Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL
(System Engineering and Information Technology)
Vol. 5 No. 3 (2021) 594 – 601 ISSN Elektronic Media: 2580-0760

Classifying Quranic Verse Topics using Word Centrality Measure

Ferdian Yulianto[1], Kemas M. Lhaksmana[2], Danang Triantoro Murdiansyah[3]
[1,2,3]Informatics, School of Computing, Telkom University
[1]ferdiany@student.telkomunivesity.ac.id, [2]kemasmuslim@telkomuniversity.ac.id, [3]danangtri@telkomuniversity.ac.id

Abstract

Muslims believe that, as the speech of Allah, The Quran is a miracle that has specialties in itself. Some of the specialties that have studied are the regularities in the number of letters, words, vocabularies, etc. In the past, the early Islamic scholars identify these regularities manually, i.e. by counting the occurrence of each vocabulary by hand. This research tackles this problem by utilizing centrality in quranic verse topic classification. The goal of this research is to analyze the effect of The Quran word centrality measure on the topic classification of The Quran verses. To achieve this objective, the method of this research is constructing the Quran word graph, then the score of centralities included as one of the features in the verse topic classification. The effect of centrality is observed along with support vector machine (SVM) and naïve Bayes classifiers by performing two scenarios (with stopword and without stopword removal). The result shows that according to the centrality measure the word "الله" (Allah) is the most central in The Quran. The performance evaluation of the classification models shows that the use of centrality improves the hamming loss score from 0.43 to 0.21 on naïve Bayes classifier with stopword removal. Finally, both of classification method has a better performance in word graph that use stopword removal.

Keywords: *The Holy Quran, centrality, topic classification, SVM, naïve Bayes, multilabel classification.*

1. Introduction

The Quran is the holy book of Muslims that is used as a guide for life. It consists of 30 parts (juz), 114 chapters, and 6236 verses. As the speech of The God which was revealed to the Prophet Muhammad (praised be upon him) through the archangel Gabriel, muslims believe that The Quran is a miracle which has many specialties. Its miracle and specialties have been studied by Islamic scholars from many aspects, especially its content as the primary source of Islam. Other Islamic scholars also studied other specialties of The Quran from linguistic aspects, such as the regularities on the number of letters, words, vocabularies, etc.

Muslims believe that The Quran has regularities in terms of the number of vocabularies and relationships between words in verses which make these verses have a different topic. In graph theory, centrality measurement identifies how important nodes are close to each other. Centrality measurement can be used to distinguish important nodes in the graph [1].

In recent years, there are various research about centrality measurement. First, on a research about graph centrality-based spam SMS detection, centrality measurement is used for SMS detection that contains spam or not using several machine learning methods such as random forest, support vector machine (SVM), and naïve Bayes. The centrality type that used on the research is degree centrality and closeness centrality. The result in the study showed that degree centrality gave the best precision and recall results with 81% and 76%, respectively [1]. In another research about implementing term weighting for text categorization, centrality is used to categorize a single label text that was implemented using term weighting (TW), which represents the term value of each centrality type. In the research, the graph is implemented as a directed graph. The graph model represents the co-occurrence of terms or words in each document or paragraph. The centrality type that used in the research is degree centrality, in-degree, out-degree, and closeness centrality with the best accuracy is 96.61% on degree centrality [2].

Accepted by editor: 31-05-2021 | Final Revision: 25-06-2021 | Online Publication: 30-06-2021

