A multilabel classification on topics of qur'anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF

G I Ulumudin¹, A Adiwijaya², M S Mubarok³
School of Computing, Telkom University, Bandung, Indonesia

E-mail: ¹ganeshaihya@student.telkomuniversity.ac.id, ²adiwijaya@telkomuniversity.ac.id, ³syahrulmubarok@telkomuniversity.ac.id

Abstract. There are so many information contained in the Qur'an, it will be difficult to bring up the information manually, moreover if someone wants to know more about the Qur'an. Therefore, there is a need to find information with a certain topic that already classified in the Qur'an, especially in one verse of the Qur'an may have more than one topic (multilabel). This research examined how to build classifier to classify multilabel data which is topics of Qur'anic verses with k-Nearest Neighbor method. In this research, there is a comparison between feature extraction, Weighted TF-IDF and TF-IDF. The result of that comparison is that Weighted TF-IDF has better performance compared to normal TF-IDF. The highest result by finding the most optimal k score is k=25 with the average score of hamming loss = 0.134875. There will be a test to measure the effect of stopword removal and lemmatization with optimal k value, for a case without stopword removal, the result is 0.136375, whereas without the lemmatization the result is 0.13537. For not using stopword removal and lemmatization the average hamming loss is 0.1373125.

1. Introduction
Al-Qur'an is a holy book for every muslim in the world that consists knowledge, law, story, history, warning, and guidance to live the life. Al-Qur'an consists of 6236 verses, 114 chapters, and 30 juz [1]. As Al-Qur'an given to Muhammad as prophet and God's messenger, it has three main purposes, guidance about beliefs that should be followed by human, guidance about deeds that has to be done by muslim, and guidance about Islamic law [14]. It can be concluded that, muslim should learn Al-Qur'an to do the deeds from every guidance that shows in Al-Qur'an. There are so many information from Al-Qur'an that we can obtain, but it's hard to find that information manually, as there are so many muslims that want to learn Al-Qur'an. Because of that, there should be a way to find information about specific topic that already classified in Al-Qur'an. That information will be found from verses in Qur'an, as one verses can be consisted one to many topics. In machine learning, that term is called multilabel. As an example, Ash-Shura [3] chapter 42, 13th verse that says “He has ordained for you of religion what He enjoined upon Noah and that which We have revealed to you, [O Muhammad], and what We enjoined upon Abraham and Moses and Jesus - to establish the religion and not be divided therein. Difficult for those who associate others with Allah is that to which you invite them. Allah chooses for Himself whom He wills and guides to Himself whoever turns back [to Him]”. That verse
has three topics such as, Arkanul Islam, Al-Qur’an, humans and its community relation [1]. K-NN as one of the lazy learning methods, works by finding the nearest neighbor from new object then classify it class by choosing class from the nearest neighbor as we set how many neighbor that can determine the class in this method called k. The performance of KNN will be work better for data with small dimension, because it will generate a better accuracy than other methods [6,7]. Regardless of the better performance, KNN has weakness such as, high computation time if it handles data with high dimensions or many attributes [8]. KNN usually handles a single label classification, but in this research, KNN will handle a multilable classification case. In this research, it will build a KNN model to handle multi label classification on topics of Qur’anic verses.

2. Multi label classification on text
Classification is a process to find a model that can classify, so that model can predict a class for a datum with no class [7]. This research will solve a multilabel classification on text, where the data are Qur’anic verses in english. The verses can have one to many topics as their class that’s why it’s called multi label classification. To handle the problem, classifier will be built by machine learning. With machine learning, classifier can be built with learning process automatically, so it can predict a class for input data that the class has already defined. There will be two set of data, train data and test data. Train data will be used to build the model, while test data will used to evaluate the prediction. There are many methods that can be used to build a classifier, one of them is k-Nearest Neighbor. There are many research about classification on topic quranic verses that already done. There is a research about that in arabic verses, in that research the researcher compares many methods such as, Decision Tree, K-Nearest Neighbor, and Support Machine Vector [4]. Pane [15] in his research did a multilabel classification on Qur’anic verses in english with Multinomial Naive Bayes. Multinomial Naive Bayes is one of the Bayes model that using multinomial distribution. Pane using preprocessing such as case folding, tokenization, and stemming, also hamming loss as evaluation method. The best performance of his classifier with average score 0.1247. Izzaty [16] also has a research about multi label classification on the same data as Pane but with different method. Izzaty using Tree Augmented Naive Bayes (TAN) which is the upgrade version of Naive Bayes. This method increases the interaction in every variable, that can increase the accuracy of prediction compared to Naive Bayes. In this research, Izzaty also using a feature selection, that is Mutual Information to select some independent variable, so it can increase the system performance.

