categorize and calcify hadith. At present, the process of categorization or classification is facilitated by the process of text mining technology. In the study of text mining itself, there are various kinds of tools and methods or algorithms that can be used and also help provide maximum results in the process of mining information from a text. An example is the Decision Tree C4.5 and K-Nearest Neighbor algorithm. Based on that, the author wants to make research and this final project to compare the performance resulting from the classification process of text documents using Decision Tree C4.5 and K-Nearest Neighbor algorithm for the classification of Imam At-Tirmidzi hadith. With this research, it is expected to be knowledgeable about the process of classifying text documents along with the performance of the two algorithms. Based on testing that has been done, the Decision Tree C4.5 algorithm produces an average accuracy value of 70.53% with an average processing time of 0.083 seconds. While the K-Nearest Neighbor algorithm produces an average accuracy value of 66.36% with an average processing time of 0,03 seconds.**

Keywords—Decision Tree C4.5, K-Nearest Neighbor, Text Mining, Classification, Hadith

I. INTRODUCTION

The Muslim scholars try to make a concept formulation that can be used as a guide in the process of selection and categorization of the Hadith so that later Muslims can distinguish which Hadith really originated from the Prophet Muhammad (*shahih*), which Hadiths have weak associations (*dhaif*) and where the Hadith has no validity at all (*maudhu'*). There are so many methods or techniques used by the previous Muslim scholars in carrying out the process of categorizing the Hadith. For example, Imam At-Tirmidzi has a method of taking the Hadith (showing the place of the Hadith from the original source, and giving an explanation related to the law) which is the practice of Fuqaha (*Fiqh* Expert), as well as providing an explanation of the quality and state of the narrated Hadith [1].

Data mining is not limited to business can be used by businesses in many ways, data mining involves statistical and/or artificial intelligence analysis, usually applied to large-scale data sets. Traditional statistical analysis involves an approach that is usually directed, in that a specific set of expected outcomes exist [2]. The categorization process is currently facilitated by information technology that makes it easy for users to obtain the information needed. One of them is categorization in a text document. Categorization in a text document is an effort that aims to divide text documents into several groups of documents in the form of text and then put into certain categories based on the suitability of the characteristics and characteristics of the theme or topic associated with the document. To study and develop the text categorization process there is a discipline that has a focus on studying the process of categorizing and grouping text documents, namely Text Mining [3].

The implementation process of text mining is facilitated by a variety of methods or algorithms that already exist. However, from several methods or algorithms that can be used in the text mining process, it is chosen that has good performance. In the process of selecting a method or algorithm, it can be done by measuring the efficiency of the time produced in one process as well as the level of accuracy that results from that process. The method or algorithm that is often used for the purposes of the classification process is the Decision Tree C4.5 algorithm [4]–[6] and K-Nearest Neighbor [7], [8]. So in this study, a process was carried out to make a comparison of the performance of the two methods or algorithms. This study also has the aims to identify the best method or algorithm between two choices, the Decision Tree C4.5 algorithm and K-Nearest Neighbor. Decision Tree C4.5 and K-Nearest Neighbor algorithm were chosen because these two algorithms are including algorithms that are quite often used for the classification process and there are no studies that explain the comparative performance of the two algorithms.

The data used in this study were sourced from the translation of Imam At-Tirmidzi's Hadith. The process of classifying text documents will only be limited to the process of classification of Hadith from the history of Imam At-Tirmidzi. In this study, what will be used as an assessment parameter is the level of accuracy in the classification process and the amount of time needed to carry out the classification process as parameter for comparison of the performance of the two algorithms. And will be classified according to the class Recommendations, Prohibitions, and Information.

II. Research Method

A. Research Stage

1) Process Steps in Research

The flowchart or sequence of steps that will be carried out in the study to compare the performance of the Decision Tree C4.5 Algorithm and K-Nearest Neighbor is as in the Figure 1.

