Journal of Soft Computing Exploration

Homepage: htttps://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Content-based filtering using cosine similarity algorithm for alternative selection on training programs

Muhammad Falah Abdurrafi[1], Dewi Handayani Untari Ningsih[2]
[1, 2]Department of Informatics Engineering, Universitas Stikubank, Indonesia

Article Info

Article history:

Received October 25, 2023
Revised November 27, 2023
Accepted November 29, 2023

Keywords:

Text mining
Recommendation system
Content-based filtering
Cosine similarity
Training program

ABSTRACT

Recommendation systems are widely applied in various fields to help make choices from the many options available. One method that can be used is the Content-Based Filtering method, which is a filtering method based on the content of an object and measuring similarity using Cosine Similarity. Some applications of recommendation systems with various methods do not get optimal results and the application of text preprocessing and weighting is still minimal. This research aims to optimize the recommendation system using Content-Based Filtering with the Cosine Similarity algorithm. The training program data from the Ministry of Manpower of the Republic of Indonesia will be applied. Training program recommendations are generated based on measuring the suitability between the description of the training program and the interests of prospective trainees using the cosine similarity distance measurement. The test results using this method can achieve an average precision value of 88%, which shows the ability of the system to provide training program recommendations that are very relevant to the interests and needs of trainees.

Corresponding Author:

Muhammad Falah Abdurrafi,
Department of Informatics Engineering,
Universitas Stikubank,
Jl. Tri Lomba Juang, Mugassari, Semarang Selatan, Semarang, 50241, Indonesia.
Email: falahabdurrafi@gmail.com
https://doi.org/10.52465/joscex.v4i4.232

1. INTRODUCTION

Performance is a work result achieved by a person in carrying out his assigned tasks based on experience skills . Education and training have an important role in career development and improving the quality or performances of human resources [1]. Training can increase goals, motivation, and encouragement for employees to improve their abilities [2]. Significant behavioral changes were seen in participants who attended the training, while the group that did not attend the training did not experience significant changes [3]. However, the increasing variety of training programs offered also raises new challenges, namely in terms of how potential trainees can choose the program that best suits their personal interests, needs, and aspirations. There are many algorithms that can be applied to solve the problem above [4]. The content-based filtering approach is one of the most effective methods in developing recommendation systems, it will recommend suitable items based on the item description and also the interests of the user [5]. Content-Based Filtering refers to a way of recommending items to users based on the characteristics or attributes associated with the items.

The similarity between these items is taken into account based on the features used in the comparison, such as content, title, or description [6]. The content-based filtering method will determine the relationship between objects by measuring the level of similarity of the content of each object that has been previously weighted so that a high level of prediction can produce strong similarities [7].

Before weighting the text, it is necessary to do Text Preprocessing to prepare the content text into a format that is easily understood by the machine, so that the algorithm can be applied [8]. In the text preprocessing stage begins with cleaning the text from punctuation marks, HTML tags, and numbers, to changing the text to lowercase which is called Case Folding [9]. This text cleaning process is done with the help of regular expressions which are tools to determine strings using predetermined patterns [10]. Then after the text is cleaned, the Stopwords Removal and Stemming stages can be carried out. Stopword removal is used to remove words that include "stopwords". Stopwords are words that often appear in documents so that these words are not useful if included in the next process [11]. Meanwhile, Stemming will return words that are derivative words to their basic words [12]. Tokenization is done to convert the content text in the training program into a series of tokens [13]. After the content text has been cleaned also converted into a series of word tokens, what is done next is to give weight to the word tokens. The weighting method that can be used is TF-IDF which extracts and evaluates the relationship of each word in a group of documents [14]. Term Frequency (TF) is used to determine how important a word is based on how often it appears in a document [15]. Then, Inverse Document Frequency (IDF) is used to calculate the distribution of words in the collection of documents concerned [16]. Furthermore, the weighted text will be measured for similarity to determine the degree of similarity between two objects, so that the relationship can be known [17]. Similarity calculation can be done with Cosine Similarity, which is a commonly used metric. This method measures the similarity between two vectors as the angle between them. The difference with distance calculation methods such as Euclidean is that as the Euclidean distance between two patterns increases, the degree of similarity decreases, but conversely, as the cosine similarity value between two patterns increases, they are considered more similar [18].

