8 (1) January 2022 50-57
Register: *Jurnal Ilmiah Teknologi Sistem Informasi*
ISSN 2502-3357 (online) | ISSN 2503-0477 (print)

Contents lists available at www.journal.unipdu.ac.id

Register

Journal Page is available to www.journal.unipdu.ac.id/index.php/register

Research article

Effect of information gain on document classification using k-nearest neighbor

Rifki Indra Perwira [a], Bambang Yuwono [b], Risya Ines Putri Siswoyo [c], Febri Liantoni [d,], Hidayatulah Himawan [e]*

[a,b,c] *Department of Informatics Engineering, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Yogyakarta, Indonesia*
[d] *Department of Informatics Education, Universitas Sebelas Maret, Surakarta, Indonesia*
[e] *Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Malacca, Malaysia*
email: [a] rifki@upnyk.ac.id, [b] bambangy@upnyk.ac.id, [c] risya.ines@gmail.com, [d,] febri.liantoni@staff.uns.ac.id, [e] p031910047@student.utem.edu.my*
** Correspondence*

ARTICLE INFO

Article history:
Received 15 April 2021
Revised 21 April 2021
Accepted 19 June 2021
Available online 5 January 2022

Keywords:
classification
feature selection
information gain
k-Nearest Neighbor
TF-IDF document

Please cite this article in IEEE style as:
R. I. Perwira, B. Yuwono, R. I. P. Siswoyo, F. Liantoni and H. Himawan, " Effect of information gain on document classification using k-nearest neighbor," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 8, no. 1, pp. 50-57, 2022.

ABSTRACT

State universities have a library as a facility to support students' education and science, which contains various books, journals, and final assignments. An intelligent system for classifying documents is needed to ease library visitors in higher education as a form of service to students. The documents that are in the library are generally the result of research. Various complaints related to the imbalance of data texts and categories based on irrelevant document titles and words that have the ambiguity of meaning when searching for documents are the main reasons for the need for a classification system. This research uses k-Nearest Neighbor (k-NN) to categorize documents based on study interests with information gain features selection to handle unbalanced data and cosine similarity to measure the distance between test and training data. Based on the results of tests conducted with 276 training data, the highest results using the information gain selection feature using 80% training data and 20% test data produce an accuracy of 87.5% with a parameter value of k = 5. The highest accuracy results of 92.9% are achieved without information gain feature selection, with the proportion of training data of 90% and 10% test data and parameters k = 5, 7, and 9. This paper concludes that without information gain feature selection, the system has better accuracy than using the feature selection because every word in the document title is considered to have an essential role in forming the classification.

1. Introduction

Higher education has a library as a facility to support the education and science of students. The primary function of a library is to unite information or knowledge from one side to the other. Libraries are traditionally managed by functional departments, catalogs, acquisitions, magazines [1]. It is estimated that more than 80% of digital documents are text-type data. So that it will force the emergence of new disciplines, namely text mining, whose role is to analyze the text from the excess of unstructured text information. The way it works starts from information extracted in an unstructured text and tries to find the patterns. In some universities, there are libraries which store various books, research results from multiple majors, and other knowledge information.

In order to facilitate the search for research references and manage research lists based on areas of interest, a text classification or document classification can be applied to categorize research titles based on the field of study. In a previous study that used k-Nearest Neighbor (k-NN) and Naïve Bayes for text classification, the result showed that k-NN performed better than Naïve Bayes over 7% [2].

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

Based on a previous study that used Information gain for feature selection, it showed better results but still must be tested for other types of data [3]. In another study, k-NN was used to classify plants based on leaf shape and used seven values from the invariant moment of leaf in classification with an accuracy rate of up to 80% [4]. Another previous study that used Naïve Bayes, k-NN, and Support Vector Machine for news classification, the result of Naïve Bayes showed stable performance, SVM performed a random result and k-NN had the best result for few data training [5].

Previous studies show that k-NN is better than other text classification methods such as Naïve Bayes and Support Vector Machine. Thus, this paper will look at the effect of information gain as feature selection on document classification based on areas of interest using k-NN [6]. The k-NN method for the text classification process is a reasonably efficient classification method [7]. This method results in a high level of accuracy in identifying or classifying data [8]. However, the precision and recall still need to be improved [9]. Information gain will calculate the number of bits with the information obtained for category prediction by looking at the presence or absence of terms in a document. This paper's primary purpose is to determine the effect of information gain feature selection on categorical data classification to determine the accuracy of the classification results.