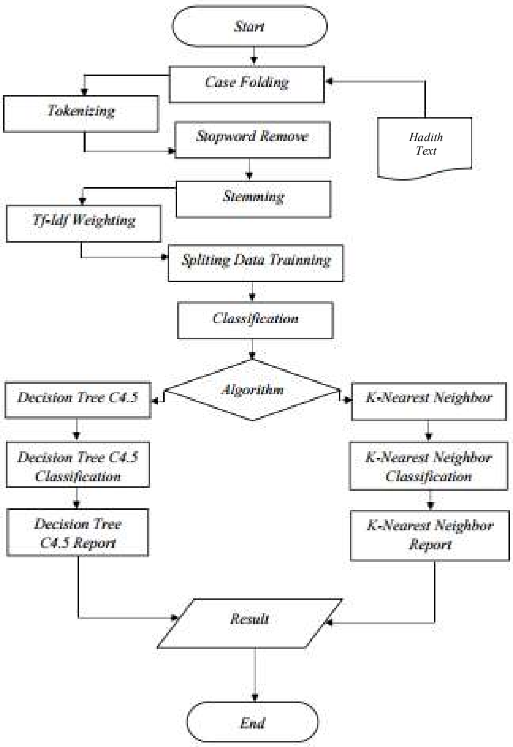

Fig. 1 Research Stages

2) Data Preparation Process

This process is carried out with the aim of ensuring the quality of the dataset used as research objects and is expected to later help provide maximum results to the classification process. This process is done by taking the contents of the Hadith text documents and giving them classes manually based on the classes that have been determined. The process is represented in Figure 2.

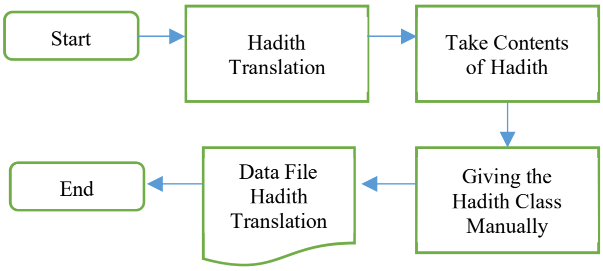
Start
Hadith Translation
Take Contents of Hadith
Giving the Hadith Class Manually
Data File Hadith Translation
End

Fig. 2 Data Preparation

B. System Architecture

The representation of the system architecture in this study was made in order to make it easier to understand the flow of the system that will work later. The system architecture is as shown in Figure 1.3 below:

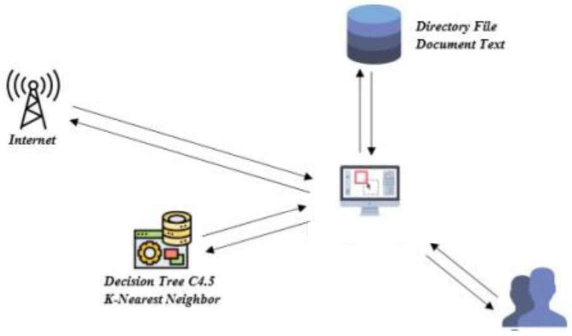

Fig. 3 Architecture System

The architecture represented has several elements or components which are connected to each other which have the aim to carry out instructions and provide information about the classification process of Imam At-Tirmidzi's Hadith. The explanation of the elements or components in the Figure 3 are as follows:

a. Users, are entities that act as subjects who will use and run software that will be created with the aim of gaining knowledge about the process of classification of Hadith text documents.

b. Computer Devices, an entity in which there is software created and developed for the classification of Hadith text documents.

c. Directory File Document Text, is an entity to ensure that the text documents that will be used as research objects can be accessed and provide useful information for the research to be conducted.

d. The Internet, is an entity that has the role of an intermediary or liaison to ensure that the necessary libraries can be available and help the process run adequately.

e. Decision Tree C4.5 dan K-Nearest Neighbor, is an entity that acts as a method or algorithm used in the process of classification of Hadith text documents. This method or algorithm will carry out the process of managing text document data through mathematical calculations aimed at producing the appropriate information. These two methods or algorithms will be seen and compared to their performance in classifying Hadith text documents.

