MACHINE LEARNING JOURNAL ARTICLE RECOMMENDATION SYSTEM USING CONTENT BASED FILTERING

Afika Rianti[1)], Nuur Wachid Abdul Majid[2)], and Ahmad Fauzi[3)]
[1, 2, 3)]Kampus UPI di Purwakarta, Universitas Pendidikan Indonesia
Jl. Veteran No.8, Nagri Kaler, Kec. Purwakarta, Kabupaten Purwakarta, Jawa Barat, Indonesia
e-mail: afika@upi.edu[1)], nuurwachid@upi.edu[2)], ahmad.fauzi@upi.edu[3)]

ABSTRACT

Indonesia is a country that hasn't studied much about artificial intelligence. This has resulted in a small number of publications related to that field including areas within such as machine learning. For that reason, it caused difficulties in finding relevant journal articles. The purpose of this study is to know the performance of the Content Based Filtering method in providing machine learning journal article recommendations. The research procedure used is CRISP-DM with algorithms used are TF-IDF and Cosine Similarity. The dataset used consists of 100 machine learning journal articles. Based on the research that has been done, it's concluded that the performance of the Content Based Filtering method in providing machine learning journal article recommendations as measured using the precision evaluation matrix showed a score of 76%, which means the result is quite good. However, the model couldn't be used properly for some data due to the small number of datasets which affects the limited recommendations.

***Keywords**: Content Based Filtering, CRISP-DM, machine learning journal article, precision, recommendation system*

SISTEM REKOMENDASI ARTIKEL JURNAL MACHINE LEARNING MENGGUNAKAN CONTENT BASED FILTERING

Afika Rianti[1)], Nuur Wachid Abdul Majid[2)], and Ahmad Fauzi[3)]
[1, 2, 3)]Kampus UPI di Purwakarta, Universitas Pendidikan Indonesia
Jl. Veteran No.8, Nagri Kaler, Kec. Purwakarta, Kabupaten Purwakarta, Jawa Barat, Indonesia
e-mail: afika@upi.edu[1)], nuurwachid@upi.edu[2)], ahmad.fauzi@upi.edu[3)]

ABSTRAK

Indonesia merupakan negara yang belum banyak mempelajari tentang kecerdasan buatan. Hal ini menyebabkan sedikitnya jumlah publikasi terkait dengan bidang tersebut termasuk bidang-bidang di dalamnya seperti machine learning. Oleh karena itu, hal ini menyebabkan kesulitan dalam menemukan artikel jurnal yang relevan. Tujuan dari penelitian ini adalah untuk mengetahui performa metode Content Based Filtering dalam memberikan rekomendasi artikel jurnal machine learning. Prosedur penelitian yang digunakan adalah CRISP-DM dengan algoritma yang digunakan adalah TF-IDF dan Cosine Similarity. Dataset yang digunakan terdiri dari 100 artikel jurnal machine learning. Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa performa metode Content Based Filtering dalam memberikan rekomendasi artikel jurnal machine learning yang diukur menggunakan matriks evaluasi presisi menunjukkan nilai sebesar 76% yang berarti hasilnya cukup baik. Namun, model tersebut tidak dapat digunakan dengan baik untuk beberapa data karena jumlah dataset yang sedikit sehingga mempengaruhi rekomendasi yang terbatas.

***Kata Kunci**: Content Based Filtering, CRISP-DM, artikel jurnal machine learning, presisi, sistem rekomendasi*

I. INTRODUCTION

Artificial intelligence (AI) is one of the work fields that is needed in the industry. AI itself is the development and integration of the fields of electronics, computer science, and mathematics that are capable of doing work like humans [1]. This field continues to grow from time to time. Quoted from the Journal Risk and Financial Management, stated that the rapid development of AI has started to occur from 2012 until now [2]. Then from the data in the 2020 World Economic Forum (WEF) Future of Jobs Survey, it said that jobs related to AI are in second place based on the ranking of jobs needed in the industry [3]. Therefore, many countries are trying to study the AI field so that it can follow AI development by having sufficient human resources.

In Indonesia, AI field hasn't learned that much. It could be seen from the few study programs which focus on AI, especially in the undergraduate program study. Until now there are only two undergraduate study programs

http://dx.doi.org/10.12962/j24068535.v22i1.a1193

that focus on it. First is Teknik Robotika dan Kecerdasan Buatan study program at Universitas Airlangga (UNAIR) which was founded in 2019 [4]. Then there's Mekatronika dan Kecerdasan Buatan (MKB) study program at Universitas Pendidikan Indonesia (UPI) which was founded in 2021 [5]. This made MKB as the second undergraduate AI study program in Indonesia. The small number of AI study programs in Indonesia has had an impact on the number of publications including in the area within it, such as machine learning. Machine learning is one of the areas in the AI field that is related to solving problems by computers through the experience during training [6]. Publications in this area are more difficult to find than the field of AI itself because the area is more specific. For the publications, people can access it through journals which contain articles in it. In SINTA website, until now there's only one journal that focuses on machine learning articles, that's MALCOM: Indonesian Journal of Machine Learning and Computer Science [7]. As it's only one, it means that it will be harder for readers to find machine learning articles even when there are other journals which also include machine learning in the scope. Therefore, there's a need for a system which can be used to provide appropriate machine learning journal articles such as a recommendation system.