Online course recommendations using Deep Convolutional Neural Networks with Negative Sequence Mining obtained an overall highest precision rate of 39% [19]. Furthermore, the implementation of the Content-Based Filtering method to provide comic selection recommendations resulted in a similarity percentage value of 76.38%, text preprocessing method is being used but the stemming steps for converting the words to their basic form are not explored [20]. Collaborative Filtering with Alternating Least Square Method and Singular Value Decomposition method is used to recommend books which result in 57% precision value for SVD method and 0.059% precision value for ALS method [21]. Then on a tour recommendation system using another method, namely Simple Additive Weighting can recommend tour packages based on cost, number of participants, and number of facilities and obtain results showing that packages with preference values above 0.7 are highly recommended, while those between 0.6 to 0.7 can still be selected because they have advantages in one of the criteria. However, packages with a preference value below 0.6 are not recommended as they are considered too expensive compared to the benefits provided. Tour packages are recommended based on criteria that are most economically suitable for prospective tourists, but have not been able to recommend tour packages that best suit the interests of prospective tourists [22]. Then the application of the content-based filtering method with a hybrid model based on chi-square feature selection and Softmax regression to recommend journals or conferences in accordance with the priority order based on the abstract produces an accuracy value of 61.37% [23], the content selection can be varied again to increase the accuracy value. Then the e-commerce product recommendation system with the content-based filtering method obtained an average precision of 78%, but this system is still less efficient in terms of power management because to create its own recommendation model requires an average time of up to 10 minutes [24].

A number of previous studies have implemented recommendation systems in various fields such as online courses, comics, books, travel, and e-commerce product recommendations. Various methods are applied to provide recommendations that match the user's wishes, but many still provide less than optimal precision results. This research aims to fill the gap by improving the optimization of the recommendation system by exploring the Content-Based Filtering method using Cosine Similarity algorithm and applying weighting to the content text with TF-IDF and adding the Stemming stage in text preprocessing which is less applied in previous studies.

2. METHOD

In this section, we will discuss the methods and stages used in this research. We use the Content-Based Filtering method to produce training program recommendation results that match the interests and needs of prospective trainees. The stages carried out include Data Collection, Text Preprocessing, TF-IDF Weighting, Cosine Similarity, Top-N Recommendation, and Evaluation. The flow of this research can be seen in Figure 1.

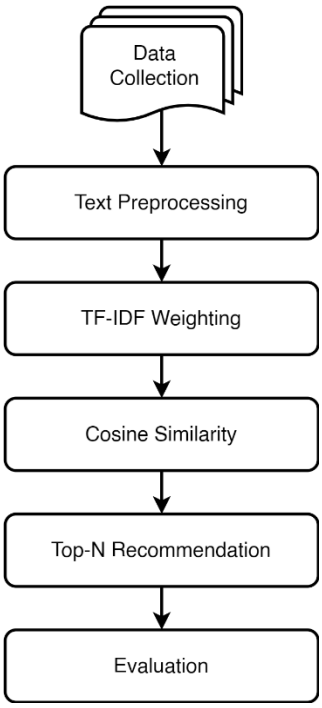

Figure 1. Research method flow

Content-Based Filtering

The Content-Based Filtering method is a filtering method that is carried out by measuring the level of similarity of the content features of each item using mathematical functions such as the Cosine Similarity measurement [6]. In this research, the similarity level between the description of the training program and the description of interests entered by prospective trainees is measured. Training programs that will be recommended are training programs with the highest level of similarity in description to the description of the interests of prospective trainees.