C. Decision Tree Algorithm C4.5

Decision Tree C4.5 algorithm is an algorithm used in making a decision tree or Decision Tree [5], [6], [9]. This decision tree is made or built by carrying out the process of dividing the values and attributes that exist into a branch for each of the 4 possibilities that can occur [6]. The following are some steps that must be passed to build a decision tree using the decision tree C4.5 algorithm:

1) Select an attribute as root

This selection is based on the highest gain value of all available attributes. To carry out the process of calculating the highest gain an equation is used below:

$$Gain\ (S,A) = Entropy(S) - \sum_{i-1}^{n} \frac{|Si|}{|S|}$$

where:
S: as a Case Set
A: as an Attribute
n: as Attribute Partition Number A
|Si|: as the Number of Cases in Partition A
|S| : as the number of cases in S

For Entropy value can be calculated with the following equation:

$$Entropy(S) = \sum_{i-1}^{n} - pi * log_2 pi$$

where:
S: as a Case Set
n: as the Number of Partitions
Pi: as a Proportion of Si to S

2) Make a branch for each value.
3) Divide existing cases into a branch.
4) Repeats the entire process for each branch until all cases in the branch have the same class

D. K-Nearest Neighbor (KNN) Algorithm

K-Nearest Neighbor algorithm or commonly known as the KNN algorithm is an algorithm that is often used to carry out the classification process of an object based on learning data that has the closest distance to the object. KNN is an algorithm that classifies based on the proximity of the location (distance) that is owned by a data with other data [10]. Several major kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques [11].

The steps in KNN itself are more concise as follows [7]:

1) Enter training data, test data, and k values.
2) Calculate the weights for each term of the training data and test data using the TF-IDF technique.
3) Calculate the similarity held by each training data that has been previously classified with the test data. The similarity of training data with test data can be calculated using equation Cosine Similarity as follows:

$$\cos(\Theta_{ij}) = \frac{\Sigma_k (d_{ik} d_{jk})}{\sqrt{\Sigma_k d_{ik}^2}\sqrt{\Sigma_k d_{jk}^2}}$$

The process can be carried out with the following equation [7]:

$$d(i,j)\sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{in} - x_{in}\right)^2}$$

Information:
$d(i,j)$ = Distance Value
x_i = The values contained in feature 1
x_j = Values contained in feature 2

4) Sorts the results of the calculation of the similarity value.
5) Taking k training data as much as k and has the highest similarity value with the data test.
6) Determine the class based on the class that has the most dominant value of the k test data that has been taken previously (Based on Euclidean values).

E. Term Frequency Invers Document Frequency (TF-IDF)

Term Frequency Invers Document Frequency (TF-IDF) is a method that is often used in the process of determining how far the relevance of a word (term) to a document by giving weight to each word that exists as a whole[12]. Calculated in the following equation as follow

$$tf_{t,d} = \log f_{t,d} + 1 \qquad if \quad f_{t,d} > 0$$

Information:
$tf_{t,d}$ = Term Frequency
$f_{t,d}$ = d Number of occurrences of words / terms t in the document d

Whereas IDF is calculated with the equation below:

$$idf_t = log \frac{N}{N_t}$$

Where:
idf_t = Inverse Document Frequency
N = Total number of documents
N_t = Number of documents containing term t

F. Confusion Matrix

Confusion Matrix is a tool that can be used as a tool to represent the performance results of Supervised Learning and is useful to provide knowledge about the actual information and prediction information generated from a learning process. In general, measurements in measurements using this confusion matrix have a benchmark on the combination of precision and recall[13]. To calculate the value of Precision the following equation is used:

$$Precision = \frac{TP}{TP + FP}$$

Where :

