According to World Health Organization (WHO), traditional medicine is the culmination of all the knowledge, abilities, and practices derived from the theories, beliefs, and experiences that are unique to various cultures and that are used to maintain health as well as to prevent, diagnose, treat, or improve physical and mental illness. recently classified traditional herbal therapy as comprised of medicinal techniques that have existed, frequently for hundreds of years, prior to the establishment of modern medicine. The lack of easily accessible information regarding the description and efficiency of traditional medicine makes it difficult for users to understand the benefits of each type of traditional medicine. Because of this, a recommendation system is needed that aims to facilitate users in finding traditional medicine that suit their preferences. This research proposes a traditional medicine recommendation system with the content-based filtering method using a combination of term frequency-invers document frequency and Word2Vec feature extraction. This method analyzes the traditional medicine description text and recommends based on word weights and semantic relationships between words. Results show optimal performance at dimensions 50–200 and window sizes 9–15 for the combination of term frequency-invers document frequency and Word2Vec, while term frequency-invers document frequency alone reaches 80% of accuracy and Word2Vec has lower performance (4–14%) across a wide range of parameter experiments. Based on optimal result above, this recommendation system can be applied to obtain information of traditional medicine that suitable with people needed by adjust the best model of dimensions and window size

Keywords: content-based filtering, dimension, feature extraction, recommendation system, semantic relationship, term frequency-invers document frequency (TF-IDF), traditional medicine, window size, Word2Vec, word weight

UDC 004.89:004.932.2:615
DOI: 10.15587/1729-4061.2025.338128

IMPLEMENTATION OF TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) AND WORD2VEC IN TRADITIONAL MEDICINE RECOMMENDATION SYSTEM BASED ON CONTENT-BASED FILTERING

Rika Yunitarini
Corresponding author
Doctor of Informatics Engineering
Department of Informatics Engineering*
E-mail: rika.yunitarini@trunojoyo.ac.id
Dwi Aqilah Pradita
Bachelor of Informatics Engineering
Department of Informatics Engineering*
Ernaning Widiaswanti
Doctor of Industrial Engineering
Department of Industrial Engineering*
*Trunojoyo University
Jl. Raya Telang, Po.Box 2 Kamal-Bangkalan,
East Java, Indonesia, 69161

Received 04.06.2025
Received in revised form 22.07.2025
Accepted 12.08.2025
Published 29.08.2025

How to Cite: *Yunitarini, R., Pradita, D. A., Widiaswanti, E. (2025). Implementation of term frequency-inverse document frequency (TF-IDF) and Word2Vec in traditional medicine recommendation system based on content-based filtering. Eastern-European Journal of Enterprise Technologies, 4 (2 (136)), 70–80. https://doi.org/10.15587/1729-4061.2025.338128*

1. Introduction

According to the World Health Organization, traditional medicine is the culmination of all the knowledge, abilities, and practices derived from the theories, beliefs, and experiences that are unique to various cultures and that are used to maintain health as well as to prevent, diagnose, treat, or improve physical and mental illness [1]. Traditional medicine is the term used to describe indigenous populations and developing nations unique health practices [2]. Patients employ traditional medicine for a variety of reasons. First of all, they can be in a far-off place where access to contemporary medical care is limited. Additionally, they might desire to try a traditional treatment after a modern one failed. Some people steer clear of modern healthcare facilities because they believe they are costly, risky, unwelcoming, or corrupt. The fact that many contemporary medications on the market are phony or counterfeit may also make patients anxious [3]. The World Health Organization (WHO) recently classified traditional herbal therapy as comprised of medicinal techniques that have existed, frequently for hundreds of years, prior to the establishment of modern medicine. The combination of generations of indigenous medical practitioners' therapeutic experiences is known as traditional herbal medicine [4]. The number of patients pursuing herbal and alternative therapies is rapidly increasing. Because they are more culturally acceptable, more compatible with the human body, and have less negative effects, traditional herbal medicines are becoming increasingly popular in the poor countries for primary healthcare, not just because they are less expensive [5]. Whether or not they can be explained, traditional herbal medicine, according to the World Health Organization, used to prevent, diagnose, treat, or improve physical and mental illness as well as to maintain health [1]. 90% of developing countries utilize

