Hadith Text Classification: A Literature Review

1st Nursyahidah Alias[1], 2nd Nurazzah Abd Rahman[1], 3rd Normaly Kamal Ismail[1], 4th Zulhimi Mohamed Nor[2], 5th Muhammad Nazir Alias[3], 6th Mohd Sham Kamis[3]
{syahidah@pahang.uitm.edu.my[1], nurazzah@tmsk.uitm.edu.my[1], normaly@tmsk.uitm.edu.my[1], zulhilmi@usim.edu.my[2], nazir@ukm.edu.my[3] mohdsham856@ukm.edu.my[3]}

Fakulti Sains Komputer dan Matematik, Universiti Teknologi MARA Pahang, Bandar Tun Abdul Razak Jengka Pahang, Selangor, Malaysia[1], Fakulti Pengajian Quran dan Sunnah, Universiti Sains Malaysia, Bandar Baru Nilai, Nilai, Negeri Sembilan[2], Fakulti Pengajian Islam, Universiti Kebangsaan Malaysia, UKM Bangi, Selangor[3], Pusat PERMATA Pintar Negara, Universiti Kebangsaan Malaysia, UKM Bangi, Selangor[3]

Abstract. Hadith text classification is a domain specific research which requires involvement from both - science computer and hadith researcher. This paper presents literature on hadith text classification from different perspective in computer science such as classification process and issues related to it. This research therefore look into develop standard data set for hadith text classification with narrator chain part of the text as the basis. It will be the basis for future research on the classification of Malay-translated hadith text.

Keywords: text classification; hadith classification; hadith authentication

1 Introduction

A hadith is a record of the sayings or traditions of the Prophet Muhammad (Peace be Upon Him), revered and received as a major source of religious law and moral guidance, next to the authority of the Quran, the holy book of Islam [1]. Hadith classification is a process that supports hadith authentication. The process in hadith classification is being carried out by computer research community. Meanwhile, hadith authentication is carried out by hadith scholars. Hadith classification plays an important role in assisting hadith scholars in carrying their task [4].

Hadith text has two structures, such as sanad and matan [21] in Fig. 1. The figure shows the Malay translated hadith text from Hadith scholarly book Shahih Bukhari. Each Hadith there are two important things, namely the first real narrative text (Content / Matan), the second chronological list of people who were the transmitters of Content / Matan, known as the Narrator Chain / Isnad [4] [22] [2]. Sanad/Narrator Chain part in Fig. 1 is "*Telah menceritakan kepada kami Abdullah bin Muhammad Al Ju'fi dia berkata, Telah menceritakan kepada kami Abu 'Amir Al 'Aqadi yang berkata, bahwa Telah menceritakan kepada kami Sulaiman bin Bilal dari Abdullah bin Dinar dari Abu Shalih dari Abu Hurairah dari Nabi shallallahu 'alaihi wasallam, beliau bersabda:*". the narrator chain found in the hadith is "Nabi shallallahu 'alaihi wasallam → Abu Hurairah Abu Shalih → Abdullah bin Dinar → Sulaiman bin Baal → Abu 'Amir Al 'Aqadi → Abdullah bin Muhammad Al Ju'fi". The hadith in the Fig. 1 is "*Iman memiliki lebih dari enam puluh cabang, dan malu adalah sebahagian*

ICONQUHAS 2018, October 02-04, Bandung, Indonesia

DOI 10.4108/eai.2-10-2018.2295445

dari iman". Between the content and the narrator chain in the diagram below are clear and marked with symbols ‘:’.