The definition of a recommendation system is a system that can study the experiences and opinions of the user's preference so that it can provide recommendations for the options that may be most relevant to the user's choices [8]. This system has been widely used in various fields such as e-commerce, health, social relations, industry, e-learning, music, the internet of things (IoT), food and nutrition information systems, and marketing [8]. In the area of machine learning, recommendation systems are grouped into semi-supervised learning. Semi-supervised learning is a machine learning method that combines supervised and unsupervised learning [9]. The four most popular methods for building recommendation systems are Collaborative Filtering (CF), Content Based Filtering (CBF), Demographic Filtering, and Hybrid Recommender Systems [10]. CF gives recommendations based on the history of previous ratings [11]. CBF gives recommendations based on user preference with the given category [12]. Demographic Filtering gives recommendations based on the setting like user profiles as it's specifically used for new users who have no history [13]. Then Hybrid Recommender System is used when we combine more than one recommendation method. As the problem above is related to user preferences, the appropriate method is Content Based Filtering. Based on the discussion, this research aims to know the performance of Content Based Filtering methods in providing machine learning journal article recommendations.

II. RELATED RESEARCH

In recent years, there have been several researches that also discussed recommendation systems using CBF method. This method is commonly used for user preferences by taking input from users and providing recommendations based on relevance [14]. This method also doesn't need a history rating so it just focused on the current input [15]. Consequently, it's widely used for recommendations in specific categories or lists of objects. However, there's no studies which discuss the recommendation system for machine learning journal articles. Despite the limited research with the same focus, the components used in general, such as algorithms and evaluation matrices, remain the same. The following are three research studies considered most relevant as they utilized the CBF method. These studies are presented in Table I.

TABLE I
RELEVANT RESEARCHES

Information	Research 1 [16]	Research 2 [17]	Research 3 [18]
Recommendation object	Scientific articles	Book	Book
Algorithm	K-Means Clustering and Cosine Similarity	Weighted Tree and Cosine Similarity	TF-IDF and Cosine Similarity
Evaluation Matrix	Precision and recall	Precision	Precision
Result	1. K-Means Clustering a. Precision: 68% b. Recall: 64% 2. Cosine Similarity a. Precision: 44% b. Recall: 64%	Precision: 88%	Precision: 85%

Table I showed three most relevant researches from different sources. The first research recommendation object

is scientific articles with features used are title, keyword, and journal scope. Then the second research recommendation object is books with features used are title, synopsis, and author. Last, the third research recommendation object is books with features used are book site, title, author, genre, description and book cover. In those three references, the recommended objects are not too different. The main problem encountered in the is the difficulty in finding relevant reading sources. Therefore, those researches utilized a recommendation system as a given solution using Content Based Filtering method. In those researches, they have similarities like the algorithm used is Cosine Similarity and precision as the evaluation matrix. The results of each research showed different scores. In the first research, the algorithms used separately, it compared both results using K-Means Clustering and Cosine Similarity, the result showed that K-Means Clustering has better performance. In the second and third research studies, researchers combined two algorithms into one process to generate recommendations. The result showed that it has good performance with precision scores of 88% and 85%. Based on these findings, it could be concluded that combining algorithms within the recommendation process enhances overall performance, making them suitable for integration with the CBF method. To get good results, one or more algorithms may be combined.

Besides combining algorithms, the CBF method itself can also be integrated with other methods, such as Collaborative Filtering (CF). One of the papers that experimented with a proposed recommendation system using CBF and CF had better results than pure CF and CBF [19]. This discovery supports the idea that using a mix of different recommendation methods in research leads to better performance. Additionally, recommendation systems can also be implemented using deep learning. However, it's important to note that deep learning requires a substantial amount of data, among other considerations [20].

In this research, the recommendation system opted for Content Based Filtering (CBF) due to the constraints of a small dataset. Then algorithms used are TF-IDF and Cosine Similarity with precision as the evaluation matrix. The reason for that is because Cosine Similarity is widely used for recommendation systems as it is used to count the relevance. Then for the TF-IDF which stands for Term Frequency Inverse Document Frequency is needed to calculate how important and how often a word appears [18]. It then will be used in the next algorithm using Cosine Similarity.

III. METHODOLOGY

A. Research Procedure

The research procedure used is the cross industry standard for data mining (CRISP-DM). CRISP-DM is a goal-oriented method through the use of data mining to explore information in data [21]. This method is known as de-facto or standard for data mining projects as this method is well structured [22]. The stages in this method are business understanding, data understanding, data preparation, modeling, evaluation, and deployment [23]. The schematic of the stages in the CRISP-DM method shown in Figure 1.