traditional herbs as their primary source of healthcare, and 88% of the world's population uses traditional herbal medicine, according to the WHO [6]. This is further supported by Indonesia's tropical climate, which contributes to its abundance of natural resources. Around 9.600 of the more than 30.000 plant species that occur in Indonesia are known to have pharmacological properties [7]. Traditional medicine is commonly used as a treatment in Indonesia [8]. Madura Island, which is part of Indonesia's East Java Province, is notable for its traditional knowledge of herbal components. An ethnobotanical study of medicinal plants can be used to investigate this indigenous knowledge about different medicinal compounds [9]. The rapid growth of traditional medicine among Indonesian has encouraged the emergence of various production and utilization of herbal medicine. Madura herbs has various types, ranging from herbs for women's health to herbs for married couples. Considering the various types of Madurese herbal medicine, it is necessary to manage Madura herbs data including aspects of location, ingredient, side effects [10].

Although, Indonesia has the potential to make herbal concoctions, there are several problems that can be identified. First, it is difficult for users who want to choose herbs that suit their health needs due to the many types of herbs and their diverse properties. Second, the lack of easily accessible information regarding the description and efficiency of traditional medicine makes it difficult for users to understand the benefits of each type of herb. Because of this, a recommendation system is needed that aims to facilitate users in finding herbs that suit their preferences. Previously, solutions existed to help users choose herbs by query searches on databases and manual guidance or consulting with traditional herbalists. However, these solutions are not always practical and can be time consuming. In the development of a recommendation system, there are two most commonly used methods, namely collaborative-filtering and content-based filtering. Recommendations on collaborative-filtering focus on items or products that other users have chosen as preferences in the past, similarity of tastes between users can be determined based on ratings, likes or user browsing history. Whereas, recommendations on content-based filtering determine a recommendation based on content that is being or has been accessed by an active user, similarities between products are compared based on features that the product has such as title, description and category [11]. The hybrid method is built by combining the advantages and minimizing the shortcomings of each method [12]. In the development of recommendation systems, problems are often found due to lack of information in the data, such as sparsity (lack of information available to determine the relationship in the data) and cold start (new data that has no interaction with other data) [13]. But in this study the data used is only traditional medicine data containing the attributes of the herbal medicine and there is no parameter data for comparison between user preferences, so the data used is not possible to use the collaborative filtering method. Therefore, this research uses the content-based filtering method to create a recommendation system. Based on the previous research, there are lack of investigation that concern to the traditional medicine combine with machine learning. Several studies have been explained that traditional medicine commonly used as treatment in several countries in the world because of have fewer adverse effects and affordable cost. This very relevant with the progress of research conducted about the traditional medicine. In order to provide suitable information about traditional medicine that required by people, so the existence of recommendation system is strongly needed.

Therefore, research on the development of the Traditional Medicine Recommendation System is strongly needed as solution option to provide more relevant recommendations based on text analysis.

2. Literature review and problem statement

The paper [14] utilizing content-based filtering in beauty product recommendations by displaying products that have a cosine similarity of more than 10%. The system is planned to be implemented in a website and will be tested through a black box test by the community. The test results show that people agree that this recommendation system helps in choosing skin care products. All this suggests that it is advisable to conduct a study on content-based filtering recommendation system. But there was unresolved issue related to another factor/attribute that doesn't consider, such as sensitivity of the skin. However, for more complex products, a recommendation system can't give relevant recommendation. A way to overcome these difficulties, it was needed justification from another expertise. The paper [15] utilizing content-based filtering for recommendations related to the potential of agricultural land that can produce food crop commodities. The test results on 10 merchant profiles with the 15 highest farmer group recommendations show an average precision rate of 78.40%. But it still needs another evaluation matrix model, like recall and accuracy so the recommendation result would be more relevant with the requirement.

