# 1. INTRODUCTION

This introductory chapter discusses the reasons for selecting the topic of hadith classification in Indonesian translation and the method to solve the problem.

## 1.1    Rationale

Islam is one of the largest religions in the world, especially in Indonesia which covers 12.7% of the total population of Muslims worldwide. For Muslims, there are two main sources in their teachings, namely the Al-Qur'an and Hadith. Hadith is a guide for Muslims that comes from the actions and words of the Prophet Muhammad. It becomes a guide for Muslims to better understand Islam in order to live their life in the world.

There have been many hadiths written by hadith scholars. One of the hadith scholars who has narrated many traditions is Al-Bukhari. Currently, there are about 6898 authentic traditions that have been narrated by Bukhari and have been translated into Indonesian. Hadith is divided into several categories including hadiths based on their level of validity such as sahih, hasan, maudo', and dhaif has been done by Kawther et al. [1]. In their research, they stated that several attributes in the data used are indicated as missing values. Therefore, they use a mechanism called Missing Data Detector (MDD) to deal with the missing data. MDD combined with the Decision Tree classifier can greatly increase the accuracy from 50.15% to 97.59% and it is concluded that the Decision Tree is a viable approach to hadith classification seen from its ease of induction rules.

Al-Kabi et al. classified hadith into eight categories based on Al-Bukhari's classification using the TF-IDF method [2]. The eight categories were "Knowledge", "Praying", "Eclipse", "Call to Prayer", "Faith", "Good manners", "Fasting", and "Almsgiving". The method used calculated the weight of each term that referred to each class. The weight of a term was higher if it often appeared in one class and rarely in another class. A total of 15 data from each class were used as training data and 5 data from each class were used as test data. By using TF-IDF, this study managed to get an average accuracy of about 83.2%.

The same thing has been done by Harrag [3] in his research, but with a larger number of classes in 14 categories, all of which were scattered into 453 documents in Arabic. They used Artificial Neural Network (ANN) as classifiers. Because the number of unique words contained in a document could be very large and various, Singular Value Decomposition (SVD) was used in research to find features that were relevant to the classification process.

The use of SVD as a feature selection succeeded in increasing performance with an accuracy of 88.33% compared to without using the feature selection process.

Hadith is also categorized into hadiths recommended, prohibited, and information as conducted by [4]–[7]. The categorization of the hadiths was carried out against the hadith of the Indonesian translation of Bukhari. Hafizh [4] classified each of these hadiths into one of three existing classes or what is commonly referred to as a multi-class classification. In their research, they applied the Latent Semantic Indexing method as feature selection and Bootstrap Aggregating (Bagging) to improve accuracy results with the Support Vector Machine (SVM) classification method. LSI was one of the best feature representation methods for semantic words. LSI aimed to find the relationship between a term and a document. LSI would calculate the semantic correlation between terms and documents on unstructured text representation [8].

Similar work also conducted by Al-Faraby et al. [7] in his research on "Classification of hadith into positive suggestion, negative suggestion, and information". They used the same dataset as the study [4]. SVM and Neural Network were selected as classification methods with TF-IDF assistance as feature extraction methods. Several simple rule-based string-matching techniques were constructed as comparisons to the proposed methods. This was done because at the time the research was being carried out, classifying the hadiths into the three groups (suggestions, prohibitions, and information) was a new method. This study obtained a good accuracy, which was around 88% using the F1-Score measurement metric.

In addition, [5], [6], [9], [10] categorized the hadith classification into the form of multi-label classification. Multi-label classification is grouping each data into more than one class [11], [12]. It applies to hadith data where a hadith may contain both suggestions and prohibitions. Due to those problems, research [5], [6], [9], [10] focused on working on the classification of the Indonesian translation hadith into a multi-label classification. They used a classification method that was assisted by a feature selection method to improve accuracy results. Backpropagation with the help of Information Gain done by [6], [10] obtained pretty good results with an accuracy of 88.42% while k-Nearest Neighbor with the help of Latent Semantic Analysis in research [5] managed to obtain an accuracy value of 90.28%.

Hadith classification is included in one of the text categorizations groups. Currently, there are many studies discussing text categorization [13]–[15]. Research on text categorization has come a long way with the existence of sophisticated methods such as Neural Networks and Deep Learning [16], [17]. Deep Learning is a method that can automatically search for features from a number of data with a learning process without having to be based on human-made rules, especially for processing very large amounts of data where it is a difficult thing to do and takes a very long time if done by humans [18].