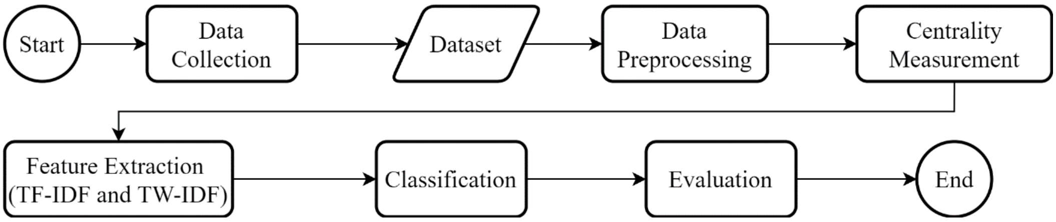

Figure 1. The Classification Process in This Research

In the study about the effectivity of graph-based term weighting (TW), a comparison about term feature such as TF-IDF, TW, TW-IDF and TW-SRW (Supervised Relevance Weighting) is conducted on different datasets. The SVM method achieves the best recall score at 98.02% [3]. Next, study about keyword extraction from Arabic documents that used four type centralities (degree, betweenness, closeness, and eigenvector centrality). In the research also use several machine learning methods for identify keywords including multilayer perception, naïve Bayes, random forest, and oneR. The result shows the performance of all machine learning method above are good results in precision of 95.5% until 96.4% [4].

Furthermore, the studies about classification on Quranic verse topics also have been done in recent years. First, the study in 2019 that was built a multilabel classification system from Quranic verses in English translation using K-Nearest Neighbor (KNN) and Weighted TF-IDF. The result shows the best average hamming loss score is 0.1349 with the k-score is 25 [5]. Next, the study in 2019 was using Quranic verse in English translation dataset and compare any feature selection with any machine learning methods such as naïve Bayes, SVM, and ANN. The result shows that the best hamming loss score is 0.0938 when naïve Bayes method is used, and the type of feature selection is called mutual information [6]. Another research about topic classification in Al-Baqarah, one of the chapters of The Quran, with three topics, named *iman* (faith), *ibadah* (rituals), and *akhlaq* (morals). Several classification methods are used in its research such as naïve Bayes, SVM, J48, and KNN classifier. The result shows SVM and J48 gives a higher accuracy of 85% [7].

Based on previous related studies, this research builds a Graph of Word (GoW) model which represents words in all documents (all verses) of The Quran. GoW model built into two models. First, the GoW model without stopword removal and second GoW model with stopword removal. In the Graph of Word model, some nodes were connected with edges. Nodes represent a word of verses and edges represent co-occurrence relationship between a word of verses and a relationship between a word from the last verse with the first word in the next verses. Centrality measurements that will be used in this research are degree centrality, betweenness centrality, and closeness centrality.

The classification methods used in this research are Naïve Bayes and SVM. Naïve Bayes is a probabilistic method based on Bayes theorem with strong independence assumption between each feature and has been successfully applied in other studies [8][9][10]. Meanwhile, Support Vector Machine (SVM) is one of the most widely used classification methods and has been successfully applied in many applications [11].

The purpose of this research is to find the central or important word from The Quran. All the centralities type combined with TW-IDF are used for classification. So, we can know the effect of centrality words on the classification system. In addition, this research also compares the performance between each feature extraction, TF-IDF and TW-IDF, with two types of GoW models. Also, this research compares two machine learning methods, Naïve Bayes and SVM.

2. Research Method

2.1. System Overview

In this research, centrality measurement on the multilabel classification of The Quran verses topic is carried out with several main processes, those are data collection, data preprocessing, centrality measurement, feature extraction, data classification, and evaluation of performance of classification system. The design scheme of the system can be look in Figure 1.

2.2. Data Collection

The dataset used in this study is The Quran data with the original The Quran language (Arabic language). The Quran data was taken from an international project called Tanzil.net which provides the verified text of The Quran in Unicode [12]. In this study, we will used eight topics such as Aqidah (faith), Akhlaq (morals), Syariah (laws), Ilmu (knowledge), Kisah (stories), Alam Dunia (universe), Alam Akhirat (hereafter), dan Alam Ghaib (the unseen). These topics are based on the topics published by the Hadith Study Center [13]. The labeling process is done during the data scraping from the website.