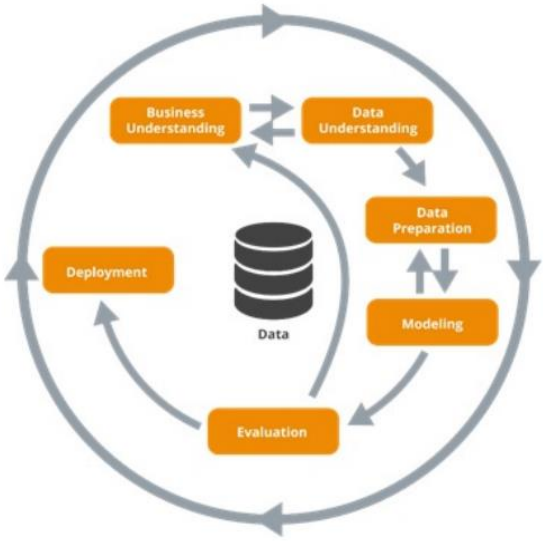

Fig. 1. CRISP-DM research procedure [23]

1) Business Understanding

1) This stage discusses what is needed by the business to solve the problem [23]. The main stages are to know the current situation and research the problem. After that, at this stage has to decide what the main goal needs to be achieved.

2) Data Understanding

2) This stage discusses the dataset used [23]. The research procedures at this stage are collecting data, describing data, verifying data quality, and exploring data. At this stage, coding has begun to process data up to exploratory data analysis (EDA). EDA is an approach to analyzing and making conclusions from data through the visualization of statistical charts to make it easier to understand data [24].

3) Data preparation

3) This stage discusses how the data is processed so that it can be used for modeling [23]. The research procedure at this stage focuses on data preparation through feature engineering. Feature engineering is an approach to prepare data by creating and processing features and data so that they are easier to use [25]. In this research, feature engineering is carried out by creating new features, removing unused features, sorting features, and filtering features for modeling.

4) Modelling

4) This stage discusses the selection of techniques used to build the model [23]. The specified technique can be in the form of approaches such as methods and algorithms. The model created later can then be used to provide recommendations.

5) Evaluation

5) This stage discusses measuring the accuracy of the model that has been made [23]. The results of these measurements are then evaluated further to find out whether the model created can answer the problems faced or not. In addition, at this stage conclusions are also drawn on the results obtained based on a review of the previous stages [26].

6) Deployment

6) Deployment is the final stage in the CRISP-DM method. This stage discusses the dissemination of the results that have been carried out in the previous stage which can be in the form of a final report or a software product [26].

B. Approach

In this study, the recommendation system was built using the Content Based Filtering (CBF) method. CBF is a recommendation system method based on items according to user individual preferences [27]. This method doesn't involve other users in determining recommendations because the choice is based on the user himself [28]. It will count the similarity of each category and represent it like distance on a vector [14]. Then this method will filter and choose items based on the relevance of the given keyword [14]. In this research, users will be asked to give input in the form of machine learning approaches. The system will then respond by generating recommendations in the form of relevant journal article titles that align with the specified machine learning approach given. The integration of this method with the CRISP-DM model process ensures a systematic and data-driven methodology, optimizing the generation of recommendations.

The algorithms used in this research are TF-IDF and Cosine Similarity.

1) TF-IDF

7) TF-IDF (Term Frequency Inverse Document Frequency) consists of two components: TF and IDF. TF represents the number of times words appear in a document, while IDF indicates the number of documents containing a specific word [29]. TF-IDF works by assigning weight to each word as a keyword. This algorithm helps calculate the significance and frequency of a word in a document. It assigns low weight to frequently occurring words and higher weight to those that appear rarely.

2) Cosine Similarity

8) This algorithm calculates the similarity of two documents so that it can help in providing relevant recommendations [30]. The result given will be a list of options that the user may like. By utilizing this algorithm, the algorithm assesses the angle between vector representations of documents, offering a measure of their similarity. In practical terms, a higher cosine similarity score indicates greater similarity between documents.

C. Dataset

In collecting journal article data for the dataset, the technique used is literature study. It includes searching,

studying, reading literature in the form of journal articles and books, and other sources relevant to the research to be carried out [31]. Journal article data is taken from Google Scholar which discusses machine learning and can be accessed freely. The period is the last ten years or from 2013 to 2023. The number of journal articles in the dataset is 100.

D. Tools

The tool used to process the data is Google Colab. Google Colab is a cloud Jupyter notebook which is used to program machine learning by writing source code [32]. This tool supports Python language. Python is a programming language that is easy to learn, interpret, and read [33]. This language will be used as the main language in data mining to produce a recommendation system model. In this process, several Python libraries are needed that can support it so that data can be processed properly. These libraries are Pandas, Numpy, Scikit-learn, and Matplotlib.

