Published *online* on the journal's web page: http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

A Comparative Analysis of Content-Based Filtering and TF-IDF Approaches for Enhancing Sports Recommendation Systems

Herimanto[1], Kevin Samosir[2], Fastoria Ginting[3].

[123]*Department of Informatics, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Laguboti, Indonesia*

[1]*pardzheri@gmail.com,* [2]*samosirkevin873@gmail.com,* [3]*riaginting03@gmail.com.*

ARTICLE INFORMATION

Article History:
Received: August 17, 2024
Last Revision: September 26, 2024
Published *Online*: September 30, 2024

KEYWORDS

Recommender System,
Content-based Filtering,
TF-IDF,
Cosine Similarity

CORRESPONDENCE

Phone: 085265124310

E-mail: pardzheri@gmail.com

ABSTRACT

Sport is an important factor in maintaining or improving one's health. People exercise for various reasons, but many still struggle to identify the type of exercise that best suits their preferences. Recommendation systems have become an integral part of modern life, offering relevant suggestions to users. The aim of this research is to develop a sports recommendation system that provides accurate exercise recommendations to users and explains how the system functions in addressing both cold-start and non-cold-start problems. The method used in this research is content-based filtering, utilizing a Term Frequency-Inverse Document Frequency (TF-IDF) vectorization matrix and a cosine similarity algorithm. When a new user logs in, the system first checks their preferences to determine whether a cold-start or non-cold-start problem occurs. If a cold-start problem arises, TF-IDF is used to generate recommendations. Conversely, in non-cold-start situations, cosine similarity is applied. The results demonstrate that using TF-IDF and cosine similarity, the system successfully provides relevant sports recommendations to users in both cold-start and non-cold-start scenarios. The optimal results from the data splitting experiment showed a Precision of 0.256, a Recall of 0.307, and an Accuracy of 0.869. The novelty of this research lies in its ability to convey an understanding of sports to users through sports-related journals, which enhances user satisfaction, trust, compliance, and provides education on engaging in sports activities.

1. INTRODUCTION

Sports, whether physical or mental, play a key role in maintaining and improving health [1]. While a balanced diet and proper sleep are essential, regular exercise is one of the most effective ways to boost individual well-being. Today, a variety of platforms offer services to help people incorporate sports into their routine's gyms with personal trainers, YouTube exercise tutorials, yoga classes, and more. However, with the wide range of sports available, finding the right activity that suits personal preferences can be challenging [2], [3]. This is where a recommendation system can be invaluable, helping individuals discover the ideal sport for their needs [4], [5].

Recommendation systems have become essential tools for personalizing user experiences, offering tailored suggestions based on individual preferences [5]. These systems filter information to present only what is most relevant to the user, ensuring a more efficient and enjoyable experience [6]. Commonly used methods in recommendation systems include content-based filtering, collaborative filtering, and hybrid filtering [7], [8]. However, like any system, recommendation methods face certain challenges, including cold-start problems, scalability issues, over-specialization, changing user preferences, changing data, and so on [9], [10], [11].

A major challenge in recommendation systems is the cold-start problem, which arises when the system lacks sufficient data on a new user's preferences [12], [13]. Without previous interactions, it becomes difficult to make relevant recommendations. To tackle this, content-based filtering specifically, techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and cosine similarity can be employed to match users with suitable sports, even in the absence of prior data [14], [15], [16]. By using these methods, a sports recommendation system can guide individuals toward the types of sports that are most aligned with their interests, helping them get started and stay engaged with activities that promote their health and well-being [15], [16], [17].

This research aims to develop a sports recommendation system that helps individuals find the type of sport that best suits their personal preferences, even when user data is limited or unavailable. The system will address the major challenge of the "cold-start" problem by employing content-based filtering methods, such as TF-IDF and cosine similarity, to suggest relevant sports based on user characteristics. The research will also explore ways to tackle other challenges, such as changing user preferences and scalability issues, ultimately creating a more effective system for promoting health and well-being through sports engagement.

2. RELATED WORK

Previous research conducted by Faisal Ramadhan and Aina Musdholifah outlines the process of creating a course and syllabus-based online video learning recommendation system to assist students in their learning activities. The research method involved collecting real-time learning video data using the YouTube API, and the recommendation system employed was content-based filtering. The study focused on learning video data and curriculum data, which were processed to generate an annotated list of recommendations for users. To evaluate the recommendations, a survey was conducted with 40 Computer Science students from UGM. This research is relevant to the study being conducted, as both involve the construction of a content-based filtering recommendation system using TF-IDF and cosine similarity [8].

Additionally, a study by Arif Huda et al. presents the development of a recommendation system for news articles. This research also employed content-based filtering, utilizing the TF-IDF (Term Frequency-Inverse Document Frequency) and cosine similarity algorithms to recommend articles based on user profiles and article features. The study's subjects were active students in the Informatics program at Amikom University Yogyakarta, and the dataset consisted of news articles news portal of the university. The relevance of this research lies in its use of the same recommendation system methodology, applying TF-IDF and cosine similarity. Furthermore, the evaluation in this study used a recall matrix, which is also likely to be used in the current research [18].