According to the stated problem, it is necessary to conduct research that combines the k-NN method with information gain feature selection in the classification of documents for the informatics field of study. In this paper, the algorithm used is k-Nearest Neighbor as the classification method and information gain as the feature selection. The first step in the classification system is to perform the preprocessing, which is the stage that aims to process the text data into ready-to-use data, and then to be carried out for further analysis processes. Preprocessing is an important step before processing it into ready-to-use data [10].

The main contributions of this paper are 1) To propose k-NN with information gain as a feature extractor to show the effect on categorical data; 2) To compare the accuracy of the k-NN with information gain and k-NN without information gain on categorical data classification.

This paper is organized as follows: Section 2 presents the material and methods, including the data, research stage, preprocessing, TF-IDF and k-NN. Section 3 discusses data training, calculations, and test result. Finally, the summary and conclusion are presented in Section 4.

2. The Material and Method

2.1. Data

Data is obtained from 276 research titles in the Department of Informatics. The data contains the complete research title, which is classified into four research fields of study. There are four areas of study: intelligent systems, multimedia, geoinformatics, and network computing. It is based on the field study in the Department of Informatics. The data attributes as predictors used are year code, date of entry, student's name, student's ID number, research title, the field of study, and lecturer's name.

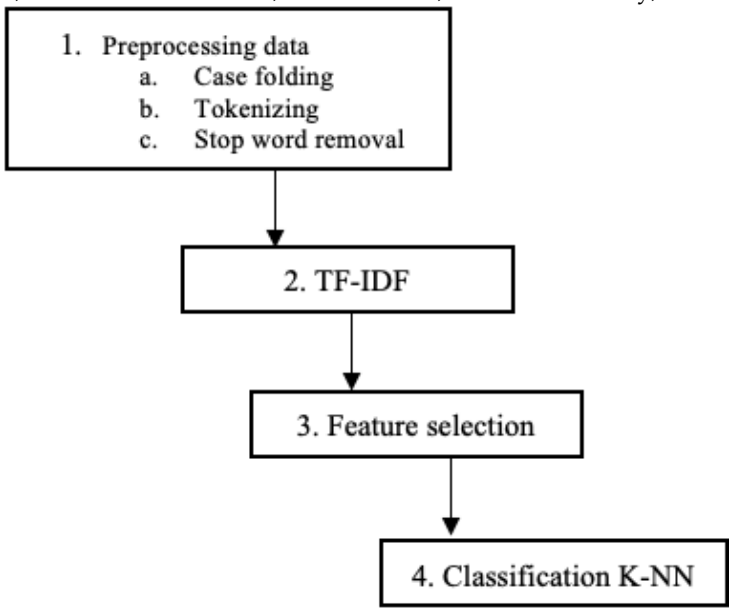

Fig. 1. Research stages

 http://doi.org/10.26594/register.v8i1.2397

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

2.2. Research stages

The stages carried out in this paper can be seen in Fig. 1. There are four main stages, including (1) preprocessing data (case folding, tokenizing, stopword removal), (2) TF-IDF, (3) information gain feature selection, and (4) k-NN classification.

2.3. Preprocessing data

Preprocessing aims to process text data into ready-to-use data for further text analysis. This stage consists of:

1). Case folding: this subprocess changes all mixed-case text to lowercase.
2). 2). Tokenizing: this subprocess divides or splits text into small pieces as collections of words or several tokens and removes punctuations [11]. Tokenization can eliminate punctuation and separate them by space.
3). Stop words removal: this subprocess is used to delete words that are not important or less relevant to the meaning of documents [12]. The most common technique to deal with these words is to remove them from the texts and documents [13] as presented in Fig. 2.