TP = True Positive
FP = False Positive

Meanwhile, to calculate the Recall value, the following equation is used:

$$Recall = \frac{TP}{TP + FN}$$

Where :
TP = True Positive
FN = False Negative

In addition, by using this confusion matrix also calculated the value of accuracy and F-Score. To calculate the accuracy value used equation as follows:

$$Accuracy : \frac{(TP + FN)}{(TP + TN + FP + FN)}$$

Where :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Meanwhile, for the calculation of the F-Score used equation (1.10):

$$Accuracy : \frac{(TP + FN)}{(TP + TN + FP + FN)}$$

G. Hadits

Hadith is everything passed down through or to the Prophet Muhammad which is in the form of words, deeds, attitudes (*Taqrir*) and so on [14]. The position of the Hadith or As-Sunnah itself in Islam has two main functions, including [15]:

1) *Mubbayyin* is anything that functions as an explanatory or specific representation of all kinds of things that are stated globally or publicly in the Qur'an. For example, it gives a specific explanation of the procedures for the prayer, fasting, pilgrimage, and other worship.
2) Separate sources of law are domiciled as sources of law for all matters that are not explained or discussed in the Qur'an in general or specifically. Examples are the law prohibiting marriage by polygamy between a niece and sister/brother of parents and the law prohibits eating all kinds of animals that have fangs, claws, and others.

H. Text Mining

Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization [16]. Many research on text mining prove multiple of words better than bag of words, among others sequence of words, among them is text mining [17]. Text data mining or text mining is a process of finding or extracting information where the data comes from a collection of texts by the user by utilizing certain tools for analysis purposes.

Text mining is the process of seeking or extracting the useful information from the textual data. It is an exciting research area as it tries to discover knowledge from unstructured texts. It is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT)[18]. Text mining adopts the processes contained in Data Mining in general which will be used in the text mining process. In addition, in-text mining there are also similar techniques used in data mining. The steps or stages contained in the text mining process include:

1) Case Folding, in this step the changes are made to all character letters contained in the data in the form of lowercase letters.
2) Tokenizing, in the tokenizing stage, a decomposition process is carried out on the description which was originally in the form of a sentence into a word form. As well as in the tokenizing process, a delimiting process in words such as periods (.), Commas (,) and spaces (non-punctuation separators) and numeric characters in words is used as data.
3) Stopword, in the stopword stage a process is carried out to eliminate words that do not have a value or are said to be not descriptive and if removed do not change the content or meaning of the sentence itself.
4) Stemming, in the stage of stemming, a process of searching for a root or root words are generated in the previous stopword stage. One library that can be used to help run the Stemming process for Indonesian translation is the Literature library that uses the Nazief and Adriani Algorithms.

III. Result and Discussion

A. Testing

In the testing process that has the objective to see the performance produced by the Decision Tree C4.5 algorithm and K-Nearest Neighbor to predict or classify the Imam At-Tirmidzi Hadith text documents, in this study several classification testing processes were carried out using 1000 translation data Hadith Imam At-Tirmidzi. In the process the classification process is carried out with 5 (five) combinations of training data and testing data which for determining the training data and testing data are carried out randomly or randomly as follows:

1) Using a combination of 900 Training data and 100 Testing data
2) Using a Combination of 800 Training data and 200 Testing data
3) Using a combination of 700 Training data and 300 Testing data
4) Using a Combination of 600 Training data and 400 Testing data
5) Using a Combination of 500 Training data and 500 Testing data

B. Discussion

The results of testing the two algorithms provide performance comparison information, for comparison of processing time values can be seen in Figure 4.