A limitation of both studies is their inability to effectively address cold-start problems. When users provide limited information or preferences, the systems struggle to generate accurate and relevant recommendations. Building on the insights from these studies, the current research aims to address the limitations related to cold-start problems by enhancing the content-based filtering approach with additional techniques, such as user profiling and the inclusion of demographic and contextual information. While the prior research by Faisal Ramadhan and Aina Musdholifah as well as Arif Huda et al. successfully utilized TF-IDF and cosine similarity for recommendation systems, they did not fully resolve the issue of providing relevant suggestions for new or inactive users. By integrating more diverse data sources, such as user behavioral patterns and preferences derived from questionnaires or initial onboarding processes, the proposed sports recommendation system seeks to improve its ability to make accurate recommendations even in the absence of significant user data. This approach will also ensure a more dynamic adaptation to user preferences over time, making the system more robust and personalized.

3. METHODOLOGY

Based on previous research, content-based filtering is a powerful method for building personalized and transparent recommendation systems. This is because it provides highly personalized recommendations by focusing solely on the individual user's preferences, based on items they have previously consumed. Additionally, content-based filtering is relatively simple and easy to understand, making its implementation easier across various platforms. However, this method also has weaknesses, particularly in dealing with cold-start problems. Therefore, this research aims to combine content-based filtering with the TF-IDF method and cosine similarity to develop a personalized sports recommendation system for improved user.

3.1 Research Stages

The research stages are a very important part to show a series of sequential steps that can help ensure the research runs from start to finish. The research stages are shown in Figure 1 below:

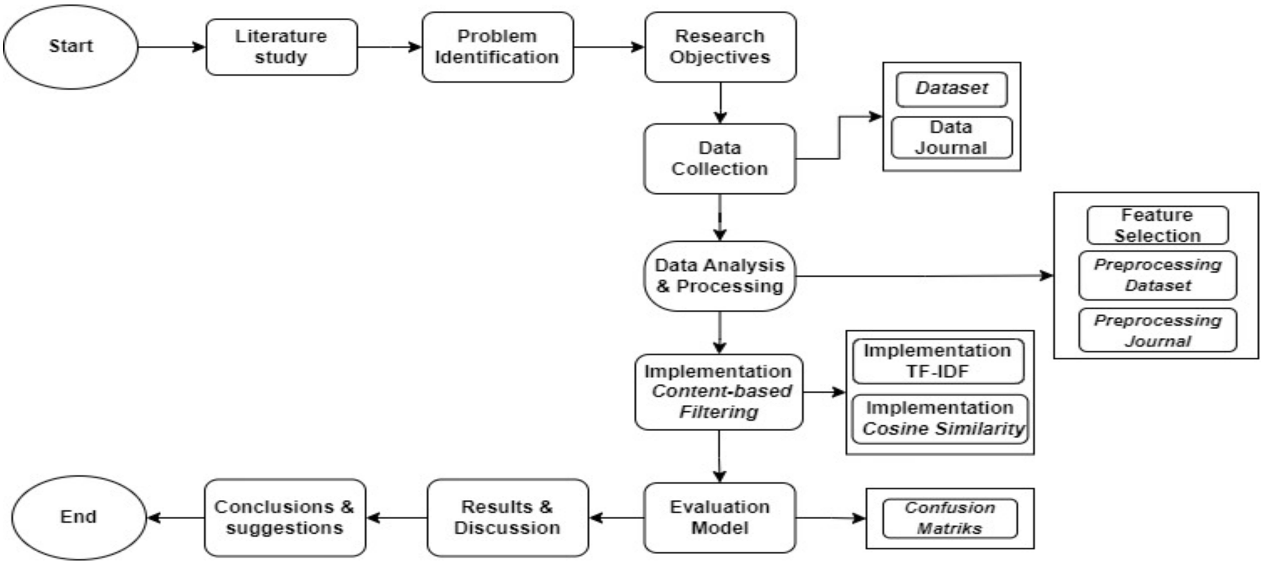


FIGURE 1. RESEARCH STAGES

This stage begins with the literature study stage, where researchers review previous research to understand existing problems and find gaps that can be filled by this research. After that, the researcher identifies the problem to determine the specific problem to be solved. Once the problem was identified, the researcher set clear and specific research objectives. The next stage is data collection which involves collecting datasets and data from sports-related journals. The collected data then went through a feature selection stage, which included dataset preprocessing and journal preprocessing to ensure the data was in a suitable format for further analysis. The next stage is data analysis and processing, where the data is analyzed and processed in preparation for implementation. Once the data is ready, the implementation of content-based filtering is done with two main approaches: TF-IDF implementation and cosine similarity implementation. TF-IDF is used to measure the importance of words in a particular document in the context of the dataset, while cosine similarity is used to measure the similarity between documents.

After the implementation is complete, model evaluation is conducted to assess the performance of recommendation accuracy when non-cold-start problem. In this evaluation, confusion matrix is used to calculate accuracy. The results of the implementation and evaluation are then presented in the results and discussion stage, where the research findings are discussed in detail. The study concludes with conclusions and suggestions that summarize the research results and provide recommendations for future research.

3.2 Content-based Filtering

Content-based filtering is a relatively common approach in the field of Information Retrieval. The basis of this method lies in the attributes, characteristics, or features possessed by an item. This method is user independent which it works on selecting items based on the correlation between item content and user preferences as opposed to a filtering system. If a user has done a certain type of sport then the system will try to recommend sports with similar preferences available in the dataset that may match the user's preferences [19].