3. K-Nearest Neighbor (KNN)
KNN is a simple classification algorithm where it will count the shortest distance as a similarity measure to classify an object [5]. KNN is a supervised learning algorithm which has learning process with data with defined class and the result of the learning process will be used to predict new data with no class [10]. This method works by finding an amount of k data object that is closest to the input object, then choose a class with the most object between k [2]
In figure 1, the case is binary classification that classify object into circle or square. The new object is triangle and the value $k$ is 7. KNN classify triangle as circle because there are more circles than square in range $k=7$. One of the weaknesses of this model is its vulnerability to noise, so it need a preprocessing to reduce the noise [5]. To classify an object by using similarity measure, this research using euclidean distance in equation (1).

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$$  \hspace{1cm} (1)

Where $x_1$ is sample data, $x_2$ is test data, $i$ is a variable index data, $d$ is distance, and $p$ is the dimension of data.

4. Weighted TF-IDF
In this research, there is a Weighted TF-IDF as feature extraction that will reduce noise on data. Weighted TF-IDF will combine Term Frequency(TF) and Inverse Document Frequency(IDF). TF will count how often a word showed on a document in equation (2), while IDF will count the documents that consists a term in equation (3). This research uses Weighted TF to avoid domination of documents with little amount of term in query but has high frequency [11].

$$wtf_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d}, & \text{if } tf_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$ \hspace{1cm} (2)

$$idf_t = \log_{10}(N/df_t)$$ \hspace{1cm} (3)

Where $tf_{t,d}$ is term frequency in one document, $N$ is total documents, $df_t$ is term frequency in all document.

5. Hamming loss
In this research, the evaluation method will be using hamming loss, because it’s relevant to the multilabel classification unlike precision, recall, and f1-measure. This method will evaluate classifier based on the false prediction from all the prediction class in equation (4).

$$Hamming Loss = \frac{1}{NL} \sum_{i=1}^{N} \sum_{j=1}^{L} [\hat{y}^{(i)}_j \neq y^{(i)}_j]$$ \hspace{1cm} (4)

Where, $N$ is total documents, $L$ is total class, $\hat{y}^{(i)}_j$ is prediction class, $y^{(i)}_j$ is actual class, and $[\hat{y}^{(i)}_j \neq y^{(i)}_j]$ is false prediction from all prediction class in this case is difference between prediction class and actual class.

6. System
6.1. Data
Data that will be used is an english translation of Quranic verses. Data consisted 6236 verses that will be split into train data and test data. There are 16 class as many as the classification topics from “Syamil Al-Qur’an – Miracle the Reference”. The 16th class consists verses that have no classification topics.
6.2. System

The design of system can be described as follows.

- Data will be distributed circularly as train data and test data with k-fold Cross Validation with 80% train data and 20% test data composition. \( k \) for \( k \)-fold cross validation is set by 5, that means there are 5 folds and data will be circulated 5 times.
- Data will be processed in the preprocessing. In this research, there are 4 phases of preprocessing, such as case folding, tokenization, stopword removal, and lemmatization [9,12,13].
- After data processed in preprocessing, there will be feature extraction with Weighted TF-IDF. First, it will count Weighted TF in equation (2), then it counts IDF in equation (3) later.
- Later, classifier will classify the data with KNN in equation (1).
- Then, the classifier will be evaluated by using hamming loss in equation (4), every fold has one result, after all fold is executed, system will count the average of all hamming loss score.

7. Evaluation

7.1. Experiment result

In first experiment, there is an observation to find the optimal \( k \) to solve the multi label classification on Qur’anic verses in english. The value of \( k \) that has been observed is start from \( k=3 \) to \( k=39 \) and \( k \) is an odd number. Based on the first experiment, \( k=25 \) is the optimal \( k \) with average hamming loss score 0.1348, and the lowest performance with \( k=3 \) with average hamming loss score 0.1528.

| Nearest Neighbor | Average HL Score |
|------------------|------------------|
| 3                | 0.1528           |
| 5                | 0.1468           |
| 7                | 0.1458           |
| 9                | 0.1428           |
| 11               | 0.14             |
| 13               | 0.1382           |
| 15               | 0.1357           |
| 17               | 0.1356           |
| 19               | 0.1356           |
| 21               | 0.1358           |
In second experiment, system will be tested by comparing the use of feature extraction in this case the use of Weighted TF-IDF and TF-IDF using the optimal $k$. The result by using TF-IDF is 0.1364, which is worse than using Weighted TF-IDF. The other experiment examined the influence of stopword removal and lemmatization in preprocessing. System without lemmatization has 0.1353 as the average hamming loss score. System without stopword removal has 0.1363 as the average hamming loss score. System without stopword removal and lemmatization has 0.1373 as the average hamming loss score.