The Fig. 2 shows the hadith text that has two narrator chains in one hadith text. The first narrator chain in the hadith is "Rasulullah shallallahu 'alaihi wasallam → Ibnu 'Abbas → Ubaidullah bin Abdullah Az Zuhri →Yunus →Abdullah → Abdan". While the second narrator chain is "Rasulullah shallallahu 'alaihi wasallam → Ibnu 'Abbas → Ubaidullah bin Abdullah → Az Zuhri → Yunus dan Ma'mar → Abdullah → Bisyir bin Muhammad-. Meanwhile for the hadith content is "*Rasulullah shallallahu 'alaihi wasallam adalah manusia yang paling lembut terutama pada bulan Ramadlan ketika malaikat Jibril 'Alaihis Salam menemuinva, dan adalah Jibril 'Alaihis Salam mendatanginya setiap malam di bulan Ramadlan, dimana Jibril 'Alaihis Salam mengajarkan Al Qur'an. Sungguh Rasulullah shallallahu 'alaihi wasallam lebih lembut daripada angin yang berhembus*”. That hadith was ambiguous and referred to hadith researchers.

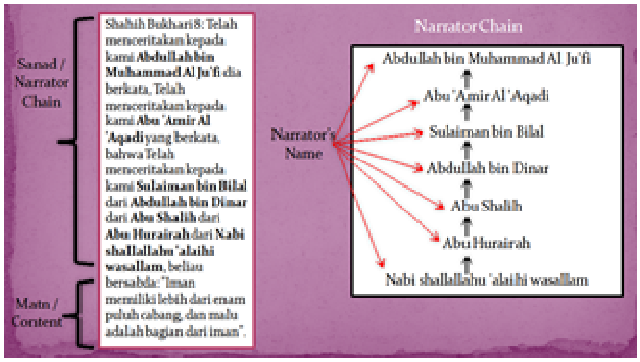

Figure 1. Hadith text in Malay text document

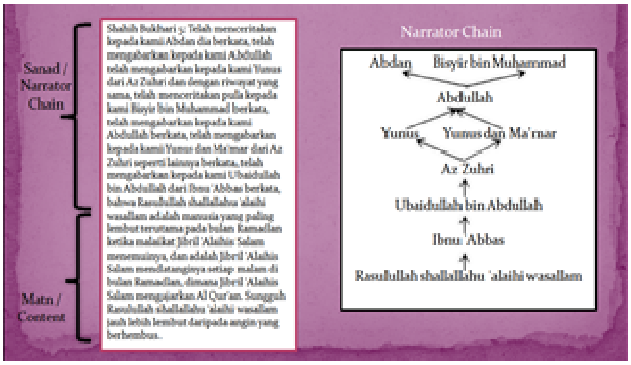

Figure 2. Hadith text with two narrator chain in Malay text document

Hadith text is a form of story-based text with no specific formatting. There is no standard numbering of the hadith. This leads to different numbering of the same text, depending on the type of publication — hardbound or paperback. The numbering of hadith is done by hadith researchers. not the original author of the hadith compilation 'Shahih Bukhari'.

2 Hadith Classification in Computer Science

2.1 Hadith Text Classification Research

[11] stated the classification of hadith is divided into seven as shown in the Fig. 3 which is: the number of the narrator in each stage, reference to authority, links of isnad, reliability and memory of narrators, nature of matn and isnad receive manner and hidden defect found [12][24].

Hadith classification research mostly focused on classification as: sahih, hasan, or da‘if [3]. Researchers that conducted the research were [2] and [3]. Therefore, there are six classifications of hadith that have yet to be explored as research.

The Table 1 shows the study which was conducted by [6] that differs hadith classification approaches. That table has several number of hadiths that have been researched, NLP tools / approaches, classification algorithm that have been used along with the results produced. The number of hadith used is different between one another and there no standard number of hadith need to be used.

In the hadith classification there is a major problem, namely how the documents being trained really belong to a particular class [10]. Most of previous study as in [6][3][18] [19][16][7][8] [10] this problem can be solved by classifying hadith which Al-Bukhari had previously classified, namely the very well-known Al-Hadith scientist.

2.2 Dataset

The compilation of Hadith AI-Bukhari has been widely published by various publishers in various languages. Shamela is a digital library which can help the process of classifying the hadith and contains details about narrators, but has not yet automated the process of classification of hadith, with the definition of the hadith validity (authenticity) and da 'if (weak) cannot be distinguished [12]. All those systems are can see in [18][15 19][17][4]. The most important thing to fulfill end user needs is information retrieval [10].