The number of data used to build the Graph of Word (GoW) model are 6236 verses, which represent all of the verses in The Quran. This is because, in constructing the GoW, we want to know which words are the most central in The Quran. However, the number of verses

DOI: https://doi.org/10.29207/resti.v5i3.3171

Table 2. Verse Samples from The Quran Dataset

Verse	Translation	Verse Number	Topics
بسم الله الرحمن الرحيم	In the name of Allah, most benevolent, ever-merciful.	1 | 1	Aqidah (faith)
الحمد لله رب العالمين	All Praise Be to Allah, Lord of all the worlds.	1 | 2	Aqidah (faith), Syariah (laws)
الرحمن الرحيم	Most beneficent, ever-merciful.	1 | 3	Aqidah (faith)
مالك يوم الدين	King of the Day of Judgement.	1 | 4	Aqidah (faith)
إياك نعبد وإياك نستعين	You alone we worship, and to You alone turn for help.	1 | 5	Aqidah (faith), Akhlak (morals)
اهدنا الصراط المستقيم	Guide us (O Lord) to the path that is straight	1 | 6	Aqidah (faith), Syariah (laws)
صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين	The path of those You have blessed, Not of those who have earned Your anger, nor those who have gone astray.	1 | 7	Aqidah (faith), Syariah (laws)

Table 3. An Example of Tokenization

Before Tokenization	After Tokenization
قل هو الله أحد	‘قل’, ‘هو’, ‘الله’, ‘أحد’

which are included in the classification process is 4747 verses, since based on our data source, these are the verses that have been labeled with the topics, whereas the remaining verses (1489) have not been labeled. Some verses from the dataset are listed in Table 1.

2.3. Data Preprocessing

Before we build the Graph of Word model and the whole classification system, the data will be preprocessed first. In this research, the preprocessing are tokenization and stopword removal. Tokenization is the initial stage of text processing which breaks one sentence into several tokens, which in this case are words. The purpose of tokenization is to create a Bag of Word model that will be used for constructing the Graph of Word models and classification process. An example of the tokenization process is illustrated in Table 2.

Following the tokenization process, the next step is stopword removal. Stopword removal is the process of removing words that oftenly appear and consider meaningless. Because this study is using the Arabic language, therefore the stopwords are also Arabic. The words ‘أنت’, ‘هو’, ‘أنتما’ that have meant ‘you’, ‘he’, ‘we’ are also removed from the dataset. This removal is performed to study the effect of stopword removal, as well as to improve the performance of the whole classification process. This process is common in text processing. The removal is not intended to change the verse itself, which is totally forbidden in Islam. An example of the stopword removal process is given in Table 3.

Table 1. An Example of Stopword Removal

Before Stopword Removal	After Stopword Removal
‘قل’, ‘هو’, ‘الله’, ‘أحد’	‘قل’, ‘الله’, ‘أحد’
Translation	
Say: He is Allah, the One and Only.	Say: Allah, the One and Only.

2.4. Centrality Measurement

Before we measure centrality, we must build a Graph model first. The graph is a collection of vertices that are interconnected with each other. If the vertices or nodes are symbolized by V and edges are symbolize by E, then the graph can be defined as G = (V, E). In this research we will build a Graph of Word model with the type of graph is an undirected graph. Graph of Word model is represented all of verses in The Quran with nodes represent the word of verse and edges represent co-occurrence relationship between a word of verses and relationship between word from last verse with first word in the next verses. If in one verse or between verses there are the same pair of nodes, it represents edge weight.

The example Graph of word model can be look in Figure 2 that build from Annas chapters with six verses. The edge weight value from this model represents the number of pair words that appear in one verse or between verses. For the example in node ‘الناس’ (human) and ‘ملك’ (king) have edge weight value is 2. Because the relationship between node ‘الناس’ (human) and ‘ملك’ (king) appears two times in annas chapter.

Graph of Word model will be built into two models. The first model is without stopword removal process and the second model is with stopword removal process. For scheme of the centrality measurement process can be look in Figure 3.

After we built the GoW model, the next step is centrality measurement process. In graph theory, centrality is an important concept to determine the important nodes in a graph model. In this research we will use three popular types of centrality in network science such as degree centrality, betweenness centrality, and closeness centrality [14].

