CONTENT

LIST OF FIGURE

LIST OF TABLE

CHAPTER 1: INTRODUCTION

Al-Hadith is the collection of words, deeds, provisions, and approvals of Rasulullah Sallallahu Alaihi Wa Salam. The Hadith is an important textual source for the laws, traditions, and teachings of Islam in the Islamic world. The Hadith is the second most fundamental law of Islam after the Holy Qur'an. Hadith requires translation from Arabic to other languages to simplify people's general understanding of the Tafseer. Tafseer is a field that touches on the content of Al-Quran, specifically the meanings of the verses that are mentioned in the Hadith.

Hadith has been studied by Muslims and non-Muslims around the world, including Indonesia Muslim majority country. Although almost all Indonesians are Muslims, most do not understand the Law of Islam in relation to their daily life and worship. They have a lot of questions about righteousness and their obligations as a Muslim. People who have no basic Islamic education find it hard to understand the Hadith in its raw form because Hadith has unique linguistic features (e.g. ancient Arabic language and story-like text) [1]. Hence, the Hadith needs to be classified to assist non-Arabic speakers to understand it, especially Indonesians who are used to learning the Hadith in Bahasa Indonesia.

This chapter discusses the rationale in Section 1.1 that explain the background of this study and related problem situation. Theories and concept used to conceptualize this study are discuses in Section 1.2, while Section 1.3 discusses the variable related to the problem and their relationship to the paradigm of this study. The intended problem within this study is explained in Section 1.4. Section 1.5 discusses the proposed approach to solving the intended problem. Besides, this study describes some assumption in Section 1.6, while Section 1.7 describes the scope of works and delimitation. Finally, the contribution of this study is described in Section 1.8.

1.1 Rationale

There are more than 200,000 hadiths with most having been translated to the Indonesian language. Nevertheless, this study used Hadith Al-Bukhari as the dataset, which contains about 7000 hadiths. This is because Hadith Al-Bukhari contains a summary of other Ahadith. Hadith Al-Bukhari is the Sahih Hadith, which was not only authenticated but also accrued by a trusted author and Imam. In this research, the Hadith was classified into three classes; information, suggestion, and prohibition. This study focused on classifying the Hadith of Al-Bukhari. Those classes determine from the expert suggestion. It can be clearly defined for guides the Moslems to follow Allah SWT orders and no less important is prohibition of Moslem which should be avoided. In this research, the Hadith class labels were annotated using expert judgment. To classify the Ahadith, the Machine Learning method was used, specifically, the Artificial Neural

Network (ANN). Before the classification process, the Hadith text underwent preprocessing using Natural Language Processing (NLP). Other preprocessing methods include Non-Alpha numeric Removal, Stopword Removal, and advanced methods such as Word2Vec for the word embedding process.

Recent studies that have utilized NLP for Hadith classification include Harrag et al. [2], who obtained the highest classification performance for Arabic Hadith. The main concern of their study was the stemming process and the reduction of word features in the Arabic Language. On the other hand, there are some different characteristic compared with this study; [1]. This study aims to develop a model for Hadith Classification in the Indonesian language. Hence, there are different challenges that are faced, especially in the language structure, as Indonesian Islamic text was used in this study instead of Arabic.

1.2 Theoretical Framework

Classification is needed for understanding hadith text. Although the hadith translated into Indonesian, Indonesian people cannot easily understand what the hadith means. Based on the rationale, this research will focus to classify the hadith text.

Harrag et. al. do some research about for hadith text classification. The best performance is on paper [2]. With the proposed method, Harrag et. al. got 94% score for hadith classification. That research does with Dictionary-lookup for stemming and dimension reduction process, then Artificial Neural Network for the classifier. The other research was conducted by Harrag in 2009 with El-Qawasmeh. In that research [3] got lower performance than research [2]. However, the performance was still high with 90% score. On this research use Artificial Neural Network with Singular Value Decomposition. With the same classifier, that research got a different score. Based on those researches, This research will use the same classification process with them. However, this research more focus on preprocessing to getting the modest model of hadith text.

1.3 Conceptual Framework/Paradigm

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

- **SUGESSTION** - “Rasulullah telah menetapkan batas miqat bagi penduduk Nejd di Qarnul Manazil, dan itu sangat jauh bila dilihat dari jalan kami, dan bila kami ingin menempuh ke sana sangat memberatkan kami. Maka dia ('Umar) berkata: Perhatikanlah batas sejajarnya dari jalan kalian. Lalu dia menetapkan miqat mereka di Dzatu 'Irqi.”.

