# Comparative Study of Machine Learning Approach on Malay Translated Hadith Text Classification based on Sanad

*Syuhairah Rahifah* Mohammad Najib[1], *Nurazzah* Abd Rahman[1], *Normaly* Kamal Ismail[1], *Nursyahidah* Alias[1,*], *Zulhilmi* Mohamed Nor[2], *Muhammad Nazir* Alias[3]

[1]Faculty of Computer and Mathematical Sciences,Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Fakulti Pengajian Quran dan Sunnah, Universiti Sains Islam Malaysia, Bandar Baru Nilai, Negeri Sembilan, Malaysia
[3]Fakulti Pengajian Islam, Universiti Kebangsaan Malaysia, Bangi, Selangor Malaysia

**Abstract.** Sanad is one of important part used to determine the authentication of hadith. However, very little research work has been found on classification of Malay translated Hadith based on sanad. There are some researches done using machine learning approach on hadith classification based on sanad but using different objective with different language. This research is to see how Machine Learning techniques are used to classify Malay translated Hadith document based on sanad. In this paper, SVM, NB and k-NN are used to identify and evaluate the performance of Malay translated hadith based on sanad. The performances are evaluated based on standard performance metrics used in text classification which is accuracy and response time. The results show that SVM has the highest accuracy and k-NN has the best response time (time taken in process for classification data) compare to other classifier. In future, we plan to extend this paper with the analysis on interclass similarity and also test on larger dataset.

## 1 Introduction

Sunnah (Hadith) is a second of fundamental sources in Islam after Qur'an [1][2] which is Muslims reference in any activities in their life [3]. Based on [2], the author said that hadith are related to actions and sayings of Prophet Muhammad by trustworthy narrators. It is essential in understanding Qur'an and Islamic jurisprudence [4]. However, [5] mentioned that hadith has been overlooked compared to Qur'an by most academicians in computer science. In the hadith, there are two main components which known as Isnador Sanad (the chain of narrators) and Matn (actual narrative or main text) [4][5][6]. The sanad contains of a chronological list of the narrators, each narrators stated the one from whom they heard the Hadith all the way to the main narrator of the Matn followed by the matn itself [4]. Sanad is essential in every hadith. It is used in the first step of checking the authentication of a

[1] Corresponding author: syahidah@pahang.uitm.edu.my

hadith [2][5][7]. To date, there is need to automatic classification Malay translated hadith [2] based on sanad.

Machine Learning is a wide area of Artificial Intelligence focused in design and development of algorithm that identify and learn patterns exist in data provided as input[8]. Text classification is an important issue which draws many researchers in machine learning and information retrieval techniques [6]. Referring to [8], the author also mentioned text classification is a key problem influenced by machine learning within information retrieval. However, very little research work has been found on classify Malay translated Hadith based on sanad. Classification hadith based on sanad has been done by [6][1][4] with different objective using different language. Review study on Malay translated hadith has been done by [9]to identify the authentication of narrator's name, improving Malay hadith retrieval system by [2] and retrieve Malay hadith text using mobile application by [3]. In this novel approach, Machine Learning techniques are used to classify Malay translated Hadith document based on sanad.

This paper is structured as follows. Section I cover on the introduction of Hadith and how the Machine Learning approach fill into the picture. Section II is focusing on some background information in Machine Learning and Malay Translation Hadith. Section III is focusing on the approach used in this paper. Section IV is a discussion on result and Section V which is contains the conclusion of the paper.

## 2.1 Research Background

### 2.1 Machine learning approach (Text classification)

In machine learning, there are three basically type algorithms used: (1) supervised; (2) semi-supervised; (3) unsupervised learning as shown in Figure 1. Supervised learning is required learning a function from training data provided as input. In the case of text classification, the training data are collected of document-class pair representing proper classes for given documents, according to human specialists (data are provided by human assistance as input data). Unsupervised learning is different from supervised learning which is no training data are provided. Semi-supervised learning combines large amount of unlabelled data with a small amount of labelled data [8].