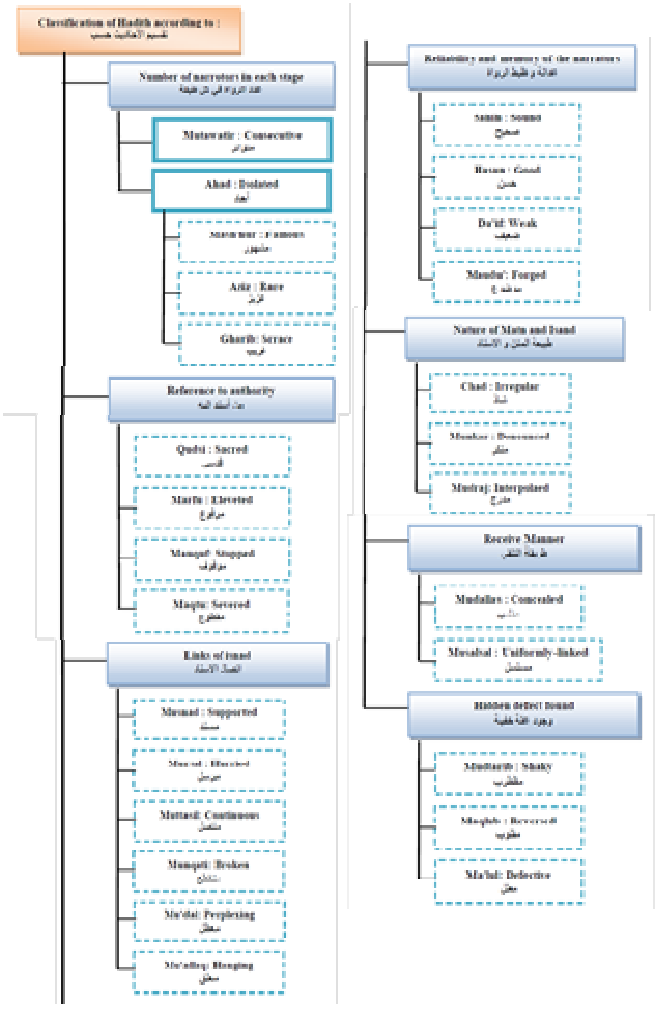

Figure 3. Classification of Hadith [11]

Limited application about Hadith Retrieval. In non-Arabic language environment, this limitation is found [29] such as Malay language. Mutiara Hadith and Lidwa Pusaka [34] are the hadith retrieval in Malay that have been used within science computer research community mainly for topical classification of the hadith text [14][20][30][31][32][33]. Nevertheless, the system lacks automatic text classification function and each researcher utilizes data from various databases. This proves a lack of standardized data among research community as far as classification of the Malay translated hadith text is concerned.

2.3 Hadith Classification Process

Fig.4 shows [10] a proposed system consists of four phases for classification of hadith. From pre-processing phase, learning database which contains the weights of features representing a class is constructed in second phase using a set of pre-classified documents. Then the third phase will be carried out a classification method where the results of the training database will be used to classify the targeted hadith. in this phase query expansion is also carried out. the final stage is data analysis and evaluation. Hadith classification process phases are slightly different than TC process that uses machine learning approach [ML] because there is no testing phase.

