http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 9, No. 2, Juni 2021 ISSN : 2302-9706

# IMPLEMENTATION OF TF-IDF AND COSINE SIMILARITY ALGORITHMS FOR CLASSIFICATION OF DOCUMENTS BASED ON ABSTRACT SCIENTIFIC JOURNALS

**Paska Marto Hasugian[1], Jonson Manurung[2] Logaraz[3], Uzitha Ram [4]**
[1,2] Software Engineering, STMIK Pelita Nusantara, Medan, Indonesia
[1]Paskamarto86@gmail.com*, [2]jonronmanro@gmail.com

**Article Info**
Received 10 May 2021
Revised 29 May 2021
Accepted 29 June 2021

**Abstract**

Research on one of the higher education dharmas is carried out by each lecturer and is a challenge for lecturers who pay attention to produce new and useful findings. Research results will be published in journals both nationally and internationally and one of the websites published by Ristekbirn is Sinta which includes all research works in Indonesia. The problem in this research is the accumulation of data that is getting bigger and it needs to be analyzed by utilizing text mining by searching for the resources contained in the abstract document and presenting part of the information. The purpose of this study is to classify the suitability of another document so that knowledge is found. and placement in groups according to existing topics. The process of these problems is by classifying documents based on abstracts from the publication of scientific papers. Solving these problems involves two mutually supporting algorithms, namely TD-IDF with Cosine Similarity with different tasks. TF-IDF ensures the weight of each document that can be read and read with Cosine Similarity. This research uses text mining as part of the search for related patterns and documents that have been tested. For the process of calculating the test data, 1 document and 15 documents were used as training data. With the calculation of TD-IDF the weight of each document from Q, D2 to D15 is 10,946, 28,050,27,176, 39,043, 36,535, 30,696, 25,612, 12,581, 42,335, 29,661, 33,867, 31,706, 22,654, 15,450, 59,832, 42,127, The similarity of the data is tested by determining the value of k = 4 which results in similarity to the Expert System and Cryptography, while with the selection of K = 5 with the highest similarity to the expert system..

***Keywords:*** Text Mining, Tf-If, Classification, Cs

## 1. Introduction

The development of technology today has an impact on the storage of data that is getting bigger and experiencing a buildup in databases with unlimited data, so many of the researchers conduct an experiment or research to extract by exploring the potential of the data stack or big data and produce information that can be used both for prediction, description and classification or often called Data Mining. The problem used as a reference in this study is how to dig information from the abstract scientific work of researchers that has been published in international or national journals with the main focus is abstract documents. The process is carried out in the determination of classification by utilizing text mining. Text Mining that is used to have a process by mining data or digging for information related to the text is an abstract lecturer publication that is involved by analyzing the relationship between documents in accordance

with the function and purpose of text mining, namely mining data text with document data sources to provide an overview of the connectedness of each word contained.. [1]–[3]

