Journal of Applied Engineering and Technological Science
Vol 2(2) 2021 : 14-21

PRODUCT CODEFICATION ACCURACY WITH COSINE SIMILARITY AND WEIGHTED TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY (TF-IDF)

Sintia[1*], Sarjon Defit[2], Gunadi Widi Nurcahyo[3]
Department of Information System, Putra Indonesia University YPTK Padang[1,2,3]
Sintiasartikaputri11@gmail.com[1]

Received : 29 March 2021, Revised: 03 May 2021 , Accepted : 03 May 2021
**Coresponding Author*

ABSTRACT
In the SiPaGa application, the codefication search process is still inaccurate, so OPD often make mistakes in choosing goods codes. We need Cosine Similarity and TF-IDF methods that can improve the accuracy of the search. Cosine Similarity is a method for calculating similarity by using keywords from the code of goods. Term Frequency and Inverse Document (TFIDF) is a way to give weight to a one-word relationship (term). The purpose of this research is to improve the accuracy of the search for goods codification. Codification of goods processed in this study were 14,417 data sourced from the Goods and Price Planning Information System (SiPaGa) application database. The search keywords were processed using the Cosine Similarity method to see the similarities and using TF-IDF to calculate the weighting. This research produces the calculation of cosine similarity and TF-IDF weighting and is expected to be applied to the SiPaGa application so that the search process on the SiPaGa application is more accurate than before. By using the cosine sismilarity algorithm and TF-IDF, it is hoped that it can improve the accuracy of the search for product codification. So that OPD can choose the product code as desired.
Keywords : *TF-IDF; Cosine Similarity; Term Frequency; Invers Document Frequency; Search Accuracy*

1. Introduction

The SiPAGA application is an information system for planning prices for goods and services managed by the Procurement Administration Bureau and BMD Management. In this application, there is a codification search feature to complete the Regional Property Needs Plan (RKBMD) as well as the Standard Price for Goods and Services (SHBJ). The codification of goods is an important part of the SiPAGA application which aims to facilitate the implementation of the management and administration of regional property.In the code of goods, there is the same description but the accounts, groups, types and objects are different, there is often the wrong choice of codification of goods, to reduce error grouping of goods We need a search technique that looks at the suitability of the codification being sought. If the search for codefication uses the keyword "bahan bangunan" and the code for the goods available for that keyword, but if the keyword is changed to "bangunan bahan" then no existing codification will appear. Search should be made with both "bahan bangunan" or "bangunan bahan" because they contain the same word. For this reason, the application of text-mining can be used in analyzing data for the codification of goods. In order to apply text-mining, existing goods coding data is used as new knowledge. The Cosine Similarity algorithm and Term Frequency and Invers Document Frequency (TF-IDF) weighting can be used to classify the codification of goods according to the search word.

The codefication of goods is based on the regulation of the Ministry of Home Affairs of the Republic of Indonesia number 108 of 2016 concerning the classification and coding of regional property with the aim of the local government doing the codification which describes the account code, group code, type code, object code, object detail code, object sub detail code and code. the sub-details of the objects belonging to the area. Codification includes 7 levels including:

1. Level 1 shows the account code
2. Level 2 shows the group code
3. Level 3 shows the type code

4. Level 4 shows the object code
5. Level 5 shows the detailed code of the object
6. Level 6 shows the sub code details of the object
7. Level 7 shows the sub-code details of the object

The number of available goods codification is 14,417 lines, if the item codification is not available in the sub-details of the object, it can be added that the goods codification is determined by the regional head's decision, so that the product codification can be more than 14,417 lines previously available.

2. Literature Review

Knowledge Discovery In Database (KDD) is the process of searching for and identifying patterns in data, the resulting patterns make data more useful and understandable (Putra et al., 2018). In KDD there are several phases, namely selection where data is changed from unstructured to structured. Text mining uses Term Frequency-Invers Document Frequency (TF-IDF) weighting and cosine similarity. The cosine similarity method is widely used in data mining and machine learning. In particular, cosine equality is most commonly used in higher dimensional spaces. For example, in information retrieval and text mining, cosine similarity provides a useful measure of how similar two documents. (Luo et al., 2018). In the research that has been carried out for testing web browsers using Ontology and TF-IDF (Hafeez & Patil, 2017), which aims to improve searches on web browsers. To assess text similarity using Cosine Similarity has also been researched and produced an application that can be used to detect text similarities (Rozeva & Zerkova, 2017).