E. Evaluation

In conducting the evaluation, the evaluation metric used is precision. Precision is the number of relevant recommendation items [34]. The formula for knowing precision shown in (1).

$$\textbf{\textit{Precision}} = \frac{\textbf{\textit{True Positive}}}{\textbf{\textit{True Positive + False Positive}}} \textbf{\textit{x}}\ \textbf{100\%} \quad (1)$$

The values needed to determine precision are true positive (TP) and false positive (FP). TP is positive data that is predicted to be true. FP is negative data but predicted as positive data. The calculations here will be done by counting each TP and FP and then assigning it to the formula (1). As there's no class target for each data, then the data isn't separated into training data and testing data. However, the evaluation can be done by checking the recommendation whether it is relevant to the dataset or not.

IV. RESULT AND DISCUSSION

A. Business Understanding

The main problem in this research is related to the difficulty in finding AI journals. In Indonesia, this field hasn't been studied much. It could be seen from the small number of AI study programs in Indonesia, especially in the undergraduate study program. Based on data obtained from PDDikti, it is stated that currently there are only 5 AI study programs in Indonesia 2 of them are undergraduate and the rest are graduate level. Then, based on the results of a survey conducted on 20 respondents who had studied AI, the following results were obtained:

1) Difficulties in finding AI journal articles

9) Based on Figure 2, it could be seen that 90% of respondents have experienced difficulties in finding Indonesian AI journal articles. One of the reasons for this difficulty is the small number of AI journal articles published in Indonesia. This makes users have to look for other references such as looking for journal articles from abroad. When the respondents have experienced difficulties in finding Indonesian AI journal articles, then for sure it will be harder to find relevant journal articles in their sub area such as machine learning.

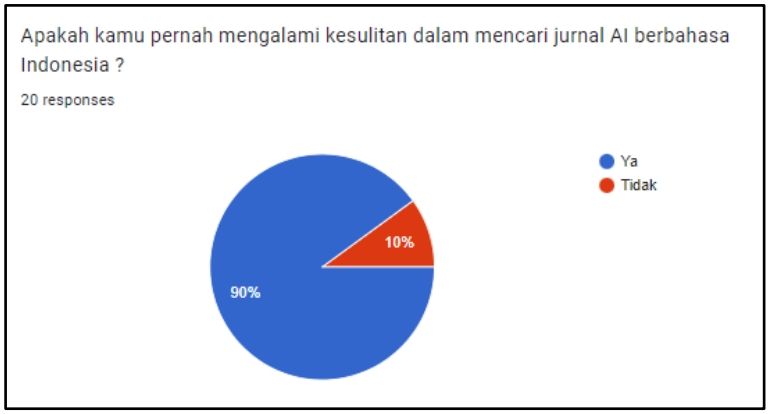

Fig. 2. Difficulties in searching for AI journal articles

2) Respondent opinion

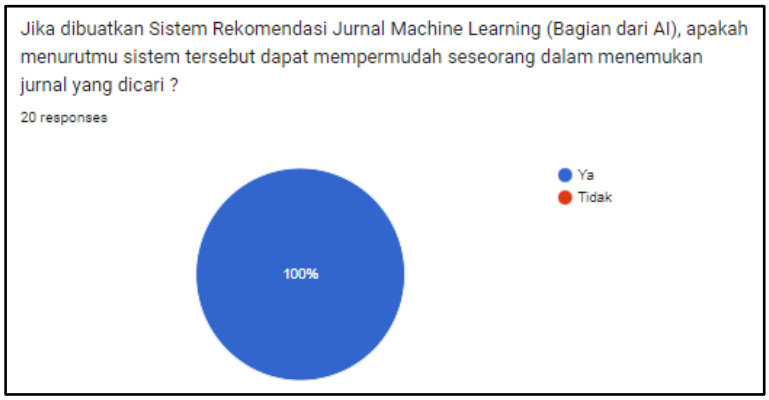

Fig. 3. Respondents opinion

10) Based on Figure 3, it could be concluded that respondents agree that the machine learning journal article recommendation system can help respondents find the journal articles they are looking for. The recommendation system here is grouped based on some features which will help to search for journal articles more specifically. For this reason, a machine learning recommendation system will be built to help users to find relevant journal articles.

B. Data Understanding

The total data in the dataset is 100 data which 50 journal articles using Indonesian and the remaining 50 using English. The machine learning issues raised are regression, prediction, classification, clustering, and recommendation systems. Then the sectors raised are economics, health, education, information technology, and socio-culture with each sector consisting of 20 journal articles. The features used are language, sector, title, author, year, method, algorithm, and journal article link. Each of those features is taken from its own journal articles, especially from metadata for the main information. Then for the features of sector, method, and algorithm are also used literature review with further analysis to make sure all the information is correct. The content of the dataset shown in Table II.