The paper [16] using content-based filtering with a dataset that has complete features to make article recommendations that match user interests. The evaluation results show that the system has a Recall@5 of about 73% and Recall@10 of about 80%, indicating that the recommendations provided are quite relevant to the user's interests. Although the data is limited, this limitation can be overcome by adding, merging, or replacing data. Data replacement is easier because the system does not depend on the dataset used. There is potential for improvement in various aspects, such as article tokenization, to improve the performance of this recommendation system. The paper [17] utilizing content-based filtering and TF-IDF to weight attribute values. As the results of this study, the accuracy is 96.5% and it can be said to be in accordance with customer needs. But it's necessary to explore more complex data and another machine learning method that relevant with the result. The method used by researcher only TF-IDF without any machine learning method, and is simple. It needs to compare with another method and adding more records dataset. The paper [18] using content-based filtering with a cosine similarity approach to overcome the problem of buyer difficulties in finding suitable products in the NFT Marketplace. The results show that the machine learning model used can provide Top-N recommendations from products that are being searched by buyers. But there was unresolved issue related to apply exploratory data analysis (EDA) in more detail, as well as the use of larger datasets that can be obtained from the results of crawling data from the NFT marketplace. However, for more complex dataset/information, a recommendation system can't give relevant recommendation. A way to overcome these difficulties was an exploratory data analysis (EDA).

The paper [19] evaluates three Word2Vec parameters on sentiment analysis of hotel reviews in Indonesian, namely model architecture, evaluation method and vector dimension. The results show that the highest accuracy value is obtained

in the combination of parameters, namely the Skip-gram model architecture because it produces the highest accuracy for words that rarely appear, for the evaluation method, namely Hierarchical Softmax because it provides better results because during the training process it uses a binary tree model to represent all words in the vocabulary and leaf nodes represent rare words so that words that rarely appear will inherit vector representations in it, and using a vector dimension parameter of 100 because it gets the optimal accuracy value. But there was unresolved issue related to the vector dimensions that also affect the average value of accuracy. There is a way to overcome these difficulties and obtain the optimal value of accuracy, then it should increase the vector dimensions and amount of data simultaneously. This approach was used in [20], however increase the vector dimensions and amount of data simultaneously can reach optimal value of accuracy. All this suggests that it is advisable to conduct a study on research about traditional medicine recommendation system.

In paper [20], this research is intended to assess the impact of using Word2Vec in extracting features on the accuracy of sentiment analysis models on YouTube comments developed using the Random Forest method. Experimental results using 1, 5, and 20 epochs, as well as window size variations of 3, 5, and 10, show that the average accuracy of the model is in the range of 90.1% to 91%. However, when tested, the accuracy of the model only ranged from 88.77% to 89.05%. There was a small decrease in the model's accuracy rate during testing, although it was not significant. In addition, the experiments also indicated that the number of epochs and window size have an influence on the accuracy rate; the higher the epoch value and window size, the higher the accuracy rate of the model tends to be, although the increase is not so significant in this experiment. A way to overcome these difficulties, it was needed to see the effect of epoch and window on Word2Vec, it is recommended to use a large number of epochs and a larger window size so that the word context has a higher semantic level.