In the field of education, the application of TF-IDF weighting and the Vector Space Model is carried out to determine the examiner lecturers (Siregar et al., 2017). The results of this study can recommend the examiner according to the topic based on the suitability of the title and abstract. Other research to detect plagiarism in scientific works has also been applied by using bibliography to find similar themes (Sejati et al., 2019), using the cosine similarity method. Other research is used to determine the supervisor (Yasni et al., 2018), This research produces supervisors who are in accordance with the final project submitted by students. The system created using Cosine Similarity Matching. In other studies discussing about check the document similarity (Naf'an et al., 2019). Documents with high similarity with a value of 50% and a low value of 40% so this method produces a similarity value from each of the comparators. The method used by Cosine Similarity to detect document similarities. Other studies to see similarities in journals (Kharismadita & Rahutomo, 2017). This system gets a similarity value that compares the entire journal content starting from the abstract, title and content. This system uses TF-IDF and Cosine Similarity.

In social media, Twitter is used to analyze sentiment among Twitter users (Deviyanto & Wahyudi, 2018), discussing sentiment using K-Nearest Neighbor in the DKI regional elections, the results obtained an accuracy of 67.2%, 56.94% precision value and 78.24% recall. In the religious field, especially Islam is used to find out the sharah hadith (Amrizal, 2018), The resulting system can find the hadith syarah accurately because of the high level of accuracy, recall and precision that applies the TF-IDF and Cosine Similarity methods. Paper grouping based on classification by combining TF-IDF and LDA to calculate the importance of each paper and grouping papers with similar subjects using the K-Means algorithm in order to get the correct classification results (Kim & Gil, 2019). Classifier-based approach to ontology alignment based on a hybrid of string-based features and semantic similarity. Word embedding is used to produce a slick feature for classification in addition to the new features being introduced (Nkisi-Orji *et al*, 2019). To detect topics that are being discussed to get the latest information from existing words, using an algorithm using TF-IDF is proposed to solve this problem, with experimental results with search accuracy reaching 78.36% (Zhu *et al*, 2019). Experiments in the IMDB dataset show that accuracy

is improved when using Cosine Similarity compared to using point products, whereas using a combination of features with weighted Naıve Bayes n-gram bags achieves a new state of the art accuracy of 97.42% (Thongtan & Phienthrakul, 2019)

3. Research Methods

Text Mining is the process of extracting information from unstructured or less structured data sources, such as from Word documents, PDFs, text citations, etc. whereas Data Mining is structured data (Siregar et al., 2017). Another definition of Text Mining is that it can be broadly defined as a process of extracting information in which a user interacts with a set of documents using analysis tools which are components in mining data, one of which is categorization (Amrizal, 2018). Text Mining generates data in the form of basic words from data sources, Each root word can appear in more than one document, the number of occurrences of each word is useful for measuring how important a word is in the document (Nurdiansyah et al., 2019). Data Mining is part of the Knowledge Discovery from Data (KDD) process (Putra, Randi Rian, 2018).

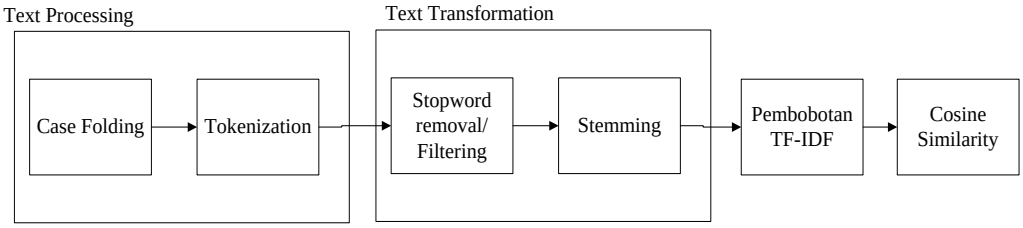

Fig. 1. Text mining stages

Text Processing

Text Preprocessing is a stage for preparing text into data to be processed (Amrizal, 2018). The text preprocessing stage also selects and removes words that have no meaning (Sejati et al., 2019). The following are the stages in the text preprocessing process:

a. Case Folding
   The Case Folding process is carried out to uniform all characters in the text to lowercase.
b. Tokenizing
   This process is carried out to separate the text into individual features (tokens) which will be processed by the system.