From the dataset, it has also been evaluated that there is no duplication of data. But for missing values, there is some data that has null values. After further investigation, most of the missing values are found in the method and algorithm features. This is because the recommendation system problems are usually differentiated based on the method first. Meanwhile, other problems are directly differentiated based on the algorithm. Apart from the description of the data and evaluation of the data, visualizations of the distribution of the data using EDA were also carried out. The data distribution based on the year feature shown in Figure 4.

TABLE II
DATASET

Language	Sector	Title	Author	Year	Method	Algorithm	Link
English	Economy	Prediction of House Price Using XGBoost Regres...	J.Avanija,Gurram Sunitha, K.Reddy Madhavi, Pad...	2021	NaN	XGBoost	https://www.turcomat.org/index.php/turkbilmat/...
English	Economy	Statistical Machine Learning Regression Models...	Yasser T. Matbouli, Suliman M. Alghamdi	2022	NaN	Multiple Linear Regression	https://www.mdpi.com/2078-2489/13/10/495
English	Education	Predictive Analysis of the Enrolment of Elemen...	Elizalde Lopez Piol, Luisito Lolong Lacatan, J...	2021	NaN	Linear Regression	https://www.researchgate.net/profile/Luisito-L...
English	Education	A Regression Model to Predict Key Performance ...	Ashraf Abdelhadi, Suhaila Zainudin, Nor Samsia...	2022	NaN	Linear Regression	https://www.researchgate.net/profile/Suhaila-Z...
English	Health	Classifying White Blood Cells Using Machine Le...	Abdullah Elen, M. Kamil Turan	2019	NaN	Multinomial Logistic Regression	https://dergipark.org.tr/tr/download/article-f...

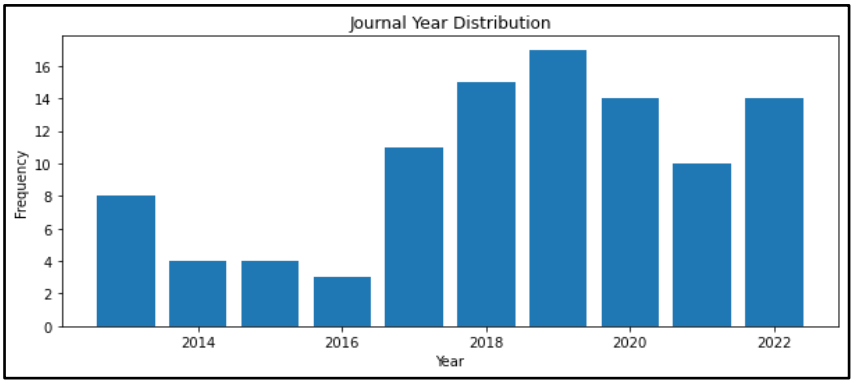

Fig. 4. Year distribution

Based on Figure 4, it could be seen that most of the journal articles were taken from 2019. The number of journal articles taken from 2019 was 17 journal articles. Then followed by journal articles from 2018 with 15 journal articles, and journal articles from 2020 and 2022 with a total of 14 journal articles. From Figure 4, it could also be seen that there are no journal articles originating from outside the specified year, that's from 2013 to 2023. Even so, there's no journal article from 2023. This happened because the data was taken around February 2023 which is still in the first second month of the year. However, it could be seen that most of the journal articles taken were from the last five years, so it could be said that the journal articles are still new and feasible to use.

C. Data Preparation

To prepare the data, several steps were carried out here. The first step was to create an index feature to assign an ID to each data starting from 1 to 100. After that, we then created approach feature. This feature is a combination of method and algorithm features. Because the method and algorithm features are no longer used, these two features were removed. After that, some writing changes were made using regex, such as changing spaces to underscores. Then the unused features are deleted and only the features that would be used for modeling are left. This aims to save time and memory so that it is faster in execution. The features used for modeling are index, title, and approach. Features title and approach will help users to get recommendations based on its approach which is usually mentioned on the title. The approach feature is specially made to make sure each journal article was classified into the correct method and algorithm so that it can help users to find the relevant journal articles easily. The final dataset information shown in Figure 5.

```
1   DF = data.filter(['Index', 'Title', 'Approach'])
2   DF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Index     100 non-null    int64
 1   Title     100 non-null    object
 2   Approach  100 non-null    object
dtypes: int64(1), object(2)
memory usage: 2.5+ KB
```

Fig. 5. Final dataset information

D. Modelling

As explained in the previous discussion, the modeling here used two algorithms, thay are TF-IDF and Cosine Similarity. The detail in those two algorithms are explained in the following discussion:

11) TF-IDF

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer()
tfidf.fit(DF['Approach'])