The way the Content-Based Filtering method works in this research is starting with Data Collection which produces files in (.csv) format containing content data from the training program. Then the data will go through Text Preprocessing to prepare the appropriate text format before proceeding to the weighting stage. Furthermore, in the weighting stage, each term from all content will be given a weight value using the TF-IDF method. After getting the weight value, the similarity measurement between each training program is then carried out using the Cosine Similarity algorithm, and to get the desired recommendation, the similarity value between the interest data and each training program data will be sorted and the top five training programs that have the highest similarity value are selected using the Top-N Recommendation method.

By implementing the Content-Based Filtering method in this research, we can see that the content data of each training program is an important part of determining the recommendation results. To evaluate the recommendation results obtained, we use Precision which will consider the relevant and irrelevant results of the recommendations obtained.

Data Collection

The initial stage of this research is Data Collection, which is the process of collecting data to gain insight related to the research topic [25]. In this research, the data collected is used as an object in applying the Content-Based Filtering method to get appropriate recommendations. The data collected is training program data from the Ministry of Manpower of the Republic of Indonesia that are publicly available on the websites https://skillhub.kemnaker.go.id and https://pelatihan.kemnaker.go.id. The data is collected by web scrapping techniques, which is a procedure for extracting data automatically and not done by copying data manually [26]. Before collecting data, it is necessary to decide in advance what attributes or columns are needed, as in this research we use the eight attributes that can be seen in Table 1.

Table 1. Attribute dataset

Attribute	Description
program_name	the name of the training program
content	the content of the training program
vocational	vocational training program
sub_vocational	sub-vocational training program
type	the type of training program (online, offline, both)
image_cover	the image link of the training program
proglat_link	the link to the Proglat site of the training program
skillhub_link	the link to the Skillhub site of the training program

Then, the data available on the website is extracted and stored according to the attributes that have been determined previously. The data obtained is then saved into a file with the format (.csv) to be processed in the next stage.

Text Preprocessing

Text Preprocessing is done before performing the weighting process on the text. The series of stages carried out in Text Preprocessing includes Case Folding, Stopwords Removal, Stemming, and Tokenization on the content text of each training program. The purpose of this stage is to prepare the content text into a format that can be more easily understood by the machine so that we can apply algorithms to the content text in the next stage [8].

At this stage, the process begins by updating the content data of the training program. This involves adding the name of the training program to its content data so that its content consists of the program's name, competencies, and description. Next, we perform Case Folding on the training program content text by cleaning the text from punctuation, HTML tags, and numbers. After cleaning the text, the next step is to remove the words that are included in the stopwords, which are no longer needed in the training program content text. Then, the content text that does not contain these stopwords will be subjected to the Stemming process by returning each word to its basic form. The last process after the words are returned to their basic form is to convert the content text that is still in the form of strings into word tokens that will then be processed to calculate the weight value at the next stage.

TF-IDF Weighting

The objective of using the TF-IDF method in this research is to weigh the terms of each training program document based on the content text which has been done by Text Preprocessing before. At this stage, the weight value of terms will be calculated and used to measure the similarity value at the next stage. Calculations to produce weighting values using the Term Frequency - Inverse Document Frequency method can be seen in Formula 1 and Formula 2 [27].

$$IDF(t) = log \frac{N}{N(t)+1} \tag{1}$$

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{2}$$

Where *TF-IDF*(*t*,*d*) is the weight of the word *t*. Then *TF*(*t*,*d*) is the frequency of occurrence of the word t in the content text d. While *N* is the total number of content texts of the training program. *N*(*t*) is the number of training program content texts that contain the word *t*. And plus 1 serves to avoid the result of 0 from *N*(*t*).

Cosine Similarity

In this research, similarity measurement is used to determine the similarity between training programs based on previously weighted content. At this stage, the measurement is done by comparing two vector documents that have been weighted previously. Vectors that are compared in this research can be in the form of content from training programs with other training programs or training programs with the interests of prospective trainees. The similarity measurement we use here is cosine similarity, which can be seen in Formula 3 [28].

$$\cos\theta = \frac{\vec{A}.\vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{(a_1 \times b_1) + (a_2 \times b_2) + \cdots + (a_n \times b_n)}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} \cdot \sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}} \tag{3}$$

Where A and B are the two vectors being compared. Cosine similarity will make measurements in vector form on the cosine of the angle between two documents [29].