3.3 Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF in this research is used to build item profiles. TF (Term Frequency) is used to determine the frequency of occurrence of words in a document. If the word frequency is high, the word is considered important and can be used to build item profiles. While IDF (Inverse Document Frequency) serves to normalize the number of occurrences of words throughout the document. IDF is the inverse of the document frequency in the entire corpus. TF-IDF weighting is used to eliminate the effects of frequently occurring words so that the importance of a feature can be determined more accurately. TF-IDF is considered to be the most appropriate method for this research due to its ability to highlight important and unique words in documents, while reducing the impact of uninformative common words [16], [20], [21].

In the context of a content-based sports recommendation system, important information such as user preferences contained in the dataset (such as motivation for doing sports) can be given higher weight in the vector representation using TF-IDF, so that the recommendations generated become more personalized and relevant to each user. The frequency of a word in a document is divided by the max of all terms in the document [14], [16].

$$TF(i,j) = \frac{freq(i,j)}{max_k f_{kj}} \qquad \text{Equation 2.1}$$

$TF(i,j) = Term\ Frequency$ of word i in document j
$freq(i,j) =$ Frequency of occurrence of i in document j
$max_k f_{kj} =$ Number of words in document.

The Term Frequency of feature i in document j is the number of times feature i appears in document j, divided by the maximum number of times feature i appears in the entire corpus.

$$IDF(i) = log \frac{N}{n(i)} \qquad \textit{Equation 2.2}$$

The number of items in the database divided by the number of words contained by the documents in the database. n(i) is the number of documents containing feature (word) i, and N is the total number of documents. If the feature appears more frequently, n(i) will be larger, and the larger n(i), the smaller the IDFi value.

$$TF - IDF(i,j) = TF(i,j) * IDF(i) \qquad \textit{Equation 2.3}$$

3.4 Cosine Similarity

Cosine similarity is a method to calculate the similarity between two vectors by finding the cosine of the angle. In this case, cosine similarity can help determine how similar or different two documents or texts are based on the words contained in them, by calculating the cosine angle between the two vectors [22]. In general, the similarity function is a function that accepts two objects in the form of real numbers (0 and 1) and returns the similarity value between the two objects, also in the form of real numbers. In addition, cosine similarity ignores scale and only considers the direction of the vector so it is suitable for measuring the similarity between documents with many features (words in the text) [14]. A value close to 1 indicates that it has a strong relationship between two variables [23]. Whereas a value close to 0 indicates that there is no correlation (independent variables).

Here is the formula for cosine similarity, namely:

$$\begin{aligned} A.B &= \left[A_{1,}A_{2,\dots}A_n\right].B_1\ B_2\ B_n \\ &= A_1 B_2 + A_1 B_2 + \cdots + A_n B_n \\ &= \textstyle\sum_{i=1}^{n} A_i B_i \end{aligned}$$

$$Similarity = \cos 0 = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ x\ \sqrt{\sum_{i=1}^{n} B_i^2}} \qquad \textit{Equation 2.4}$$

Desc:

$\Sigma\ = used\ to\ add\ up\ many\ things.$
$i\ = index\ used\ to\ keep\ track\ of\ the\ word$
$we\ are\ working\ on$
$n\ = the\ number\ of\ words\ in\ the\ sentence$
$A\ dan\ B\ = refer\ to\ the\ two\ sentences\ available.$

3.5 Dataset

The dataset consists of 709 users' personal data containing name, age, gender, description of level of exercise, time spent exercising per week and per day, dietary responses, and exercise goals or motivations. User preference data that will be processed and used as data to provide recommendations to users through a vectorization process.

3.6 Journal Data

Journal data was collected from Google Scholar with the help of Search Engine Result Page (SERP API). Information from the journals will then be entered into the database, in the form of sport types, journal links and journal summaries. The journal summary is taken from the journal snippet section available in the journal search results on Google Scholar, the fetch API will be used as a tool to retrieve the journal summary. All data collected through a data preprocessing process before it will be used in the implementation stage. Next, the experiment process is shown in Figure 4 below:

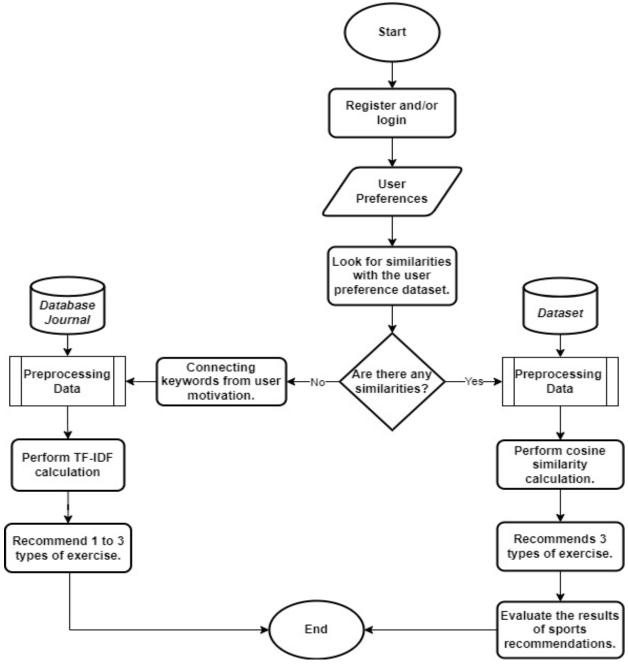

Start
Register and/or login
User Preferences
Look for similarities with the user preference dataset.
Database Journal
Dataset
Preprocessing Data
Connecting keywords from user motivation.
No
Are there any similarities?
Yes
Preprocessing Data
Perform TF-IDF calculation
Perform cosine similarity calculation.
Recommend 1 to 3 types of exercise.
Recommends 3 types of exercise.
End
Evaluate the results of sports recommendations.