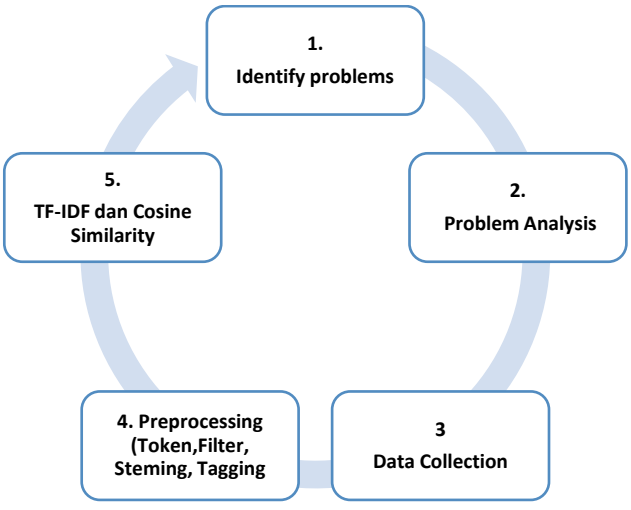
Publications related to the utilization of text mining and TF-IDF algorithms in the publication Musfiroh Nurjannah et al have conducted text mining research by utilizing TF-IDF and concluded that the application of term frequency inverse-document frequency algorithm for text mining is very helpful to obtain information on document sets. With a txt file format based on keywords entered by users on the system. [4] Bahruni et al research with a general topic is to determine the topic of student thesis based on publications on sinta and wos has analyzed with percentage of the trend of sinta used as a thesis topic, from the scope that has beenaried that the concept of text mining is to analyze cases that have the same field. And according to the principle of data mining publication sinta and directed towards the topic of students.[5] Research Ahmad Fathan Hidayatullah and Muhammad Rifqi Ma'arif conducted text mining research with the main topic is the Application of Text Mining in Thesis Title Classification and provides conclusions to compare the accuracy of the mining process by utilizing two algorithms. .[6]. Windu gata research results, purnomo has conducted text mining accuracy testing by utilizing the K-NN algorithm and concluded that the precision of content with accuracy with kategoriiya reached 98%. [7].

Based on the description, researchers conducted a classification of researchers with a combination of TD-IDF algorithm, Cosine Similarity. The analysis stage specifically utilizes the TF-IDF Algorithm, where this algorithm is a support in the management and weighting of the advantages between documents or between available texts. The process of work carried out by utilizing the TF-IDF principle, TF supports the Determination of Frequencies against the weighting of each Term, while the IDF (Inverse Document Frequency) which serves to ensure or reduce the weight value of a term if there are many occurrences. [6]–[11].

## 2. Method

### 2.1 Research Work Steps

The work steps usedto solve problems in this study and answer the phenomena defined in the introduction, then this research was developed according to the rules and procedures in figure 1 of the following:




**Gambar 1.** Working Steps of Research Methods

1. Identify problems

The problem found in this research is that lecturer research does not have the expertise or uniqueness of each lecturer in exploring topics related to Informatics.

2. Problem Analysis
   From the identification of the problem will be done analysis of the publication of lecturers before and perform the calculation process to provide a description of the most dominant trend or type of research on lecturers so as to solve the problem.
3. Data Collection
   The data used in this study is an abstract of the scientific work of lecturers that have been published in various publishers by downloading in an overall manner the paper from the lecturer concerned.
4. Preprocessing
   This stage is the stage of data preparation before analyzing and determining the trend of preprocessing description research with the following description:
   1) The Tokenizing stage is the stage of separation or cutting of each input string based on each word series in the abstract document of the lecturer's scientific publication.
   2) Filtering stage is a condition to ensure and make selection of words contained in the document based on token results in support of this process utilizing the stop list algorithm.
   3) The stemming stage is the stage of finding the basic word of each word that has been selected at the filtering stage.
   4) The tagging stage is the stage of finding the raw form or involving past words or stemming word results.
5. The analyzing stage is the determining stage of how far the connection between words between existing documents by applying the TF-IDF algorithm
6. K-NN algorithm
   At this stage, classification is carried out with the working steps of the K-NN Algorithm to ensure the similarity of test documentsand training documents.

**2.2 TF-IDF Work Process**

TF-IDF's work by following the description of Term Frequency is the frequency of the appearance of term *i* in document *j*, Document frequency (df) is the number of documents where a term (t) appears. The IDF serves as part of determining the weight of a *term* if its appearance is widely spread throughout the document. *N* is the total number of documents in the corpus, $N = |D|.|\{d \in D : t \in d\}| = df(t)$, is the number of documents containing the term *t*. The process of adding the number 1 is a determination to avoid dividing against 0 if *df(t)*is not found in the corpus .[11] The formula used to determine idf values is

$$\text{Idf (t,d)}=\log \frac{N}{|\{d \epsilon D : t \epsilon d\}|} \quad \text{Idf (t,d)}=\log \left[\frac{N}{df(t)+1}\right]$$