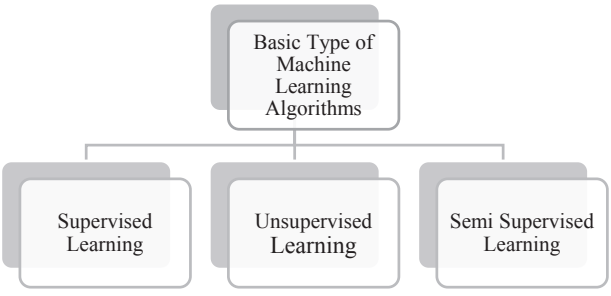


**Fig. 1.** Basic Type of Machine Learning Algorithm [8].

To classify document into a fixed number of predefined categories is the main reason of text classification. Each document can be categorised in multiple, exactly one, or no category at all. The classification of documents is recognised as a supervised learning task because the purpose is to use machine learning to automatically classify documents into categories based on previously labelled documents [10]. Based on [10], the author also

mentioned that Naive Bayes, k-NN and Support Vector Machines (SVM) are popular techniques and common approach [11][12] in automatic text classification. This paper is focused on supervised classification using three most popular technique as stated by [10][11][12] as Figure 2.

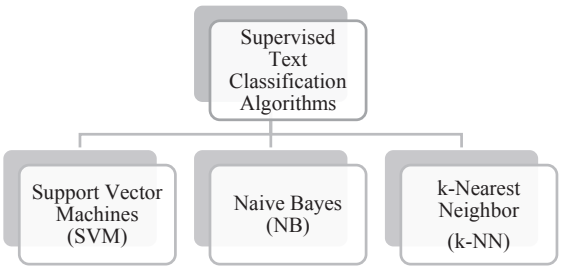


**Fig. 1.** Supervised Text Classification Algorithms [8].

### *2.1.1 Support Vector Machine*

The Support Vector Machine (SVM) is a classification technique that was introduced by Vapnik and was first applied by Joachims for text classification [8][13][14]. It is a supervised learning algorithm that analyse the data and identify patterns used for classification [15]. The core fundamental of SVM is to determine the most appropriate border line in separating hyper plane [14]. In the set of training data, SVM creates a hyper plane for separating data in two categories (positive and negative) and classification in which data must be placed in this two categories [13][15]. Referring to [13], the author believed SVM is a powerful classifier based on the lowest structural risk principle. SVM also popular in text classification with better performance than other methods [14][16][17].

### *2.1.2 Naive Bayes*

Naive Bayes algorithm is one of classification technique that makes exploit of statistical approach and based on the conditional probabilities for the problems of pattern recognition [16]. Naive Bayes uses Bayes Theorem concept [15][16][18] with strong independence assumptions. The classifier are named "naive" because the algorithm assumes that all terms occur independent from each other [19]. The independence assumptions of features do not depend and effect on each other in classification tasks. Although it is severely limited in its applicability, the computation of Bayesian classification approach is more efficient. It can be trained efficiently to estimate parameters for classification without requiring large amount of training data. The naive Bayes classifiers often work much better in many complex real-world situations than one might expect due to its apparently oversimplified assumptions. Under some specific conditions, naive Bayes classifier has been reported to perform surprisingly well for many real world classification applications [18].

### *2.1.3 K-Nearest Neighbor*

K-Nearest Neighbor (k-NN) classifier is an on-demand (lazy) classifier. The classification is done only at the moment a new document is given to the classifier. The classification decision is computed based on the classes of the k "nearest" neighbour of the new document using a distance function in a predefined metric space. This is accomplished as

follows: a) Determine the k nearest neighbours of the new document in a given document training set, b) Use the classes of the nearest neighbour to determine a class for the new document. This algorithm concentrate on specific features of the document to be classified [8]. It is effective and easy to implement [18] and also the most accepted algorithms for pattern recognition [16].