DOI: https://doi.org/10.29207/resti.v5i3.3171

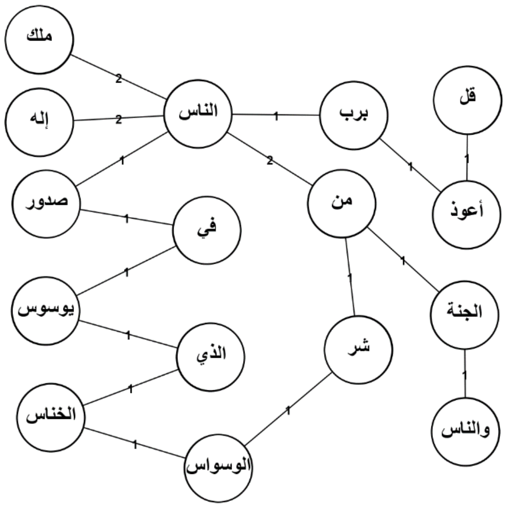

Figure 2. A Portion of The Graph of Word of The Quran

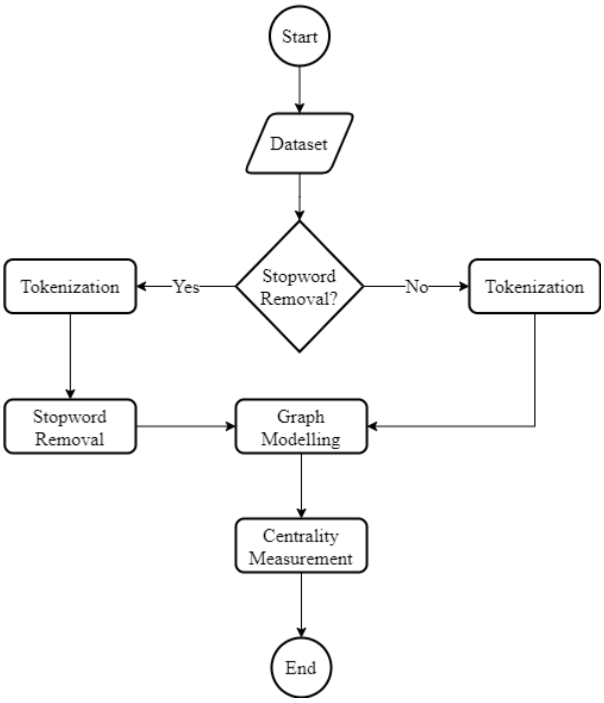

Figure 3. The Centrality Measurement Process

2.4.1. Degree Centrality

Degree centrality is calculated based on the degree of the nodes which represent the number of edges connected to the nodes. In GoW, the degree centrality of the node V_i represents the number of words that co-occur with words that have a relationship with node V_i [15]. In our model, the node V_i is considered to have a relationship with another node V_j, if the word that is represented by V_j appears before or after the word represented by V_i in a verse. Therefore, the degree centrality can be computed with,

$$C_{Deg}(V_i) = \frac{|N(V_i)|}{|V|-1} \qquad (1)$$

where $C_{Deg}(V_i)$ is the degree centrality value of node V_i and $N(V_i)$ is the set of nodes connected to V_i and V is the total of nodes number in the graph.

2.4.2 Betweenness Centrality

Betweenness centrality is defined as how many times a node acts as a bridge or connector in a graph network [16]. We can compute the betweenness centrality with formula 2 [15],

$$C_{Between}(V_i) = \frac{\sum_{V_i \neq V_j \neq V_k \in V} \frac{\sigma(V_j,V_k|V_i)}{\sigma(V_j,V_k)}}{(|V|-1)(|V|-2)\,/\,2} \qquad (2)$$

where $\sigma(V_j, V_k|V_i)$ is the number of paths that pass node V_i and $\sigma(V_j, V_k)$ is the number of shortest paths from node V_j to node V_k.

2.4.3. Closeness Centrality

Closeness centrality is defined as the number of shortest paths between one node to all the other nodes. We can compute closeness centrality with the following formula [15],

$$C_{Close}(V_i) = \frac{|V|-1}{\sum_{V_j \in V} distance(V_i,V_j)} \qquad (3)$$

where $distance(V_i, V_j)$ is the shortest paths between node V_i and node V_j.