FIGURE 2. EXPERIMENT FLOWCHART

The process begins with the user registering and/or logging into the system. After login, the system receives the user's preferences regarding his/her motivation to exercise. Next, the system looks for similarities in the new user's preferences with previous users in the dataset. If there are no similarities, the data will be further processed by linking keywords from the user's motivation to the scientific journal database. The data then undergoes preprocessing, where the keywords from the user's motivation are processed to perform TF-IDF calculations. The results of the TF-IDF calculation will be used to recommend 1 to 3 types of exercise that match the user's preferences, in which case the following case belongs to the cold-start problem condition.

However, if there are similarities in user preferences with the dataset, the system will preprocess the available data and then perform a cosine similarity calculation. Based on the results of the cosine similarity calculation, the system will recommend 3 types of sports to the user. The last step is to evaluate the results of sports recommendations provided by the system. In evaluating this sports recommendation system, the important thing to measure is how accurate the system is in providing recommendations to users. The following is an explanation of the matrix that will be used in the following recommendation system:

a. *Precision*

Precision measures the effectiveness of recommendation systems by calculating the proportion of relevant items (true positives) among those suggested, with false positives being irrelevant items [24], [25].

$$Precision = \frac{TP}{(TP + FP)} \qquad \text{Equation 2.5}$$

b. *Recall*

Recall is used as a measure of relevant documents generated by the system. False negatives are all relevant items that are not generated by the system. Calculations made to measure the suitability and success of finding more information [24].

$$Recall = \frac{TP}{(TP + FN)} \qquad \text{Equation 2.6}$$

c. *Accuracy*

Accuracy in recommendation systems reflects how well the recommendations align with user preferences and needs, indicating user satisfaction [24]. Mathematically, accuracy can be calculated as follows:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \qquad \text{Equation 2.7}$$

As such, the system aims to provide relevant exercise recommendations that match the user's preferences, both in the presence of cold-start problems and non-cold-start problems[24], [25].

4. RESULT AND DISCUSSION

In this study, we used 3 types of data splitting, the aim of which was to see which value produced the most optimal performance based on the confusion matrix.

4.1 Experiment Splitting Data

The first experiment of splitting data carried out was data splitting with a value of 70-30, the results of which can be seen through the confusion matrix in Figure 3 below:

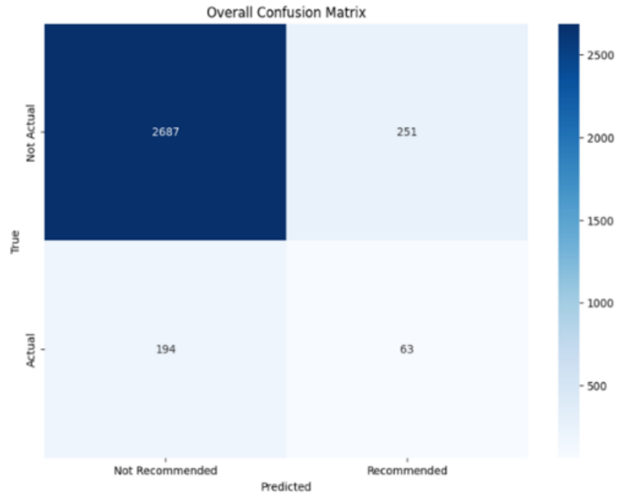


FIGURE 3. EXPERIMENT SPLITTING DATA I

Figure 3 above is the result of splitting the data into 70% and 30%. The values of TP, FP, FN, TN are 63, 251, 194, and 2687, respectively. By using the formula in the previously existing equation, the value is obtained:

a) $Precision = \frac{TP}{FP+TP}$

$$Precision = \frac{63}{(63+251)}$$

$$Precision = 0{,}201$$

b) $Recall = \frac{TP}{FN+TP}$

$$Recall = \frac{63}{(63+194)}$$

$$Recall = 0{,}245$$

c) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

$$Accuracy = \frac{63+2687}{(63+2687+251+194)}$$

$$Accuracy = 0{,}861$$

The next experiment used the value 80-20, the results of which are shown in Figure 4.

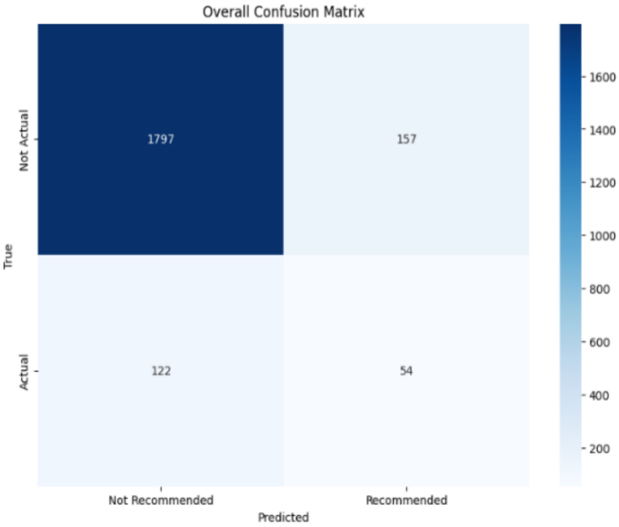


FIGURE 4. EXPERIMENT SPLITTING DATA II

Figure 4 above is the result of splitting the data into 80% and 20%. The values of TP, FP, FN, TN are 54, 157, 122, and 1797 respectively. By using the formula in the previously existing equation, the value is obtained:

a) $Precision = \frac{TP}{FP+TP}$

$$Precision = \frac{54}{(54+157)}$$

$$Precision = 0{,}256$$

b) $Recall = \frac{TP}{FN+TP}$

$$Recall = \frac{54}{(54+122)}$$

$$Recall = 0{,}307$$

c) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

$$Accuracy = \frac{54+1797}{(54+1797+157+122)}$$

$$Accuracy = 0{,}869$$

While the last experiment used a value of 90-10, the results of which are shown in Figure 5 below:

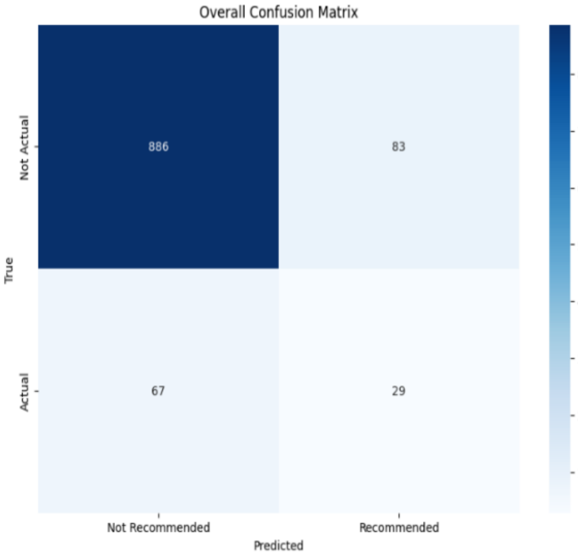


FIGURE 5. EXPERIMENT SPLITTING DATA III

Figure 5 above is the result of splitting the data into 90% and 10%. The values of TP, FP, FN, TN are 29, 83, 67, and 886 respectively. By using the formula in the pre-existing equation, we get the value of:

a) $Precision = \frac{TP}{FP+TP}$

$$Precision = \frac{29}{(29+83)}$$

$$Precision = 0{,}259$$

b) $Recall = \frac{TP}{FN+TP}$

$$Recall = \frac{29}{(29+67)}$$

$$Recall = 0{,}302$$

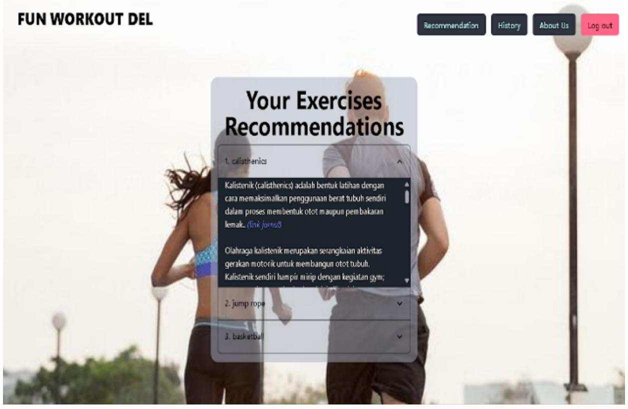
c) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

$$Accuracy = \frac{29+886}{(29+886+83+67)}$$

$$Accuracy = 0{,}859$$

4.2 Experiment Recommendation Result

Figure 6 is the website display when there is no cold-start problem, where the system can utilize all user preferences to provide more precise and personalized recommendations.



Your Exercises Recommendations

1. calisthenics

Kalistenik (calisthenics) adalah bentuk latihan dengan cara memaksimalkan penggunaan berat tubuh sendiri dalam proses membentuk otot maupun pembakaran lemak. (link jurnal)

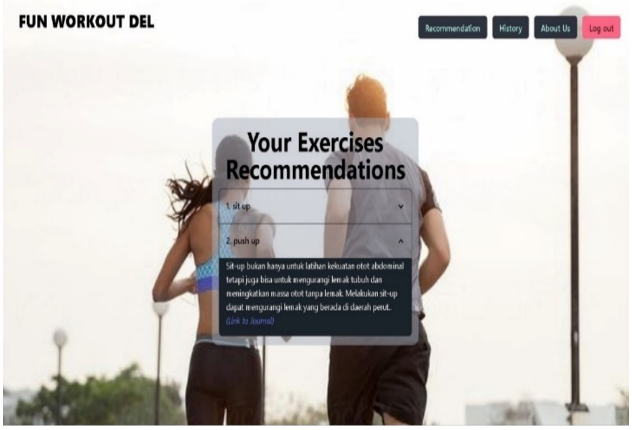
Olahraga kalistenik merupakan serangkaian aktivitas gerakan motorik untuk membangun otot tubuh. Kalistenik sendiri hampir mirip dengan kegiatan gym;

2. jump rope

3. basketball

FIGURE 6. DISPLAY WHEN COLD-START PROBLEM OCCURS

The user preferences include various aspects such as gender, age range, health level, frequency of exercise, duration of exercise, response to diet, and motivation to exercise. In this case, the user has preferences in the age range of 31 to 39 years old, medium/average/good enough health level, exercise 3 to 4 times a week, exercise duration of 30 minutes, not always on a diet, and have the goal of exercising to reduce weight. Based on these preferences, the system provides appropriate exercise recommendations, as shown in Figure 6. In non-cold-start situations, the system can recommend three types of exercises most suited to the user, along with multiple relevant journals offering detailed information about each exercise. This allows the system to provide a richer experience, helping users achieve their fitness goals more effectively.