### 2.2 Malay translation hadith

The text of Malay hadith comes from translated version of Arabic hadith [9]. According to [7], there are six books contain reliable collection of Hadith text: Bukhari, Muslim, Abu Dawud, Termizi, Nasai and ibn Majah. At first, writing narration of Hadith was prohibited due to some religious reasons. However after the death of the prophet, Umar Ibn Abd al-Aziz initiated the writing project of Hadith to guarantee an integrity and uniformity of the text upon fearing that some of hadiths are being lost [6].

## 3 Methodology

Figure 3 shows the process of text classification in this research is referred to the framework based on [8]. Malay translated hadith dataset are used for this experiment. This dataset is focused only in sanad and the matn are removed.

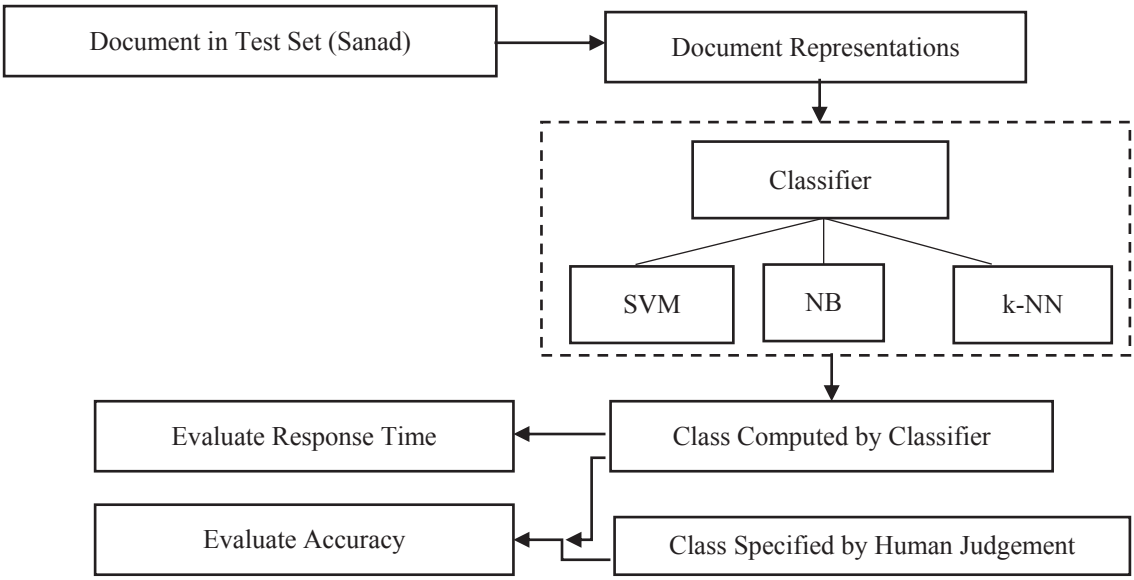


**Fig. 2.** Text Classification Process.

### 3.1 Document in test set (sanad)

The data used in experiment are data from Lidwa Pusaka [20] website for dataset hadith Sahih Bukhari and Mutiara Hadis [21] website for dataset hadith Sunan At-Termizi. 100 Hadiths are choose randomly for this experiment which divided by two categories: 50 hadiths from Sahih Bukhari and another 50 from Sunan At-Termizi. The data are labelled manually and the labels belong to two different category classes: 1. Sahih Bukhari (SB) and 2. Sunan At-Termizi (ST). Figure 4 shows a sample of sanad from Sahih Bukhari (SB). Figure 5 shows a sample of sanad from Sunan At-Termizi (SB).