When the weight of each document is found, it is continued by staging the vector length for each document and calculations are carried out using Cosine Similarity. The main task of the CS algorithm is to compare similarities between test documents and training documents?.[14] so that there will be a value that can determine the percentage of each document with the provision $CosSim(Dj, Qk)$is the similarity value of each document, $tdij$is the term to I limited vector j and n is a unique term in the document mandating the formula.[15]–[18]

$$CosSim(Dj, Qk) = \frac{\sum_{i=k}^{n}(tdij * tqik)}{\sqrt{\sum_{i=1}^{n} tdij^2 * \sum_{i=1}^{n} tdik^2}}$$

The results of similarities calculated based on Cos with K number testing are as much as 4 and 5. And grouping the appearance of the highest values in accordance with the topics that have been described.

## 3. Results And Discussions

### 3.1 Presentation of data

Data that is used as part of text mining in determining the classification of research publication documents refers to publications that have been listed on the sinta page. As a sample in the calculation process involving 16 Abstract data with 15 data being training data and 1 data as the test data description in table 1 below:

**Table 1. Text Documents Based on Journal Abstracts**

| Term | Publication Abstract |
|---|---|
| Q | Market Basket Analysis is an analysis of a customer's buying habits by looking for associations and correlations between different items that customers put in their grocery carts.…. |
| D2 | An expert system is a system that uses human knowledge where it is fed into a computer and then used to solve problems that require human expertise.…. |
| D3 | The process of selecting outstanding lecturers carried out from year to year has been carried out well and has produced a ranking in accordance with the criteria set by the Ministry of Technology Research and Higher Education.…. |
| D4 | The study aims to make predictions of students' learning achievement based on parents' socioeconomic status, motivation, student discipline and past achievement using data mining methods with the J48 algorithm.……. |
| D5 | Home wallpaper or wallpaper is a wall decoration with various motifs and colors. Wallpaper is used to change the appearance of a space to make it more beautiful and have added value. Plain walls tend to make the occupants of the house feel bored because of the monotonous appearance of the walls.… |
| D.. | Computer-based flower type recognition system is the process of entering information in the form of flower type imagery into the computer…... |
| D15 | Expert Systems is a computer system designed with special abilities to solve problems and decisions that are generally made by an expert.… |

From table 1, grouping for each document is The Topic of Data Mining with document description D4, D5, Decision Support System on D3, Expert system with documents D2, D15, D16, Artificial Neural Network D12,D13, D14, Image Processing D9,D10, D11 and Cryptography with data D6,D7,D8. The description of the data grouping is described in the following graph:

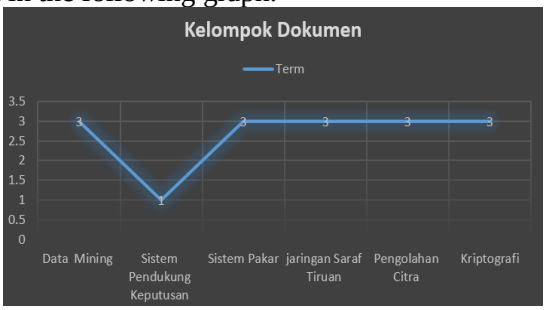



**Figure 2.** Document Group According to Topic

Based on data with 16 abstract documents carried out the process by following the rules in text mining, namely by determining tokens as many as 862 words, followed by the filtering process with the number of words as many as 705, with stemming stages as many as 644 words with word selection used in testing as many as 213 words, the description of word utilization with the following graph :

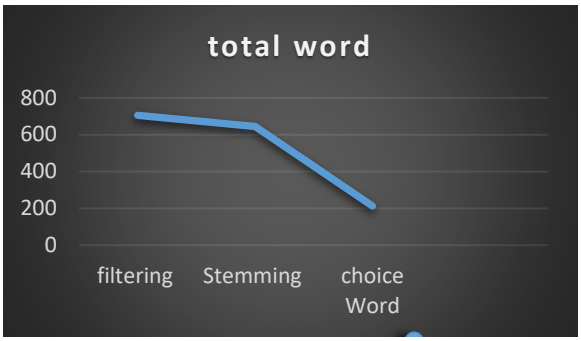



**Figure 3.** Text Mining Utilization Graph

### 3.2 Implementation of TF-IDF

Based on the data presented, the calculation process is carried out by applying the TF-IDF algorithm as outlined in table 2. WITH DF (DocumentFrequency) and IDF (Invers Document Frequency) values.

