Improving Chi-Square Feature Selection using a Bernoulli Model for Multi-label Classification of Indonesian-Translated Hadith

Fahmi Salman Nurfikri, Adiwijaya
School of Computing
Telkom University
Bandung, Indonesia

***Abstract*—Hadith is the foundational knowledge in Islam that must be studied and practiced by Muslims. In the Hadith, several types of teachings are beneficial to Muslims and all of mankind. Some Hadith serve as advice, while others contain prohibitions that Muslims should adhere to. There are yet others that do not belong to these categories and serve only as information. This study focuses on increasing the performance of Chi-Square feature selection to obtain relevant features for multilabel classification of Indonesian-translated Bukhari Hadith data. This study proposes a Chi-Square-based Bernoulli model to improve Chi-Square feature selection which is appropriate for short-text data such as Hadith. The findings of this study show that the proposed method can select relevant features based on data classes; thereby improving Hadith classification performance with an error value of 9.38% compared to that (9.91%) obtained using the basic Chi-Square feature selection.**

Keywords*—Bernoulli model; Chi-Square; feature selection; hadith classification

I. Introduction

Hadith is an important textual source of law, tradition, and teachings in the Islamic world [1]. Following the advancement in technology, several research studies have been conducted on Hadith including the application of natural language processing to classify Hadith based on its content. Hadith classification is a method of categorizing Hadith based on its content [2]; the structure of an Hadith is different from other textual representations. A Hadith comprises three components: Matn, Isnad, and Taraf [1]. Matn is the central text, Isnad is the chain of narrators, and Taraf is the beginning phrase(s) of the Hadith. In addition, some Hadith, for example, the Hadith provided in the book of Sahih Al-Bukhari, belong to more than one label (i.e., the data is multilabel) [1], and therefore, a multi-label classification approach is required.

Multilabel classification is a type of supervised learning where a classification algorithm needs to learn from datasets and classify data into multiple classes; in single-label classification, data can only be classified into one class. For example, a movie is multilabel data as it can simultaneously be categorized as action, crime, and/or thriller [3]. However, the generality of multilabel data makes it more difficult to classify it compared to other data.

In text classification, the features are terms or words contained in the text. A document or textual data contain a considerable number of words that can cause high computational complexity and decrease accuracy as irrelevant features may be considered during the classification [4]. To overcome this limitaion, feature reduction must be applied. One method of feature reduction is feature selection [5] wherein only relevant features to be used for classification are selected. An example of a feature selection method that has been proved to produce good results is the Chi-square [6]. However, one of the limitations of the Chi-Square is that all measured participants must be independent, i.e., one individual cannot fit into more than one class or a single label. Further, its other disadvantage is that the data must have multinomial data frequency. This is a limitation in our case because the text in the Hadith is short. The Bernoulli model has been proved to work effectively with few features [7] and therefore, it is worth exploring.

II. Related Work

A. Hadith Classification

A considerable amount of research has been conducted on Indonesian-translated Hadith, with Faraby et al. [8] being a notable work in this area. Their study categorized Sahih Al-Bukhari Hadith data into three classes: advice, prohibition, and information. The study compared the classification results using artificial neural networks (ANNs) and support vector machine (SVM), and they applied term frequency-inverse document frequency. The results of the study showed that the SVM method performed better than the ANN method, with an F1-Score of 88% to 85%.

Furthermore, Afianto et al. [9] used a dataset similar to Faraby et al. [8]; however, they used random forest as the classification method. The study obtained an F1-Score of 90%, which is better than that of previous study [8]. The most significant process in this research study was the determination of the bootstrap method used where the bootstrap sample was set to 100.

Bakar et al. [10] conducted multi- and single-label Hadith classification using 1064 data points. The multilabel classification comprised three classes (advice, prohibition, and information), while the single-label classification comprised five (faith, knowledge, ablution, prayer, and prayer times). The

study used information gain (IG) as the selection feature technique and the backpropagation neural network (BNN) as the classifier. The study obtained an F1-score of 65.275% for single-label classification, while the Hamming Loss value for multilabel classification was 0.1158. Hence, using IG as the selection feature technique significantly improved the classification performance of the model.

B. Multi-Label Classification

Classification of multilabel data can be problem as such data can be categorized into two or more classes. Research on multilabel classification is motivated by medical diagnosis and text categorization problems. Two approaches can be used for multilabel classification: problem transformation and algorithm adaptation [3].