> *"Telah menceritakan kepada kami Al Humaidi Abdullah bin AzZubair dia berkata, Telah menceritakan kepada kami Sufyan yang berkata, bahwa Telah menceritakan kepada kami Yahya bin Sa'id Al Anshari berkata, telah mengabarkan kepada kami Muhammad bin Ibrahim At Taimi, bahwa dia pernah mendengar Alqamah bin Waqash Al Laitsi berkata; saya pernah mendengar Umar bin Al Khaththab diatas mimbar berkata; saya mendengar Rasulullah shallallahu 'alaihi wasallam bersabda;"*

**Fig. 3.** Sample of sanad [20].

> *"Muhammad bin Basysyar menceritakan kepada kami, Ibrahim bin Abul Wazir memberitahukan kepada kami, Muhammad bin Musa memberitahukan kepada kami (yang berasal) dari Sa'd bin Ishaq bin Ka'b bin 'Ujurah dari ayahnya dari datuknya dimana ia berkata: "Nabi s.a.w. mengerjakan solat Maghrib di masjid Bani 'Abdil Asyhal kemudian orang-orang mengerjakan solat sunat, lantas Nabi s.a.w. bersabda;"*

**Fig. 4.** Sample of sanad [21].

### 3.2 Document representations

The dataset is represented as Figure 6. Row data referred to hadith document which contain 100 hadith from Sahih Bukhari and Sunan At-Termizi. The Column data referred to sanad and there are 272 data used for it. The documents are represented in two classes: Sahih Bukhari (SB) and Sunan At-Termizi (ST).

| | 2 | 3 | ... | 271 | 272 | class |
|---|---|---|---|---|---|---|
| | umarbinalkhat | alqamahbinw | ... | islbinsufy | atha' | |
| | | | | | | |
| 1 | 1 | 1 | ... | 0 | 0 | SB |
| 2 | 0 | 0 | ... | 0 | 0 | SB |
| 3 | 0 | 0 | ... | 0 | 0 | SB |
| 4 | 0 | 0 | ... | 0 | 0 | SB |
| 5 | 0 | 0 | ... | 0 | 0 | SB |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 96 | 0 | 0 | ... | 0 | 0 | ST |
| 97 | 0 | 0 | ... | 0 | 0 | ST |
| 98 | 0 | 0 | ... | 0 | 0 | ST |
| 99 | 0 | 0 | ... | 0 | 0 | ST |
| 100 | 0 | 0 | ... | 1 | 1 | ST |

**Fig. 5.** Document representation.

### 3.3 Classifier

The data were entered into classifier to predict the class of hadith. There are three most popular classifier are used in this experiment which are SVM, NB and k-NN.

### 3.4 Evaluation

Finally, the performances of classifiers are evaluated. Evaluation is a very important step to determine the performance of the classifiers used [8][22]. In this experiment, the evaluation are covered based on standard performance metrics: Accuracy [1][14][22]and response time [14]. Concisely, accuracy is related to the fraction of correct classifications [22]. Weka tool[23] is used to evaluate the response time (time taken in process for classification data) and for the accuracy, we compare the Class Computed by Classifier and Class Specified by Human Judgement.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where:

True Positive value (TP)
True Negative value (TN)
False Positive value (FP)
False Negative value (FN) [14]

## 4 Experiment and Result

Three classifiers are used to compare the performance of Malay translated hadith based on sanad. There are Support Vector Machines (SVM), Naive Bayes (NB) and k-Nearest Neighbor (k-NN). The results of experiment are shown in Table 1. Results are evaluated based on accuracy and response time.

**Table 1. Result of experiment**

| Classifier | Accuracy (%) | Response Time (second) |
|---|---|---|
| NB | 81 | 0.01 |
| k-NN | 62 | 0.00 |
| SVM | 82 | 0.16 |

Figure 7 shows the results of accuracy are 82% for SVM, 81% for NB and 62% for k-NN. We found that SVM shown the highest accuracy compare to other classifier.