Text Transformation

The text transformation stage consists of the stemming process and the term stopword or filtering

a. Stopword Removal
   This stage will take words that are considered important from the results of tokenization or discard words that are considered not too important in the text mining process.
b. Stemming
   Stemming aims to transform a word into a root word by removing all word affixes.

TF-IDF

TF-IDF weights are calculated locally from the Post editing dataset using word frequency as TF. The word stop is included in the model (Arroyo-Fernández et al., 2019). TF-IDF is a way to give weight to a word relationship (term) to a document. This method combines two concepts for weight calculation, namely the frequency of occurrence of a word in a particular document and the inverse frequency of documents containing that word. The frequency with which the word appears in a given document shows how important it is in the document.

Cosine Similarity

Cosine similarity merupakan kesamaan tekstual yang mengandalkan pengenalan matriks relasi dalam menghitung kesamaan kosinus antara 2 vektor yang berhubungan langsung dengan jumlah dari kata-kata yang sama dalam kedua teks (Charlet & Damnati, 2018).

4. Results and Discussions

The input data used as an example is only 4 can be seen in table 1 as follows

Table 1 - Input Data

No	Code	Description	Information
1	q	Bahan bangunan	Keyword
		Bahan Bangunan	Teks Pembanding
2	t1	Dan Konstruksi	
3	t2	Electro Dalas	Teks Pembanding
4	t3	Patok Beton	Teks Pembanding
5	T4	Bahan Kimia Padat	Teks Pembanding

All the data that has been obtained are then processed and analyzed for the problems that occur so as to produce useful information to overcome the problems and can propose an improvement. This data processing uses text mining theories with Cosine Similarity and TF-IDF weighting. can be seen in table 2 as follows

Table 2 - Weigthing TF-IDF

Term	TF						d	D/Df	IDF
	q	T1	T2	T3	T4	T5			log(D/df)
bahan	1	1	0	0	0	1	2	2.5	0.398
bangun	1	1	0	0	0	0	1	5	0.699
beton	0	0	0	1	1	0	2	2.5	0.398
dalas	0	0	1	0	0	0	1	5	0.699
electro	0	0	1	0	0	0	1	5	0.699
kimia	0	0	0	0	0	1	1	5	0.699
konstruksi	0	1	0	0	0	0	1	5	0.699
Padat	0	0	0	0	0	1	1	5	0.699
patok	0	0	0	1	0	0	1	5	0.699
tiang	0	0	0	0	1	0	1	5	0.699

After getting the TF and IDF values, multiply the TF and IDF values as in table 3.

Table 3 - TF-IDF

Term	Wg=tf x Idf					
	q	T1	T2	T3	T4	T5
bahan	0.398	0.398	0.000	0.000	0.000	0.398
bangun	0.699	0.699	0.000	0.000	0.000	0.000
beton	0.000	0.000	0.000	0.398	0.398	0.000
dalas	0.000	0.000	0.699	0.000	0.000	0.000
electro	0.000	0.000	0.699	0.000	0.000	0.000
kimia	0.000	0.000	0.000	0.000	0.000	0.699
konstruksi	0.000	0.699	0.000	0.000	0.000	0.000
Padat	0.000	0.000	0.000	0.000	0.000	0.699
patok	0.000	0.000	0.000	0.699	0.000	0.000
tiang	0.000	0.000	0.000	0.000	0.699	0.000

After TF-IDF calculations are carried out, Cosine Similarity calculations are then carried out manually with the sample data on the "material" token, with the following formula wqi x tij.

t1 = 0.398 x 0.398 = 0.158

t2 = 0. 398 x 0 = 0

t3 = 0. 398 x 0 = 0

t4 = 0. 398 x 0 = 0

t5 = 0. 398 x 0, 398 = 0, 158

The results of the above calculations can be seen in table 4 below.