Your Exercises Recommendations

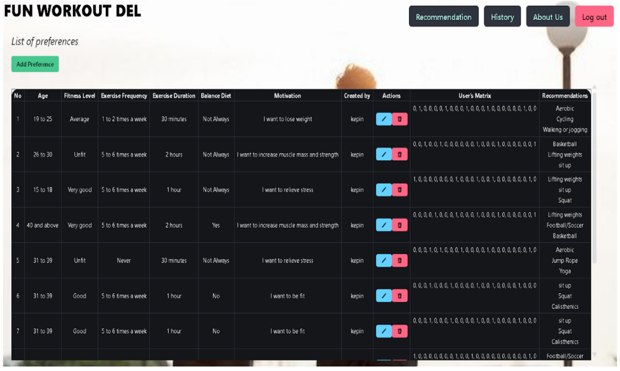
1. sit up

2. push up

Sit-up bukan hanya untuk latihan kekuatan otot abdominal tetapi juga bisa untuk mengurangi lemak tubuh dan meningkatkan massa otot tanpa lemak. Melakukan sit-up dapat mengurangi lemak yang berada di daerah perut.. *(Link to Journal)*

FIGURE 7. DISPLAY WHEN NON-COLD-START PROBLEM

Figure 7 shows the website display during a cold-start problem, where exercise recommendations are based solely on the user's main purpose, without considering other factors like gender, age, health, exercise habits, or diet. Each motivation generates unique keywords, which are used to find relevant exercises from the system's list. For example, a user with a goal preference of 'I want to lose weight' will get exercise recommendations as shown in Figure 7. In a cold-start problem situation, the number of recommended exercises ranges from one to three types of exercise. Each recommendation is accompanied by a summary and links to relevant journals. The number of recommended sports can be less than three if the highest Term Frequency-Inverse Document Frequency (TF-IDF) value obtained leads to the same sport, thus reducing the diversity of recommended sports.

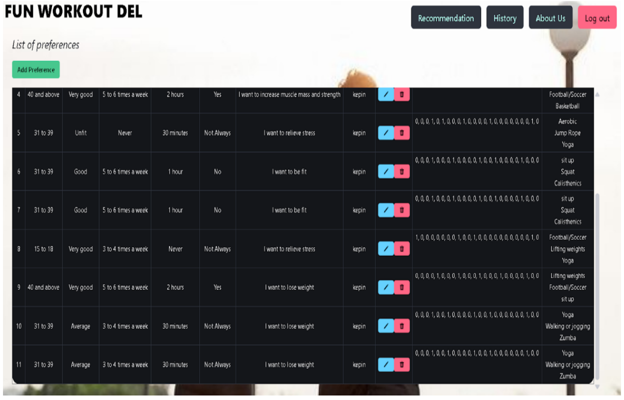


List of preferences

No	Age	Fitness Level	Exercise Frequency	Exercise Duration	Balance Diet	Motivation	Created by	Actions	User's Matrix	Recommendations
1	19 to 25	Average	1 to 2 times a week	30 minutes	Not Always	I want to lose weight	kepin	✏ 🗑	0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0	Aerobic Cycling Walking or jogging
2	26 to 30	Unfit	5 to 6 times a week	2 hours	Not Always	I want to increase muscle mass and strength	kepin	✏ 🗑	0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1	Basketball Lifting weights sit up
3	15 to 18	Very good	5 to 6 times a week	1 hour	Not Always	I want to relieve stress	kepin	✏ 🗑	1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0	Lifting weights sit up Squat
4	40 and above	Very good	5 to 6 times a week	2 hours	Yes	I want to increase muscle mass and strength	kepin	✏ 🗑	0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1	Lifting weights Football/Soccer Basketball
5	31 to 39	Unfit	Never	30 minutes	Not Always	I want to relieve stress	kepin	✏ 🗑	0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0	Aerobic Jump Rope Yoga
6	31 to 39	Good	5 to 6 times a week	1 hour	No	I want to be fit	kepin	✏ 🗑	0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0	sit up Squat Calisthenics
7	31 to 39	Good	5 to 6 times a week	1 hour	No	I want to be fit	kepin	✏ 🗑	0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0	sit up Squat Calisthenics
									1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0	Football/Soccer

FIGURE 8. RECOMMENDATION RESULT WHEN PREFERENCES DIFFER

Figure 8 above is a display when users enter different preferences. When a user enters different preference data, the user will get different recommendations. This can happen because different preferences can produce different levels of similarity, resulting in different recommendations.