- **PROHIBITION** - “LAA-ILAAHA ILLALLAAH, WAHDAHU LAA SYARIIKA LAHU, LAHUL MULKU WALAHUL HAMDU WAHUWA 'ALAA KULLI SYAI'IN QADIIR, (Tiada sesembahan yang hak selain Allah, tiada sekutu bagi-Nya, Milik-Nya lah segala kerajaan dan bagi-Nya segala puji dan Dia maha berkuasa atas segala sesuatu). Beliau mengucapkannya hingga tiga kali. Dan beliau juga melarang desas desus (ghosip), banyak bertanya dan menghambur-hamburkan harta, beliau juga melarang mendurhakai ibu, menghalangi orang lain memperoleh kemanfaatan dan mengubur hidup-hidup anak perempuan serta. Dan dari Husyaim telah mengabarkan kepada kami Abdul Malik bin Umair, dia berkata; saya mendengar Warrad menceritakan hadits ini dari Al Mughirah dari Nabi .”.
- **INFORMATION** - “Dan atas kalian juga.' Kemudian Aisyah berkata; 'As Saamu 'alaikum wala'anakumullah wa ghadziba 'alaikum”.

The basic concept of the proposed method is classified Indonesian translated hadith text along with word embedding process to get the best model of word embedding. Word embedding is supported mechanism based on Word to Vector (W2V) concept to help the classifier got the features for the classification process. These models help to improve the performance of classification and more consistency.

1.4 Statement of the Problem

The first problem faced in this study is that the dataset consists of Indonesian Hadith as well as some Arabic words, hence it has a high dimensional feature space. In other words, the feature space consists of a collection of every unique word in the dataset. High dimensionality reduces the performance of the classification process, because the data may contain non-informative terms (i.e., features of low discriminative power) [2]. The most common feature selection methods are filters, wrappers, embedded, and hybrid methods [4]. In one study [5], several papers using different feature selections were examined. Instead of using feature selection for reducing the input of classifier, to optimize classifier performance, this research used word embedding, which not remove any terms of the dataset.

The second identified problem is that a lot of data from a dataset contains very few words. Hence, it can be sparse data problem. The similarity method was needed to contained non-zero numbers of features. This research used a word embedding learning method known as Skip-Gram (SG). Word2Vec is also a popular method. Word embedding needs other documents to build the Indonesian corpus for the word modeling process. In a previous study [6], much of the superior performance of Word2Vec or similar embedding processes in downstream tasks was not a result of the model from the main dataset. However, model results got from specific hyper-parameters dataset. Specific hyper-parameters dataset is dataset which not only has a

very big parameter, but also specific terms. Hence, the problem of words with bilingual could be solved using this method.

1.5 Objective

Based on statement of the problem and rationale stated, the objectives of this research are using the new approach of classification to handle Hadith text translated into Indonesian along with Arabic words dataset. The model must be able to handle Hadith text bilingual dataset and solved the sparse problem with performing the ANN specific on BP algorithm and SVM for classifier along Word Embedding with SGNS.

1.6 Hypotheses

According to previous Hadith research, ANN performed better than SVM [2], but ANN has some limitations. ANN is not good at handling high dimensional data because it still needs other mechanism to obtain better performance [1] [2]. Additionally, ANN cannot handle words with multiple meanings with just the traditional weighting process because of a sparse data problem. Hence, by adding supporting mechanisms to reducing dimension especially, ANN will achieve higher performance. These processes are expected to solve this problem.

1.7 Assumption

Taking into consideration the gap in previous research, this research aims to conduct the classification of Bahasa Indonesia Hadith texts via Machine Learning and Natural Language Processing. The Artificial Neural Network (ANN) was used for classifying Hadith in Bahasa Indonesia because ANN has a good performance record when used in the TC Hadith text [2]. In one paper [2], some classifier methods, especially BP in the case of this study, still required others methods such as feature extraction or feature selection to obtain more accurate results [5] [7]. The similarity process needs to be performed again to reduce the spare data problem [8]. Similar to previous research, the similarity process that produces good accuracy was used with a word embedding approach, in which one is the word to vector (Word2Vec). To overcome sparse data problem on bilingual data and reduce the input of classifier in this research, one of the Word2Vec methods i.e. a similarity algorithm called Skip-Gram Negative-Sample (SGNS) was used [9]. The sparse data problem can be addressed using SGNS, as it does not solve the problem using traditional weighting approaches. SGNS is an algorithm that uses the Word Embedding approach.