TfidfVectorizer()

tfidf_matrix = tfidf.fit_transform(DF['Approach'])
tfidf_matrix.shape

(100, 69)
```

Fig. 6. TF-IDF algorithm

12) To call the TF-IDF algorithm, only needs to use the Tfidfvectorizer which is a library from Scikit-learn. This library is used in the feature approach because this feature is used as a user choice. After entering these features into tfidf.fit, the next step is to do a fit_transform which aims to learn vocabulary. Then it could be seen that tfidf_matrix.shape showed the results (100, 69) which means the table consists of 100 rows and 69 columns. What

is meant by 100 rows is a dataset consisting of 100 data or 100 different journal articles. While 69 are all columns in the journal articles, which means there are 69 different combinations of approaches.

13) Cosine Similarity

```
cosine_sim = cosine_similarity(tfidf_matrix)
cosine_sim

array([[1.        , 0.54395248, 0.        , ..., 0.60547894, 0.
        0.        ],
       [0.54395248, 1.        , 0.        , ..., 0.        , 0.
```

Fig. 7. Cosine Similarity algorithm

14) After using the TF-IDF algorithm, the next step is to use the Cosine Similarity algorithm. Cosine Similarity used on the tfidf_matrix variable which is data from the approach features that have been processed using the Tfidfvectorizer. The results of Cosine Similarity are displayed in the form of an array.

15) To display recommendations, a function needs to be created that can take user input and display recommendations that are considered relevant. In this case, the user must enter input according to the existing approach in the dataset. The total number of recommendations given is limited to 5. This is because of the limited number of datasets owned. An example of the recommendation results shown in Figure 8.

```
get_recommendation(user_keyword)

Top 5 recommended journals based on K-Means approach are:

1) Prediksi Nilai Mata Kuliah Mahasiswa Menggunakan Algoritma K-Apriori
2) K-Means Algorithm for Clustering of Learners Performance Levels Using Machine Learning Techniques
3) Characterization Clustering of Educational Technologists Achievement in Higher Education Using Machine Learning Analysis
4) Covid-19 Pandemic Datasets Based on Machine Learning Clustering Algorithms: A Review
5) Applications of Clustering Techniques in Data Mining: A Comparative Study
```

Fig. 8. Recommendation result

E. Evaluation

```
Evaluation of the recommendation journal based on K-Means approach :

1) K-Apriori = False
2) K-Means = True
3) K-Means = True
4) K-Means = True
5) K-Means = True