List of preferences

4	40 and above	Very good	5 to 6 times a week	2 hours	Yes	I want to increase muscle mass and strength	kepin	✏ 🗑		Football/Soccer Basketball
5	31 to 39	Unfit	Never	30 minutes	Not Always	I want to relieve stress	kepin	✏ 🗑	0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0	Aerobic Jump Rope Yoga
6	31 to 39	Good	5 to 6 times a week	1 hour	No	I want to be fit	kepin	✏ 🗑	0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0	sit up Squat Calisthenics
7	31 to 39	Good	5 to 6 times a week	1 hour	No	I want to be fit	kepin	✏ 🗑	0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0	sit up Squat Calisthenics
8	15 to 18	Very good	3 to 4 times a week	Never	Not Always	I want to relieve stress	kepin	✏ 🗑	1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0	Football/Soccer Lifting weights Yoga
9	40 and above	Very good	5 to 6 times a week	2 hours	Yes	I want to lose weight	kepin	✏ 🗑	0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0	Lifting weights Football/Soccer sit up
10	31 to 39	Average	3 to 4 times a week	30 minutes	Not Always	I want to lose weight	kepin	✏ 🗑	0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0	Yoga Walking or jogging Zumba
11	31 to 39	Average	3 to 4 times a week	30 minutes	Not Always	I want to lose weight	kepin	✏ 🗑	0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0	Yoga Walking or jogging Zumba

FIGURE 9. RECOMMENDATION RESULT WHEN PREFERENCES ARE SAME

Figure 9 above is a display when users enter the same or similar preferences. Another thing is when users enter the same preference data, for example preferences in numbers 8 and 9, then the user will get the same recommendation. This can happen because the user's previous and current preferences are the same, so the similarity value will be the same and produce the same exercise.

In the recommendation experiment, the recommendation results are different when there is a cold-start problem and when there is no cold-start problem. This difference occurs because during the cold-start problem, the system only provides recommendations based on user motivation which is converted into keywords and matches them with journals related to these keywords. Whereas during the non-cold-start problem, the system uses all user preferences including their motivations to provide recommendations. In the non-cold-start problem, the system calculates the cosine similarity between the current user and the existing user data in the dataset. In contrast, during the cold-start problem, the system uses TF-IDF calculation between keywords and journal summary data.

5. CONCLUSIONS

This research succeeded in implementing the content-based filtering method in a sports recommendation system, producing a web-based application that provides accurate sports recommendation results to users with an accuracy of 86.90%. The performance of the content-based filtering method by applying the TF-IDF vectorization matrix can handle the cold-start problem by providing 3 types of sports with the top calculated values. The weakness and limitation of this research is that, although it addresses the cold-start issue, the method used is TF-IDF, where the user's motivations are matched with summaries of sports journal documents that have been collected. This approach is not entirely accurate, thus further studies and collaboration with fitness and health experts are needed to achieve better results.

REFERENCES

[1] M. R. Abdullah, “The Importance of Sports for Public Health: The Importance of Sports for Public

Health," *IJHMS*, vol. 1, no. 4, pp. 25–28, Sep. 2023, doi: 10.46336/ijhms.v1i4.20.

[2] N. Draper, C. Williams, and H. Marshall, *Exercise Physiology: for Health and Sports Performance*, 2nd ed. London: Routledge, 2024. doi: 10.4324/9781003109280.

[3] F. B. Moghaddam and M. Elahi, "Cold Start Solutions For Recommendation Systems," 2019, doi: 10.13140/RG.2.2.27407.02725.

[4] M. H. A. Hassan *et al.*, Eds., *Proceedings of the 8th International Conference on Movement, Health and Exercise: MoHE 2022—Refocusing on Sports and Exercise for a Post-pandemic World*. in Lecture Notes in Bioengineering. Singapore: Springer Nature Singapore, 2023. doi: 10.1007/978-981-99-2162-1.

[5] A. Majumder, J. L. Sarkar, and A. Majumder, Eds., *Artificial Intelligence and Data Science in Recommendation System: Current Trends, Technologies and Applications*. BENTHAM SCIENCE PUBLISHERS, 2023. doi: 10.2174/97898151367461230101.

[6] S. K. Raghuwanshi and R. K. Pateriya, "Collaborative Filtering Techniques in Recommendation Systems," in *Data, Engineering and Applications*, R. K. Shukla, J. Agrawal, S. Sharma, and G. Singh Tomer, Eds., Singapore: Springer Singapore, 2019, pp. 11–21. doi: 10.1007/978-981-13-6347-4_2.

[7] F. Mansur, V. Patel, and M. Patel, "A review on recommender systems," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore: IEEE, Mar. 2017, pp. 1–6. doi: 10.1109/ICIIECS.2017.8276182.

[8] F. Ramadhan and A. Musdholifah, "Online Learning Video Recommendation System Based on Course and Sylabus Using Content-Based Filtering," *Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 3, p. 265, Jul. 2021, doi: 10.22146/ijccs.65623.

[9] M. M. Talha, H. U. Khan, S. Iqbal, M. Alghobiri, T. Iqbal, and M. Fayyaz, "Deep learning in news recommender systems: A comprehensive survey, challenges and future trends," *Neurocomputing*, vol. 562, p. 126881, Dec. 2023, doi: 10.1016/j.neucom.2023.126881.

[10] N. Mishra, S. Chaturvedi, A. Vij, and S. Tripathi, "Research Problems in Recommender systems," *J. Phys.: Conf. Ser.*, vol. 1717, no. 1, p. 012002, Jan. 2021, doi: 10.1088/1742-6596/1717/1/012002.

[11] D. Jannach and G. Adomavicius, "Recommendations with a Purpose," in *Proceedings of the 10th ACM Conference on Recommender Systems*, Boston Massachusetts USA: ACM, Sep. 2016, pp. 7–10. doi: 10.1145/2959100.2959186.