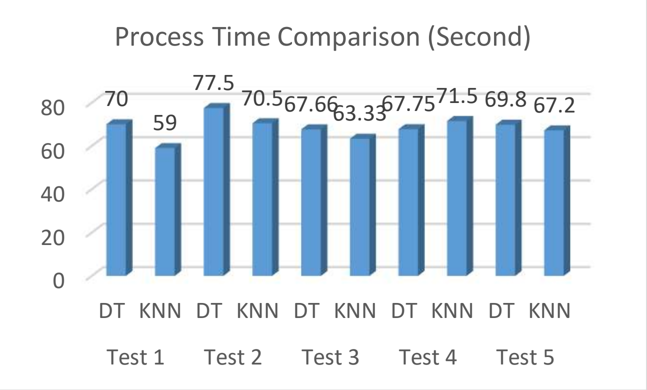

Fig. 4 Process Time Comparison

Based on Figure 4, the processing time generated by the Decision Tree C4.5 and K-Nearest Neighbor algorithm shows that the resulting processing time is very diverse. These factors may be influenced by the pattern of the Decision Tree C4.5 algorithm process which must make a decision tree from the learning process conducted on training data. So, if there is a large amount of training data, it will also take additional time to compute the decision tree making to determine the Class prediction or classification of the testing data.

Whereas the processing time generated by the K-Nearest Neighbor algorithm is influenced by the number of learning processes that must be carried out by the K-Nearest Neighbor algorithm in making predictions or classifications, which is the process of calculating the similarity of training data documents and testing data which if the testing data is used has a considerable amount, it will add to the resulting processing time. In addition, the resulting processing time is also inseparable from the supporting device factors. One of them is the hardware used in this study. As for the comparison of the level of accuracy can be seen in Figure 5.

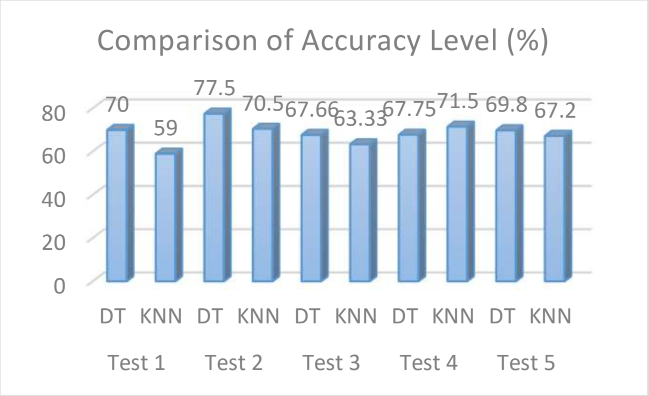

Fig. 5 Comparison of Accuracy Level

Figure 5 shows the level of accuracy of all the tests that have been done. The level of accuracy generated by the Decision Tree C4.5 algorithm itself produces an average accuracy rate of 70.53% of 5 (Five) Times of testing with different combinations of training data and testing data. Whereas the level of accuracy produced by the K-Nearest Neighbor algorithm itself produces a lower average with an accuracy rate of 66.36% of the number of tests and the combination of training data and testing data is the same as the Decision Tree C4.5 algorithm. The resulting accuracy value has an average result that is not a too significant difference. The accuracy value can be influenced by errors resulting from the process of providing Label Class at the Data Preparation stage so that it influences the results of the predictions produced. Resulting in a reduction in the value of the resulting accuracy value.