**Table 2. DF Values and IDF Documents**

| No | Term | Term Frequency | | | DF | IDF |
|---|---|---|---|---|---|---|
| | | Q | D.. | D16 | | |
| 1 | data | 2. | 0 | 1 | 13 | 0.2 |
| 2 | research | 0 | 0 | 1 | 11 | 0.4 |
| 3 | system | 0 | 7 | 6 | 9 | 0.7 |
| 4 | method | 0 | 1 | 1 | 13 | 0.2 |
| 5 | cryptography | 0 | 0 | 0 | 4 | 1.9 |
| 6 | network | 0 | 0 | 0 | 5 | 1.5 |
| 7 | flower | 0 | 0 | 0 | 3 | 2.3 |
| 8 | information | 0 | 1 | 1 | 9 | 0.7 |
| 9 | algorithm | 1. | 0 | 0 | 6 | 1.3 |
| 10 | transaction | 0 | 0 | 0 | 2 | 2.9 |
| 11 | disease | 1 | 0 | 10 | 3 | 2.3 |
| .. | Analyze | 1 | 0 | 0 | 3 | 2.3 |
| .. | Next | 0 | 0 | 0 | 2 | 2.9 |
| .. | vision | 0 | 0 | 0 | 2 | 2.9 |
| 200 | user | 0 | 0 | 0 | 2 | 2.9 |
| 212 | achievement | 0 | 0. | 0 | 2 | 2.9 |
| 213 | differentiate | 0 | 0 | 0 | 2 | 2.9 |

With the calculation of IDF Values, weights are determined for each document that is poured in table 3. Here it is:

Table 3. Weight of Each Document

| No | Term | W=TF*IDF | | | |
|---|---|---|---|---|---|
| | | Q | D2 | D4 | D16 |
| 1 | data | 0.4 | 0 | 0.206 | 0.2 |
| 2 | research | 0 | 0 | 0.447 | 0.4 |
| 3 | system | 0 | 5.1 | 0.736 | 4.4 |
| 4 | method | 0 | 0.2 | 0.206 | 0.2 |
| 5 | cryptography | 0 | 0 | 1.906 | 0 |
| 6 | network | 0 | 0 | 1.584 | 0 |
| 7 | flower | 0 | 0 | 2.321 | 0 |
| 8 | information | 0 | 0.7 | 0.736 | 0.7 |
| 9 | algorithm | 1. | 0 | 2.643 | 0 |
| 10 | transaction | 0 | 0 | 2.906 | 0 |
| 11 | disease | 0 | 0 | 2.321 | 23 |
| .. | Analyze | 2. | 0 | 4.643 | 0 |
| .. | Next | 0 | 0 | 2.906 | 0 |
| .. | vision | 0 | 0 | 2.906 | 0 |
| 200 | user | 0 | 0 | 2.906 | 0 |
| 212 | achievement | 0 | 0 | 2.906 | 0 |
| 213 | differentiate | 0 | 0 | 2.906 | 0 |

The weight description of the document tested with the value of each Document from Q, D2 to D15 is 10,946, 28,050,27,176, 39,043, 36,535, 30,696, 25,612, 12,581, 42,335, 29,661, 33,867, 31,706, 22,654, 15,450, 59,832, 42,127, vector data description on the following graph: description of vector data on the following graph: description of vector data on the following graph: description of vector data on the following graph: description of vector data on the following graph: description of vector data on the following graph:

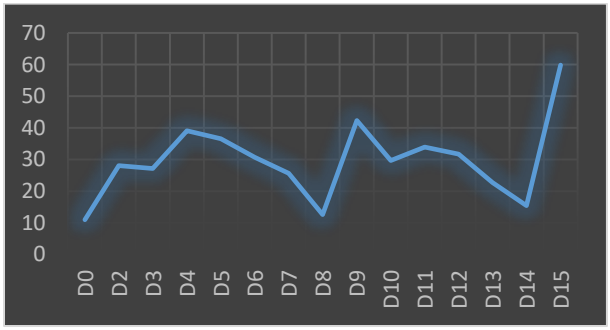



**Figure**3. Highest Weight Graph of Each Document

### 3.3 Vector Length Determination

The vector lengths for each document are outlined with the following table:

Table 4. Length of Vector Test Data Against Training Data

| Q | D2 | D3 | D4 | D15 | D16 |
|---|---|---|---|---|---|
| 0.170 | 0.000 | 0.170 | 0.043 | 0.043 | 0.043 |
| 0.000 | 0.000 | 1.802 | 0.200 | 1.802 | 0.200 |
| 0.000 | 26.61 | 0.000 | 0.543 | 4.888 | 19.55 |
| 0.000 | 0.043 | 0.043 | 0.043 | 0.000 | 0.043 |
| 0.000 | 0.000 | 0.000 | 3.636 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 2.512 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 5.391 | 0.000 | 0.000 |
| 0.000 | 0.543 | 0.000 | 0.543 | 2.172 | 0.543 |
| 1.747 | 0.000 | 0.000 | 6.990 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 8.450 | 1428.0 | 0.000 |
| 0.000 | 0.000 | 0.000 | 5.391 | 0.000 | 539.2 |
| 5.391 | 0.000 | 0.000 | 21.56 | 5.391 | 0.000 |
| 0.000 | 0.000 | 0.000 | 8.450 | 8.450 | 0.000 |
| 0.000 | 0.000 | 0.000 | 8.450 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 8.450 | 0.000 | 0.000 |
| 0.000 | 0.000 | 33.80 | 8.450 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 8.450 | 0.000 | 0.000 |

### 3.4 Stages of Cosine Similarity

At the Cosine Similarity stage is a process to test the similarity of test data that has been provided to ensure the 1st document (First) is more dominant on the research topic. For the determination of Cosine Similarity with the main focus is the determination of vector length. In table 4 is a description of the Cosine similararity value of the document to be tested

Table 5. Value Cosine Similarity

| Dokumen | *cosine similarity* |
|---|---|
| D2 | 1 |
| D8 | 0.9999 |
| D6 | 0.9999 |
| D14 | 0.9999 |
| D11 | 0.9998 |
| D3 | 0.9998 |
| D9 | 0.9997 |
| D10 | 0.9997 |
| D5 | 0.9993 |
| D12 | 0.9992 |
| D15 | 0.9991 |
| D7 | 0.9984 |
| D13 | 0.9982 |
| D4 | 0.9582 |

From the description of table 5 taken data with the number of K = 4 with the highest values, namely D2, D8, D6 and D14 From table 1 is grouping for each document, namely Data Mining Topic with document description D4, D5, Decision Support System on D3, Expert system with documents D2, D15, D16, Artificial Neural Network D12, D13, D14, Image Processing D9,D10, D11 and Cryptography with data D6,D7,D8. The description of the data grouping is described in the following graph:

Table 6. Similarity of Test Document to Value K=4

| Document | Description of cosine similarity |
|---|---|
| D4, D5 | The document does not support the group. |
| D3 | No M meets therequirements for the document. |
| D2, D15, D16 | Only D2 supports similarities. |
| D12,D13,D14 | No M meets therequirements for the document. |
| D9,D10,D11 | No M meets therequirements for the document. |
| D6,D7,D8 | Only D6 supportssimilarities. |