[12] G. Sharma, K. Pragada, P. Deb Purkayastha, and Y. Vajpayee, "Research Paper on Exploring the Landscape of Recommendation Systems: A Comparative Analysis of Techniques and Approaches," *int. jour. eng. com. sci*, vol. 13, no. 06, pp. 26196–26218, Jun. 2024, doi: 10.18535/ijecs/v13i06.4827.

[13] H. Yuan and A. A. Hernandez, "User Cold Start Problem in Recommendation Systems: A Systematic Review," *IEEE Access*, vol. 11, pp. 136958–136977, 2023, doi: 10.1109/ACCESS.2023.3338705.

[14] R. Priskila, Nova Noor Kamala Sari, and Putu Bagus Adidyana Anugrah Putra, "IMPLEMENTASI CONTENT-BASED FILTERING MENGGUNAKAN TF-IDF AND COSINE SIMILARITY UNTUK SISTEM REKOMENDASI RESEP MASAKAN," *JTI*, vol. 18, no. 1, pp. 43–51, Jan. 2024, doi: 10.47111/jti.v18i1.12543.

[15] Ms. T. Sharad Phalle and Prof. S. Bhushan, "Content Based Filtering And Collaborative Filtering: A Comparative Study," *JAZ*, pp. 96–100, Mar. 2024, doi: 10.53555/jaz.v45iS4.4158.

[16] M. T. Mohammed and O. F. Rashid, "Document retrieval using term term frequency inverse sentence frequency weighting scheme," *IJEECS*, vol. 31, no. 3, p. 1478, Sep. 2023, doi: 10.11591/ijeecs.v31.i3.pp1478-1485.

[17] C. G. Reswara, J. Nicolas, I. M. D. Widyatama, D. David, and P. Arisaputra, "Book recommendation system using TF-IDF and cosine similarity," presented at the THE 1ST INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING, SYSTEMS, AND APPLICATIONS (InCASA) 2023, Bali, Indonesia, 2024, p. 020003. doi: 10.1063/5.0212477.

[18] A. A. Huda, R. Fajarudin, and A. Hadinegoro, "Sistem Rekomendasi Content-based Filtering Menggunakan TF-IDF Vector Similarity Untuk Rekomendasi Artikel Berita," *bits*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2511.

[19] M. S. Negara and A. Z. Mardiansyah, "Implementasi Machine Learning dengan Metode Collaborative Filtering dan Content-Based Filtering pada Aplikasi Mobile Travel (Bangkit Academy): Implementation of Machine Learning with Collaborative Filtering and Content-Based Filtering Methods in Mobile Travel Application (Bangkit Academy)," *JBegaTI*, vol. 5, no. 1, pp. 126–136, Mar. 2024, doi: 10.29303/jbegati.v5i1.1193.

[20] S. Chhipa, V. Berwal, T. Hirapure, and S. Banerjee, "Recipe Recommendation System Using TF-IDF," *ITM Web Conf.*, vol. 44, p. 02006, 2022, doi: 10.1051/itmconf/20224402006.

[21] A. H. J. P. Juni Permana and Agung Toto Wibowo, "Movie Recommendation System Based on Synopsis Using Content-Based Filtering with TF-IDF and Cosine Similarity," *ijoict*, vol. 9, no. 2, pp. 1–14, Dec. 2023, doi: 10.21108/ijoict.v9i2.747.

[22] A. Javed, F. Rehman, N. Sarfraz, H. Sharif, R. Khan, and A. M. Khan, "Movie Recommendation System with Sentimental Analysis Using Cosine Similarity Technique," in *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, Karachi, Pakistan: IEEE, Dec. 2022, pp. 1–8. doi: 10.1109/ICONICS56716.2022.10100512.

[23] M. S. Reddy, P. T. R. Kumar, L. M. Siddarth, and R. Mothukuri, "Designing Recommendation System for Hotels Using Cosine Similarity Function," in *Soft*

Computing for Security Applications, vol. 1449, G. Ranganathan, Y. El Allioui, and S. Piramuthu, Eds., in Advances in Intelligent Systems and Computing, vol. 1449. , Singapore: Springer Nature Singapore, 2023, pp. 1–15. doi: 10.1007/978-981-99-3608-3_1.

[24] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, Nov. 2015, doi: 10.1016/j.eij.2015.06.005.

[25] M. Yin, J. Wortman Vaughan, and H. Wallach, “Understanding the Effect of Accuracy on Trust in Machine Learning Models,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–12. doi: 10.1145/3290605.3300509.

AUTHORS

Herimanto

He received his S.Kom. (Bachelor of Computer Science) degree in Computer Science from Universitas Sumatera Utara in 2017 and his M.Kom. (Master of Computer Science) degrees in Informatics from Universitas Utara in 2021. His academic pursuits focused on computer science and artificial intelligence, particularly object detection, computer vision, machine learning, and deep learning. He currently contributes his expertise as a lecturer and researcher at the Institut Teknologi Del, Toba, Indonesia.

Kevin Willys Nathaneil Samosir

He recently graduated with a Bachelor's degree in Computer Studies from Del Institute of Technology in 2024. Currently, he is specializing in web development, with a strong focus on frontend technologies, including frameworks, using tools like HTML, CSS, JavaScript, and modern frameworks.

Fastoria Ginting

She has just completed her Bachelor of Computer Studies in 2024 from Del Institute of Technology. Born in Porsea on January 03, 2002. She is currently looking for opportunities to continue her education to a higher level or work in the technology field.