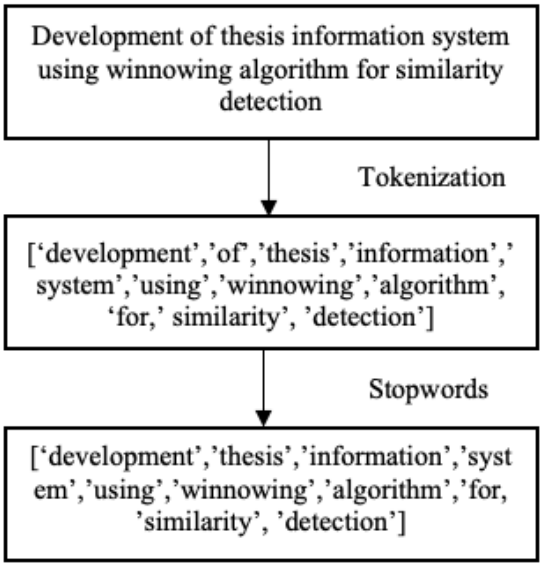

Fig. 2. Tokenization and stopword process

2.4. TF-IDF

After data preprocessing, the next step is the calculation of TF-IDF for term weighting. TF-IDF calculates the relative frequency of terms/words in a particular document by means of the inverse proportion of words throughout the document [14]. In order to complete this stage, it is necessary to take steps to find the term frequency (TF) and Inverse Document Frequency (IDF). The TF dan IDF is calculated using the formula in Eq. 1 [15],

$$TFIDF = tf \times log\left(\frac{N}{df_t}\right) \quad (1)$$

where tf is term frequency, N is the number of documents, df_t is the document frequency or the number of documents containing the term t.

2.5. Information gain

The following process is feature selection using information gain. The information gain is applied in the classification to speed up the classification process, reducing less relevant features. The Entropy and Information Gain is formulated in Eq. 2 – Eq. 3 [16],

$$Entropy_{(S)} = \sum_{i=1}^{c} -p_i \, log_2 p_i \quad (2)$$

$$IG_{(S,A)} = Entropy_{(S)} - \sum values_A \frac{|S_v|}{|S|} Entropy_{(S_v)} \quad (3)$$

where A is attribute, v is possible values in the attribute A, $values_A$ is the set of possible values for A, $|S_v|$ is number of samples for value v, $|S|$ is the sum of all data samples, and $Entropy_{(S_v)}$ is value entropy value in the attribute A.

2.6. k-Nearest Neighbor

http://doi.org/10.26594/register.v8i1.2397

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

k-Nearest Neighbor (k-NN) is a supervised learning algorithm also known as category classification algorithm [17]. After the information gain process is complete, the classification process uses the k-NN to classify documents. In this process, distance and similarity calculations are calculated between the training data and the tests data using cosine similarity. This equation is shown in Eq. 4 [18],

$$Cos_{(A,B)} = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \, x \, B_i}{\sqrt{\sum_{i=1}^{n}(Ai)^2} \, x \sqrt{\sum_{i=1}^{n}(Bi)^2}} \quad (4)$$

where A is first vector A to be compared for similarity; B is the second vector B to be compared; $A \cdot B$ is is the cross product between vector A and vector B, $|A|$ is vector length of A, $|B|$ is vector length B, and $|A||B|$ is the cross product between $|A|$ and $|B|$.

2.7. Multiclass Confusion Matrix

The accuracy of the classification system can be computed using the confusion matrix for multiclass. The example of a multiclass confusion matrix with 3 classes is represented in Fig. 3 [19]. First, calculate the value of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The equation to find accuracy, precision, and recall is formulated in Eq. 5, Eq. 6, Eq. 7, respectively [20].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Confusion Matrix		Predicted		
		Class 1	Class 2	Class 3
Actual	Class 1	A	B	C
	Class 2	D	E	F
	Class 3	G	H	I

True positives True Negatives Misclassified cases.

Fig. 3. Multiclass confusion matrix with 3 classes [19]

3. Results and Discussion

After all of the processes are completed, the result is obtained in the form of fields of study in accordance with the title of the document. The data used is the data title of the research field of informatics that has been labeled in accordance with the interest topics concerned. The total data used is 276 data with four categories and the year of data used ranges from 2015-2018. The data distribution is depicted in Fig. 4. Some portion of data is used for testing the accuracy of the system. The following subsection is the detailed data on each topic of interest for training and test.