The paper [21] aims to detect document similarity by calculating the cosine similarity of the word vector. The similarity result using TF-IDF method is smaller than using Word2Vec, this is because TF-IDF cannot detect paraphrases. This unresolved issue can solve by apply Word2Vec method to analyze the semantic of word. This approach was used in [22], however by applying another method to find semantic of word can affect the similarity result. All this suggests that it is advisable to conduct a study on research about traditional medicine recommendation system. In paper [22], the purpose of this research is to improve library management by using the support vector machine (SVM) method in order to understand the patterns and characteristics of book titles. Tests were carried out with various feature extraction methods: TF-IDF only, Word2Vec only, and a combination of TF-IDF and Word2Vec. The most accurate book classification results were obtained through the combination of TF-IDF and Word2Vec. Therefore, the library book classification method using SVM can be applied because it provides the highest accuracy of 73% with a precision of 83%, recall of 72%, and F1-scores of 76%. But there was unresolved issue related to some errors in the classification results that are not appropriate due to the imbalance in the number of datasets in each classification class. A way to overcome these difficulties is adding some larger dataset with more and balanced categories. This is to determine and prove the effectiveness of the support vector machine algorithm with a combination of TF-IDF and Word2Vec feature extraction methods in classifying complex text.

The paper [23] aims to combine deep learning approaches with modifications to the TF-IDF algorithm and the use of Word2Vec as a word embedding layer to improve text classification accuracy. The results of this study show that the developed method successfully improves text classification accuracy and reduces the cost of manual feature extraction. Nevertheless, there are still possibilities for further improvement of the model. The possible directions for optimization are the probability distribution of the input window weights of the bidirectional LSTM could be further improved by using the laws of word distribution and knowledge of probability statistics to obtain a new probability distribution that better fits the human habit of reading text. From the analysis of literature, it follows that it still leaves several problems, including whether TF-IDF alone can provide the best information recommendation accuracy results or is it necessary to combined TF-IDF method with other feature extraction.

3. The aim and objectives of the study

The aim of the study is to develop traditional medicine recommendation system based on content-based filtering method by combining TF-IDF feature extraction with Word2Vec in order to improve the results.

To achieve this aim, the following objectives were accomplished:

– to preprocess data involves a series of steps to clean and transform raw text into a form more suitable for further analysis;

– to perform term frequency-inverse document frequency (TF-IDF) feature extraction on the traditional medicine dataset for calculating the weight of words in a document;

– to apply the word2vec method to the traditional medicine dataset for calculating the semantics between words in a document;

– to search the similarity between documents through feature similarity search using cosine similarity;

– to evaluate the model by using precision.

4. Materials and methods

4. 1. The object and hypothesis of the study

This study was taken the traditional medicine as object of study. Based on some of the problems mentioned earlier, the many types of traditional medicine require a recommendation system that can provide advice on the right traditional medicine and according to consumer complaints. In developing a recommendation system, an appropriate method is needed, so that the results are accurate. One of the methods proposed by the researcher is content-based filtering which combines TF-IDF and Word2Vec extraction features. The hypothesis of this research is that the combination of TF-IDF and Word2Vec can produce traditional medicine recommendations that suit user needs. Some of the assumptions that researchers use are related to the use of several attributes from the traditional herbal medicine dataset as well as setting the dimensions and window size on the word2vec extraction feature.

4. 2. Research method

In developing the traditional medicine recommendation system, the use of research methods is an important foundation. The research method acts as a foundation in designing

an effective and accurate system. With a good research method, it can ensure that this system will provide useful traditional medicine recommendations according to the user's health needs. The research method of this traditional medicine recommendation system, it is explained through Fig. 1. Fig. 1 consists of several stages which can later explain how the system was developed.