Table 4 Cosine Similarity

Term	wqi x tij				
	wt1	wt2	wt3	wt4	wt5
bahan	0.158	0.398	0.000	0.000	0.158
bangun	0.489	0.000	0.000	0.000	0.000

beton	0.000	0.000	0.000	0.000	0.000
dalas	0.000	0.000	0.000	0.000	0.000
electro	0.000	0.000	0.000	0.000	0.000
kimia	0.000	0.000	0.000	0.000	0.000
konstruksi	0.000	0.000	0.000	0.000	0.000
Padat	0.000	0.000	0.000	0.000	0.000
patok	0.000	0.000	0.000	0.699	0.000
tiang	0.000	0.000	0.000	0.000	0.000
Total	0.647	0	0	0	0,158

After using the formula above. The results of the TF-IDF weighting are squared with the sample on the "material" token data as follows:

Wq2= $0{,}398^2$=0,158

W12= 0^2=0

W22= 0^2=0

W32= 0^2=0

W42= 0^2=0

W52= 0, 398^2=0,158

The manual results can be seen in table 5 below.

Table 5 - Vector Length

Term	wqi x tij					
	q	T1	T2	T3	T4	T5
bahan	0.158	0.158	0.000	0.000	0.000	0.158
bangun	0.489	0.489	0.000	0.000	0.000	0.000
beton	0.000	0.000	0.000	0.158	0.158	0.000
dalas	0.000	0.000	0.489	0.000	0.000	0.000
electro	0.000	0.000	0.489	0.000	0.000	0.000
kimia	0.000	0.000	0.000	0.000	0.000	0.489
konstruksi	0.000	0.489	0.000	0.000	0.000	0.000
Padat	0.000	0.000	0.000	0.000	0.000	0.489
patok	0.000	0.000	0.000	0.489	0.000	0.000
tiang	0.000	0.000	0.000	0.000	0.489	0.000
Jumlah	0.647	1.135	0.977	0.647	0.647	1.135
Akar	0.804	1.066	0.988	0.804	0.804	1.066

As for calculating Cosine Similarity manually, only 5 data samples were taken based on the results in tables 4 and 5, as follows

$$Cos\ t1 = \frac{\sum_{i=1}^{n} wq_i x\ wt2}{\sqrt{\sum_{i=1}^{n}(wq_i)^2}\ x\ \sqrt{\sum_{i=1}^{n}(wt2)^2}} = \frac{0{,}647}{0{,}804\text{x}\ 1{,}066} = 0{,}755$$

$$Cos\ t2 = \frac{\sum_{i=1}^{n} wq_i x\ wt2}{\sqrt{\sum_{i=1}^{n}(wq_i)^2}\ x\ \sqrt{\sum_{i=1}^{n}(wt2)^2}} = \frac{0}{0{,}804\ \text{x}\ 0{,}988} = 0$$

$$Cos\ t3 = \frac{\sum_{i=1}^{n} wq_i x\ wt3}{\sqrt{\sum_{i=1}^{n}(wq_i)^2}\ x\ \sqrt{\sum_{i=1}^{n}(wt3)^2}} = \frac{0}{0{,}804\ \text{x}\ 0{,}804} = 0$$

$$Cos\ t4 = \frac{\sum_{i=1}^{n} wq_i x\ wt4}{\sqrt{\sum_{i=1}^{n}(wq_i)^2}\ x\ \sqrt{\sum_{i=1}^{n}(wt4)^2}} = \frac{0}{0{,}804\ \text{x}\ 0{,}804} = 0$$

$$Cos\ t5 = \frac{\sum_{i=1}^{n} wq_i x\ wt5}{\sqrt{\sum_{i=1}^{n}(wq_i)^2}\ x\ \sqrt{\sum_{i=1}^{n}(wt5)^2}} = \frac{0{,}158}{0{,}804\text{x}\ 1{,}066} = 0{,}185$$

Based on the results of manual calculations, the results of the codification search are obtained as in table 6 below

Table 6 - Cosine Similarity Result

No	Code	Description	Result
1	t1	Bahan Bangunan Dan Konstruksi	0.755
2	t5	Bahan Kimia Padat	0.185
3	t3	Patok Beton	0.000
4	t4	Tiang Beton	0.000
5	T2	Electro Dalas	0.000

The results of the text mining system can be seen using software that has been built using the Codeigniter Framework, following the appearance of a text mining system that applies Cosine Similarity and TF-IDF weighting, can be seen in Figure 3.