The problem transformation approach solves multilabel problems by transforming multilabel data into single-label data, while the algorithm adaptation approach classifies multilabel data using algorithms designed for multilabel classification. Multilabel classification using the problem transformation approach achieved better performance than that using the algorithm adaptation approach [3].

Several studies on multilabel classification have been conducted; however, only a few such as those of Bakar et al. [10], Mediamer et al. [11], and Kabi et al. [12] focused on the multilabel classification of Hadith data. Liu et al. [13] conducted multilabel classification using a correlation function; this was effective for overfitted and noisy data. However, such methods are not designed to obtain optimal parameters, and therefore, this can affect classification performance. Soleimani et al. [14] used semi-supervised learning methods and latent Dirichlet allocation for learning topic classes, and used a small number of labeled training data for multilabel classification. However, this method also had a limitation; it had a high time complexity given the large amount of data used in the study. Huang et al. [15] combined a feature selection technique and a classifier for multilabel classification thereby providing an advantage for selecting relevant features of each label and training the classifier to increase the effectiveness of the model. However, the drawback of this method was that it requied a high computational time to obtain optimal parameters.

C. Feature Selection

A problem with text classification is that textual data contain a considerable number of words that can cause high computational complexity and decrease the accuracy of classification results [16], [6], [17]. One approach to tackle this problem is applying feature selection to the data. Yang et al. [6] investigated document frequency (DF), IG, Chi-Square, mutual information (MI), and term strength as feature selection methods for the Reuters corpus. The experiment found that IG and Chi-Square were the most effective feature selection methods as they could remove 98% of irrelevant features without compromising classification performance. However, Chi-Square and IG showed a limitation in that they incurred high computational cost, whereas DF had the lowest computational cost but strong correlation with Chi-Square and IG.

Forman [18] presented an extensive comparative study of feature selection metrics for text classification of high-dimensional data focusing on SVM for the two-class problems. Forman found that the new feature selection metric—Bi-normal separation—achieved better performance compared to other feature selection methods. Xu et al. [19] compared DF, IG, MI, and pointwise MI and found that MI and IG achieved the same performance. Another study used a Bernoulli model as the feature selection method [7] and found that it worked best for documents with short texts, while a multinomial model was better for handling documents with long texts.

III. Design Process

The steps followed by the proposed method (Fig. 1) are described in this section.

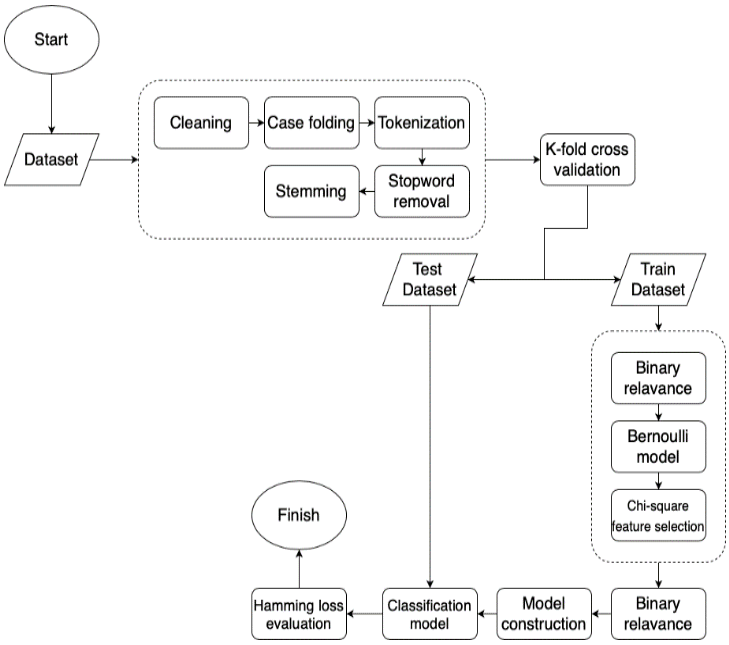

Fig. 1. System Design.

The initial stage involves collecting text-based data from the Hadith in Bahasa Indonesia from the book of Hadith Sahih Al-Bukhari; the book consists of 1066 data points and is divided into three class labels (Advice, Prohibition, and Information). An example of the data representation is listed in Table I.

