CHAPTER

1 INDTRODUCTION

This chapter discusses the underlying background of the research and the method to solve the problem. This chapter includes the following subtopics, namely: rationale, conceptual framework/paradigm, statement of the problem, objective and hypothesis, assumption, scope and delimitation, and significance of the study.

1.1 Rationale

The classification text problem is a common problem in machine learning used to predict class labels by studying the input data, which predicts class accuracy using the output data [1]. The imbalanced data is one of the problems found in text classification. Imbalanced data significantly affects the evaluation result on classification. The imbalanced data means the data in which the major class is much higher than the minority class [2].

Imbalanced data can be found in various real-world cases, including credit card fraud detection [3], fault diagnosis of induction motors [4], intelligent fault diagnosis of machines [5], disease prediction problems [6], fault prediction category hadith and al-Qur'an [1][7]. For example, in the case of credit card fraud detection, the data regarding this case is a rare occurrence where only a 1-5% [3] chance of credit card detection fraud occurring. Most machine learning algorithms need to add specific effort to give better work results when working with imbalanced datasets. So, the problem of imbalanced data is important to resolve because it will significantly impact the evaluation results.

Imbalanced data dramatically influences accuracy results, as explained in the research conducted by Abu Bakar et al. [8], that imbalanced data has quite an effect on accuracy. Several researchers have proven that handling imbalanced data can improve accuracy using specific methods. A study conducted by Hong-Chan Cang et al. [4]using the Deep Convolutional Generative Adversarial method produced a diagnostic accuracy of 90% using an additional oversampling method. This is close to the accuracy achieved using a balanced data set. Then using the additional oversampling method increases the accuracy by 5%. At the same time, the research conducted by Bassam Arkok et al. [7] using ensemble learning can increase accuracy by 1-2%.

This thesis discusses handling imbalanced datasets in Bukhari's hadith translated into Indonesian. Research on hadith datasets has been carried out by Kustiawan et al. [1] and Abu Bakar et al. [8]. However, previous research only built a multi-label text classification system. This study optimized the classification system to obtain better evaluation results and handle imbalanced data.

1.2 Conceptual Framework/Paradigm

Research on classification generally consists of several stages, including preparing the dataset to be used, preprocessing, classification model, and evaluation. These stages can be seen in the image below.

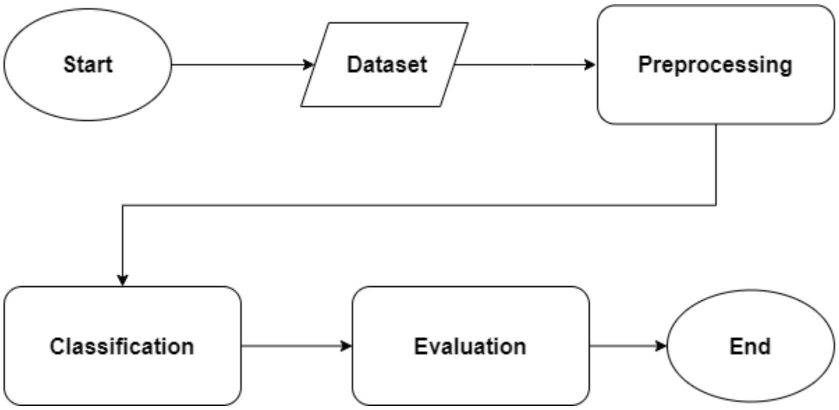

Figure 1 Classification Process in General

In classification research, many types of datasets were used. In the research conducted by Kustiawan et al. [1] and Abu Bakar [8] used a multi-label hadith dataset in Indonesian translation. At the same time, in research conducted by Renaldi Pane et al. [9] using a multi-label Al-Qur'an dataset in English translation. The research conducted by Amit Singh et al. [3] used a credit card fraud detection dataset with very imbalanced data. The research was conducted by Yousif A.Alhaj et al. [10] using a novel dataset in Arabic.

In the preprocessing stage, the most common are data cleaning, stop word removal, stemming, and tokenization. Usually, after the data is tokenized, the words are weighted or converted into numbers to be processed in machine learning. For example, in the research conducted by Choirulfikri et al. [11] at the preprocessing stage using case folding, punctual removal, stemming, stop word removal and tokenization methods.

In addition, the next stage is the classification method used in text classification. In this research, the data used is a multi-label dataset that is imbalanced data. Many researchers have researched imbalanced data using various machine-learning techniques [12]. Using the Graph Convolutional Network technique, Masha Grohbani et al. [6] studied disease prediction problems on imbalanced data. Then Hong-Chan Chang et al. [4] conducted a Fault Diagnosis of Induction Motors with Imbalanced Data study using the Deep Convolutional Generative Adversarial Network technique. Meanwhile, in the review paper on imbalanced data conducted by Adane et al. [13], there are four techniques for handling imbalanced data: resampling methods, classifier adaptation, ensemble methods, and cost-sensitive approaches. All the techniques used by the researchers above have their advantages and disadvantages.