2.5. Feature Extraction

In text classification, a text needs to be extracted into numeric form because, a computer just processes a numeric number. Therefore, it is necessary to extract text into numeric so that it can be processed by a computer and extract it with the best pattern that may have a huge impact on the classification technique capability [17]. In this research, feature extraction will be giving weight to words in the Bag of Word model. One of the popular feature extraction techniques is TF-IDF which is the combination between term frequency (TF) and inverse document frequency (IDF). TF is the number of terms, i.e. generally words, that appear in a document. IDF is the number of documents that contain the certain terms. We can compute TF-IDF with formula 6 [5].

$$tf_{t,d} = \begin{cases} 1 + log_{10}count(t,d) & if\ count(t,d) > 0 \\ 0 & otherwise \end{cases} \qquad (4)$$

$$idf_i = log_{10}\left(\frac{N}{df_i}\right) \qquad (5)$$

$$w_{t,d} = tf_{t,d}\ x\ idf_i \qquad (6)$$

where $count(t, d)$ represent the number of terms in one document, df_i represent the number of documents that

Table 4. The Top Ten Words According to Degree Centrality on The Graph of Word Model without Stopword Removal

Words	Translate	Degree Centrality Score
من	From	0.1312
الله	Allah	0.0837
في	At	0.0665
ما	What	0.0600
إن	That (conjunction)	0.0545
ولا	Nor (conjunction)	0.0543
لا	No	0.0477
أن	That (conjunction)	0.0473
وما	And what	0.0445
على	On	0.0414

Table 5. The Top Ten Words According to Betweenness Centrality on The Graph of Word Model without Stopword Removal

Words	Translate	Betweenness Centrality Score
من	From	0.1919
الله	Allah	0.1092
في	At	0.0813
ما	What	0.0606
إن	That (conjunction)	0.0556
ولا	Nor (conjunction)	0.0541
أن	That (conjunction)	0.0444
لا	No	0.0443
على	On	0.0406
وما	And what	0.0360

Table 6. The Top Ten Words According to Closeness Centrality on The Graph of Word Model without Stopword Removal

Words	Translate	Closeness Centrality Score
من	From	0.4860
الله	Allah	0.4690
ما	What	0.4439
في	At	0.4402
إن	That (conjunction)	0.4370
إلا	except	0.4284
لا	No	0.4250
ولا	Nor (conjunction)	0.4240
لهم	Their	0.4230
به	With	0.4207

contains term-i, N represent total all of document and $w_{t,d}$ represent TF-IDF value of word t in document d.

After we obtain the centrality value from formula 1, 2, and 3, we can implement that value as weight in term weighting (TW). The process of calculating term weighting is same as TF-IDF where term weighting value will be combined with IDF to get TW-IDF [2]. In undirected graph, TW-IDF can be computed by the following formula [18],

$$w_{t,d} = tw_{t,d} \; x \; idf_i \tag{7}$$

where $tw_{t,d}$ is a value of centrality of term t in document d.

2.6. Classification

In this research, the classification method to be used is naïve Bayes and SVM. Naïve Bayes is one of the supervised learning methods that calculate probability based on Bayes Theorem, under the strong assumption of independence between each features [9]. Besides that, naïve Bayes have endurance to face the missing value and efficient in computationally [19]. Naïve Bayes can be computed with this following formula,

$$P(Class \mid Document) = P(Class) \prod P(Word_i|Class) \tag{8}$$

where $P(Class)$ is the probability of independence class, $P(Word_i|Class)$ is the probability of $Word_i$ in document in class C, and $P(Class \mid Document)$ is the probability class of the document.

SVM is one of the supervised learning methods [9]. This method separates data between two or more classes by maximizing the margin between hyperplane with the closest data from each class. We can compute the distance between data with hyperplane using the following formula,

$$y_i(w.x_i + b) \geq 1 \;\; ; i = 1, 2, \dots n \tag{9}$$

where y_i represents the class of the i-th data whose value is given by x_i. The parameter w and b is the weight value of the hyperplane that need to be found.