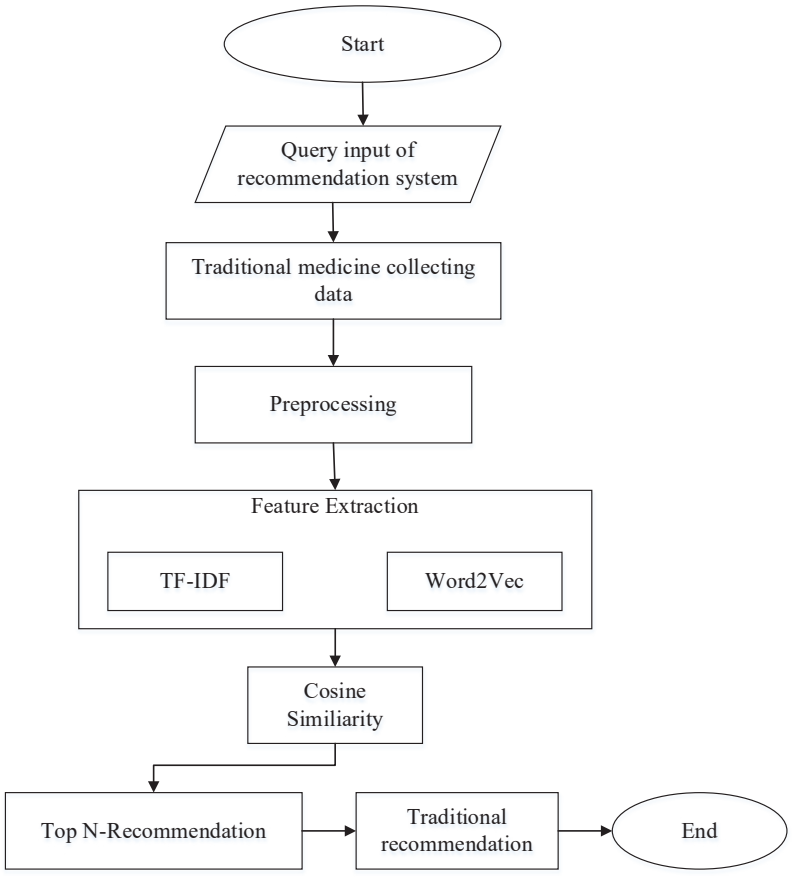

Fig. 1. Research method

Fig. 1 describes flowchart about traditional medicine recommendation system based on TF-IDF and Word2Vec. First, the researcher prepared a dataset of Madura traditional herbal medicine. Then the dataset is analyzed and processed at the preprocessing stage so that the data is structured and uniform. In this method there are several steps, the first is preprocessing, in the preprocessing stage there are steps in it, namely stopword removal, stemming, case folding and tokenizing. The second step is feature extraction in this study using TF-IDF feature extraction, Word2Vec, and a combination of TF-IDF Word2vec. After that, a similarity distance calculation is carried out using cosine similarity. Next, the value ranking is carried out by sorting the highest to lowest values and then the top-N recommendation value is taken. The resulting output is a recommendation result for Madurese herbal medicine in accordance with user needs, which is arranged based on the top-N recommendation value from the cosine similarity process.

4. 3. Dataset

This research used primary and secondary data as data collection techniques. Primary data collection was carried out by interviewing Madurese herbal medicine small and medium enterprises (SMEs). Each product is equipped with information such as properties and composition, which is then recorded or photographed. The information is then saved in Microsoft excel format along with a complete description of each product. Meanwhile, secondary data collection is carried out by taking data that already exists or has been previously collected from previous research. Indonesian traditional medicine dataset was described in Table 1. Table 1 presents the composition of Indonesia traditional medicine dataset that was used as resources data in order to process with several methods in traditional medicine recommendation system.

Table 1

Sample of Indonesian traditional medicine dataset

Traditional medicine	Efficiency	Ingredients
Jamu kecantikan	adolescent care, reduce body odor, and to overcome vaginal discharge	Betel leaf, turmeric, mangosteen skin, kunci, pinang muda
Jamu darah tinggi	lowers high blood pressure	Salam leaf, temu ireng, turmeric, garlic
Pegal linu	reduces fatigue, joint pain, and improve blood circulation	lempuyang, laos, chili, red ginger, kencur, pepper, coriander, sentok peok
Jamu galian montok	to increase appetite	curcuma, ginger
Pluntur	Improves blood flow, eliminates aches and pains, maintains body stamina, reduces birth rate	Kubeba, Galangal, Java pepper, ginger, white caraway

Table 1 was the example of dataset that consists of several attributes, namely the name of the herbal medicine, efficiency, and ingredients. This research used 206 dataset of traditional herbal medicine which various of efficiency and ingredients. For method computation example, it was used several data below. This research also used examples of keyword "pegal linu" that will be entered and checked for similarity values in existing data.