Tabel 1. Differentiating Hadith Classification Approaches [6]

Reference	#hadith	Linguistic tools / approaches	Classification algorithm	Results
[26][27]	453	stop-word removal and rule-based, morphological, stemming	Decision trees, Bayesian, Entropy and Vector space models	F1-mesaure = 0.70 with decision trees
[25]		Three stemming approaches: rule-based, root- based and light stemming	ANN vs. SVM	F-measure = 0.5 with ANN + light or rule-based stemming
[28]	1500	Removing chains of narrators, stop words and affixes	Rocchio, K-NN, Naïve Bayes and SVM	Recall=100% Precision =63.36% (SVM) and 67.11% (Rocchio)
[24]	80 (for testing)		Vector Space Models with several similarity measures	F-measure: from 0.42 (Dice Factor) to 0.85 (Naïve Bayesian)
[10]	1321	Removing chains of	The cosine coefficient	49% and 37% of

Reference	#hadith	Linguistic tools / approaches	Classification algorithm	Results
[26][27]	453	stop-word removal and rule-based, morphological, stemming	Decision trees, Bayesian, Entropy and Vector space models	F1-mesaure = 0.70 with decision trees
		narrators, stop words and affixes Stem-based, Word- based and hybrid representation.		improvement in F-measure for the hybrid method compared to the word-based and the stem-based methods

2.4 K-Nearest Neighbor (K-NN)

K-NN was named as one of the ten most popular and most important algorithms. K-NN is known to be simple. K-NN is an example-based learning group of classification. K-NN is done by searching for the group of K objects in the closest training data to objects in new data or testing data [6]. Generally the Euclidean distance formula is used to define the distance between two training objects and testing [11], [18].

$$dxy = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

2.5 Confusion Matrix

Confusion matrix used to evaluate classification models to estimate the true or false objects [10]. A matrix of prediction which will be compared with the original class of inputs or in other words contains information of actual and predicted value on classification [12].

3 Result and Analysis

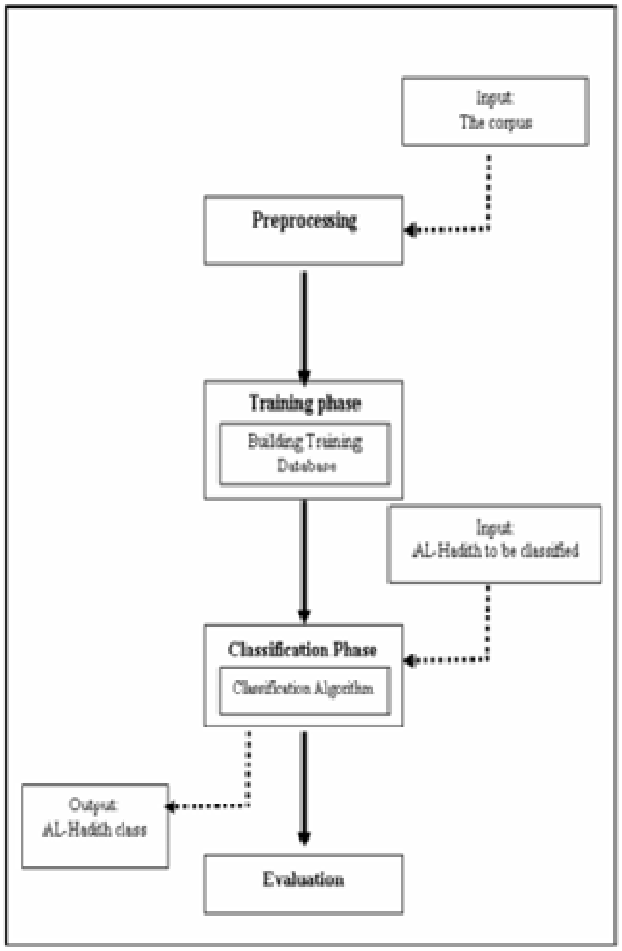

Figure 4. An system consist of four phases for classification of hadith [10]