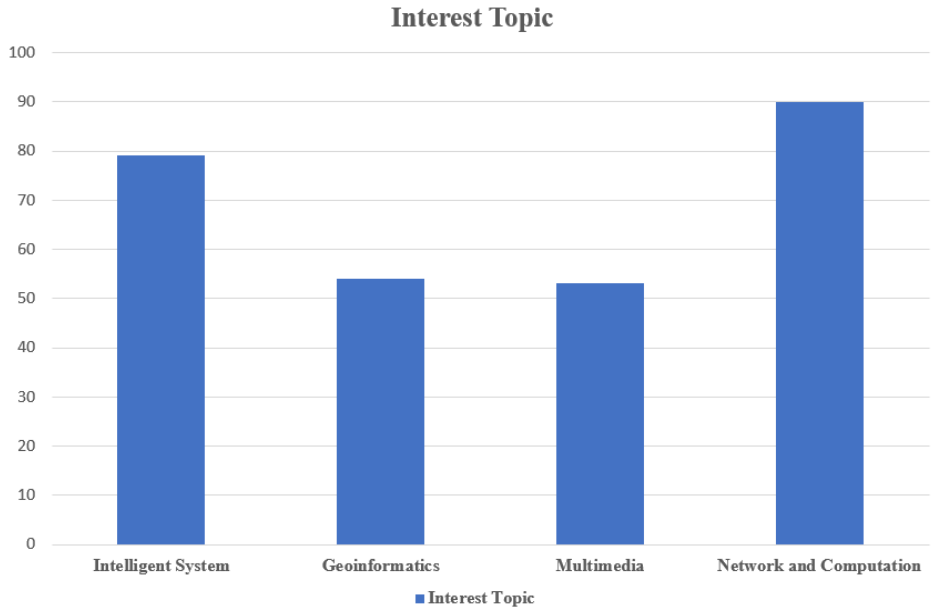

Fig. 4. Amount of interest topics

http://doi.org/10.26594/register.v8i1.2397

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

3.1. Data training

In the classification process, the field of study or labels from training data are converted into numbers to simplify the classification process. The converted labels and training data are illustrated in Table 1.

Table 1. Topic label

No	Interest topic	Label
1.	Intelligent system	0
2.	Geoinformatics	1
3.	Networking and computation	2
4.	Multimedia	3

As an example of processing of data labeling, five sample of data is given as training data (Table 2) to be carried out later in the preprocessing stage and label conversion.

Table 2. Training data

No	Data transformation: Title and corresponding topic into label		
	Title of Research	Topics	Label
1.	Application of decision making system for the selection of advanced oil production method x field enhanced oil recovery using fuzzy logic	Intelligent system	0
2.	Mapping of groundwater recharge calculations based on rainfall data from Kabupaten Sleman	Geoinformatics	1
3.	Android-based doctor appointment schedule information system at the DK clinic	Networking and Computation	2
4.	The development of thesis information system applies to detect the similarity of document case study techniques of UPN Veteran Yogyakarta	Intelligent system	0
5.	Multimedia application basic introduction to Indonesian traditional culture elementary students	Multimedia	3

3.2. Calculation TF-IDF

After each document is calculated for its weight, the next step is to calculate the document's ranking based on the level of similarity of the document to the query. If the document has a higher weight, it contains more similarity to the queried category [21], then calculates each term, followed by iterating the calculation process for each term. After calculating the TF-IDF for each term, calculate for each entropy average in each term in each field of study by using the information gain.

3.3. Cosine similarity

In the next step, cosine similarity is used to compare similarities between documents [22]. After getting the results of the calculation of the multiplication of test weights and training data as well as the quadrate of test weights and training data, the next step is to calculate cosine similarity. For each document and query, calculate the similarity of the two by using the formula of cosine similarity.

Table 3 is the result of descending cosine similarity calculations. For the value of $k = 5$, it is found that the result of the test data is classified into the field of study of Intelligent System.

Table 3. Cosine similarity result

	D1	D2	D3	D4	D5
Topics	Intelligent system	Geoinformatics	Networking and computation	Intelligent system	Multimedia
Result	0.52	0	0	0	0

3.4. Result testing

After all the process is completed, started from data preprocessing, TF-IDF calculation, feature selection using information gain, and classification with k-NN, the result of the system is evaluated. In order to properly evaluate the system effectiveness, the multiclass confusion matrix is utilized to yield accuracy, precision, and recall.