Relevant Journal = 4/5

True Positive(TP) = 4
False Positive(FP) = 1

Precision = TP/(TP+FP)
Precision = 4/(4+1)
Precision = 4/5
Precision = 80%
```

Fig. 9. Evaluation result

The evaluation is carried out on the five most widely used approaches, that are K-Means, Naive Bayes, K-Nearest Neighbor, Multiple Linear Regression, and Linear Regression. Each approach generated five recommendations, with the system providing journal article titles as suggestions. Subsequently, an evaluation is conducted to determine the relevance of each recommendation, employing an evaluation matrix to quantify performance. This comprehensive analysis not only measures the effectiveness of individual recommendations but also offers insights into the overall performance of each approach. The results of the evaluation are shown in Table III.

TABLE III
EVALUATION RESULT

Approach	True Positive	Precision
K-Means	4	80%
Naive Bayes	4	80%
K-Nearest Neighbor	3	60%
Multiple Linear Regression	4	80%
Linear Regression	4	80%

From Table III, it could be seen that as many as four evaluations gave precision scores of 80% and the remaining one is 60%. The average is 76%. It showed that precision has quite good results. Even so, evaluation couldn't be carried out on several approaches because of the small number of datasets which made the system unable to provide enough recommendations. If testing is carried out on all approaches, it is possible to produce poor performance and even get errors.

F. Deployment

One of the literature reviews stated that 17 out of 24 studies didn't include the deployment stage [18]. The reason is that the number of false positives should be as little as possible, models with poor performance shouldn't be implemented and some say that deployment will be carried out later. Meanwhile, the other 7 studies still include the deployment stage with only 3 studies producing new products. While the rest only describe the results of the research and some aren't too technical. Based on that literature, the model here won't be implemented at the deployment stage because the model couldn't be used properly in some data. As the model couldn't be used at this stage, then at the deployment stage it only comes to making decisions and producing output in the form of a research article so that it can be shared to others.

V. CONCLUSION

Based on the results and discussion that has been presented, it could be concluded that the performance of the Content Based Filtering method using algorithms TF-IDF and Cosine Similarity in providing machine learning journal recommendations showed a score of 76%, which means the result is quite good. However, this performance only applied to a few widely used approaches. This happens because there are many variations of the approach without being accompanied by a large amount of data. For that reason, recommendations given are also limited which results in models that couldn't be used directly on the deployment. In future research, it is crucial for researchers to curate more comprehensive datasets, encompassing both a larger total number and diverse data groups per feature. Additionally, researchers can explore alternative machine learning approaches, such as hybrid recommender systems or delving into deep learning methods so that it can enhance the recommendation model's versatility and applicability.

REFERENCES

[1] Y. Devianto and S. Dwiasnati, "Kerangka Kerja Sistem Kecerdasan Buatan dalam Meningkatkan Kompetensi Sumber Daya Manusia Indonesia," *Jurnal Telekomunikasi dan Komputer*, vol. 10, no. 1, pp. 19–24, 2020. Accessed: Mar. 20, 2023. [Online]. Available: https://doi.org/10.22441/ incomtech.v10i1.7460

[2] R. S. Santos and L. Qin, "Risk Capital and Emerging Technologies: Innovation and Investment Patterns Based on Artificial Intelligence Patent Data Analysis," *Risk Financial Manage.*, vol. 12, no. 4, 2019. Accessed: Mar. 20, 2023. [Online]. Available: https://doi.org/ 10.3390/jrfm12040189

[3] R. I. Harbani. "3 Jurusan Ini Langka, tapi Bagaimana Prospek Kerjanya?" detikedu. https://www.detik.com/edu/perguruan-tinggi/d-5755063/3-jurusan-ini-langka-tapi-bagaimana-prospek-kerjanya (Accessed: Mar. 20, 2023)

[4] V. Natasha. "Jurusan Teknik Robotika dan Kecerdasan Buatan di Indonesia," myskill. https://blog.myskill.id/masa-kuliah/fakta-jurusan-robotika-kecerdasan-buatan/ (Accessed: Mar. 21, 2023).

[5] PDDikti. "Mekatronika dan Kecerdasan Buatan Kampus Purwakarta," https://pddikti.kemdikbud. go.id/data_prodi/MDVFODM1N0YtOUZEMi00NzIyLUE2QTctOTc4MTVERUYyRkJF/20211 (Accessed: Mar. 20, 2023)

[6] A. Rianti, S. Widodo, A. D. Ayuningtyas, and F. B. Hermawan, "Next Word Prediction Using LSTM," *Inf. Technol. Its Utilization*, vol. 5, no. 1, 2022. Accessed: Mar. 20, 2023. [Online]. Available: https://doi.org/10.30818/jitu.5.1.4748

[7] "SINTA - Science and Technology Index," https://sinta.kemdikbud.go.id/journals/index/?q=machine+learning (: Jan. 24, 2024)

[8] P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei and A. Darwesh, "A Systematic Study on the Recommender Systems in the E-Commerce," in IEEE Access, vol. 8, pp. 115694-115716, 2020, doi: 10.1109/ ACCESS.2020.3002803

[9] A. H. Khan, J. Siddqui, and S. S. Sohail, "A Survey of Recommender Systems Based on Semi-Supervised Learning," in Advances in Intelligent Systems and Computing, 2021, pp. 319–327. doi: 10.1007/978-981-16-3071-2_27. Available: https://doi.org/10.1007/978-981-16-3071-2_27

[10] S. K. Jaiswal and S. Agarwal, "Recommendation Systems: A Deep Survey for New Insights and Directions," *Dogo Rangsang*, vol. 12, no. 10, pp. 154–160, Oct. 2022. Accessed: Mar. 20, 2023. [Online]. Available: https://www.journal- dogorangsang.in/no_2_Online_22/45_oct.pdf