Top-N Recommendation

After obtaining the results of measuring the level of similarity of each training program using Cosine Similarity, what is done next is to determine the results of recommendations using the Top-N rule based on the ranking of the recommendation score [30]. In this research, the N value used is 5, which means that the top five training programs with the closest similarity value to the interests of prospective trainees will be selected as the result of the recommendations given. To sort the recommendations of the top five training programs, the similarity value between the interests of prospective trainees and each training program calculated in the previous stage is required.

Evaluation

To test the performance of the recommendation system, it is necessary to evaluate the system. The evaluation here is done to find out how relevant the results obtained are, so the appropriate evaluation metric to use is Precision [31]. Precision calculates the percentage ratio between the number of items that are truly positive and the number of positive items in the recommendation results [32]. To calculate Precision, it can be seen in Formula 4 below [33].

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Where TP is True Positive, and FP is False Positive. True Positive here is the number of items that are truly positive, while the number of positive items can be found by summing True Positive with False Positive.

In this research, we use five samples of trainee interest data, each of which will be tested to obtain recommendation results. Each recommendation result will be divided into two groups, namely a group of recommendations relevant to the interests of prospective trainees, which is a True Positive, and a group of recommendations that are not relevant to the interests of prospective trainees, which is a False Positive.

3. RESULTS AND DISCUSSIONS

Data Collection

The data obtained is stored in (.csv) format which consists of columns of program name, content, vocational, sub-vocational, type, link to the Proglat site, link to the Skillhub site, and cover image link. The data in the content column consists of a combination of data on program competencies and program descriptions. The data collection successfully obtained 778 training programs, which can be seen in Table 2.

Table 2. Training program dataset

Program Name	Content	Vocational	Sub Vocational	Type	Image Cover	Proglat Link	Skillhub Link
Sewing Children'S Clothes	Maintain sewing tools. | Sewing by machine…	Garment Apparel	Sewing (Knitting, Woven)	both	https://proglat-assets.kemnaker.go.id/programs...	https://proglat.kemnaker.go.id/programs/f4f2ad...	https://skillhub.kemnaker.go.id/pelatihan/f4f2a...
Video Editing	Implement Occupational Safety and Health…	Information And Communication Technology	Multimedia	offline	https://proglat-assets.kemnaker.go.id/programs...	https://proglat.kemnaker.go.id/programs/d76606...	https://skillhub.kemnaker.go.id/pelatihan/d766...
Audio Video Technician	Implementing Communication in the Workplace…	Technical Electronics	Audio Video	both	https://proglat-assets.kemnaker.go.id/programs...	https://proglat.kemnaker.go.id/programs/800a8e...	https://skillhub.kemnaker.go.id/pelatihan/800a...
…	…	…	…	…	…	…	…

Text Preprocessing

The training program content text before Text Preprocessing can be seen in Table 3. This example shows the initial content of the "Sewing Children's Clothes" program which is a combination of program name, program competencies, and program description.

Table 3. Result content before Text Preprocessing

Program Name	Content
Sewing Children's Clothes	Maintain sewing tools. | Sewing by machine | Soft skills | Making decorations on clothes | Cutting materials | Follow health, safety and security procedures at work | Productivity | Sewing Children's Clothes After attending this training, participants are competent in making children's clothes according to applicable standards

Next is the Case Folding stage, at this stage the content text of each training program will be combined first with the name of the training program, then cleaned from punctuation, HTML tags, numbers, to change the text to lowercase. The results of the content text after Case Folding can be seen in Table 4.

