CHAPTER 1

INTRODUCTION

This chapter discusses the reasons for selecting the topic of multilabel classification of Alquran verses in English translation and the method to solve the problem.

1.1 Rationale

The rapid growth of technology in the digital world is producing enormous amounts of data and causes unstructured data [2]. Text classification is one of the fields for processing unstructured text data into informative. Many researchers have conducted the research related to text classification. Every data used in text classification requires different treatment to obtain the best performance. Nowadays, contents of Islam are widely used in text classification research, including Hadith and Al-Quran. Hadith and Al-Quran are used as the sources of law for Muslim in the world. Both are also used as the guidelines for daily life by Muslim [3]. Previous studies related to text classification on hadith have been done by Mediamer [23] and Bakar [4]. Both of them discussed the multilabel classification of Bukhari Hadith into three classes, namely prohibition, advice, and information. Mediamer proves that rule-based feature extraction and SVM as a classifier could improve accuracy. While Bakar used Information Gain as a feature selection in classification process with Back Propagation as a classifier, and the system able to reduce running time and even increase the accuracy.

Alquran is the main source of law for Muslim [3]. Alquran consists of 6236 verses which are grouped into 114 surah [33]. The Alquran readers are devided into two, (1) Muslim and (2) non-Muslim. Muslim use Alquran as the guide and way of life. However, based on the population report that has been conducted by Pew Research Center Religion Public Life in 2015 [35], the number of Muslim population predicted to be 1.9 billion in 2020. It means, the Muslim population is 24.9% of the total population in the world. Meanwhile, to understand the Alquran is obligated to all Muslim. Nevertheless, not every reader can understand Alquran easily, cause in each verse of Alquran might contain more than one topic. So as the readers it is difficult to determine the topics of Alquran verses. On the other Hand, non-Muslim also read the Alquran to expand their insight about the main source of law in Islam. Therefore, this research is proposed to relieve the Alquran readers, especially for Muslim since it is an obligation to learn the Alquran.

Verses of Alquran can be classified into multi-label topics. Subsequently, the topics that contained in Alquran can be grouped into 15 topics, such as (1) Pillars of Islam, (2) Faith, (3) Alquran, (4) Science and its Branches, (5) Working, (6) Call to God, (7) Jihad, (8) Human and Social Relations, (9) Morals, (10) Regulations Related to Property, (11) Legal

Matters, (12) State and Society, (13) Agriculture and Trade, (14) History and Stories, (15) Religions [27]. Meanwhile, the system to be built needs a process which is able to classify each verse of Alquran into multi-label topics precisely.

Pane [27] has conducted research related to the multi-label classification of verses of Alquran. The research uses bag-of-words as feature extraction and Multinomial Naive Bayes as a classifier. However, bag-of-words does not focus on the semantics relationship between each term. Other than that, Izzaty [16] dan Ulumudin [34] both have also conducted study related to multilabel classification on the Alquran verses. Izzaty proposed a Tree Augmented Naïve Bayes as a classifier method, while Ulumudin proposed KNN with tf-idf weighted. On the other hand, Huda [14] proposed the Neural Network approach as a classifier with Adam Optimizer in handling multilabel data on Alquran verses. Meanwhile, Nurfikri [26] has done with the study to compare the performance between Neural Network approach and SVM on multilabel classification of Alquran verses.

Previous research related to the text classification has also been done by Kim [18]. The study classified news in English translations. One of the features used in the study is semantic concept features. The feature weight is obtained using the Lesk Algorithm which focuses on semantics for each term [6]. This study describes a document into 2^{nd} order-tensor (matrix) that called as semantic features. So as the semantic features described as independent space in 3^{rd} order-tensor of dataset. TSM is also called multilinear algebra which is a generalization of the matrices [9]. Kim [18] proposed the Semantic Naive Bayes classification, the method is a development of conventional Naive Bayes in order to be used on 3-order tensors. Accuracy results from the study reached up to 98%. Even so, the Lesk Algorithm used in the 3-order tensor still has weaknesses which is depending on word definitions that exist in dictionary. However, the number of vocabulary increases fastly and sometimes the dictionary cannot cover the new vocabulary.