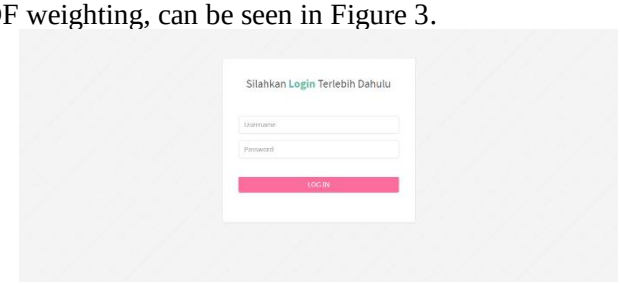

Fig. 3. Login Page

Next, to test the system, input the sample data that has been provided previously as much as 300 sample data by clicking the Choose file button, then selecting the file with the .xls extension to be input, this page can be seen in the figure, can be seen in Figure 4

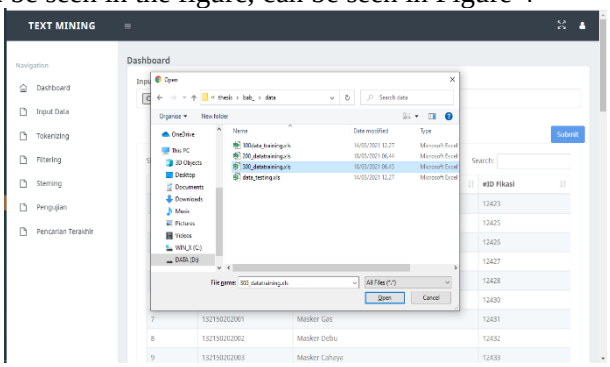

Fig. 4. Input Data

After the data is uploaded successfully, it will appear as shown in Figure 5

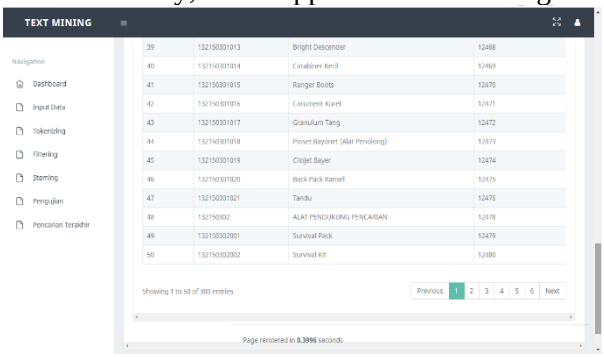

Fig. 5. Data Input Display

Furthermore, the test is carried out using the keyword "Masker Gas", the display of the test can be seen in Figure 6

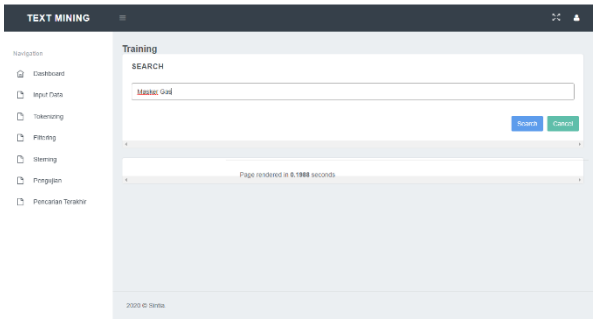

Fig. 6. Testing Data

After the search button is clicked, the test results will appear which can be seen in Figure 7

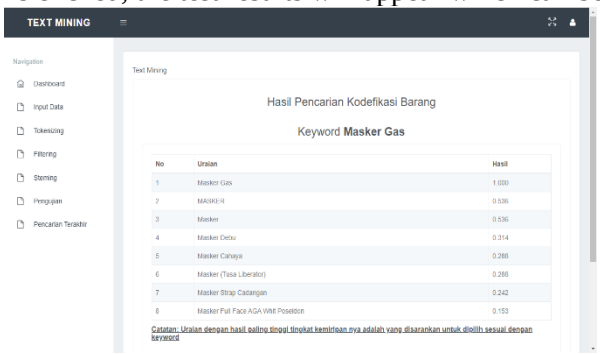

Fig. 7. Test Result

5. Conclusion

Based on the results obtained, the application of TF-IDF weighting and cosine similarity has succeeded in increasing the accuracy in the search for goods codification with Cosine Similarity calculations and TF-IDF weighting after entering keywords as keywords for search. Based on the keyword, there are several product codifications with a Cosine Similarity value of more than zero (0), the Codification of the item is the result of the search so that the highest value is the codefication of the item that is most similar to the keyword.