The test result shows that the use of information gain feature selection combined with the k-NN, where the number of selected features is 800 features and the information gain threshold value is above -99999, 70% of training data and 30% of test data (70:30) and the value of $k = 5$, produces the value of accuracy of 0.867 and both precision and recall of 0.735. Information gain combined with k-NN, where the percentage of 80% training data, 20% of test data (80:20), results in the highest value at $k = 5$ with the accuracy of 0.875 and precision and recall of 0.75. Information gain combined with the k-NN, where the percentage of 90% training data and 10% test data (90:10), produces the highest value at $k = 3$ and $k = 9$

http://doi.org/10.26594/register.v8i1.2397

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

with the accuracy of 0.804 and precision and recall of 0.607. Table 4 shows the result of k-NN combined with information gain.

Table 4. The result with information gain

Training Data: Test Data	Value of *k*	Accuracy	Precision	Recall
70:30	3	0.819	0.639	0.639
	5	0.867	0.735	0.735
	7	0.837	0.675	0.675
	9	0.813	0.627	0.627
	11	0.80	0.602	0.602
80:20	3	0.839	0.679	0.679
	5	0.875	0.75	0.75
	7	0.866	0.732	0.732
	9	0.866	0.732	0.732
	11	0.848	0.696	0.696
90:10	3	0.804	0.607	0.607
	5	0.75	0.5	0.5
	7	0.768	0.536	0.536
	9	0.804	0.607	0.607
	11	0.786	0.571	0.571

The experiment of k-NN without using the information gain feature selection, where 70% of training data and 30% of test data (70:30), and k = 5, 7, and 9, produces the accuracy of 0.873, precision and recall of 0.747. The experiment of k-NN without information gain, where 80% of training data and 20% of test data (80:20), k = 5, 7, and 9, produces the accuracy 0.902, precision and recall of both 0.804. The last, the experiment without using the information gain feature selection, where the percentage of training data and test data is 90%, 10% respectively (90:10), k = 5, 7, and 9, produces the highest accuracy of 0.929, precision and recall of 0.857. The results of the experiment without information gain are detailed in Table 5.

Table 5. The result without information gain

Training Data:Test Data	Value of *k*	Accuracy	Precision	Recall
70:30	3	0.843	0.687	0.687
	5	0.873	0.747	0.747
	7	0.873	0.747	0.747
	9	0.873	0.747	0.747
	11	0.867	0.735	0.735
80:20	3	0.884	0.768	0.768
	5	0.902	0.804	0.804
	7	0.902	0.804	0.804
	9	0.902	0.804	0.804
	11	0.866	0.732	0.732
90:10	3	0.893	0.786	0.786
	5	0.929	0.857	0.857
	7	0.929	0.857	0.857
	9	0.929	0.857	0.857
	11	0.911	0.821	0.821

From the computational point of view, computation using the information gain feature selection is faster than without using the feature selection. With information gain, the computation time is 8s, whilst without information gain, the computation time is 15s. It's because the data is smaller due to reduced features. However, this application, by leaving out information gain feature selection, gains better results than using information gain.

Table 6. The conclusion result

Method	Accuracy	Precision	Recall
k-NN with Information Gain	87.5	75	75
k-NN without Information Gain	92.9	85.7	85.7

http://doi.org/10.26594/register.v8i1.2397

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

4. Conclusion

In this paper, the combination of the k-NN method and the information gain feature selection can be used to classify research documents, but not effectively by judging the level of accuracy that is less than that without information gain, due to the features contained in the research document including the important features is removed. The highest accuracy result is 87.5% when using the information gain feature selection and the highest accuracy result is 92.9% when not using the information gain. In this study, k-NN without information gain is able to produce higher accuracy because all terms in the data are considered important. The conclusion result can be seen in Table 6.

This paper shows that information gain can reduce important features instead of not important ones, especially for the research title. For future work, it is intriguing to develop this system so that it is able to use more than 276 data and categories and use other feature extractors to compare and improve the accuracy.

Author Contributions

R. I. Perwira: Conceptualization, methodology, and writing. B. Yuwono: Methodology and resources. R. I. P. Siswoyo: Conceptualization, software, and resources. F. Liantoni: Supervision and validation. H. Himawan: Editing and testing.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1] M. B. Line, "The Functions Of The University Library," in *University and Research Library Studies*, W. L. Saunders, Ed., Pergamon, The University of Sheffield, 1968, pp. 148-158.

