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.