TABLE I. Multi-Label Data Representation of Indonesia-Translated Hadith

Data	Class
'*Janganlah kalian berdusta terhadapku (atas namaku), karena barangsiapa berdusta atas namaku dia akan masuk neraka*.'	Prohibition Information
'*Kami pernah shalat Maghrib bersama Nabi ketika matahari sudah tenggelam tidak terlihat*.'	Information

TABLE II. COMBINATION OF CLASSES IN THE DATASET

No.	Advice	Prohibition	Information	Count
1	0	0	0	0
2	0	0	1	777
3	0	1	0	6
4	0	1	1	53
5	1	0	0	10
6	1	0	1	181
7	1	1	0	5
8	1	1	1	34

The class combination in the dataset is listed in Table II.

The dataset consisted of 230 Advice, 98 Prohibition, and 1045 Information data points. Based on these data points, it can be seen that the number of data points in the Advice and Prohibited classes is very small compared to that in the Information class; hence, the data is unbalanced. This can be a problem because unbalanced data can lead to less optimum classification results.

The first step to handle unbalanced data is preprocessing. This study used cleaning, case folding, tokenization, stopword removal, and stemming as the preprocessing steps to eliminate some sentences that are not used in the classification process. An Indonesian stopword list from a study conducted by Tala [20] was used and modified to match the Hadith dataset in this study. In addition, the Nazief–Andriani stemming algorithm [21] was also used. Next, the dataset was split into training data and test data using 5-fold cross-validation to make all observations in the dataset are nicely distributed in a way that the data are not biased.

Feature extraction was performed using the bag-of-words representation. In this study, a term frequency method was used to extract the feature. This method counts each word in the vocabulary list obtained from the training dataset for each data point.

Two general approaches for multi-label classification are problem transformation and algorithm adaptation. Problem transformation converts multi-labeled data into single-labeled data, while the algorithm adaptation uses algorithms specifically adapted to handle multilabel classification. Based on the research conducted by Irsan et al. [3], the problem transformation approach achieved better performance results compared to algorithm adaptation.

Binary relevance uses problem transformation approach [22] [23]. Binary relevance creates a number of k datasets (k = |L|, the total number of classes). Each dataset has the same instance as the original data; however, each dataset contains only one class. Using this method, class data representation must first be changed into one-hot encoding.

The next step involved duplicating the dataset of q, where q is the number of classes in the training data so that each dataset only has 2 classes, namely. 0 and 1.

Next, the extracted features are selected using Chi square. In this study, a Bernoulli model was used for Chi square feature selection. This model checks for the presence or absence of a word, and therefore, it only has two possible outcomes: yes or no. The Bernoulli model was used because every Hadith contains an average of 20 words, and hence, a small number of features is the type of data that the Bernoulli model can process effectively [7]. The algorithm of the Chi square Bernoulli model is presented in Algorithm 1 below.

Algorithm 1. Chi-Square Bernoulli model algorithm

```
function Bernoulli-Chi-Square-FS()
    Input: Array of attribute and its class C
    Output: Array of Chi value for each class
Initialize
arrayofchivalue (array)
arrayofclasschivalue (array)
Begin
Change class data representation into one-hot encoding
Break the class into k classes
for each c in class do
  for each a in attributes do
    for each row in a do
      if row >= 1 then
        row ← 1
      else
        row ← 0
      end if
    end for
    Calculate ChiSquare(a, c_i) and append it to arrayofchivalue
  end for
  Sort descending arrayofchivalue
  Get top-n attributes and append it to arrayofclasschivalue
end for
return arrayofclasschivalue
```

Based on Algorithm 1, each feature row is transformed into 0 and 1. In this model, words with three occurrences are the same as words with only one occurrence.

Then, feature selection is performed for each class. Feature selection is the process of selecting a subset of relevant features for training a classification model; it is used to select relevant features that will be included in the classification process, thereby efficiently and effectively improving the process [16]. Chi square feature selection is used in this study [6]; it is expressed by

$$X^2(t,c) = \frac{N*(AD-CB)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (1)$$

A Chi square statistic measures the lack of independence between term t and class c, and it can be compared to Chi square distributions with one degree of freedom to evaluate extremeness [6], where A is the number of times t and c occur; B is the number of times t occurs without c; C is the number of times c occurs without t; D is the number of times neither t nor c occurs; and N is the total number of documents.