One method that is known to be very good in extracting features from data is the Convolutional Neural Network (CNN). CNN can extract and select features using layers called convolution and pooling layers. These two layers are used to capture the interrelated and correlated spatial information [19], [20]. Meanwhile, the Recurrent Neural Network (RNN) has long been known as a deep learning method that works well in processing sequential data such as text classification. RNN has a 'memory' which functions to remember the context of the previous input (word) [21]. This memory also enables the RNN to learn long-term dependencies such as text and voice. However, in doing text classification RNN has a fatal weakness called vanishing and exploding gradient. This problem made RNN take a very long time to train its network [22]. Currently, this problem can be overcome, one of which is to use CNN's assistance as a method of feature extraction and selection so that RNN only processes important features, thereby increasing system performance.

This study attempts to combine the CNN method as feature extraction with RNN as a classification method (CRNN) to perform multi-label classification on the dataset of Indonesian translation of Bukhari hadith totaling 6898 text data. The effect of using CNN as a feature extraction method on the classification results will be compared without using CNN.

## 1.2    Problem Formulation

Based on the background, the problems that can be formulated in this research are:

1. The process of using the deep learning methods to improve the performance of the hadith classification.
2. The effect of adding word embedding and regularization to system performance, especially solving problems that occurs when classifying hadith using deep learning methods.
3. The effect of using CNN as extraction and selection feature to the system performance in solving vanishing and exploding gradient (weaknesses) in RNN to improve the performance/accuracy.

## 1.3    Objective

Based on the formulation of the above problems, the objective of this study is to:

1. classify hadith Indonesian translation using some deep learning methods such as multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), and bidirectional long short-term memory (Bi-LSTM) to improve the performance of hadith classification.
2. classify hadith Indonesian translation with addition of word embedding & regularization and compare the system performance without using word embedding

& regularization to solve the problems called overfitting problem in deep learning methods.

3. classify hadith Indonesia translation using CNN as the extraction and selection feature to reduce computational time and solve the vanishing and exploding gradient in RNN.

## 1.4　Hypotesis

Combining convolutional neural network for selection and extraction the features and recurrent neural network for classifier, the time needed for the classification process is shorter than the time required if using state-of-the-art feature selection and classification process (e.g. information gain and TF-IDF). While maintaining the classification results well and the training time will be shorter because there are no features that can make the network difficult to learn (features that cause vanishing and exploding gradients).

## 1.5　Problems Limitation

In order to ensure that the scope of this issue does not extend to the unrelated aspect, the scope limitation of the problem needs to be determined. The scope of the problem in this study is as follows:

1. The hadith data used are shahih Bukhari in Indonesian translation taken from Lidwa application[1] as many as 6898 data.
2. The hadith data belongs to the multi-label data which consist of three (3) classes, namely "suggestion", "prohibition", and "information".
3. Data input format in this system uses the Comma Separated Values (.csv)
4. The programming language used is Python with Google Colaboratory.

## 1.6　Research Methodology

The methodology used in this research is as follows:

1. Problem Identification

   In the problem identification stage, a literature study was conducted from previous research related to the classification text on the hadith to find out the problems that might be used as the material in this study. This stage also explored and identified the problems of various length of hadith and finding solutions that could be given.

2. Requirement Identification

---

[1] www.lidwa.com

The requirement identification was then carried out consisting of research needs and system requirements. Identification of needs related to materials and methods needed in the classification text on multi-label hadith.

The research needs included data of hadith Bukhari in Indonesian translation that needed to be collected. Identification of system requirements related to software and hardware specifications were needed to support the research.

3. System Design

   The system design was done to define the stages in creating a system classifying hadith. The system built was a system that was able to classify hadith into three classes (suggestion, prohibition, information).

4. Data Collection and Hand Labelling

   Datasets were collected from Lidwa application. Furthermore, the data collected were carried out hand-labeling process and validated by Bandung State Islamic University.

5. System Implementation

   The implementation of a system designed previously was carried out in several stages: preprocessing data, dataset spliting, feature extraction using convolutional neural network, classifier using recurrent neural network.

6. System Testing & System Evaluation

   At this stage, the system built was tested using data testing to determine system performance.

7. Results Analysis

   The last stage, the results of previous tests were analyzed and conclusions from this study were made.

## 1.7     Systematics Writing

The systematics of the final report writing in this study consists of five chapters as follows:

1. INTRODUCTION

   This chapter describes the background of research, problem formulation, objectives, hypotesis, problem limitation, research methodology and systematic research writing.

2. LITERATURE REVIEW

   This chapter discusses previous research related to classification texts on hadith Indonesian translations and discusses methods that support this research consisting of preprocessing data, feature extraction, classification methods and evaluation system methods.

3. RESEARCH METHODOLOGY AND SYSTEM DESIGN

   This chapter describes the research methodology and system design that is described by the flow chart with descriptions of each stage.

4. TESTING AND RESULTS ANALYSIS

   This chapter contains the results of system testing and analysis of the results obtained in this study by comparing several testing conditions.

5. CONCLUSIONS AND SUGGESTIONS

   This chapter contains conclusions obtained from this study and suggestions for further research.