[11] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms," Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021), pp. 15–20, Jan. 2021, doi: 10.5220/0010727500003101
[12] C. Channarong, C. Paosirikul, S. Maneeroj and A. Takasu, "HybridBERT4Rec: A Hybrid (Content-Based Filtering and Collaborative Filtering) Recommender System Based on BERT," in IEEE Access, vol. 10, pp. 56193-56206, 2022, doi: 10.1109/ACCESS.2022.3177610
[13] A. Pramarta and Z. K. A. Baizal, "Hybrid Recommender System Using Singular Value Decomposition and Support Vector Machine in Bali Tourism," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 7, no. 2, pp. 408–418, May 2022, doi: 10.29100/jipi.v7i2.2770. Available: https://doi.org/10.29100/jipi.v7i2.2770
[14] P. Nastiti, "Penerapan Metode Content Based Filtering Dalam Implementasi Sistem Rekomendasi Tanaman Pangan", teknika, vol. 8, no. 1, pp. 1-10, Jun. 2019
[15] F. B. A. Larasati and H. Februariyanti, "Sistem Rekomendasi Produk EMINA Cosmetics dengan Menggunakan Metode Content-Based Filtering," MISI (Jurnal Manajemen Informatika dan Sistem Informasi), vol. 4, no. 1, p. 45, Jan. 2021, doi: 10.36595/misi.v4i1.250. Available: https://doi.org/10.36595/misi.v4i1.250
[16] R. Andriani, "Fitur Rekomendasi Artikel Ilmiah pada Open Journal System Menggunakan Content Based Filtering," Undergraduate Dissertation, Universitas Sebelas Maret, 2019
[17] M. Alkaff, H. Khatimi, and A. Eriadi, "Sistem Rekomendasi Buku pada Perpustakaan Daerah Provinsi Kalimantan Selatan Menggunakan Metode Content-Based Filtering", *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 193-202, Sep. 2020
[18] M. R. A. Zayyad, "Sistem Rekomendasi Buku Menggunakan Metode Content Based Filtering," Undergraduate Dissertation, Universitas Islam Indonesia, 2021
[19] S. Sharma, R. Vijay, and M. Malhotra, "Automatic Recommendation System based on Hybrid Filtering Algorithm," Education and Information Technologies, vol. 27, no. 2, pp. 1523–1538, Jul. 2021, doi: 10.1007/s10639-021-10643-8. Available: https://doi.org/10.1007/s10639-021-10643-8
[20] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep Learning Applications and Challenges in Big Data nalytics," Journal of Big Data, vol. 2, no. 1, Feb. 2015, doi: 10.1186/s40537-014-0007-7. Available: https://doi.org/10.1186/s40537-014-0007-7
[21] J. A. Solano, D. J. L. Cuesta, S. F. U. Ibanez, and J. R. Coronado-Hernandez, "Predictive Models Assessment on CRISP-DM Methodology for Students Performance in Colombia - Saber 11 Test," *Procedia Computer Science*, vol. 198, pp. 512–517, 2022. Accessed: Mar. 20, 2023. [Online]. Available: https://doi.org/10.1016/j.procs. 2021.12.278
[22] A. Rianti, N. W. A. majid, and A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," Jul. 25, 2023. Available: https://ojs.udb.ac.id/index.php/Senatib/article/view/3015
[23] A. Zernig, A. Pandeshwar, R. Kern, and M. Rauch, "Machine Learning and Automated Decision Making," in *SemI40 Project Prospective: Industry 4.0 Evolution Revolution*, Austria. SemI40 Consortium, 2019, pp. 58–75. Accessed: Mar. 20, 2023. [Online]. Available: https://www. researchgate.net/publication/337592264_A_SemI40_Project_ Prospective_-_Industry40_from_Evolution_to_Revolution
[24] N. T. M. Sagala and F. Y. Aryatama, "Exploratory Data Analysis (EDA): A Study of Olympic Medallist," *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 3, pp. 578–587, 2022. Accessed: Mar. 20, 2023. [Online]. Available: http://sistemasi. ftik.unisi.ac.id/index.php/stmsi/article/view/1857
[25] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Machine Learning and Feature Engineering-based Study Into Sarcasm and Irony Classification with Application to Cyberbullying Detection," *Inf. Process. & Manage.*, vol. 58, no. 4, 2021. Accessed: Mar. 20, 2023. [Online]. Available: https://www. sciencedirect.com /science/article/abs/pii/ S0306457321000984
[26] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. Accessed: Mar. 20, 2023. [Online]. Available: https://doi.org/10.1016/j.procs.2021.01.199
[27] C. A. Melyani, "Sistem Rekomendasi Hotel dengan Pendekatan Content Based Filtering," bachelor's thesis, Universitas Islam Indonesia, 2022. Accessed: Mar. 20, 2023. [Online]. Available: https://dspace.uii.ac.id/handle/ 123456789/39737
[28] M. Alkaff, H. Khatimi, and A. Eriadi, "Sistem Rekomendasi Buku pada Perpustakaan Daerah Provinsi Kalimantan Selatan Menggunakan Metode Content-Based Filtering", *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 193-202, Sep. 2020
[29] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor", *INSYST*, vol. 1, no. 1, pp. 43–49, Aug. 2019
[30] F. A. Nugroho, F. Septian, D. A. Pungkastyo, and J. Riyanto, "Penerapan Algoritma Cosine Similarity untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian Kepada Masyarakat," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, pp. 529–536, 2020. Accessed: Mar. 20, 2023. [Online]. Available: https://doi.org/10.32493/informatika. v5i4.7126
[31] A. L. Rihani, A. Maksum, and N. Nurhasanah, "Studi Literatur : Media Interaktif Terhadap Hasil Belajar Peserta Didik Kelas V Sekolah Dasar," *Jurnal Kajian Pendidikan dasar*, vol. 7, no. 2, pp. 123–131, 2022. Accessed: Mar. 20, 2023. [Online]. Available: https://journal.unismuh.ac.id/index.php/ jkpd/article/view/7702/5030
[32] M. Canesche, L. Bragança, O. P. V. Neto, J. A. Nacif and R. Ferreira, "Google Colab CAD4U: Hands-On Cloud Laboratories for Digital Design," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 2021, pp. 1-5, doi: 10.1109/ISCAS51556.2021.9401151
[33] J. Enterprise, *Python untuk Programmer Pemula*. Jakarta: PT Elex Media Komputindo, 2019. Accessed: Mar. 20, 2023.
[34] Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, and R. Kashef, "Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities," *Applied Sciences*, vol. 10, no. 21, p. 7748, Nov. 2020, doi: 10.3390/app10217748