The output of Chi square is the Chi value, which is between a feature and a class; the greater the Chi value, the greater is the relationship between the feature and the class. Each feature is calculated for each class. Once the Chi values are obtained, they are sorted in the descending order for each class, where greater the Chi value, the greater is the effect of a feature on a class [8]. Finally, the top-n features are obtained and used as inputs for the classifier.

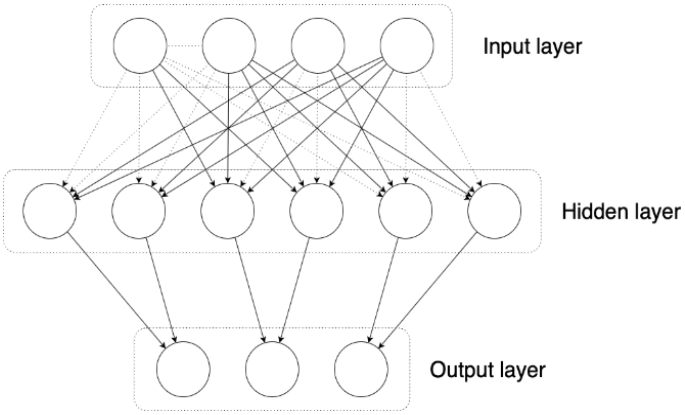

Fig. 2. Structure of the Neural Network.

The classifier is trained using the BNN. This algorithm was selected because it can process a wide variety of features to obtain a high classification performance [24], [4], [1], [10]. The classifier was trained using a modified BNN with the binary relevance approach, and therefore, the training process was conducted three times, which is equal to the number of classes. The selected features were used to train the classifier. Each class had different input data because of the different selected features. Therefore, in this the BNN was modified to tackle this problem, as shown in Fig. 2.

Fig. 2 shows that two main lines connect the input layer and the hidden layer, i.e., the bold and dotted lines. The bold line indicates that input and hidden neurons are connected, while the dotted line indicates that the input does not pass the feature selection for the class; however, it can pass the feature selection for other classes. Finally, the evaluation results of the classifier are expressed in terms of Hamming Loss. The Hamming loss is used because this method is appropriate for multilabel classification and assigns equal weight to each label [25].

IV. Results and Discussion

A. Effect of the Bernoulli Model on Chi-Square Feature Selection

The performance of the Bernoulli model was compared by varying the number of dimensions from 10% to 100%. This allows determining if the use of feature selection can help improve the performance of the classifier and to obtain the best dimensionality for optimal classification performance. The BNN input nodes are equal to the dimension of the document vector.

The results of the proposed model are compared with those of the typical BNN and Chi-square-based BNN feature selection models. The results are listed in Tables III and IV.

Based on the results listed in Tables III and IV, the proposed method achieves the best average result of 0.0938, while the CSBNN produced the best average result of 0.0991. A comparison chart of the three methods is shown in Fig. 3.

TABLE III. Hamming Loss Result of Classification using Chi-Square Bernoulli Model and Backpropagation Neural Network (BCSBNN)

Number of dimension	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
10%	0.1106	0.1064	0.0955	0.0955	0.0814	0.0979
20%	0.1308	0.1221	0.0955	0.0939	0.0861	0.1057
30%	0.1324	0.1252	0.0970	0.0908	0.0829	0.1057
40%	0.1121	0.1142	0.0970	0.1033	0.0829	0.1019
50%	0.1153	0.1111	0.0939	0.1049	0.0798	0.1010
60%	0.1075	0.0986	0.0892	0.0955	0.0782	0.0938
70%	0.1168	0.1095	0.1033	0.1002	0.0923	0.1044
80%	0.1293	0.1158	0.0939	0.1017	0.0876	0.1057
90%	0.1184	0.1064	0.1064	0.1017	0.0923	0.1051
100%	0.1137	0.1127	0.0923	0.0986	0.0814	0.0997

TABLE IV. Hamming Loss Result of Classification using Chi-Square and Backpropagation Neural Network (CSBNN)