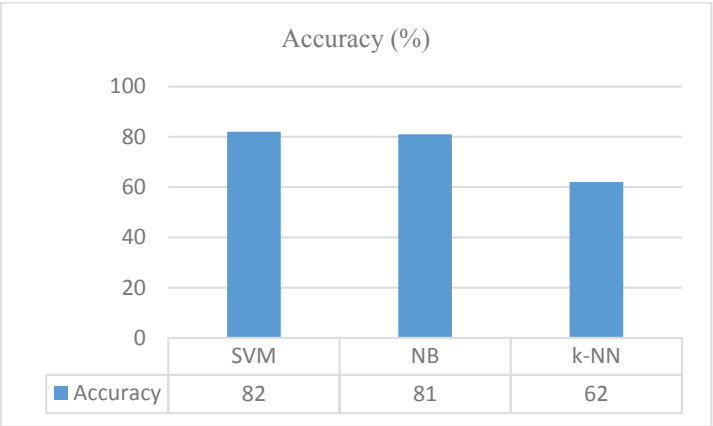


**Fig. 6.** Comparison the accuracy of classifiers.

Figure 8 shown the results of response time in second(s) are 0.16 s for SVM, 0.01 second for NB and 0.00 second for k-NN. We found that k-NN shown the best response time (time taken in process for classification data) compare to other classifier.

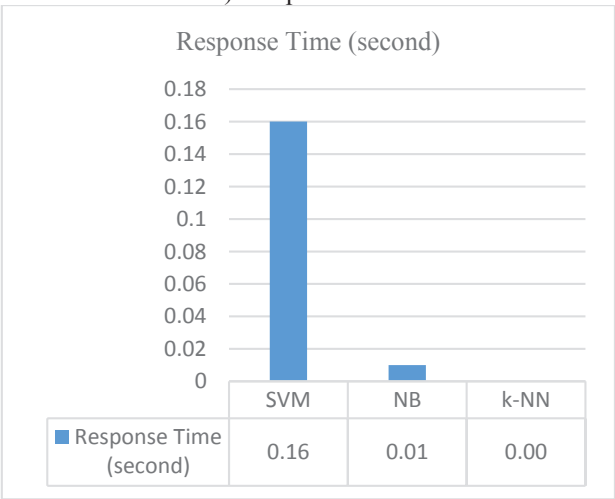


**Fig. 7.** Comparison the response time of classifiers.

## 5 Conclusions

The experimental results done on the Malay translated Hadith document based on sanad shown the best performance for accuracy is SVM classifier and response time is k-NN classifier. However, this paper only covers the accuracy and response time of the classification used without details explanation in analysis on interclass similarity. So in the future work, we plan to extend this paper with the analysis on interclass similarity and also test on larger dataset.