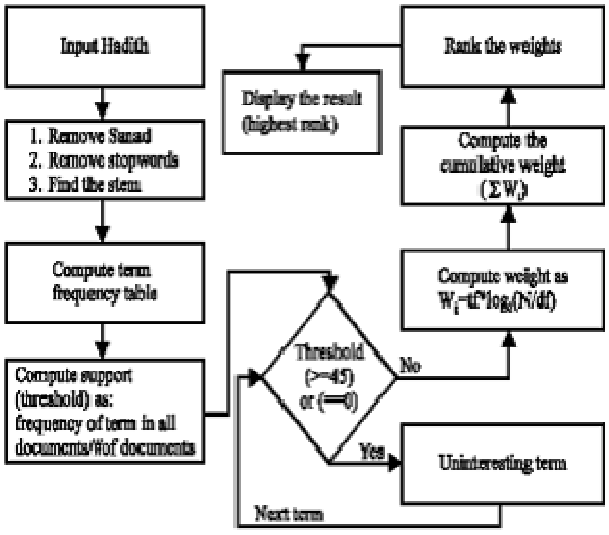

Figure 5. Hadith classification process for content [1]

The Figure 5 shows hadith classification process for content which conducted by [1]. Hadith classification process for the content is started with hadith input. Afterwards

sanad/narrator chains, stop-words were removed from hadith text and find the stein for the term. Then, the term frequency and weight were counted to produce the result for the users.

Hadith classification process for content is appropriate when using BoW text representation model in TC. Hadith classification process for content is usually classifying the hadith according to the topic in the hadith books [31][32]. Research in hadith classification that used BoW text representation can see in [6][7][20][10][14][1]. In addition, [9] used BoW to classify narrators in hadith text to trustworthy and untrustworthy.

Excluding for searching, the entire hadith classification process is done manually [2] [9] consequently it is not only prone [21] to subjectivity but also takes a lot of time [9] and expensive [13]. Therefore, automatic text (or document) categorization becomes an important attempt to save human effort required in performing manual categorization [10]. TC is a necessity due to a large amount of text documents that users must deal with [10] whose accuracy and time efficiency is much better than manual TC [13].

3.1 Performance Measure

The performance of classification model is important in determining the performance of the model with new cases [2]. [6][1] research on different methods of TC evaluates the quality of this classification in terms of performance or accuracy metrics. The assessment of F-measure, recall and precision were used as classification of hadith based on content as in [6][8][10].

Meanwhile, the assessment for classification of hadith based on sanad uses Success Rate of the Judgment and Error Rate of the Judgment respectively [3]. [2] performed four measurement which is 1) Correct Classification Rate (CCR). CCR is the number of correctly predicted scores by the classifier. 2) Error Rate(ER) represents the mathematical form of the number of incorrect prediction. 3) Sensitivity is The True Positive Rate. 4) Specificity is The True Negative Rate.

Despite the great efforts and the variety of the algorithms which have been tested, it is hard to select the best model without unifying the assessment framework [6]. The assessment accuracy in hadith classification need 9 be done because the data used did not go through the complicated method and require extensive checking and verification [3].

The F-measure values cannot interpret and compare objectively in hadith classification because of their works did not use the same datasets [6]. Besides that, the quality of the data source may affect the performance of a classification algorithm, the irrelevant and redundant features of data may reduce the quality of the result [5]. As we can see from the above works, the interest of studying Hadith science was increased, and the attempts of utilizing the new techniques to serve it and give optimistic results [12].

4 Conclusion

At present, research on hadith classification mostly focuses on the content classification of hadith text using BoW technique. Classification of the hadith narrator chain can still be considered insufficient proving that there are many more types of hadith text classification left unexplored. The absence of standard data set has undoubtedly impeded potential researches in this field. This lack makes studies on performance measurement between researches impossible as researchers' use of different databases leads to inaccurate comparison. This

literature research will therefore look into develop standard data set for hadith text classification with narrator chain part of the text as the basis.

Acknowledgements. This research was funded by the Malaysian Government under Fundamental Research Grant Scheme (FRGS) (FRGS/1/2015/ICT01/UITM/03/1) in Universiti Teknologi MARA, Shah Alam.

References

[1] A. K. Cragg, Hadith, Encyclopædia Britannica, Inc, https://www.britannica.com/topic/Hadith. Online Accessed 11th December 2017.

[2] M. N.Al-Kabi. Kanaan, G., and R. Al-Shalabi, Al-Hadith Text Classifier. Journal of Applied Sciences, 5(3), 584-587, 2005.