Number of dimension	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
10%	0.1106	0.1111	0.0970	0.0986	0.0782	0.0991
20%	0.1231	0.1142	0.1049	0.1017	0.0782	0.1044
30%	0.1075	0.1299	0.1017	0.1002	0.0782	0.1035
40%	0.1199	0.1189	0.0955	0.0955	0.0829	0.1025
50%	0.1184	0.1127	0.0970	0.0845	0.0892	0.1004
60%	0.1168	0.1174	0.0955	0.0892	0.0782	0.0994
70%	0.1215	0.1142	0.0939	0.1033	0.0845	0.1035
80%	0.1199	0.1299	0.0939	0.0986	0.1158	0.1116
90%	0.1246	0.1189	0.1064	0.0970	0.0845	0.1063
100%	0.1153	0.1111	0.0939	0.1049	0.0798	0.1010

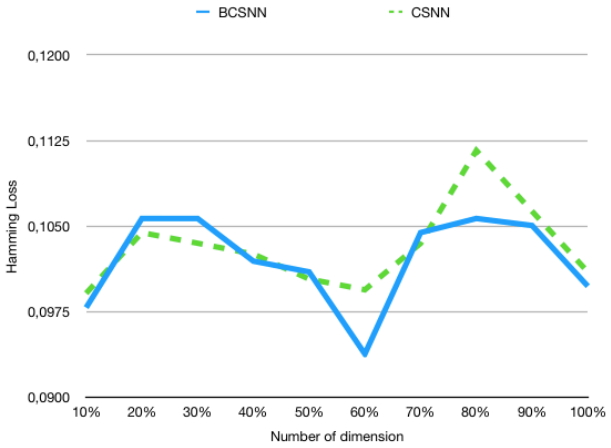

Fig. 3. Categorization Performance of BCSBNN and CSBNN according to the Number of Dimensions

As shown in Fig. 3, the performance of BCSBNN and CSBNN are not considerably different; however, on average, BCSBNN outperformed CSBNN. This is because Bernoulli distribution can select relevant features as inputs for the BNN and the Bernoulli distribution only has two possibilities (yes or no). For example, consider the word "hendak;" in the Bernoulli distribution, one word is enough to represent the word in the class to produce the probability p (hendak = 'yes' | class) and p (hendak = 'no' | class). Meanwhile, by using multinomials, the number of occurrences of each word has its respective probabilities such as p (want = 0 | class) and p (want = 1 | class). Therefore, this decreases the occurrence probability of each word.

Further, Fig. 3 shows that when using 60% of the data dimensions, the smallest Hamming Loss value is obtained. This is because the features used as inputs for the classification models match the test data. However, this can change depending on the data used. In addition, using feature selection produced better results than when the whole data (using 100% dimension) was used. This is because the feature selection technique removed irrelevant words/features from the dataset used in training and testing the classification model. However, it is necessary to determine the best parameters for choosing the number of feature dimensions to use.

B. Comparison of the Modified Backpropagation Neural Network and the Typical Backpropagation Neural Network Classification Performance

Table V lists the classification performance of the modified BNN using the binary relevance approach compared to that of the typical BNN. The performance of the networks was first compared using the Chi-Square Bernoulli Model (BCS) and then using the Chi-Square (CS) test.

TABLE V. PERFORMANCE COMPARISON BETWEEN THE MODIFIED BACKPROPAGATION NEURAL NETWORK (BINARY RELEVANCE) AND THE TYPICAL BACKPROPAGATION NEURAL NETWORK

Number of dimensions	Modified Neural Network		Original Neural Network	
	BCS	**CS**	**BCS**	**CS**
10%	**0.0979**	0.0991	0.1126	0.1129
20%	0.1057	**0.1044**	0.1135	0.1253
30%	0.1057	**0.1035**	0.1123	0.1263
40%	**0.1019**	0.1025	0.1110	0.1151
50%	0.1010	**0.1004**	0.1094	0.1119
60%	**0.0938**	0.0994	0.1094	0.1135
70%	0.1044	**0.1035**	0.1204	0.1126
80%	**0.1057**	0.1116	0.1088	0.1132
90%	**0.1051**	0.1063	0.1132	0.1163
100%	**0.0997**	0.1010	0.1094	0.1132

As shown in Table V, the modified BNN outperformed the typical BNN. This is because, in the typical BNN, the classifier must remember more patterns in the class, whereas, in the binary relevance, one classifier is focused on only remembering one pattern. For example, for the typical BNN, the classifier must remember eight different class patterns and a combination of unbalanced data, as listed in Table II. Meanwhile, the binary relevance method only focuses on each class, i.e., Advice, Prohibition, or Information.

Further, the BNN following the binary relevance approach requires less computational complexity than the typical BNN because the number of neurons connected in the former were reduced, thereby reducing the matrix computation. However, using binary relevance slightly increased time complexity because the classifier had to learn as many class patterns as possible. This can be a problem if the number of classes to be trained becomes very large.

C. Model Prediction

Samples of the classification results are listed in Table VI.

The conducted experiments and results listed in Table VI indicate that there are three types of predictions: correct prediction, partially correct prediction, and wrong prediction. Further, in the "correct prediction" row, the predictions and targets achieved the same results because the words that appear in the Hadith were relevant, and thus, only correct results were obtained.

In the "partially correct prediction" row, the system predicted only the information class, while the target classes were Advice and Information. This is because the number of Advice data points was so small that the probability of the system in retrieving the Advice class was trivial compared to that in retrieving the Information class, which has a very large number of data points. In the future, further processing of unbalanced data must be performed.

TABLE VI. SAMPLE PREDICTION

	Data	**Predicted**	**Target**
Correct prediction	*'Jika salah seorang dari kalian meludah maka janganlah ia membuangnya kearah depan atau sebelah kanannya, tetapi hendaklah ia lakukan kearah kirinya atau di bawah kaki (kirinya).'*	Advice Prohibition Information	Advice Prohibition Information
Partially correct prediction	*'Jika salah seorang dari kalian mengantuk saat salat, hendaklah tidur (dahulu) hingga ia mengetahui apa yang ia baca.'*	Information	Advice Information
Wrong prediction	*'Janganlah salah seorang dari kalian sengaja salat ketika matahari sedang terbit atau ketika saat terbenam..*	Advice Information	Prohibition

TABLE VII. SAMPLE OF ZERO PREDICTION

Data	Predicted	Target
'Luruskanlah shaf, sesungguhnya aku dapat melihat kalian dari balik punggungku.'	-	Advice Information

In the "wrong prediction" row in Table VI, which shows a sample of data that has been manually labeled before, many of the datasets used are still ambiguous when viewed in a meaningful way per word. For example, the data can be categorized into the Advice class as well. Hence, further validation of the dataset needs to be performed to achieve better performance.

A limitation of the binary relevance approach is the occurrence of zero predictions or data that cannot be categorized into any category. This happens because by following the binary relevance approach, each model is independent and so are the classes. Examples of these phenomena are summarized in Table VII.

In this research, 16 datasets could not be classified when adopting the binary relevance approach, while only 7 datasets from the overall 214 test dataset could not be classified when adopting the algorithm adaptation approach. This is attributed to unbalanced data. For example, in this study, there are only 98 data points for the Prohibition class with a total of 1066 data points, where the ratio of the Prohibition class and non-Prohibition class is 1:10, thereby making the classifier classify data as non-Prohibition class and so on for the other classes. By adopting the algorithm adaptation approach, a combination of classes connects the classes to reduce the possibility of zero predictions. Further, the use of feature selection may have an effect on the occurrence of zero predictions, as irrelevant words in documents are not selected, which causes a lack of features to sufficiently represent data. This is because the binary relevance approach entails that each model be independent and that no dependence exists among classes.

V. CONCLUSION

This research proposed a Chi-Square Bernoulli Model and a BNN model to classify Hadith into specific categories. The Bernoulli model was used as the feature selection method and was found to improve the classification performance, achieving the best average Hamming Loss result of 9.38%. This is because, in the Bernoulli distribution, one word is sufficient to represent the total number of occurrences of the word in a class, and therefore, the Bernoulli distribution can choose the relevant features as inputs for the BNN.

Furthermore, the binary relevance approach outperformed the algorithm adaptation approach. This is because when using algorithm adaptation, the classifier must remember most of the patterns in a class, whereas, in problem transformation (binary relevance), the classifier is only focused on remembering one pattern in a class.

For further research in this regard, more attention should be given to processing unbalanced data. Further, the future work should explore other methods such as the recurrent neural network, which works with data sequences or similar methods and determines their effectiveness in classifying Hadith data.

REFERENCES

[1] M. A. Saloot, N. Idris, R. Mahmud, S. Jaafar, D. Thorleuchter and A. Gani, "Hadith Data Mining and Classification: A Comparative Analysis," in Artif Intell Rev 46, pp. 113–128, 2016.