Table 4. Result content after Case Folding

Program Name	Content
Sewing Children's Clothes	sewing childrens clothes maintain sewing tools sewing by machine soft skills making decorations on clothes cutting materials follow health safety and security procedures at work productivity sewing childrens clothes after attending this training participants are competent in making childrens clothes according to applicable standards

Then the Stopwords Removal stage is carried out to remove words that are included in "stopwords", which are words that appear frequently so they are not useful if included in the next process. Words such as “by”, “on”, “and”, “at”, “after”, “this”, “are”, “in” and “to”. The results of the content text after Stopwords Removal can be seen in Table 5.

Table 5. Result content after Stopwords Removal

Program Name	Content
Sewing Children's Clothes	sewing childrens clothes maintain sewing tools sewing machine soft skills making decorations clothes cutting materials follow health safety security procedures work productivity sewing childrens clothes attending training participants competent making childrens clothes according applicable standards

After doing Stopwords Removal, the next stage is Stemming to convert existing words into their basic form. For example, the word "sewing" after Stemming will change to "sew". The results of the content text after Stemming can be seen in Table 6.

Table 6. Result content after Stemming

Program Name	Content
Sewing Children's Clothes	sew children cloth maintain sew tool sew machin soft skill make decor cloth cut materi follow health safeti secur procedur work product sew children cloth attend train particip compet make children cloth accord applic standard

The last stage of Text Preprocessing is Tokenization to separate the content text into word tokens so that the weight can be calculated at a later stage. The results of tokenization of the content text can be seen in Table 7.

Table 7. Result content after Tokenization

Program Name	Content
Sewing Children's Clothes	['sew', 'children', 'cloth', 'maintain', 'sew', 'tool', 'sew', 'machin', 'soft', 'skill', 'make', 'decor', 'cloth', 'cut', 'materi', 'follow', 'health', 'safeti', 'secur', 'procedur', 'work', 'product', 'sew', 'children', 'cloth', 'attend', 'train', 'particip', 'compet', 'make', 'children', 'cloth', 'accord', 'applic', 'standard']

TF-IDF Weighting

The results of weighting the content text of the training program using *TF-IDF* can be seen in Table 8. In this table, there are weights of all existing terms for each document or training program. There are 1705 terms, while there are 778 training programs.

Table 8. Example of TF-IDF weighting results

Program Name	thread	high	period	prevent	…
Sewing Children's Clothes	0.0	0.0849	0.0	0.0970	…
Sew Clothes According To Style	0.0	0.0	0.0999	0.1084	…
Sewing Adult Women'S Clothing	0.4539	0.0698	0.0	0.0797	…
…	…	…	…	…	…

In the weighting results using TF-IDF, each term of the training program content can be known for its weight value in each document, which in this case is the training program. For example, the term "thread" has a TF-IDF weight value of 0.4539 in the document "Sewing Adult Women's Clothing".

Cosine Similarity

At this stage, the similarity value between each training program or the interest of prospective trainees is obtained using the cosine similarity calculation. The results of the cosine similarity calculation between training programs can be seen in Table 9.

Table 9. Example of Cosine Similarity calculation results

	Sewing Children's Clothes	Sew Clothes According To Style	Sewing Adult Women'S Clothing	…
Sewing Children's Clothes	1	0.632332	0.638564	…
Sew Clothes According To Style	0.632332	1	0.654843	…
Sewing Adult Women'S Clothing	0.638564	0.654843	1	…
…	…	…	…	…

In the results of similarity calculations using cosine similarity, it can be seen that all data between training programs can be known based on the distance value of similarity between one training program and another. For example, the similarity score between the training program "Sewing Children's Clothes" and "Sew Clothes According To Style" is 0.632332.

Top-N Recommendation

In this research, the top five training program recommendations that have the highest level of similarity with the user's interests will be selected. In this example, the interest of the potential trainee inputted is "Sewing clothes". The recommendation results with Top-5 Recommendation based on the cosine similarity value can be seen in Table 10.