### Table 2. Second experiment.

| Case | Nearest Neighbor | Average HL Score |
|------|-----------------|------------------|
| optimal $k$ | 25 | 0.134875 |
| TF-IDF | 25 | 0.136371 |
| without Lemmatization | 25 | 0.1358 |
| without Stopword Removal | 25 | 0.1356 |
| without Lemmatization and without Stopword Removal | 25 | 0.13843 |

#### 7.2. Analysis result

On table 1, the best performance is while $k=25$ with average hamming loss score 0.1348. There is no method to define the best $k$ for KNN. To find the best $k$, it has to be set manually and observed carefully. The greater the $k$, the performance will be better at first, but when the system reach the optimal performance, the average hamming loss score will decrease gradually. In the second experiment, the influence of using Weighted TF-IDF as feature extraction has better performance than using TF-IDF and the influence of using stopword removal and lemmatization affect the accuracy of classifier, because two of the preprocessing will reduce noise so it can increase the system performance and accelerate the classification process.

#### 8. Conclusion

For the conclusions, the system that built can classified multilabel data on topics Qur’anic verses in English. The best performance of the system is 0.1348 as average hamming loss score. From the comparison between Weighted TF-IDF and TF-IDF, system has better performance when using Weighted TF-IDF as feature extraction. The system performance when reducing stopword removal and lemmatization can affect the accuracy because the noise which is the weakness of KNN. Because the
Qur’anic verses has many sub class as their label, the next research can classify verses into sub label as many as it is.

References
[1] Syaamil Qur’an 2004 Cordova Al-Qur’an and Translation
[2] Suyanto 2017 Data Mining for Data Classification and Data Clustering (Bandung: Informatika)
[3] Shakir M H 2009 Al-Qur’an English January 12th 2018
[4] Al-Kabi M N, Ata B M A, Wahsheh H A and Alsmandi I M 2013 A topical classification of Quranic Arabic text Proc. of the Taibah University International Conf. on Advances in Information Technology for the Holy Quran and its Sciences Dec 2013 pp 22-25
[5] Fayyad U, Piatetsky-Shapiro G and Smyth P 1996 From data mining to knowledge discovery in databases AI magazine pp 37-54
[6] Kumar S V K and Kiruthik P 2015 An overview of classification algorithm in data mining 2015 Int. J. of Advanced Research in Computer and Communication Engineering 4 no 12 pp 255-257
[7] Setiawan B S, Mubarok M S and Adiwijaya A 2018 A topical classification of news in Bahasa using weighted K-Nearest neighbour 2018 eProceedings of Engineering 5 no 1
[8] Aulia M N, Mubarok M S, Novia W U and Nhita F 2017 A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system 2017 Information and Communication Technology (ICoIC7)
[9] Henny L 2013 An implementation of K-Nearest neighbor for determining vehicle ownership credit 2013 PIKSEL (Penelitian Ilmu Komputer Sistem Embedded dan Logic)
[10] Salton G and Buckley C 1988 Term-Weighting Approaches in Automatic Text Retrieval (Cornell University: Departement of Computer Science)
[11] Wu Xi-Zhu and Zhi-Hua Zhoi 2016 A Unified View of Multi-Label Performance Measures
[12] Darlmani A A and Pribadi A W 2009 Automatic news articles classification in indonesian language by using naive bayes classifier method 2009 Proc. of the 11th Int. Conf. on Information Integration and Web-based Applications & Services ACM
[13] Mansur M 2006 Analysis of N-Gram based text categorization for Bangla in a newspaper corpus 2006 Bangladesh: Center of Research on Bangla Language Processing BRAC University Dhaka
[14] Shihab M Q 1992 Membumikan Al-Quran (Bandung: Mizan)
[15] Pane R A, Mubarok M S and Adiwijaya A 2018 A multi label classification on topics of Qur’anic verses in English using multinomial Naive Bayes 2018 eProceedings of Engineering 5 no 1
[16] Izzaty A M K, Mubarok M S and Adiwijaya A 2018 A multi label classification on topics of Qur’anic Verses in English using Tree Augmented Naive Bayes eProceedings of Engineering 5 no 1
[17] Refaeelizadeh P, Tang L and Liu H 2009 Cross Validation in Encyclopedia of database systems Springer P532-538
[18] Zhang M L and Zhou Z H 2014 A review on multi-label learning algorithms IEEE Transactions on knowledge and data engineering 26 no 8 pp1819-1837