[2] M. Azam, T. Ahmed, F. Sabah and M. Hussain, "Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 18, p. 95–101, 2018.

[3] B. Azhagusundari and A. S. Thanamani, "Feature Selection based on Information Gain," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 2, no. 2, pp. 18-21, 2013.

[4] F. Liantoni, R. I. Perwira, S. Muharom, R. A. Firmansyah and A. Fahruzi, "Leaf classification with improved image feature based on the seven moment invariant," *IOP Conf. Series: Journal of Physics: Conf. Series.*, vol. 1175, 2019.

[5] F. Fanny, Y. Muliono and F. Tanzil, "A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification," *Jurnal Informatika: Jurnal Pengembangan IT,* vol. 3, no. 2, pp. 157-160, 2018.

[6] R. Jodha, S. B. C. Gaur, K. R. Chowdhary and A. Mishra, "Text Classification using KNN with different Features Selection Methods," *International Journal of Research Publications,* vol. 8, no. 1, 2018.

[7] A. Moldagulova and R. B. Sulaiman, "Using KNN Algorithm for Classification of Textual Documents," in *8th International Conference on Information Technology (ICIT)*, 2017.

[8] R. Andrian, D. Maharani, M. A. Muhammad and A. Junaidi, "Butterfly identification using gray level co-occurrence matrix (glcm) extraction feature and k-nearest neighbor (knn) classification," *Register: Jurnal Ilmiah Teknologi Sistem Informasi,* vol. 6, no. 1, pp. 11-21, 2020.

[9] H. C. Rustamaji, O. S. Simanjuntak, S. F. Luhrie, B. Yuwono and J. Juwairiah, "Categorical Data Classification based on Fuzzy K-Nearest Neighbor Approach," in *5th International Conference on Science in Information Technology (ICSITech)*, 2019.

[10] V. Kalra and R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," in *The First International Conference on Information Technology and Knowledge Management*, 2018.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
8 (1) January 2022 50-57

[11] L. A. Mullen, K. Benoit, O. Keyes, D. Selivanov and J. Arnold, "Fast, Consistent Tokenization of Natural Language Text," *Journal of Open Source Software,* vol. 3, no. 23, p. 655, 2018.

[12] N. Chandra, S. K. Khatri and S. Som, "Anti social comment classification based on kNN algorithm," in *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2017.

[13] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, "Text Classification Algorithms: A Survey," *Information,* vol. 10, p. 150, 2019.

[14] B. Trstenjak, S. Mikac and D. Donko, "KNN with TF-IDF based Framework for Text Categorization," *Procedia Engineering,* vol. 69, pp. 1356-1364, 2014.

[15] Y. Doen, M. Murata, R. Otake, M. Tokuhisa and Q. Ma, "Construction of concept network from large numbers of texts for information examination using TF-IDF and deletion of unrelated words," in *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)*, Kitakyushu, Japan, 2014.

[16] W. Zhang, T. Yoshida and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Systems with Applications,* vol. 38, no. 3, pp. 2758-2765, 2011.

[17] R. Andrian, M. A. Naufal, B. Hermanto, A. Junaidi and F. R. Lumbanraja, "k-Nearest Neighbor (k-NN) Classification for Recognition of the Batik Lampung Motifs," *IOP Conf. Series: Journal of Physics: Conf. Series,* vol. 1338, 2019.

[18] R. T. Wahyuni, D. Prastiyanto and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *Jurnal Teknik Elektro,* vol. 9, no. 1, pp. 18-23, 2017.

[19] M. Ali, D.-H. Son, S.-H. Kang and S.-R. Nam, "An Accurate CT Saturation Classification Using a Deep Learning Approach Based on Unsupervised Feature Extraction and Supervised Fine-Tuning Strategy," *Energies,* vol. 10, no. 11, p. 1830, 2017.

[20] T. M. Mohamed, "Pulsar selection using fuzzy knn classifier," *Future Computing and Informatics Journal,* vol. 3, no. 1, 2018.

[21] C.-z. Liu, Y.-x. Sheng, Z.-q. Wei and Y.-Q. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," in *International Conference of Intelligent Robotic and Control Engineering (IRCE)*, 2018.

[22] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," *Journal of King Saud University - Computer and Information Sciences,* vol. 29, no. 2, pp. 189-195, 2017.