Table 10. Example of top-5 recommendation results

No.	Program Name	Cosine Similarity Score
1	Sew Clothes According To Style	0.8067361779313234
2	Sewing Basic Clothes For Men & Women	0.7426776184397177
3	Sewing Adult Women'S Clothing	0.7420612898545749
4	Clothing Maker'S Assistant	0.7252129700205565
5	Sewing Children'S Clothes	0.7222742541555704

The results successfully determined the five training program options that had the closest similarity distance to the interest "Sewing clothes" entered by the prospective trainees.

Evaluation

In this study, tests were conducted to measure the performance of the Cosine Similarity algorithm in performing Content-Based Filtering. This system will provide five training program recommendations based on content that has the closest similarity distance to the interests entered. The test conducted here uses five samples of data on the interests of prospective trainees. The results of the training program recommendations given to each interest from these five samples along with the results of the similarity distance calculation can be seen in Table 11.

Table 11. Recommendation testing results

No.	Need and Interest	Training Program Recommendations	Cosine Similarity Score
1	Sewing clothes	Sew Clothes According To Style	0.8067361779313234
		Sewing Basic Clothes For Men & Women	0.7426776184397177
		Sewing Adult Women'S Clothing	0.7420612898545749
		Clothing Maker'S Assistant	0.7252129700205565
		Sewing Children'S Clothes	0.7222742541555704
2	Video editing and design	Video Editing	0.6967531254949098
		Making Video Clips	0.6694263461365894
		Video Editor	0.6369325729940079
		Multimedia Design	0.6286700283894981
		Multimedia	0.5950752604915115
3	Studying foreign language	Intermediate Arabic Language Skills	0.1188328382518855
		Japanese For Cpmi (Blkk)	0.1181765084549629

		New Productive Entrepreneurship Training	0.1098766325370635
		Entrepreneurship	0.1042737550439776
		Basic Arabic Language Skills	0.1036087628869277
4	Data processing and data analysis	Basic Data Science	0.6605900838056222
		Associate Data Scientist	0.6298117054445297
		Associate Data Scientist (Blended)	0.5354793142448234
		Data Scientist	0.5088814547868130
		Data Annotator Junior	0.5040896621959168
5	Interior design	Interior Work Executor	0.7679034473209891
		Interior Work Executor	0.7494844302421693
		Design Interior	0.571526296302452
		Interior Design Expert	0.5423890643770504
		Product Design Dasar	0.3607420557897158

Based on these results, it can be seen which training programs are relevant and irrelevant to be recommended based on their interest data. The results of the Precision calculation based on the suitability of the recommendation can be seen in Table 12.

Table 12. Precision calculation results

Need and Interest	Relevant (TP)	Irrelevant (FP)	Total Recommendation	Precision
Sewing clothes	5	0	5	1.0
Video editing and design	5	0	5	1.0
Studying foreign language	3	2	5	0.6
Data processing and data analysis	5	0	5	1.0
Interior design	4	1	5	0.8

Table 13. Average precision calculation results

Total Precision	Average Precision	Precision Percentage
4.4	0.88	88%

In this test, the average Precision result is 0.88 or equivalent to 88%, which is obtained from the total number of precision values from the five samples divided by the number of samples used. The highest Precision value of the interest data tested is 1.0 from the interests of "Sewing clothes", "Video editing and design", and "Data processing and data analysis", where the five recommendations from each of these interests are included in True Positive because they provide relevant recommendations. Meanwhile the lowest Precision value obtained is 0.6 from the interest "Studying foreign language", which in this interest obtained three recommendations that are included in True Positive and two recommendations that are included in False Positive because they provide recommendations that are not relevant.

The application of the Content-Based Filtering method using the Cosine Similarity algorithm by maximizing Text Preprocessing and TF-IDF Weighting used to recommend training programs has an average precision value of 88%, which is higher when compared to the precision value obtained in previous studies with different methods. The online course recommendation system using the Deep Convolutional Neural Network method with Negative Sequence Mining has an average precision value of 39% [18]. Whereas the book recommendation system using Collaborative Filtering with Alternating Least Square Method has a precision value of 0.059% and Collaborative Filtering with Singular Value Decomposition method has a precision value of 57% [20].