[3] K. Aldhaln, A. Zeki, A. Zeki, and H. Alreshidi, Improving Knowledge Extraction Of Hadith Classifier Using Decision Tree Algorithm 1. International Conference on Information Retrieval & Knowledge Management (CAMP), 148-152, 2012.

[4] A. M Azmi., and N. bin. Badia. e-Narrator - An Application For Creating An Ontology Of Hadiths Narration Tree Semantically And Graphically. The Arabian Journal for Science and Engineering, 35(2), 51-68, 2010.

[5] R. Elhassan, and M. Ahmed, Arabic Text Classification review. International Journal of Computer Science and Software Engineering (IJCSSE), 4(1), 1-5, 2015.

[6] K. Faidi, R. Ayed, L. Bounhas, and B. Elayeb, Comparing Arabic NLP tools for Hadith Classification. International Journal on Islamic Applications in Computer Science And Technology. 3(3), 1-12, 2015.

[7] F. Harrag, A. Alothaim, and A. Abanmy, Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules. International Journal on Islamic Applications in Computer Science And Technology, 1(2), 17-26, 2013.

[8] F. Harrag, and E. El-Qawasmeh, and A. M. S. Al-Salman, Extracting Named Entities from Prophetic Narration Texts (Hadith). Communications in Computer and Information Science, 180, 289- 297, 2011. http://doi.org/10.1007/978-3-642-22191-026

[9] Helmy, T., and Daud, A. Intelligent Agent for Information Extraction from Arabic Text without Machine Translation. In Proceedings of the 1st International Workshop on Cross-Cultural and CrossLingual Aspects of the Semantic Web. Retrieved from http://people.csail.mit.edu/pcm/templSWC/workshops/C3LSW2010/pap er2.pdf,2010.

[10] K. Jbara, Knowledge Discovery in Al-Hadith Using Text Classification Algorithm. Journal of American Science, 6(11), 409-419, 2010.

[11] M. Najeeb, A. Abdelkader, M. A1-Zghoul, and A. Osman, A Lexicon for 1-ladith Science Based on a Corpus. International Journal of Computer Science and Information Technologies, 6(2), 1336-1340, 2015.

[12] M. M. Najeeb, Towards Innovative System for Hadith Isnad Processing ISSN : 2231-2803. International Journal of Computer Trends and Technology (IJCTI), 18(6), 257-259, 2014.

[13] K. Nalini, and D. L. J. Sheela, Survey on Text Classification. International Journal of Innovative Research in Advanced (IJIRAE), 1(6), 412-417, 2014.

[14] N. A. Rahman, Z. A. Bakar, and T. M. T. Sembok, Query expansion using thesaurus in improving Malay Hadith retrieval system. In Proceedings 2010 International Symposium on Information Technology - System Development and Application and Knowledge Society, Vol. 3, pp. 1404-1409, 2010. http://doi.org/10.1109/ITSIM.2010.5561518

[15] S. B. Rebhi, and Y. M. Dalloul, Building Hadith Ontology to Support the Authenticity of Isnad. International Journal on Islamic Applications in Computer Science And Technology, 2,25-39, 2014.