References

Amrizal, V. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *Jurnal Teknik Informatika*. https://doi.org/10.15408/jti.v11i2.8623

Arroyo-Fernández, I., Méndez-Cruz, C. F., Sierra, G., Torres-Moreno, J. M., & Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. *Computer Speech and Language*. https://doi.org/10.1016/j.csl.2019.01.005

Charlet, D., & Damnati, G. (2018). *SimBow at SemEval-2017 Task 3: Soft-Cosine Semantic Similarity between Questions for Community Question Answering*. https://doi.org/10.18653/v1/s17-2051

Deviyanto, A., & Wahyudi, M. D. R. (2018). PENERAPAN ANALISIS SENTIMEN PADA PENGGUNA TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR. *JISKA (Jurnal Informatika Sunan Kalijaga)*. https://doi.org/10.14421/jiska.2018.31-01

Hafeez, S., & Patil, B. (2017). Using Explicit Semantic Similarity for an Improved Web Explorer with ontology and TF-IDF. *International Journal Of Advance Scientific Research And Engineering Trends Using*.

Kharismadita, P., & Rahutomo, F. (2017). Implementasi Tokenizing Plus Pada Sistem Pendeteksi Kemiripan Jurnal SkripsI. *Jurnal Informatika Polinema*, *2*(1), 24. https://doi.org/10.33795/jip.v2i1.50

Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*.

https://doi.org/10.1186/s13673-019-0192-7

Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., & Yang, Q. (2018). Cosine normalization: Using cosine similarity instead of dot product in neural networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-01418-6_38

Naf'an, M. Z., Burhanuddin, A., & Riyani, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *Jurnal Linguistik Komputasional (JLK)*. https://doi.org/10.26418/jlk.v2i1.17

Nkisi-Orji, I., Wiratunga, N., Massie, S., Hui, K. Y., & Heaven, R. (2019). Ontology alignment based on word embedding and random forest classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-10925-7_34

Nurdiansyah, Y., Andrianto, A., & Kamshal, L. (2019). New book classification based on Dewey Decimal Classification (DDC) law using tf-idf and cosine similarity method. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1211/1/012044

Putra, Randi Rian, C. W. (2018). IMPLEMENTASI DATA MINING PEMILIHAN PELANGGAN POTENSIAL MENGGUNAKAN. *IEEE Communications Surveys and Tutorials*. https://doi.org/10.1109/COMST.2015.2457491

Putra, R. R., Wadisman, C., Sains, F., Teknologi, D., Pembangunan, U., & Medan, P. B. (2018). IMPLEMENTASI DATA MINING PEMILIHAN PELANGGAN POTENSIAL MENGGUNAKAN ALGORITMA K-MEANS IMPLEMENTATION OF DATA MINING FOR POTENTIAL CUSTOMER SELECTION USING K-MEANS ALGORITHM. *Journal of Information Technology and Computer Science*.

Rozeva, A., & Zerkova, S. (2017). Assessing semantic similarity of texts - Methods and algorithms. *AIP Conference Proceedings*. https://doi.org/10.1063/1.5014006

Sejati, F. B., Hendradi, P., & Pujiarto, B. (2019). Deteksi Plagiarisme Karya Ilmiah Dengan Pemanfaatan Daftar Pustaka Dalam Pencarian Kemiripan Tema Menggunakan Metode Cosine Similarity (Studi Kasus: Di Universitas Muhammadiyah Magelang). *Jurnal Komtika*. https://doi.org/10.31603/komtika.v2i2.2594

Siregar, R. R. A., Sinaga, F. A., & Arianto, R. (2017). Aplikasi Penentuan Dosen Penguji Skripsi Menggunakan Metode TF-IDF dan Vector Space Model. *Computatio : Journal of Computer Science and Information Systems*. https://doi.org/10.24912/computatio.v1i2.1014

Thongtan, T., & Phienthrakul, T. (2019). Sentiment classification using document embeddings trained with cosine similarity. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*. https://doi.org/10.18653/v1/p19-2057

Yasni, L., Subroto, I. M. I., & Haviana, S. F. C. (2018). Implementasi Cosine Similarity Matching Dalam Penentuan Dosen Pembimbing Tugas Akhir. *Transmisi*. https://doi.org/10.14710/transmisi.20.1.22-28

Zhu, Z., Liang, J., Li, D., Yu, H., & Liu, G. (2019). Hot Topic Detection Based on a Refined TF-IDF Algorithm. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2893980