1.8 Scope and Delimitation

This study focuses on the implementation of Skip-Gram Negative-Sample into the Classification process. The output from this system is a model of word embedding and model of classification.

The data will use Hadith Al-Bukhari translated into Indonesian. Hadith Al-Bukhari has 7008 hadiths. Data has been labeled as much as 1651 hadiths. That data is the main dataset that uses for training and testing on the classification process. The rest was used for modeling dataset along with Indonesian Wikipedia dump data.

1.9 Importance of the Study

Taking into consideration from previous research [2], this research aims for the classification of Bahasa Indonesia Hadith texts Machine Learning and Natural Language Processing. This study can get the best model of Word Embedding model for Indonesian hadith text classification. That model is never built before this research. Furthermore, word embedding needs a lot of data to cover all bag of words. This research is expected to output new word embedding model for Indonesian hadith translated text that can use for future research.

CHAPTER 2: REVIEW OF LITERATURE AND STUDIES

Text Classification (TC) is a field of research under text mining. However, TC is more specific than text mining. TC aims to classify or categorize texts that have the same characteristics into a class that has been determined using a supervised method. The field of Computing and Informatics are not the only fields utilizing TC to find new information; other fields such as Social Sciences, Political Science, and Religious Sciences also utilize TC, among which include Hadith text classification.

There are many kinds of research 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 [10], Dataset used Matn and Sanad of Sahih Bukhari Hadith with 7275 hadith which classified Bukhari traditions into 8 classes i.e. "Pray", "Eclipse", "Faith", "Call to Prayer", "Good Manners", "Knowledge", "Fasting", and "Almsgiving".

Table 1 Result of Al-Kabi research [10]

Class	SVM	NB
Ablutions (Wudu')	0.257	0.736
Fasting	0.567	0.828
Almsgiving (Zakat)	0.568	0.836
Prayers	0.409	0.718
Call to Prayers (Adhaan)	0.757	0.802

In this research, great results were obtained as seen as Table 1 and an average accuracy of 83.2% was achieved with Naïve Bayes, influenced by stopword and stemming preprocessing and TF-IDF techniques. This study aims to identify the best classification algorithm to classify Arabic text of Prophet Mohammed (PBUH) sayings among four algorithms under study. Therefore, Bagging, LogiBoost, SVM and Naïve Bayes classification algorithms were tested.

Other research such as that of Afianto [11], conducted text classification on the Indonesian-translated Bukhari Hadith text using multiclass 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 obtaining the weight of values of each word. From the test, 90% performance was achieved. This research is the research closest to our study, but our study used single class instead of multiclass. The single class was used because a hadith should behave the main class or more inclined class to got specific information to Moslems.

In 2007, Harrag et al. [12] began researching Hadith texts in the original Arabic language. This study tried several methods to solve the problem of text classification. The first successful research [13] was published in 2008. The study used Information Retrieval (IR) by combining the Vector Space Model (VSM), Cosine Similarity, and Enriched Query. The research obtained a fairly high F1-score of 88%.

After that, Harrag et al. [14] pioneered the TC research for Hadith texts in 2009. This study used VSM because the previous research (IR) had achieved good results using the same model. This research used two datasets for this research. The first Corpus is a set of Arabic texts from different domains collected from the Arabian scientific encyclopedia (Hal Tâalam "Do You Know"). It contains 373 documents distributed over 8 categories. The Second Corpus is a set of prophetic traditions or "Hadiths" (Say's of the prophet 'Peace Be Upon Him') collected from the Prophetic encyclopedia (Alkotob Altissâa, "The Nine Book"). It includes 453 documents distributed over 14 categories. This study used the different corpus for comparing with hadith corpus. In this study, VSM was compared with the machine learning classification methods i.e. Decision Tree and Naïve-Bayes. This paper focuses on decision trees technique to classify Arabic documents. It shows that this approach outperforms some other existing systems. The comparison has been done with a probabilistic system based on the Naive Bays algorithm, statistic system based on the maximum entropy algorithm and linear system based on the vector space model using Cosine, Dice, Jaccard and the Euclidian measure. This comparison shows that our classifier is one of the best systems in term of global performance, it reports better values for the F1 and precision measures what proves that our system is more accurate.