4. CONCLUSION

In conclusion, this research has successfully optimized a recommendation system using the Content-Based Filtering method with similarity measurement using the Cosine Similarity algorithm applied to the training program selection process that provides recommendations according to the interests and needs of prospective trainees. The results of this research show that this method can achieve an average precision value of 88%. The higher precision value compared to previous research suggests the system's ability to provide optimal training program recommendations by providing highly relevant recommendations to the interests and needs of prospective trainees.

REFERENCES

[1] S. Lin and C. Hsu, "A Study of Impact on—Job Trading on Job Performance of Employees in Catering Industry," *Int. J. Organ. Innov.*, vol. 9, 2017.

[2] N. Gibran and D. Ramadani, "The Effect of Training and Career Development on Employee Performance," *Almana J. Manaj. dan Bisnis*, vol. 5, no. 3, pp. 407–415, Dec. 2021, doi: 10.36555/almana.v5i3.1680.

[3] O. Sunardi, M. Widyarini, and J. H. Tjakraatmadja, "The Impact of Sales Forces Training Program to Employees Behaviour

Styles (A Quasi-experimental Case Study In a Medium Sized Enterprise)," *Procedia Econ. Financ.*, vol. 4, pp. 264–273, 2012, doi: 10.1016/S2212-5671(12)00341-3.
[4] R. Muzayanah and E. A. Tama, "Application of the Greedy Algorithm to Maximize Advantages of Cutting Steel Bars in the Factory Construction," *J. Student Res. Explor.*, vol. 1, no. 1, pp. 41–50, Dec. 2022, doi: 10.52465/josre.v1i1.112.
[5] J. Son and S. B. Kim, "Content-based filtering for recommendation systems using multiattribute networks," *Expert Syst. Appl.*, vol. 89, pp. 404–412, Dec. 2017, doi: 10.1016/j.eswa.2017.08.008.
[6] Y. Afoudi, M. Lazaar, and M. Al Achhab, "Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network," *Simul. Model. Pract. Theory*, vol. 113, p. 102375, Dec. 2021, doi: 10.1016/j.simpat.2021.102375.
[7] S. H. Nallamala, U. R. Bajjuri, S. Anandarao, D. D. D. Prasad, and D. P. Mishra, "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 981, no. 2, p. 022008, Dec. 2020, doi: 10.1088/1757-899X/981/2/022008.
[8] A. Tabassum and D. R. R. Patil, "A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:235211496
[9] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J. Big Data*, vol. 8, no. 1, p. 26, Dec. 2021, doi: 10.1186/s40537-021-00413-1.
[10] . G. K., "USAGE OF REGULAR EXPRESSIONS IN NLP," *Int. J. Res. Eng. Technol.*, vol. 03, no. 01, pp. 168–174, Jan. 2014, doi: 10.15623/ijret.2014.0301026.
[11] J. Kaur and P. Buttar, "A Systematic Review on Stopword Removal Algorithms," vol. 4, pp. 207–210, Apr. 2018.
[12] H. Dwiharyono and S. Suyanto, "Stemming for Better Indonesian Text-to-Phoneme," *Ampersand*, vol. 9, p. 100083, 2022, doi: 10.1016/j.amper.2022.100083.
[13] R. Friedman, "Tokenization in the Theory of Knowledge," *Encyclopedia*, vol. 3, no. 1, pp. 380–386, Mar. 2023, doi: 10.3390/encyclopedia3010024.
[14] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, p. 30, Dec. 2019, doi: 10.1186/s13673-019-0192-7.
[15] F. Alzami, E. D. Udayanti, D. P. Prabowo, and R. A. Megantara, "Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, pp. 235–242, Aug. 2020, doi: 10.22219/kinetik.v5i3.1066.
[16] A. Ridho Lubis, M. K. M. Nasution, O. Salim Sitompul, and E. Muisa Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, p. 976, May 2021, doi: 10.11591/ijeecs.v22.i2.pp976-984.