## Acknowledgement

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. K. A. Aldhlan, A. M. Zeki, H. A. Alreshidi, Novel mechanism to improve hadith classifier performance, *Proc. Int. Conf. Adv. Comput. Sci. Appl. Technol,* pp. 512–517 (2012)
2. N. A. Rahman, Z. A. Bakar, T. M. T. Sembok, Query expansion using thesaurus in improving Malay Hadith retrieval system, *Proc. nt. Symp. Inf. Technol. Syst. Dev. Appl. Knowl. Soc.,* **3,** pp. 1404–1409 (2010)
3. M. K. A. B. Zainudin, R. M. Rias, M-Hadith: Retrieving Malay Hadith text in a mobile application, *Proc 2012 IEEE Symp. Comput. Appl. Ind. Electron.,* pp. 60–63 (2012)
4. A. Azmi, N. Badia, iTree - Automating the construction of the narration tree of Hadiths (prophetic traditions), *Proc. 6th Int. Conf. Nat. Lang. Process. Knowl. Eng.* (2010)
5. M. A. Saloot, N. Idris, R. Mahmud R, S. Ja'afar, D. Thorleuchter, A. Gani, Hadith data mining and classification: A comparative analysis, *Artif. Intell. Rev.,* **46,** pp. 113–128 (2016)
6. M. Ghanem, A. Mouloudi, M. Mourchid, Classification of Hadiths using LVQ based on VSM Considering Words Order, *Int. J. Comput. Appl.* (0975 – 8887)*,* **148,** pp. 25–28 (2016)
7. A. Bimba, M. A. Ismail, N. Idris, S. Jaafar, R. Mahmud, Towards Enhancing the Compilation of Al-Hadith Text in Malay, *Int. Proc. Econ. Dev. Res.,* **83,** pp. 76–81 (2015)
8. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search, *Inf. Retr. Boston.*, **82**, pp. 281–335 (2011)
9. N. A. Rahman, N. Alias, N. K. Ismail, Z. M. Nor, M. N. Alias, An identification of authentic narrator's name features in Malay hadith texts, *2015 IEEE Conf. Open Syst.,* pp. 79–84 (2015)
10. J. Lilleberg, Y. Zhu, Y. Zhang, Support Vector Machines and Word2vec for Text Classification with Semantic Features, pp. 136–140 (2015)
11. A. Harisinghaney, A. Dixit, S. Gupta, A. Arora, Text and Image Based Spam Email Classification using KNN, Naive Bayes and Reverse DBSCAN Algorithm, *Proc. 2014 Int. Conf. Reliab. Optim. Inf. Technol.,* pp 153–155 (2014)
12. Y. Lin, J. Wang, Research on text classification based on SVM-KNN, *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci.,* pp 842–844 (2014)
13. S. M. H. Dadgar, M. S. Araghi, M. F. Mastery, A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification*2nd IEEE Int. Conf. Eng. Technol.,* pp. 16–20 (2016)
14. J. Watthananon, The relationship of text categorization using Dewey Decimal Classification techniques, *12th Int. Conf. ICT Knowl. Eng.,* pp. 72–77 (2014)
15. B. Y. Pratama, R. Sarno, Personality classification based on Twitter text using Naive Bayes, KNN and SVM, *Proc. 2015 Int. Conf. Data Softw. Eng.,* pp. 170–174 (2016)
16. S. Brinda, K. Prabha, S. Sukumaran, A Survey on Classification Techniques for Text Mining, *2016 3rd Int. Conf. Adv. Comput. Commun. Syst.,* pp. 1–5 (2016)
17. C. Yin, J. Xiang, H. Zhang, J. Wang, Z. Yin, A New SVM Method for Short Text Classification Based on Semi-Supervised Learning, *Proc. - 2015 4th Int. Conf. Adv. Inf. Technol. Sens. Appl.,* pp. 100–103 (2015)
18. A. Khan, B. Baharudin, L. H. Lee, K. Khan, A Review of Machine Learning Algorithms for Text-Documents Classification, *J. Advance Inf. Technol.,* **1,** pp. 4–20 (2010)
19. V. Bijalwan, V. Kumar, P. Kumari, J. Pascual, KNN based Machine Learning Approach for Text and Document Mining, *Int. J. Database Theory Appl.,* **7,** pp. 61–70 (2014)

20. Perusahaan software Indonesia dan SIMRS Saltanera Software Kitab Hadits Online Terjemah Indonesia [Online] Accessed: 06-Oct-2016 http://app.lidwa.com/ (2016)
21. N. K. Ismail, N. A. Rahman, and Z. A. Bakar, Mutiara Hadis. [Online] Accessed: 06-Oct-2016, http://sigir.uitm.edu.my/webhadis/ (2007)
22. T. R. P. M. Rúbio, C. A. S. J. Gulo, Enhancing academic literature review through relevance recommendation: Using bibliometric and text-based features for classification, *2016 11th Iber. Conf. Inf. Syst. Technol.,* pp. 1–6 (2016)
23. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update *SIGKDD Explor.,* (2009)