Before the classification process is carried out using machine learning methods, we must divide the dataset into training data and test data. In this research, several test scenarios have been tested with different training data and the result showed the best performance of classification is achieved when the training data is 70%. Therefore, in this research we use the training and testing data at 70:30 composition.

2.7. Evaluation

Hamming loss is used to measure the classification performance. Hamming loss is an evaluation method that can be used for multilabel classification process that

DOI: https://doi.org/10.29207/resti.v5i3.3171

Table 8. The Top Ten Words According to Degree Centrality on The Graph of Word Model with Stopword Removal

Words	Translate	Degree Centrality Score
الله	Allah	0.1234
قال	He Said	0.0372
قل	Say	0.0267
الأرض	Earth	0.0263
كان	Take place	0.0256
قالوا	They (will) say	0.0244
آمنوا	Believed	0.0206
ربك	Your God	0.0203
عليهم	on Them	0.0198
كفروا	Disbelieve	0.0182

Table 9. The Top Ten Words According to Betweenness Centrality on The Graph of Word Model with Stopword Removal

Words	Translate	Betweenness Centrality Score
الله	Allah	0.3441
قال	He said	0.0713
كان	Take Place	0.0486
الأرض	Earth	0.0447
قل	Say	0.0412
قالوا	They (will) say	0.0399
عليهم	on Them	0.0316
ربك	Your God	0.0294
آمنوا	Believed	0.0279
يوم	Day	0.0275

can calculate prediction error (wrong predicted label) and missing error (unpredictable relevant labels) [20]. Hamming loss can be computed with the following formula,

$$HL = \frac{1}{N}\sum_{i=0}^{N-1} 1(L_i^C[j] \neq L_i^D[j]) \qquad (10)$$

where N is total label, $L_i^C[j]$ is result label that generate from classification system, and $L_i^D[j]$ is label in dataset. So, the smallest value of hamming loss is shows good performance system. Because there are only a few prediction errors or missing errors.

3. Result and Discussion

There are two testing scenarios. First, to study the effect of stopword removal in this classification process, we compare the classification with and without the stopword removal process. Second, to compare the classifier, feature extraction methods, and the effect of centrality on classification, naïve Bayes and SVM are implemented with different feature extraction methods such as TF-IDF and TW-IDF with each type of centrality measurement.

Table 7. The Top Ten Words According to Closeness Centrality on The Graph of Word Model with Stopword Removal

Words	Translate	Closeness Centrality Score
الله	Allah	0.4459
قال	He said	0.3816
كان	Take Place	0.3782
قالوا	They (will) say	0.3747
قل	Say	0.3706
الأرض	Earth	0.3664
آمنوا	Believed	0.3647
ربك	Your God	0.3643
والله	And to Allah	0.3632
كفروا	Disbelieved	0.3630

3.1. Quranic Word Centrality

The effect of centrality is observed in two GoW models: using stopword removal and without stopword removal. Different GoW models demonstrate different word centrality values. After the centrality values of each word are obtained, they are included as feature in the classification process.

Table 4, 5, and 6 listed the top ten result of centrality measurement using GoW model without stopword removal. Most of the top 10 words are stopwords, such as “from”, “what”, and “at”, except that of the word “الله” (Allah). The word “من” (from) has the highest degree, betweenness, and closeness centrality values at 0.1312, 0.1919, and 0.4860.

Table 7, 8, and 9 listed the top ten result of centrality measurement using GoW model with stopword removal. The tables show that the most central word is “الله” (Allah) with the degree, betweenness, and closeness centrality value at 0.1234, 0.3441, and 0.4459. All words in the tables are more meaningful, compared to the that in Table 4, 5, and 6, such as “Earth”, “Believe”, and “Disbelieved”. This is because the stopwords have been filtered out.