[16] A. R. Saced, and S. W. Jaffry, Information Mining from Muslim Scriptures. The 4th Workshop on South and Southeast Asian NLP, International Joint Conference on Natural Language Processing. 66-71, 2013.
[17] Z. Shukur, N. Fabil, S. A. M. Noah, and J. Salim, Formalisma Hubungan Logikal dan Visualisasi Hubungan Pcrawi Hadith. Jurnal Teknologi, Universiti Teknologi Malaysia. 56.83-99, 2011.
[18] M. A. Siddiqui, M. E. Sala. and A. A. Bagais, Extraction and Visualization of the Chain of Narrators from Hadiths using Named Entity Recognition and Classification. International Journal of Computational Linguistics Research. 5(1). 14-25, 2014.
[19] M. D. Yahya, An Ontology-Based Approach to Support the Process of Judging Hadith Isnad. Islamic University of Gaza, 2013.
[20] M. K. A. B. Zainudin, and R. M. Rias, M-Hadith: Retrieving Malay Hadith Text in a Mobile Application. In 2012 International Symposium on Computer Applications and Industrial Electronics pp. 60-63. IEEE, 2012.
[21] F. Zaraket, and J. Makhlouta, Arabic Cross-Document NLP for the I ladith and Biography Literature. Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference Arabic, 1,256-26, 2013.
[22] M. S. Al-Munajjid, Conditions of a Sahih (sound) hadith. Islamic Propagation Office in Rabwah, Riyadh. islamhouse.com, 2009.
[23] A. Mahmood, F.K. Alarfaj, H.U. Khan, M. Ramzan, and M. A. Ilyas, Multilingual Datasets Repository of the Hadith Content. International Journal of Advanced Computer Science and Applications, 9(2), 165 – 172, 2018.
[24] M. Al-Kabi, and S. I. Al-sinjilawi, A comparative study of the efficiency of different measures to classify arabic text, Journal of Pure & Applied Sciences Volume 4, No. 2 June 2007.
[25] Harrag, F., El-Qawasmah, E., and Al-Salman, A. M. S. Stemming as a feature reduction technique for Arabic Text Categorization. In 10th International Symposium on Programming and Systems (ISPS), pp. 128- 133, 2011.
[26] Harrag, F., El-Qawasmah, E., and Pichappan, P. Improving Arabic text categorization using decision trees. The First International Conference on Networked Digital Technologies NDT '09, Ostrava, Czech Republic, 28-31 July, 2009.
[27] F. Harrag, and A. Hamdi-Chrif, Classification des Textes Arabes Basée sur l'Algorithme des Arbres de Décision, International Conference on Web and Information Technologies ICWIT'08, Sidi Bel Abbes, Algeria, 29-30 June, 2008.
[28] M. Al-Khatib, Classification of Al-Hadith Al-Shareef Using Data Mining Algorithm, the European Mediterranean & Middle Eastern Conference on Information systems (EMCIS'2010) Abu-Dhabi United Arab Emirates, 2010.
[29] I. Rasyidi, A. R., and A. T. Wibowo. Indonesian Hadith Retrieval System Using Thesaurus. 2013 International Conference on Computer, Control, Informatics and Its Applications. IEEE, 2013.
[30] N. N. Amirah, T. M. Rahim, Z. Mabni, H. M. Hanum and N. A. Rahman. A Malay Hadith translated document retrieval using parallel Latent Semantic Indexing (LSI). 2016 Third International Conference on Information Retrieval and Knowledge Management. IEEE, 2016.
[31] H. M. Hanum, Z. A. Bakar, N. A. Rahman, M. M. Rosli and N. Musa. Using Topic Analysis for Querying Halal Information on Malay Documents. Journal of Procedia-Social and Behavioral Sciences. Vol. 121, p: 214-222. Elsevier, 2014.
[32] N. A. Rahman, A. B. M. Soom and N. K. Ismail. Enhancing Latent Semantic Analysis by Embedding Tagging Algorithm in Retrieving Malay Text Documents. Advanced Topics in Intelligent Information and Database Systems. 309-319. Springer International Publishing, 2017.
[33] N.A. Rahman, N.K.Ismail, Z.A. Bakar and T.M.T Sembok. Mutiara Hadis: A Web Based Retrieval System For Malay Hadith Text. http://sigir.uitm.edu.my/webhadis/, 2007.
[34] Lidwa Pusaka, Ensiklopedia Hadits Kitab 9 Imam, Indonesia, 2010- 2011.

[35] N. K.Ibrahim, M. F. Noordin, S. Samsuri, M.S.A Seman, and A.E. B. Ali¸ Isnad Al-Hadith Computational Authentication: An Analysis Hierarchically. 2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M). IEEE, 2016.