This study attempts to use word embedding as semantic concept features extraction that used for multilable classification for the data of Alquran verses in English translation. The effect of classification results using word embedding as a semantic concept feature extraction is compared with the baseline extraction method using the Lesk Algorithm.

1.2 Problem Formulation

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

1. The Lesk algorithm relies on dictionary in the feature extraction process.
2. The Lesk algorithm produced a high feature dimension, since the model used wikipedia pages to extract semantic concept feature.
3. The Lesk algorithm takes a lot of learning time, since it needed the overlapping calculation on the feature extraction process.

1.3 Objective

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

1. Propose a feature extraction method that not relies on dictionary to overcome lack of Lesk algorithm.
2. Reducing the size of feature dimensions that produced by Lesk algorithm.
3. Reducing the learning process on building the features.

1.4 Hypothesis

The hypothesis to this research are as follows:

Premise 1 : Naive Bayes can be used in multilabel classification based on verses of Alquran [27].

Premise 2 : 3^{rd} order-tensor based on semantic concept features and Semantic Naive Bayes as a classifier produced a best accuracy result [18].

Premise 3 : Word Embedding as feature extraction has been proven to produce the best accuracy in text classification [23], [31].

Hypothesis : The Word Embedding used in semantic concept features and Semantic Naive Bayes as classifier on 3^{rd} order-tensor could increase the accuracy result for multilabel classification based on topics of Alquran verses in English translation.

In this research, semantic concept features using a word embedding on 3^{rd} order-tensor. The use of this feature aims to reduce the dimensions of wikipedia corpus, and could overcome the lacks of Lesk Algorithm, because the algorithm much depends on the dictionary. Different from [18]. the research uses Lesk Algorithm as a semantic concept feature. However, this study also uses Wikipedia as external corpus to produce the semantic concept features. Hopefully using word embedding in semantic concept features produce best accuracy on multi-label text classification for Alquran verses in English translation. In addition, the innovation result proposed through this study is to develop a system that able to relieve Alquran readers in understanding the topics of Alquran verses more easily and appropriately.

1.5 Problem Limitation

The limitation of the problem in this study are as follows:

1. The word embedding vector is obtained only from pretrained data which is built by the English corpus from Wikipedia using standard English language.
2. The process of splitting the train and test data uses only one file division.
3. One classifier used in this research is Semantic Naive Bayes.

1.6 Research Methodology

The methodology used in this research is as follows:

1. Problem Identification

   In the problem identification stage, the literature study was conducted by reviewing previous research with similar research fields to this study such as text classification on the Alquran verses. This section is also to find problems that exist in previous research to produce the solutions that can overcome them.

2. Requirement Identification

   Requirement Identification consists of two parts, namely research needs and system requirements. This process is carried out to identify the materials and methods needed in this study.

3. System Design

   The system design was conducted to define the processing steps in creating a system. The purpose of the system is to classify Alquran verses in English translation using various methods.

4. Data Collection

   Datasets were collected from Pane's research [27], which is previous research as main reference of this study.

5. System Implementation

   The system implementation of this research was carried out in several stages: preprocessing data, dataset spliting, collecting wikipedia pages, feature extraction, classification, and system performance.

6. System Testing System Evaluation

   At this step, the system built was tested using data testing to measure the system performance in classifying the Alquran verses.

7. Results Analysis

   Finally, the results of the system in this study were analyzed and concluded in this step of research.

1.7 Systematic Writing

The systematic 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, hypothesis, problem limitation, research methodology and systematic research writing.

2. LITERATURE REVIEW

   This chapter discusses previous research related to classification texts on Alquran verses in English translation and its method that support this study consisting of preprocessing data, feature extraction, wikipedia pages, word embedding, classification methods, and system evaluation methods.

3. RESEARCH METHODOLOGY AND SYSTEM DESIGN

   This chapter describes the research methodology and system design in building the system of this research.

4. TESTING AND RESULTS ANALYSIS

   This chapter describes the results of system testing and analysis of the results based on the testing scenarios in this study.

5. CONCLUSIONS AND SUGGESTIONS

   This chapter describes the conclusions obtained from this study and suggestions for further studies.