It is also worth to mention that, the word “الله” (Allah) is always among the top-two words on both scenarios. To be precise, it is the second most central word in the first scenario (Table 4, 5, 6) and the top most central in the second scenario (Table 7, 8, 9). Therefore, the word “الله” (Allah) is the most central in The Quran according to the centrality measure. The words having the highest

Table 10. The Hamming Loss Score of
The Topic Classification using Naïve Bayes Method

Feature Extraction	GoW Model 1 (without Stopword Removal)	GoW Model 2 (with Stopword Removal)
TF-IDF	0.4385	0.4385
TW-IDF Deg	0.2078	0.2078
TW-IDF Bet	0.2078	**0.2077**
TW-IDF Close	0.2925	0.2789

Table 11. The Hamming Loss Score of
The Topic Classification using SVM Method

Feature Extraction	GoW Model 1 (without Stopword Removal)	GoW Model 2 (with Stopword Removal)
TF-IDF	0.1540	**0.1540**
TW-IDF Deg	0.1739	0.1716
TW-IDF Bet	0.1770	0.1723
TW-IDF Close	0.1606	0.1605

centrality mean they are having the most relation to the other words (degree centrality), connecting the most subgraphs (betweenness centrality), and the closest with the other words in the graph (closeness centrality).

3.2. Quranic Verse Topic Classification

Quranic verse topic classification is implemented using two machine learning methods: naïve Bayes and SVM. Since this research is also comparing feature extraction methods, the classification system is implemented with four kinds of feature extraction methods: TF-IDF, TW-IDF Deg (degree centrality), TW-IDF Bet (betweenness centrality), and TW-IDF close (closeness centrality). Term weight (TW) value in this research is obtained from the centrality score as illustrated in Table 4, 5, 6, 7, 8, and 9. In this study, the classification system also used two types of GoW models which have been described in the previous subsection. Finally, the performance of classification system is measured using hamming loss.

The evaluation result of the topic classification using naïve Bayes is presented in Table 10. The table shows that the lowest (best) hamming loss score is obtained using TW-IDF Bet at 0.2077. This shows that the topic classification using naïve Bayes method achieves the best performance with the feature extraction method TW-IDF Bet. The table also shows that the hamming loss score between the two GoW models is very similar for each feature extraction method. This means that the stopword removal does not significantly affect classification performance.

Table 11 presents the evaluation result of topic classification using SVM. The table shows the lowest hamming loss score is achieved using TF-IDF at 0.1540. Similar to Table 10, the use of stopword removal also does not significantly affect classification performance. However, in general, both scenarios show that the GoW model 2 (with stopword removal) has slightly better performance than the other (without stopword removal).

Next, to compare naïve Bayes and SVM, we can compare the hamming loss scores between the two tables. The hamming loss scores in Table 10 (naïve Bayes) are between 0.2077 and 0.4385, whereas in Table 10 are between 0.1540 and 0.1770. Therefore, it is clear that SVM outperforms naïve Bayes in Quranic verse topic classification.

4. Conclusion

In this research, we investigated how word centrality affects multilabel topic classification on The Quran verses. SVM and naïve Bayes classifiers are implemented with TW-IDF feature extraction to construct classification models and evaluate the effect of word centrality on classification performance. According to the evaluation result, we found that the use of word centrality improves classification performance. The classification model with centrality and stopword removal process shows the best result on both classifiers. When the betweenness centrality is used with naïve Bayes classifier, the best performance is achieved with hamming loss score of 0.21. While in using SVM, the best hamming loss score is achieved at 0.15, and thus it is better than the former. Another study also explain that SVM is better than naïve Bayes method when it is implemented without using any additional features [21].

For the future work, different dataset will be considered, such as the translation of The Quran in Indonesia or English. Identifying the most central words in The Quran translation is interesting, because one Arabic word can be translated into multiple words or phrases. For example, the word "ربك" has two words with means "Your God". Some other possible future directions are to employ different classifiers, feature extractions, etc.

References

[1] A. Ishtiaq, M. A. Islam, M. Azhar Iqbal, M. Aleem, and U. Ahmed, "Graph Centrality Based Spam SMS Detection," *Proc. 2019 16th Int. Bhurban Conf. Appl. Sci. Technol. IBCAST 2019*, no. March, pp. 629–633, 2019, doi: 10.1109/IBCAST.2019.8667174.

[2] F. D. Malliaros and K. Skianis, "Graph-Based Term Weighting for Text Categorization," *Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015*, pp. 1473–1479, 2015, doi: 10.1145/2808797.2808872.