IV. CONCLUSION

1) A Comparison of the performance of the Decision Tree C4.5 algorithm and K-Nearest Neighbor for the classification of Imam At-Tirmidzi's Hadith has been successfully carried out. The way to do this is to prepare a Hadith translation text document that first goes through the Data Preparation process to better ensure the quality of the text documents used as well as giving Class labels manually. Before entering the prediction or classification stage then the text document passes the Preprocessing stage to retrieve terms or words that are considered to have value. Then the weighting by weighting Tf-Idf so that each term or word in the document can be processed at the stage of prediction or classification.
2) In the process, the text documents used will be divided into training data and testing with some predetermined combination of compositions. After the process is successfully passed, then the text document will then go through a prediction or classification process using the Decision Tree C4.5 algorithm and K-Nearest Neighbor. And the end result is testing data that already has a prediction class label and detailed performance results generated by the Decision Tree C4.5 algorithm and K-Nearest Neighbor.
3) The accuracy produced by the Decision Tree C4.5 algorithm has an average of 70.53% and the accuracy rate produced by the K-Nearest Neighbor algorithm has an average of 66.36%. Whereas the average processing time produced by the Decision Tree C4.5 algorithm is 0.083 Seconds, while the average processing time produced by the K-Nearest Neighbor algorithm is 0.033 Seconds.
4) For the further works, the classification of hadith can use the others data mining algorithm, even use the deep learning, with more dataset of hadith.

REFERENCES

[1] H. Su'aidi, "Mengenal Kitab Sunan Al-Tirmidzi (Kitab Hadits Hasan)," *Religia*, vol. 13, no. 1, pp. 123–137, 2017, doi: 10.28918/religia.v13i1.178.

[2] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[3] S. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-Nearest Neighbor," *J. Sains dan Seni ITS*, vol. 5, no. 2, 2016, doi: 10.12962/j23373520.v5i2.16643.

[4] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," *J. Online Inform.*, 2017, doi: 10.15575/join.v2i1.71.

[5] E. Darmawan, "C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education," *J. Online Inform.*, vol. 3, no. 1, p. 22, 2018, doi: 10.15575/join.v3i1.171.

[6] D. Setiawati, I. Taufik, J. Jumadi, and W. B. Zulfikar, "Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile," *J. Online Inform.*, vol. 1, no. 1, p. 24, 2016, doi: 10.15575/join.v1i1.7.

[7] D. Syahid, J. Jumadi, and D. Nursantika, "Sistem Klasifikasi Jenis Tanaman Hias Daun Philodendron Menggunakan Metode K-Nearest Neighboor (KNN) Berdasarkan Nilai Hue, Saturation, Value (HSV)," *J. Online Inform.*, vol. 1, no. 1, p. 20, 2016, doi: 10.15575/join.v1i1.6.

[8] A. Ali, "Sentiment Analysis on Twitter Data using KNN and SVM," vol. 8, no. 6, pp. 19–25, 2017.

[9] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017, doi: 10.15575/join.v2i1.71.

[10] E. Prasetyo, "Data mining mengolah data menjadi informasi menggunakan matlab," *Yogyakarta Andi Offset*, 2014.

[11] T. Phyu, "Survey of Classification Techniques in Data Mining," *Lect. Notes Eng. Comput. Sci.*, vol. 2174, Mar. 2009.

[12] S. Andayani and A. Ryansyah, "Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen," *JuSiTik J. Sist. dan Teknol. Inf. Komun.*, vol. 1, no. 1, pp. 53–62, 2017.

[13] F. Gorunescu, *Data Mining: Concepts, models and techniques*. 2011.

[14] N. Yusuf, "HADIS SEBAGAI SUMBER HUKUM ISLAM (Telaah Terhadap Penetapan Kesahihan Hadis Sebagai Sumber Hukum Menurut Syafi'iy)," *Potret Pemikir.*, 2015, doi: 10.30984/pp.v19i1.714.

[15] K. H. M. M. Zein, *Ilmu Memahami Hadits Nabi; Cara Praktis Menguasai Ulumul Hadits & Mustholah Hadits*, vol. 2. PUSTAKA PESANTREN, 2017.

[16] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," *Ann. Phys. (N. Y).*, vol. 54, p. 770, 2006, doi: 10.5860/CHOICE.49-3305.

[17] D. S. A. Maylawati, M. A. Ramdhani, A. Rahman, and W. Darmalaksana, "Incremental technique with set of frequent word item sets for mining large Indonesian text data," in *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*, 2017, doi: 10.1109/CITSM.2017.8089224.

[18] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015, [Online]. Available: http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf.