IJoICT
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 7, No. 2, Dec 2021. pp. 11-22
doi: 10.34818/ijoict.v7i2.573

Classification of Hadith Topic of Indonesian Translation Using K-Nearest Neighbor and Chi-Square

Ghinaa Zain Nabiilah[1], Said Al Faraby[2], Mahendra Dwifebri Purbolaksono[3]

[123]*School of Computing, Telkom University*
Bandung 40257, Indonesia

[1]ghinaazain@student.telkomuniversity.ac.id, [2]saidalfaraby@telkomuniversity.ac.id,
[3]mahendradp@telkomuniversity.ac.id

Abstract

Hadith is the main way of life for Muslims besides the Qur'an whose can be applied in everyday life. Hadith also contains all the words or deeds of the Prophet Muhammad which are used as a source of the law of Islam. Therefore, many readers, especially Muslims, are interested in studying hadith. However, the large number of hadiths makes it difficult for readers or those who are still unfamiliar with Islam to read them. Therefore, we conducted a study to classify hadith textually based on the type of teaching, so that readers can get an overview or other reference in reading and searching for hadith based on the type of teaching more easily. This study uses KNN and chi-square methods as feature selection. We also carried out several test scenarios, including implementing stopword removal modifications in preprocessing and experimenting with selecting k values for KNN to determine the best performance. The best performance was obtained by using the value of k = 7 on KNN without implementing chi-square and with stopword removal modification with a hammer loss value of 0.1042 or about 89.58% of the data correctly classified.

Keywords: chi-square, hadith, text classification, k-nearest neighbor.

I. Introduction

HADITH is something that comes from the Prophet Muhammad in the form of words, deeds, or approval and serves as an explanation and strengthening of the meaning of the content of the verses of the Qur'an so that its position in Islam is the source of basic law [1]. Various kinds of knowledge, and the teachings contained in the hadith are something interesting to learn because they can apply in everyday life. These teachings can be grouped textually into several types such as recommendations, prohibitions, and information. However, the large number of hadiths makes readers or those who are still unfamiliar with Islam find it difficult to read it. So, we need a system that can classify hadith based on the type of teaching so that readers get imagine or other references in reading and looking for hadiths that contain suggestions, prohibitions, and information easier.

Multi-label classification is a text categorization process, where each document may include several predetermined topics [2]. For example, in this research, multi-label classification to classify hadith according to the group of teachings such as recommendations, prohibitions, information, a combination of the two, or a combination of the three. There are various classification methods one of them is K-Nearest Neighbor (KNN). The KNN method has been widely used because of its implementation simplicity and excellent performance

Received on 15 July 2021. Revised on 28 Sep 2021. Accepted and Published on 24 Dec 2021

[3]. KNN has good accuracy and is effective in text classification. In addition, the KNN method is also suitable for multi-label cases because it produces a superior classification value than other classification methods [4]. The application of the KNN algorithm is a classification method that has also been carried out for multi-label data and produces good accuracy, which is almost 90% of the data is classified correctly as in the research conducted by [4,5], and the research conducted by [6] resulted in the value of accuracy above 90% the data is classified correctly.

However, the performance of KNN classification is highly dependent on the selection of k values, and KNN also has a weakness in handling data with high dimensions or data with many variables [5]. Therefore, additional methods are needed to improve the performance of KNN by reducing dimensions [7]. Feature selection is an effective method for reducing feature dimensions [8]. Feature selection can reduce the number of features that are less influential for the classification process. The commonly used method is chi-square, information gain, and mutual information. Chi-square gets the best accuracy for multi-class and multi-label [9]. This study combines the KNN classification and chi-square methods because both have good accuracy for multi-label data where the classification results in previous studies can be classified more than 80% of correctly classified data.

This study focuses on knowing how the performance of the KNN algorithm in classifying hadith with multi-label data, how the effect of choosing k on the KNN algorithm in the classification process, how the effect of using the chi-square selection feature on the results of KNN classification, and how the effect of using stopword removal on the results. classification. This study also aims to provide an overview of the performance of the KNN algorithm in classifying multi-label data, the effect of k on the KNN selection algorithm, the effect of using the chi-square selection feature on the classification results, and the effect of using stopword removal on the classification results.

The next discussion in this journal is a related study in part 2, containing an explanation of the theory or literature that supports and is related to the research carried out, part 3 methodology explain the design of the system that was built, part 4 contains an evaluation of the test results along with an analysis of the test results, and section 5 contains conclusions from test results and analysis of test results.

II. LITERATURE REVIEW

Research related to the classification of multi-label texts has previously been carried out using various methods, one of which is research conducted by [4] focusing on the process of the classification of the multi-label text of Indonesian news topics using KNN which produce the optimal hamming loss is 0.1116 or about 88.84% of data that is classified correctly. This study also explained that the selection of the value of k, and the type of distance proximity measurement on KNN can affect the results of the KNN classification performance.

Other text classification studies using KNN have also been carried out by [5] who classified the topics of the Qur'anic verses in English by producing the most optimal hamming loss is 0.1348 or about 86.52% of the data classified correctly. In this study, it is also explained that the selection of the k value greatly affects the accuracy results, because there is no best method to determine the accuracy of KNN based on the selection of the k value. In addition, research conducted by [6] using KNN and Latent Semantic Analysis (LSA) to classify the topic of Indonesian Translated Hadith resulted in an LSA-KNN performance value of 90.28%.

Research on text classification has also been carried out by [10] using the classification method Support Vector Machine (SVM) which focuses on comparing results using two feature selections, namely information gain, and chi-square. The best result in this study is to use the chi-square with the f1-score value is 95.29% and with information gain it produces an f1-score is 89.43%.

Meanwhile, research on text classification use data Hadith Shahih Bukhari in Indonesian translation has previously been carried out by [11] using the Backpropagation Neural Network (BPNN) method with stemming modifications. The optimal result in this study is not using stemming which produces a performance value is

84.63%, the results of research using modified stemming are better than using ordinary stemming but not better than not using stemming when doing preprocessing.

Text classification is a process that aims to determine one or more classes of each document in a text data set based on previous learning [12]. Multi-label classification is different from single labels. In single-label classification, a document is only classified into one category class only [13]. While in the multi-label classification, each document can be classified into several classes. In general, text classification has several stages, namely data collection, preprocessing, feature extraction, feature selection, classification, and performance evaluation.

Term Frequency and Inverse Document Frequency (TF-IDF) are methods used for the feature extraction process. TF-IDF is used to assess the importance of a word to a category or category in a file set. TF-IDF represents a matrix where each row matrix is data, and the existing column is a word or feature [14].

Feature selection is used to select the best features used in building the model and reduce the number of irrelevant features in the classification process. Chi-square is a feature of one of the selection methods commonly used for the text classification process. Chi-square uses statistical theory to test the independence of a term with its category [15].

The multi-label classification used in this study is KNN. KNN is a classification method that is included in supervised learning (a method that is supervised learning where the expected results are known beforehand). The KNN algorithm performs classification by predicting the test sample according to the training sample of the k closest neighbors of the test sample. The KNN algorithm is relatively simple and easy to implement [16].

In this study, the process of splitting data or dividing the dataset into a collection of training data is also carried out and test data sets. Training data is used for the model learning process on the machine, while test data is used to evaluate the results of the model learning process. The final process carried out in this study is to evaluate the results of the classification that has been done. Evaluation method one of them can be done by using a hamming loss. Hamming loss counts the amount errors in the classification process of the test data. The smaller hamming loss value, the less error in classification or means accuracy of the classification system is getting higher.

III. Research Method

This study builds a system that can classify multiple labels on the Hadith Shahih Bukhari of Indonesian Translation textually into several teaching topics, namely recommendations, prohibitions, and information using K-Nearest Neighbor and Chi-Square methods. The general description of the system is described in Figure 1.

A. Dataset

This study uses 7007 data of Hadith Sahih Bukhari in Indonesian translation sourced from research [17] which has been manually labeled according to the characteristics of each topic, then validation is carried out on the sample data by parties who are experts in their fields. This data is multi-label consisting of three classes, namely recommendations, prohibitions, and information. Each data has one or two or even all the three types of classes that exist. Examples of multi-label data in this study can be seen in Table I.

1) Recommend: This label shows the hadith that textually gives advice. An example of this label is “If any of you falls asleep while praying, let him sleep (first) until he knows what he is reading.”
2) Prohibition: This label shows a hadith that textually describes a form of prohibition against a thing or an action. An example of this label is “Leave what you cannot afford.”
3) Information: This label shows a hadith that textually describes the form of information that must be known related to the Islamic religion. An example of this label is “Allah will not be bored until you

are bored yourself, and the religion most loved by Him is the one that is always practiced regularly and continuously."

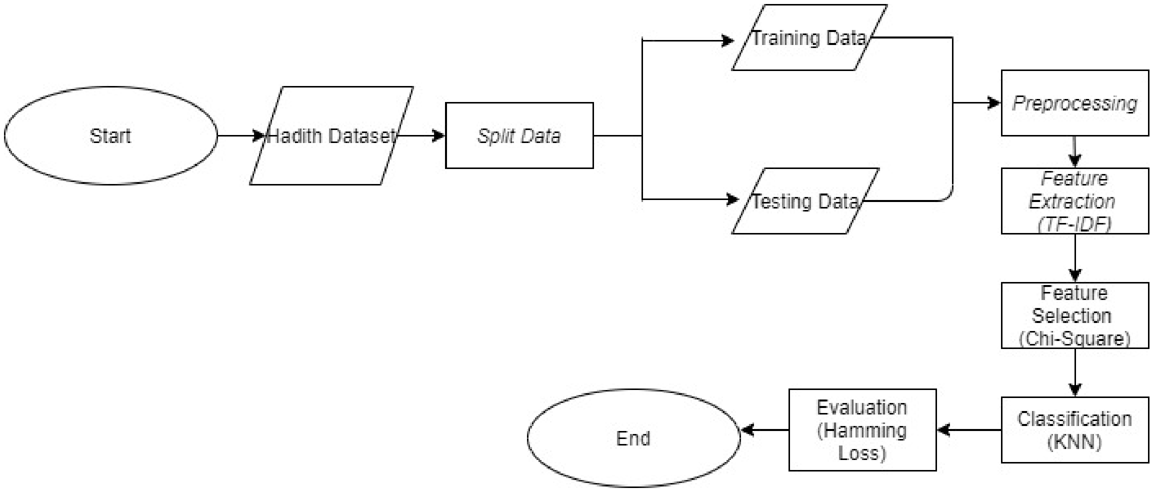

Fig. 1. Flowchart of the system built

The diagram of the hadith topic of Indonesian translation classification system can be seen in Fig. 1. There are 5 main processes carried out including: split data, preprocessing, feature extraction, feature selection, classification and evaluation.

TABLE I
MULTI-LABEL DATA REPRESENTATION

Text of Hadith	Recommend	Prohibition	Information
If any of you falls asleep while praying, let him sleep (first) until he knows what he is reading.	Yes	No	Yes
Leave what you cannot afford, for the sake of Allah, Allah will not be bored until you yourself are bored, and the religion most loved by Him is what is always done regularly and continuously.	Yes	Yes	Yes

B. Preprocessing

Preprocessing is the stage used to process raw data that has a lot of noise. In preprocessing, several stages are carried out to eliminate parts and words that are not needed in the classification process. The stages in preprocessing can be seen in Fig 2.

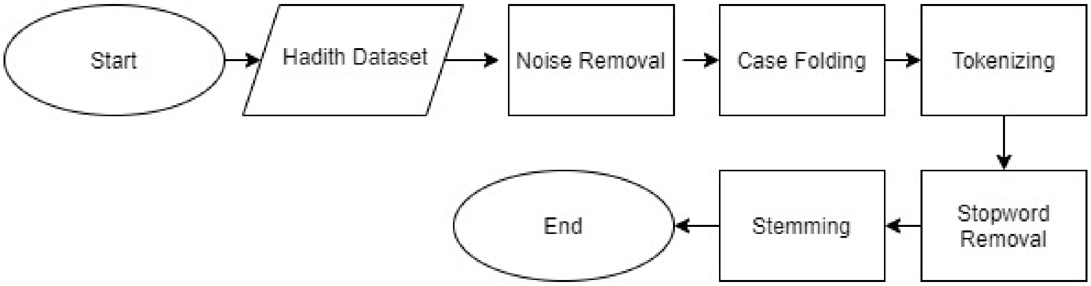

Fig. 2. Data preprocessing stages

The process carried out at the preprocessing stage is as follows:

1) Noise Removal. This process can remove spaces, punctuation marks, and characters or numbers.
2) Case Folding. This process can change letters to the same capitalization.
3) Tokenizing. This process can break sentences into words.
4) Stopword Removal. This process can remove words that are considered less influential and can annoying in the classification process. Like words that are not unique words that usually do not have meaning of specific words, such as conjunction or other adverbs.
5) Stemming. This process can remove affixes until every word in dataset contains only basic word.

C. Feature Extraction

After the data is cleaned in the preprocessing stage, Feature Extraction process is carried out using TF-IDF method. TF-IDF is a weighting matrix that measures word relationships. Words with high weights will be saved, and words with low weights will be removed. The process of giving weights is by calculating the value of TF-IDF which is a combination of two parts, namely Term Frequency (TF) and Inverse Document Frequency (IDF). TF is the number of occurrences of each word in the document and IDF is the number of related documents containing certain words [15]. TF-IDF calculation can be done using the formula as in equation (1).

$$W_{i,j} = tf_{i,j} \times idf = tf_{ij} \times log\left(\frac{N}{df_i}\right) \tag{1}$$

Information:
$W_{i,j}$ = weight of the word tj to the document d_i
$tf_{i,j}$ = number of occurrences of the word t_j in the document d_i
$tf_{i,j}$ = number of documents
df_i = the number of occurrences of the word in the document

D. Feature Selection

Next is an additional step, namely feature selection using chi-square. In this process, features that have a high influence will be used for the process classification, and remove features that are considered less influential in the classification process. Chi-square or also called kai square is used to determine the level of dependence on two events, namely if in this study the intended event is on the number of feature occurrences, and label, where the chi calculation will be performed between each feature x of label y. if y is independent of x then the feature will be discarded. The equation to perform the chi-square calculation can be seen in equation (2).

$$X^2 = \sum_{i=1}^{k} \frac{(fo - fe)^{2}}{fe} \tag{2}$$

Information:
X^2 = Chi Square Value
fo = Frequency of observations on the same label
fe = Expected frequency according to the label
k = Number of samples.

E. Classification

KNN is an algorithm that performs data classification based on training data and test data, which pays great attention to the distance of the k closest neighbors. The classification process is carried out based on the comparison of the k values of the nearest neighbors. The parameter k has a large influence on the prediction results. KNN also does measurement of data similarity, and measurement of distance. The steps taken by KNN algorithm are as follows:

1) Determine the number of k.
2) Calculating the similarity of words to documents in the training data.
3) The calculation results are sorted by the highest.
4) Choose the closest k (high level of similarity) to the word that is processed after that can be determined class of the word.

F. Evaluation

The last process carried out is evaluating the results of the classification that has been carried out to determine the level of accuracy of the classification results. The evaluation process is carried out using the hamming loss method which measures the amount of incorrect data from the classification results compared to the total amount of data. Hamming loss calculation can be done according to equation (3). The smaller or closer to 0 Hamming Loss results, the better the accuracy of the classification method built.

$$h = \left(\frac{1}{P}\right)\sum_{i=0}^{p}\frac{1}{Q}\ \left|h\ (x_i\ )\Delta Y_i\right| \tag{3}$$

Information:
p = Total number of data
Q = Number of classes
$|h\ (x_i\ )\Delta Y_i|$ = The number of errors in the classification that occur.

IV. EXPERIMENT RESULTS AND DISCUSSION

In this study, several tests were carried out such as the first test, namely at the preprocessing stage by looking at the effect of using stopword removal, the next test was at the feature selection stage, namely by looking at the effect of the large number of features used on the chi-square, the next test was at the classification stage, namely by looking at the effect of selecting k on the accuracy results.

A. Testing the Effect of Using Stopword Removal

The first test scenario was conducted to compare the effect of using modification on stopword removal, without modification of stopword removal, and without using stopword removal in the preprocessing stage. By using the TFIDF feature extraction process, the chi-square feature selection with the number of features is 1138 features and using the KNN classification process with the number of k = 7, then the hamming loss value is obtained in Table II.

The test results in first scenario in Table II produce an optimal hamming loss value is 0.11714 or about 88.29% of the data correctly classified, by using a stopword removal modification in preprocessing. This is because when using stopword removal modifications, the words that will be removed are adjusted according to the needs of the hadith data used.

TABLE II
FIRST SCENARIO RESULTS

Parameter	Hamming Loss Score
Using Modified Stopword Removal	**0.117142857**
Without Using Stopword Removal	0.121666666
Using Stopword Removal Without Modification	0.121190476

This means that the word that was omitted alreadyare words that are not useful according to the data used. In Table III is a further explanation regarding the difference between stopword removal and modification of stopword removal carried out along with words that are removed in the stopword removal process using modification and without modification.

TABLE III
DIFFERENCES STOPWORD REMOVAL WITH MODIFIED AND WITHOUT MODIFICATION

Parameter	Stopword Removal Without Modification	Stopword Removal with Modification
Explanation	The use of stopword removal without modification means using all the words available in the NLTK library, without making any changes to the existing word list.	The use of stopword removal with modifications made in this study is to delete some words contained in the NLTK library and add some words as new dictionaries in the NLTK library.
Removed words from list library NLTK stopword removal	-	no, most, self, more, besides, do not, often, do, no.
Word added from list library NLTK stopword removal	-	Whatever, whoever, he, said, anyone, anyone, really.

The effect of the test in the first scenario can be seen in Table IV and Table V which presents the results of the classification of the system that has been carried out by comparing the use of stopword removal, the use of modified stopword removal and without the use of stopword removal.

TABLE IV
DIFFERENCES STOPWORD REMOVAL WITH MODIFIED AND WITHOUT MODIFICATION

Hadith	Results Preprocessing Without Stopword Removal	Results Preprocessing Without Stopword Removal Modifications	Results Preprocessing with Stopword Removal Modifications
Don't buy and don't take your sadaqah back.	don't you buy and don't you take back your sadaqah	buy take your sadaqah	don't buy don't take your sadaqah.
Whoever leaves the 'Asr prayer, his deeds have indeed been erased.	whoever leaves the asr prayer his deeds have indeed been erased.	whoever leaves asr prayer really erased his deeds	leaves asr prayer really erases his deeds

Based on the results in Table IV and Table V, it can be seen that the classification results with the modification of stopword removal in the first hadith give correct predictions. This is because the previous modification was done by deleting the word "don't" from the stopword removal library list so that the word "don't"was used in the classification process. Meanwhile, word removal in the preprocessing process without using stopword removal, and without modification of stopword removal gives a different meaning so that the prediction results are also different. So it gives a different value of hamming loss and the optimal way is to use a modified stopword removal. However, for words that do not have ambiguous meanings, system can still correctly classify each parameter being tested because these words only contain noise that does not have ambiguous meanings.

TABLE V
THE EFFECT OF THE FIRST SCENARIO TEST ON THE CLASSIFICATION RESULTS

Hadith	Actual Classification			Classification Prediction								
				Without Stopword Removal			Without Stopword Removal Modifications			With Stopword Removal Modifications		
	A	L	I	A	L	I	A	L	I	A	L	I
Don't buy and don't take your sadaqah back.	0	1	0	0	0	1	0	0	1	0	1	0
Whoever leaves the 'Asr prayer, his deeds have indeed been erased.	0	0	1	0	0	1	0	0	1	0	0	1

Information:
A stands for recommendation
L stands for prohibition
I stands for Information

B. *Testing the Effect of Number Features on Chi-Square*

The next scenario is to test the effect of the number of selected features in feature selection process using chi-square on classification results. Using all data preprocessing processes such as noise removal, case folding, tokenizing, modification of stopword removal, and stemming, the feature extraction process with TFIDF, and using the value of k = 7 on KNN, the hamming loss value is obtained as shown in Table VI.

TABLE VI
SECOND SCENARIO RESULTS

Parameter	Hamming Loss Score
2277 features	0.120714285
1138 features	0.117142857
683 features	0.121428571
Not Using chi-square	**0.104285714**

Based on the evaluation results in the second test scenario, it can be seen that the number of features used affects the evaluation value of the hamming loss, but it is not certain that many or at least features will increase or decrease accuracy in the classification process, because in this test the value of hamming loss has an increase and decrease accuracy that does not according to many or least number of features. As in the number of features as many as 2277 and 1138, the largest value of hammering loss is to use 1138 features. But the optimal result in this study is not to use chi-square which produces a hammer loss value of 0.1042 or about 89.58% of the data classified correctly. This is because when using the chi-square there are features that do not have a specific meaning but are ambiguous enough for the classification process to not be selected and affect the classification process. In table 7 are some of the selected and unselected features on the chi-square.

TABLE VII
LIST OF SELECTED AND UNSELECTED FEATURES IN CHI-SQUARE PROCESS

Examples of Available Features	Selected Features on Chi-Square With 2277	Selected Features on Chi-Square With 1138	Selected Features on Chi-Square With 683
Aabiduuna, aajili, aajilihi, aamantu, addin, ayyub, zulaekha, anta, ignore, afterlife, expert, good, seek, wretched, close, listen, self, calm, vile, no, kind, deny, torment, besides, say, many times, say, don't, most, tell	Ignore, afterlife, wretched, close, listen, self, peaceful, vile, kind, deny, say, many, say, don't, tell	wretched, peaceful, vile, kind, deny, tell, don't	afterlife, expert, vile, tell, don't

TABLE VIII
EFFECT OF UNSELECTED FEATURES OF CHI-SQUARE ON CLASSIFICATION RESULTS

Unselected Features	**Hadith**	**Preprocessing Result**	**Data Actual**	**Data Prediction**			
				Without Chi-Square	**Chi-Square with 2277 Features**	**Chi-Square with 1138 Features**	**Chi-Square with 683 Features**
Most	The most sins among the major sins are associating partners with Allah, killing, disobeying one's parents, telling lies, or he said; false testimony.	the most sins are associating partners with Allah, kill disobeying one's parents, say lies, lie, lie witnesses,	011 (Types of prohibition and information)	011	001 (Types of Information Only)	001	001

TABLE IX
LIST OF SELECTED AND UNSELECTED FEATURES IN CHI-SQUARE PROCESS

Feature Not Selected	**Hadith**	**Preprocessing Result**	**Data Actual**	**Data Prediction**			
				Without Chi-Square	**Chi-Square with 2277 Features**	**Chi-Square with 1138 Features**	**Chi-Square with 683 Features**
Deny	There are three signs of hypocrisy; when he speaks a lie, when he makes a promise he breaks it, and when he is given a message he betrays.	signs of hypocrisy speaking lies, promises to break the mandate, betrayal	001 (Types of Information)	001	001	001	001

The chi-square process in this study many chooses words such as pieces of verses from the Qur'an or the names of the Prophet's companions, the Prophet's family, or stories that mention someone's name at that time. The word often appears in sentences but is not an important feature that is related, and if the feature is omitted it does not affect the classification results. However, there are some words that do not have a specific word meaning but if they are not selected it will affect the classification results as shown in Table VIII. The system that uses chi-square does not choose the words listed in Table VIII so it is unable to distinguish the hadith sentences tested as multi-label sentences that have two hadith topics. This affects the value of the hamming loss and causes without chi-square has an optimal result. Although the results are not optimal, the use of chi-square can eliminate features that are less influential without affecting the classification results for words that do not have ambiguous word meanings as shown in Table IX.

C. Testing the Effect Number of k in KNN

The next scenario is done by comparing the selection of different numbers k in KNN. By using all data preprocessing processes such as noise removal, case folding, tokenizing, modification of stopword removal, and stemming, the feature extraction process with TFIDF, and feature selection with chi-square which uses 1138 features, the results obtained by using the hamming loss according to the table 10.

TABLE X
THIRD SCENARIO RESULTS

Number of k in KNN	Hamming Loss Score
2	0.13547619
3	0.12547619
5	0.11928571
7	**0.11714285**
11	0.11738095
13	0.11761904
15	0.11809523

Based on the overall test results, it can be seen that the KNN algorithm has good performance for multi-label classification by using the combination of the previously described formulas in the third scenario, where the classification process error rate or the overall hamming loss value of each k amount is not more than 0.1354 or correctly classified data about 86.64%. The optimum performance in this scenario is 7 in the number of k with a hamming loss value is 0.1171 or about 88.29% of data correctly classified. The KNN algorithm will look for the best k value and if it finds the best k value then the Hamming loss value will slowly increase from the previous value along with the number of k which is also enlarged or it can be said that KNN performance slowly decreases if it finds the best k value and the selection of the value right k greatly affect the performance of KNN.

TABLE XI
THE EFFECT NUMBER OF K KNN FOR CLASSIFICATION RESULT

Hadith	**Classification Actual**			**Classification Prediction**																				
				***k* =2**			***k* =3**			***k* =5**			***k* =7**			***k* = 11**			***k* = 13**			***k* = 15**		
	A	**L**	**I**	**A**	**L**	**I**	**A**	**L**	**I**	**A**	**L**	**I**	**A**	**L**	**I**	**A**	**L**	**I**	**A**	**L**	**I**	**A**	**L**	**I**
Do not insult the dead, for they have faced what they did.	0	1	1	1	0	0	1	0	0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	1

Information:
A stands for recommendation
L stands for prohibition
I stands for Information

In Table IX it can be seen that the number of k = 2, and k = 3 made the wrong classification, while the other k besides k = 7 did the classification correctly but could not classify the hadith into two teaching topics. This is because a high number of k will reduce the effect of noise in the classification, but too many numbers of k also make the boundaries between labels less clear or blurred. The resulting optimal k value also depends on the data used for the classification process and the data cleaning process that was carried out previously.

V. Conclusion

KNN multi-label classification using Sahih Bukhari Hadith data Indonesian translation with the chi-square feature selection method which has been described in the test scenario section has been successfully built with performance for each number of k selected in the KNN algorithm having an overall hamming loss value of each number k is not more than 0.1171 or correctly classified data is around 88.29%, and the most optimal hamming loss value is to use the number of k in KNN as much as 7. The selection of the right k value in KNN greatly affects the classification results. This is because a high number of k will reduce the effect of noise in the classification, but too many numbers of k also make the boundaries between labels less clear or blurred so that an optimal number of k is needed to overcome this problem.

The use of the chi-square method provides less than optimal performance compared to not using the chi-square. This is because when using chi-square by limiting the number of selected features, there are features that are ambiguous in the data and affect the results of the classification is not selected. The feature is less influential and does not describe a class in the hadith but if it is not selected it will affect the classification results. So that in some hadith data that uses the chi-square system, it has not been able to classify hadith into two teaching topics. Although the results are less than optimal, the use of chi-square can eliminate features that are less influential without affecting the classification results on words that do not have ambiguous word meanings.

Meanwhile, preprocessing process using stopword removal results in less than optimal performance compared to using modified stopword removal. This is because when using stopword removal modifications, adjustments are made to the words to be removed according to the needs of the hadith data used. So the words that are omitted are words that are not useful according to the data used. Stopword removal can affect the classification results because it can remove noise in words that have specific word meanings in certain data and cause an inaccurate classification, so modifications are needed to stopword removal so that the noise removed in the data is in accordance with the data used.

The suggestions that can be applied for future research are conducting research and testing using methods that can overcome data imbalances because this study has not used methods to deal with these problems. If possible, the next research can add hadith data that has been labeled in the data group that is less balanced so that it can be used to reduce data imbalance problems.

References

[1] Abdul Majid Khon. Ulumul Hadis. Jakarta: Amzah, 2012.

[2] Ling Zhang Min, Hua Zhou Zhi, “A k-Nearest Neighbor Based Algorithm for Multi-label Classification” in IEEE International Conference on Granular Computing. 2005. pp. 718-721.

[3] Shichao Zhang, et al., “Learning *k* for KNN Classification” in CM Transactions on Intelligent Systems and Technology. 2017.

[4] Nikmah Isnaini, Adiwijaya, Mohamad Syahrul Mubarok, Muhammad Yuslan Abu Bakar, “A multi-label classification on topic of Indonesian news using K-Nearest Neighbor” in International Conference on Data and Information Science. 2019.

[5] G I Ulumudin, Adiwijaya, M S Mubarok, “A multilabel classification on topics of qur’anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF” in International Conference on Data and Information Science. 2019.

[6] Dian Chusnul Hidayati, Said Al Faraby, Adiwijaya, "Classification of Multi Label Topics in Sahih Bukhari Hadith Using K-Nearest Neighbor and Latent Semantic Analysis", in JURIKOM (Journal of Computer Research) Vol. 7 No. 1, 2020.

[7] Adiwijaya, et al., "A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system" in International Conference on Information and Communication Technology (ICoICT). 2017.

[8] Yan Hong Li, "Text feature selection algorithm based on Chi-square rank correlation factorization" in Journal of Interdisciplinary Mathematics Vol. 20 No. 1 pp. 153–160. 2017.

[9] Syair Audi Sacra, Said Al Faraby, Danang Triantoro M, "Classification of Recommendations, Prohibitions, and Information on Sahih Bukhari Hadith Using Naïve Bayes Classifier", in e-Proceeding of Engineering: Vol.4 No.3 PP. 4794 – 4802, 2017.

[10] Yujia Zhai, et al., "A Chi-square Statistics Based Feature Selection Method in Text Classification" in 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). 2018.

[11] Mahendra Dwifebri Purbolaksono, et al., "Indonesian text classification using backpropagation and sastrawi steming analysis with Information Gain for selection feature" in International Journal on Advanced Science Engineering and Information Technology Vol. 2 pp. 60-65, 2020.

[12] Khitam Jbara, "Knowledge Discovery in Al-Hadith Using Text Classification Algorithm" Vol. 6. No. 11 pp. 409–419, 2010.

[13] Masoumeh Zareapoor, K. R. Seeja, "Feature extraction or feature selection for text classification: A case study on phishing email detection" in International Journal of Information Engineering and Electronic Business, 2015.

[14] Said Al Faraby, et al., "Classification of hadith into positive suggestion, negative suggestion, and information" in International Conference on Data and Information Science, 2018.

[15] Juen Ling, I Putu Eka N Kencana, Bagus Tjokorda Oka., "Analysis Sentiment Using Naïve Bayes Classifier With Feature Selection Chi Square" in E-Journal Mathematic Vol. 3 pp. 92-99, 2014.

[16] Huijuan Li, et al., "An Improved KNN Algorithm for Text Classification" in International Conference on Instrumentation and Measurement Computer Communication and Control, 2018.

[17] Muhammad Yuslan Abu Bakar, Adiwijaya, Said Al 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.