CHAPTER 1

INTRODUCTION

1.1 Rationale

Hadiths are important textual sources of law, tradition, and teaching in the Islamic world [21]. Along with the development of the era, there's a lot of research on the hadiths such as with the application of Natural Language Processing (NLP) and one of them is the Classification of hadith [21].

Hadith classification is a way to categorize hadits into a particular category based on its contained information [20]. Hadith structure is different from any other text representation. [21] explain that hadith consist of 3 components, that is Matn, Isnad and Taraf. Matn is the central text, Isnad is chain of narrators and Taraf is the beginning phrase(s) of the Hadith. In addition, some hadiths belong to more than one label, for example in the Hadiths in the books of Sahih Al-Bukhari [21]. Thus, to deal with this, a multilabel classification approach is needed.

Multilabel classification is a form of supervised learning where the classification algorithm is required to learn from a set of data, where each data can belong to multiple classes, different from single label classification where one data only belong to one class. For example a movie can simultaneously belong to action, crime and thriller categories [14]. However, the generality of multi-label problems makes it more difficult than the others.

Generally in text classification, features are terms or words contained in the text. Usually a document or textual data contains considerable amount of words that can cause high computational complexity and decrease accuracy because some attributes may be irrelevant [12].

To overcome these problem, it takes a feature reduction. One way to perform the feature reduction is to use the feature selection [4]. Feature selection is a process to select the relevant features to be used in the classification process.

One feature selection method that can produce good results is Chi-Square [26]. However, Chi-Square has a problem, one of which is that all measured participants must be independent, meaning that one individual cannot fit into more than one class, or single label. In addition, another disadvantage is that the data used must be data frequency (multinomial). This is a limitation because the text in the hadith is a short text, which according to [17] that the Bernoulli model can work well with a small number of features.

1.2 Theoretical Framework

Chi-square is one method to perform a selection feature well, so that it can improve performance from classification performance. This is because Chi-Square can choose attributes that have high relevance to the class. However, the chi-square has limitations, one of which is one attribute cannot fit into more than one class, or single label. In addition, another disadvantage is that the data used for multinomial data and it can cause to overfitting and based on [17] that multinomial data is not suitable for short text, so it does not fit the data in this study, which only has an average of 20 attributes per document.

1.3 Conceptual Framework

The basic concept of the proposed method is to addition Bernoulli model to improve the Chi-Square feature selection. This study observes the effect of the feature selection and classification approach to the performance of the multi-label Hadith classification.

1.4 Problem Statements

Based on some research that has been done, that using chi-square feature selection can outperform other methods [9, 21, 26]. However, the Chi-Square is not without problems, one of which is that all measured participants must be independent, i.e. one individual cannot fit into more than one class, or a single label. In addition, another disadvantage is that the data used must have multinomial data frequency. This is a limitation because the text in the Hadith is short and based on Manning et. al. [17] not suitable for using multinomial distribution.

1.5 Objective

The objective of this research is to improve the performance of the multi-label hadith classification by making improvements in the chi-square feature selection.

The specific objectives are as follows:

- to improve the performance of multi-label hadith classification,
- to make chi-square suitable with Bernoulli distribution, and
- to make chi-square feature selection can be used for multi-label classification.

1.6 Hypotheses

The use of the Bernoulli model can improve performance from the chi-square feature selection so that it can improve classification performance. This is because Bernoulli cares about counts for a single feature that do occur and counts for the same feature that do not occur [17]. In addition, binary relevance is used to overcome chi-square problems that cannot fit into more than one class and improve classification performance because each classifier only focuses on one class [27].

1.7 Scope and Delimitation

The scope of this research is focused on feature selection that can improve classification rate of multi-label hadith Classification on Hadith Sahih Al-Bukhari on Indonesian translation.

1.8 Importance of the Study

The importance of this research is to improvise in a feature selection especially for hadith data, and also as a solution to the feature selection in multi-label classification.