With testing with a value of K = 4, it is obtained group of documents 1 (D1) with similarities to expert system documents, and cryptography. The process continues with the determination of the value of K = 5 with the following similarity explanation:

Table7. Similarity of Test Document to Value K=5

| Document | Description of cosine similarity |
|---|---|
| D4, D5 | Represented by 1 document |
| D3 | There are no documents representing |
| D2, D15, D16 | Represented by two documents |
| D12,D13,D14 | Represented by a single document |
| D9,D10,D11 | Represented by a single Similarity document |
| D6,D7,D8 | Represented by a single similarity document |

Similar to these calculations, the dominant test data against the D2, D15 and D16 data groups is in the expert system cluster.

## 4. Conclusion

From the results of calculations found that the weighting order of each document with a value of 10,946, 28,050,27,176, 39,043, 36,535, 30,696, 25,612, 12,581, 42,335, 29,661, 33,867, 31,706, 22,654, 15,450, 59,832, 42,127 and similarities between test data and all training data withthe highest rated description D2, D8, D6, D14,D11, D3, D9, D10, D5,D12, D15, D7, D13, D4 with calculation values are1, 0.9999, 0.9999, 0.9999, 0.9998, 0.9998, 0.9997, 0.9997, 0.9993, 0.9992, 0.9991, 0.9984, 0.9982, 0.9582. The testing process is used by determining the value of K = 4 and providing similar information The test document is Q dominant to the topic of Expert Systems, and Cryptography while at the time the value of K = 5 topic Q is dominant against the Expert System cluster. The calculation value of Q's similarity to all documents tested has a significant proximity because the test data term is too low so that the frequency is identical to the existing document.

## REFERENCE

[1] S. Kurniawan, W. Gata, D. A. Puspitawati, N. -, M. Tabrani, and K. Novel, “Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, 2019.

[2] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, 2019.
[3] H. S. Utama, D. Rosiyadi, B. S. Prakoso, and D. Ariadarma, "Analisis Sentimen Sistem Ganjil Genap di Tol Bekasi Menggunakan," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, 2019.
[4] M. Nurjannah, Hamdani, and I. Fitri Astuti, "Penerapan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) Untuk Text Mining," *J. Inform. Mulawarman*, 2013.
[5] B. Bahruni and F. Fathurrahmad, "Analisis Trend Topik Penelitian pada Web Of Science dan SINTA untuk Penentuan Tema Tugas Akhir Mahasiswa AMIK Indonesia Banda Aceh," *J. SAINTEKOM*, 2020.
[6] A. Fathan Hidayatullah, M. Rifqi Ma, and arif Program Studi Manajemen Informatika STMIK Jenderal Achmad Yani Yogyakarta Jl Ringroad Barat, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," *Semin. Nas. Apl. Teknol. Inf. Agustus*, 2016.
[7] W. Gata, "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS," vol. 6, pp. 1–13, 2017.
[8] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Syst. Appl.*, 2011.
[9] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in *Proceedings - 2014 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology Through University-Industry Collaboration, ICITEE 2014*, 2014.
[10] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognit. Lett.*, 2016.
[11] I. Yahav, O. Shehory, and D. Schwartz, "Comments Mining With TF-IDF: The Inherent Bias and Its Removal," *IEEE Trans. Knowl. Data Eng.*, 2019.
[12] A. I. Kadhim, Y. N. Cheah, and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," in *Proceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014*, 2015.
[13] H. Niemann, M. G. Moehrle, and J. Frischkorn, "Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application," *Technol. Forecast. Soc. Change*, 2017.
[14] Imam Riadi, Sunardi, and P. Widiandana, "Investigating Cyberbullying on WhatsApp Using Digital Forensics Research Workshop," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, 2020.
[15] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Comput. y Sist.*, 2014.
[16] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.
[17] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *Proceedings of 2016 4th International Conference on Cyber and IT Service Management, CITSM 2016*, 2016.
[18] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, 2018.