Table 2 Result of Harrag & El-Qawasmah research [14]

System	Precision	Recall	F1
Decision Tree	**73.00**	**70.00**	**70.00**
Naïve Bayes	67.88	71.96	67.83
Maximum Entropy	50.00	84.20	62.70
VSM (Dice)	41.00	44.00	42.00
VSM (Jaccard)	54.00	61.00	57.00
VSM (Euclidian)	54.00	57.00	55.00

Yet, the results of the study were not as successful as that of previous studies, with the highest score obtained using the ID3 Decision Tree yielding an F1-score of 81% as shown in Table 2. This was because the process of TC differs from IR. IR focuses on queries built for searching using similarity from other documents. However, TC requires the learning of every attribute prior to classification. For the future works, researchers will consider passing to the multiclass classification by the use of the fuzzy decision trees or the topic segmentation technique. The last perspective is to lead other experiences to

study the impact of using root based stemmers and externals Arabic thesauruses or ontology on the performances of our classification system.

In the same year (2009), Harrag and El-Qawasmah [3] conducted a study using VSM and compared it to ANN. Data set is a set of prophetic traditions or “Hadiths’ (Sayings of The Prophet Mohammad 'Peace Be Upon Him') collected from the Prophetic encyclopedia (Alkotob Altissâa, “The Nine Book”). However, the research compares VSM and SVD to obtain word similarity. Instead, a method similar method called Singular Value Decomposition (SVD) has better performance. VSM has been proven less powerful than SVD in performing Hadith text classification. As the number of unique words in the collection set is big, the Singular Value Decomposition (SVD) has been used to select the most relevant features for the classification. SVD produces a new representation space of the observations starting from the initial descriptors by preserving the proximity between the examples [15].

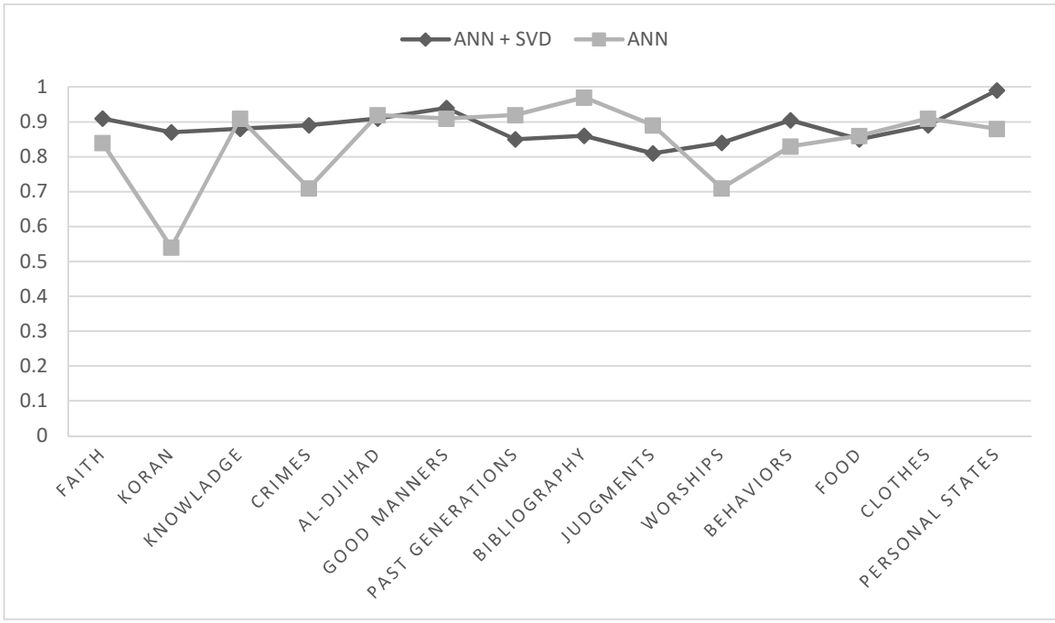

Figure 1 Result of Harrag & El-Qawasmah research [3]

This study used 453 documents categorization into 14 categories. Amount of features are 739 features with VSM and 10 - 200 features using SVD. Best accuracy used SVD on 70 and 130 features. In comparison, Harrag and El-Qawasmah [3] obtained significantly better results, with the highest score obtained using SVD and ANN, resulting in an F1-score of 90% as shown in Figure 1. The results indicate that the ANN model using the SVD method is more able to capture the non-linear relationships between the input document vectors and the document categories than that of basic ANN model. However, this research use VSM for getting the pattern from similarity. VSM has a big