[3] N. Shanavas, H. Wang, Z. Lin, and G. Hawe, "Supervised graph-based term weighting scheme for effective text classification," *Front. Artif. Intell. Appl.*, vol. 285, pp. 1710–1711, 2016, doi: 10.3233/978-1-61499-672-9-1710.

[4] W. Al Etaiwi, A. A. Awajan, and D. Suleiman, "Keywords Extraction from Arabic Documents Using Centrality Measures," *2019 6th Int. Conf. Soc. Networks Anal. Manag. Secur. SNAMS 2019*, pp. 237–241, 2019, doi: 10.1109/SNAMS.2019.8931808.

[5] G. I. Ulumudin, A. Adiwijaya, and M. S. Mubarok, "A multilabel classification on topics of qur'anic verses in English translation

using K-Nearest Neighbor method with Weighted TF-IDF," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012026.

[6] F. S. Nurfikri and Adiwijaya, "A comparison of Neural Network and SVM on the multi-label classification of Quran verses topic in English translation," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012030.

[7] M. I. Rahman, N. A. Samsudin, A. Mustapha, and A. Abdullahi, "Comparative analysis for topic classification in Juz Al-Baqarah," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 406–411, 2018, doi: 10.11591/ijeecs.v12.i1.pp406-411.

[8] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," *2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc.*, pp. 900–903, 2017, doi: 10.1109/UKRCON.2017.8100379.

[9] A. H. Mohammad, T. Alwada'n, and O. Almomani, "Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network," *Glob. Sci. Technol. Forum J. Comput.*, vol. Volume 5, no. 1, pp. 108–115, 2016, doi: 10.7603/s40601-016-0016-9.

[10] N. F. Hardifa and K. M. Lhaksmana, "Topic Classification of Islamic Question and Answer Using Naive Bayes Classifier," vol. 4, no. August, pp. 199–204, 2019, doi: 10.21108/indojc.2019.4.2.346.

[11] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "Comparative analysis of text classification algorithms for automated labelling of Quranic verses," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 4, pp. 1419–1427, 2017, doi: 10.18517/ijaseit.7.4.2198.

[12] H. Zarrabi-Zadeh, "Tanzil Documents," 2007. http://tanzil.net/docs/home (accessed Nov. 19, 2020).

[13] A. L. Fathullah, "Indeks Tematik Al-Qur'an," *Pusat Kajian Hadist*. https://alquranalhadi.com/ (accessed Nov. 27, 2020).

[14] M. Ahmadi, E. Khadangi, S. P. Shariatpanahi, and M. H. Foroughmand-Araabi, "Presenting a computing method for finding the central verse of Quranic surahs," *2018 8th Int. Conf. Comput. Knowl. Eng. ICCKE 2018*, no. Iccke, pp. 308–313, 2018, doi: 10.1109/ICCKE.2018.8566366.

[15] F. Boudin and L. U. M. R. Cnrs, "A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction," *Ijcnlp*, no. October, pp. 834–838, 2013.

[16] E. Mailoa, "Analisis Node dengan Centrality dan Follower Rank pada Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 937–942, 2020, doi: 10.29207/resti.v4i5.2398.

[17] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.

[18] K. Skianis, F. D. Malliaros, and M. Vazirgiannis, "Fusing document, collection and label graph-based representations with word embeddings for text classification," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Student Res. Work.*, pp. 49–58, 2018, doi: 10.18653/v1/w18-1707.

[19] R. Irmanita, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, "Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naïve Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 10–16, 2021, doi: 10.29207/resti.v5i1.2770.

[20] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Inf. Process. Manag.*, vol. 54, no. 3, pp. 359–369, 2018, doi: 10.1016/j.ipm.2018.01.002.

[21] Sharazita Dyah Anggita and Ikmah, "Algorithm Comparation of Naive Bayes and Support Vector Machine based on Particle Swarm Optimization in Sentiment Analysis of Freight Forwarding Services," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 362–369, 2020, doi: 10.29207/resti.v4i2.1840.