[17] D. Liu, X. Chen, and D. Peng, "Cosine Similarity Measure between Hybrid Intuitionistic Fuzzy Sets and Its Application in Medical Diagnosis," *Comput. Math. Methods Med.*, vol. 2018, pp. 1–7, Oct. 2018, doi: 10.1155/2018/3146873.
[18] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Inf. Sci. (Ny).*, vol. 307, pp. 39–52, Jun. 2015, doi: 10.1016/j.ins.2015.02.024.
[19] M. Gao, Y. Luo, and X. Hu, "Online Course Recommendation Using Deep Convolutional Neural Network with Negative Sequence Mining," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1–7, Aug. 2022, doi: 10.1155/2022/9054149.
[20] A. Kurniaji and R. C. N. Santi, "Implementasi Metode Content Based Filtering Pada Pemilihan Komik," *J. Inform.*, vol. 10, no. 2, pp. 109–117, Oct. 2023, doi: 10.31294/inf.v10i2.16113.
[21] H. A. Adyatma and Z. K. A. Baizal, "Book Recommender System Using Matrix Factorization with Alternating Least Square Method," *J. Inf. Syst. Res.*, vol. 4, no. 4, 2023, doi: 10.47065/josh.v4i4.3816.
[22] E. Y. Utomo, "Recommendation of Yogyakarta tourism based on simple additive weighting under fuzziness," *J. Soft Comput. Explor.*, vol. 2, no. 1, Mar. 2021, doi: 10.52465/joscex.v2i1.13.
[23] D. Wang, Y. Liang, D. Xu, X. Feng, and R. Guan, "A content-based recommender system for computer science publications," *Knowledge-Based Syst.*, vol. 157, pp. 1–9, Oct. 2018, doi: 10.1016/j.knosys.2018.05.001.
[24] A. Nurcahya and S. Supriyanto, "Content-based recommender system architecture for similar e-commerce products," *J. Inform.*, vol. 14, no. 3, p. 90, Sep. 2020, doi: 10.26555/jifo.v14i3.a18511.
[25] H. Taherdoost, "Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects," Aug. 2021.
[26] V. Singrodia, A. Mitra, and S. Paul, "A Review on Web Scrapping and its Applications," in *2019 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, Jan. 2019, pp. 1–6. doi: 10.1109/ICCCI.2019.8821809.
[27] Y. Fu and Y. Yu, "Research on Text Representation Method Based on Improved TF-IDF," *J. Phys. Conf. Ser.*, vol. 1486, no. 7, p. 072032, Apr. 2020, doi: 10.1088/1742-6596/1486/7/072032.
[28] N. Febriyanti, D. P. Rini, and O. Arsalan, "Text Similarity Detection Between Documents Using Case Based Reasoning Method with Cosine Similarity Measure (Case Study SIMNG LPPM Universitas Sriwijaya)," *Sriwij. J. Informatics Appl.*, vol. 3, no. 2, Aug. 2022, doi: 10.36706/sjia.v3i2.47.
[29] Ylber Januzaj and Artan Luma, "Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 12, pp. 258–268, Jun. 2022, doi: 10.3991/ijet.v17i12.30375.
[30] N. Hu, "Application of Top-N Rule-based Optimal Recommendation System for Language Education Content based on Parallel Computing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.01406110.
[31] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, IEEE, Jul. 2020, pp. 237–242. doi: 10.1109/IWSSIP48289.2020.9145130.
[32] X. Li, D. Bian, J. Yu, M. Li, and D. Zhao, "Using machine learning models to improve stroke risk level classification methods of China national stroke screening," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 261, Dec. 2019, doi: 10.1186/s12911-019-0998-2.
[33] S. Seo *et al.*, "Predicting Successes and Failures of Clinical Trials With Outer Product–Based Convolutional Neural Network," *Front. Pharmacol.*, vol. 12, Jun. 2021, doi: 10.3389/fphar.2021.670670.

.