[2] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," Applied Computational Intelligence and Soft Computing, 2018.

[3] I. C. Irsan and M. L. Khodra, "Hierarchical Multilabel Classification for Indonesian News Articles," in 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016.

[4] F. Harrag, E. El-Qawasmah and A. M. S.Al-Salman, "Stemming as a Feature Reduction Technique for Arabic Text Categorization," in 10th Interna- tional Symposium on Programming and Systems, 2011.

[5] M. D. Purbolaksono, F. D. Reskyadita, Adiwijaya, A. A. Suryani and A. F. Huda, "Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature," International Journal on Advance Science, Engineering and Information Technology, vol 10, pp. 234-238, 2020.

[6] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning, 1997.

[7] C. D. Manning, P. Raghavan and H. Schtze, Introduction to Information Retrieval, Cambridge University Press, 2009.

[8] S. A. Faraby, E. R. R. Jasin, A. Kusumaningrum and Adiwijaya, "Classification of Hadith into Positive Suggestion, Negative Suggestion, and Information," Journal of Physics: Conference Serie, 2018.

[9] M. F. Afianto, Adiwijaya and S. A. Faraby, "Text Categorization on Hadith Sahih Al-Bukhari using Random Forest," Journal of Physics: Conference Series, 2018.

[10] M. Y. A. Bakar, Adiwijaya and S. A. Faraby, "Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language translation) using Information Gain and Backpropagation Neural Network," in International Conference on Asian Language Processing (IALP), 2018.

[11] G. Mediamer, Adiwijaya and S. A. Faraby, "Development of Rule-based Feature Extraction in Multi-label Text Classification," J. Adv. Sci. Eng. Inf. Technol, 9(4), 2019.

[12] M. N. Al-Kabi, G. Kanaan, R. Al-Shalabi, S. I. Al-Sinjilawi and R. S. Al-Mustafa, "Al-hadith text classifier," Journal of Applied Sciences, vol. 5, no. 3, pp. 584-587, 2005.

[13] H. Liu, X. Li and a. S. Zhang, "Learning Instance Correlation Functions for Multilabel Classification for Multilabel Learning," IEEE TRANSACTIONS ON CYBERNETICS, pp. 2168-2267, 2016.

[14] H. Soleimani and D. J. Miller, "Semisupervised, Multilabel, Multi-Instance Learning for Structured Data," Neural Computation, vol. 29, p. 150, 2017.

[15] J. Huang, G. Li, Q. Huang and X. Wu, "Joint Feature Selection and Classification for Multilabel Learning," IEEE TRANSACTIONS ON CYBERNETICS, pp. 2168-2267, 2017.

[16] F. S. Nurfikri, M. S. Mubarok and Adiwijaya, "News Topic Classification using Mutual Information and Bayesian Network," in 6th International Conference on Information and Communication Technology (ICoICT), 2018.

[17] H. Aydadenta and Adiwijaya, "A Clustering Approach for Feature Selection on the Microarray Data Classification using Random Forest," Journal of Computer Science, 2016.

[18] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Journal of Machine Learning Research 3, pp. 1289-1305, 2003.

[19] Y. Xu, G. Jones, J. Li, B. Wang and C. M. Sun, "A Study on Mutual Information-Based Feature Selection for Text Categorization," Journal of Computational Information Systems, vol. 3, no. 3, pp. 1007-1012, 2007.

[20] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," in Inst. Log. Lang. Comput. Univ. Van Amst. Neth., 2003.

[21] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi and H. E. Williams, "Stemming Indonesian: A Confix-Stripping Approach," in ACM Transactions on Asian Language Information Processing (TALIP), 2007.

[22] M.-L. Zhang, Y.-K. Li, X.-Y. Liu and X. Geng, "Binary Relevance for Multi-Label Learning: An Overview," Frontiers of Computer Science, 2018.

[23] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," IEEE Transactions on Knowledge and Data Engineering, 2014.

[24] F. Harrag and E. El-Qawasmah, "Neural Network for Arabic Text Classification," in Second International Conference on the Applications of Digital Information and Web Technologies, 2009.

[25] M. S. Sorower, "A Literature Survey on Algorithms for Multi-label Learning," in Oregon State University, Corvallis , 2010.