Moreover, the last is the evaluation stage of the classification. Generally, there are two evaluations in classification, namely by using a confusion matrix and hamming loss. However, research conducted by Pereira et al. [14], there are several evaluations used in the case of multi-label datasets, including Hamming Loss, Accuracy, F-micro-AUC Precision, Recall, micro-f-measure, macro f-measure, subset accuracy, average precision, rating loss, single fault coverage, macro precision, micro precision, 0/1 subset loss, micro draw, macro draw, micro AUC, and macro AUC.

The accuracy results obtained vary depending on the variable to be used. Usually, in data cleaning, if one of the data cleaning processes is omitted, for example, stop word removal and stemming, it affects the results of the evaluation accuracy. Then even in the classification model, if the variables used are different, it affects the accuracy of the results.

1.3 Statement of the Problem

This research focuses on handling imbalanced multilabel data. Based on the practical problems conducted by experts discussed in 1.1. Research on the classification of imbalanced data text still needs to be done because there are still cases of imbalanced data whose accuracy and f1-score still need to be improved, especially on multi-label datasets [1], [8], [9]. Since the classification results in imbalanced data are often misclassified, optimal accuracy cannot be achieved. There are several formulations of the problem in this research, including:

1. How to handle imbalanced data against multi-label classification.
2. How to use ensemble stacking methods can improve the hadith classification performance.

1.4 Objective and Hypotheses

This research uses the ensemble stacking method to build a multi-label text classification system using an Indonesian translation of Bukhari hadith data whose data is imbalanced. The system built using the ensemble stacking method is expected to provide better performance and accuracy than previous studies. The built system can also provide evaluation results (F1-Score) for handling imbalanced data and can also provide a relatively short computation time. Stacking is used because it can combine different algorithms with improving the classification accuracy of imbalanced data classification.

Several studies have proven that using the ensemble method can increase the accuracy by 1-2 % [1], [7]. Several studies handle imbalanced data using the ensemble method. Kustiawan et. al [1] and Arkok [7], by using the ensemble learning method (bagging and boosting), can provide good accuracy. The research by Ziang et. al [15] using the Self-Adaptive Stacking Ensemble Model method provided very high accuracy results. Nevertheless, using ensemble stacking in terms of computational time requires a longer time compared to a single classifier. At the same time, research conducted by Rout et al. [2] found that using the ensemble method was very good at handling cases of imbalanced data.

1.5 Assumption

Hadith are guidelines for Muslims in the world which are used as a way of life. The hadith dataset consists of 3 classes: recommendations, prohibitions, and information. To make it easier to study hadith, a system is needed to categorize them.

1.6 Scope and Delimitation

To ensure that the scope of this issue does not extend to the unrelated aspect, it is necessary to determine the boundaries of the problem scope. The limitations of the scope of the problem in this study are as follows:

1. The dataset used is the Bukhari hadith dataset in Indonesian translation.
2. Hadith Bukhari datasets are multi-labelled and imbalanced data.
3. The basic model used in the ensemble stacking model experiment has several models, including Random Forest, KNN, Gaussian NB, SVM, Decision Tree and AdaBoost.
4. The meta-learner used in the ensemble stacking is logistic regression.
5. In this research, ensemble stacking is used, which has the disadvantage that the computation time required is long.
6. In previous studies, the evaluation only used accuracy/hamming loss, but this study added an evaluation method, namely F1-Sore. This is because this research focuses on handling imbalanced data.
7. Data input format in this system uses the Comma Separated Values (.csv).
8. The programming language used is Python with Google Collaboratory Pro.

1.7 Significance of the Study

This research hopes to contribute to the field of machine learning, especially the topic of imbalanced data multi-label text classification. There are two primary significances, namely practical and theoretical significance. In its practical significance, it can provide a solution for handling imbalanced multi-label data using an ensemble stacking approach. At the same time, the theoretical significance is that this study uses a hadith that is expected to make it easier for Muslims to study it according to their class or category correctly.

Considering the problems of classification multi-label and imbalanced dataset, research on multi-label text classification on an imbalanced dataset needs to be done. Using stacking is expected to improve the results of multi-label classification on an imbalanced dataset. Stacking is used because it can give the result of a combination of different algorithms to improve the classification accuracy of an imbalanced dataset.