4. 4. Preprocessing

Preprocessing data is the process of processing raw data into more structured data. Preprocessing in this study uses natural language processing (NLP) to provide understanding and convenience to computers regarding human language. Before preprocessing the data, the herbal name attributes, efficiency attributes and content of the data are combined into 1 attribute called "metadata" then carry out the data cleaning process. Cleaning the data involves several steps, including case folding, tokenization, removing stop words, and stemming. In the case folding process, the text is changed from uppercase to lowercase. Tokenization involves breaking a document or sentence into a number of words with spaces as separators, while removing punctuation and special characters. Removal of stop words is done to remove irrelevant words,

such as 'from', 'me', 'and', etc. Meanwhile, stemming involves removing word affixes such as – an, me-, ke-, ber-, etc [24].

4. 5. Term frequency-inverse document frequency (TF-IDF)

Term frequency-inverse document frequency (TF-IDF) is a method of weighting each word by calculating the frequency of occurrence of words in each document and the frequency of occurrence of words in all documents [25]. TF-IDF consists of term frequency (TF) and inverse document frequency (IDF). Term frequency is a value that shows the frequency of terms that often appears in a document. The greater the number of occurrences of a term in the document, the greater its weight [19]. Inverse document frequency (IDF) aims to reduce the weight of the term if it exists in all documents. In contrast to TF, the less the frequency with which words appear in the document, the greater the value [26]. TF-IDF is done to change news titles in textual form to numeric ones so that they can be understood and processed by computers. The author first calculates the term frequency (TF), which is to calculate the frequency of occurrence of words in all news titles in the dataset. Previously, the writer calculated the frequency of occurrence of a word that had been divided into a basic form.

The formula of TF-IDF is

$$IDF = \log \frac{N}{df}, \tag{1}$$

N – the number of documents; df – the number of occurrences of the words in all documents

$$W = tf \times idf, \tag{2}$$

W – TF-IDF weight; tf – occurrence of the word in a document; idf – invers result.

4. 6. Word2Vec

This research uses the Word2Vec Skip-gram model to calculate the semantics between words in documents. The process of this model can be seen in the following steps:

1. Data preparation.

For example, a sentence in document 5 is used so that the data corpus used consists of 5 words/vocabs from the pre-processing process, namely jamu, gali, plump, add, lust, eat, temulawak, ginger. In this corpus, parameters $d = 10$ and $c = 2$ were used, after which one-hot vector coding of the corpus was performed. This encoding changes the target word to 1 and the other words to 0.

2. Forward pass.

Weighting in word2vec occurs in the hidden layer and consists of two weights, $W1$ and $W2$. In the standard word2vec configuration, both are initialized as random numbers according to the $m \times d$ number. However, in this calculation, $W1$ and $W2$ are initialized in the range between –1 to 1. The first step is to find the value of h or the hidden layer of a *wt*.

3. Softmax output.

The softmax value is calculated from the output layer to get y_pred.

4. Finding the error value.

Next, find the error value with the total amount of SUM of Different on wt or Σ*I*. For example, on the word "jamu", it has y_pred "dig" and "plump". Therefore, the word "jamu" will count the number of errors derived from y_pred "dig" and "plump". The error finding process itself involves a reduction between one-hot and softmax encoding.

5. Backpropagation.

In this step, a readjustment of each weight and bias is performed based on the error obtained. In this phase, the delta value of each W is identified. The delta for W2 is calculated by multiplying h and ΣI.

6. Update weight.

After getting the delta value for each weight, the next step is to update the value of each weight by subtracting the learning rate multiplied by the Delta Weight from the weight value. Here are the values of the weights associated with the word "jamu".

7. Combination of TF-IDF and Word2Vec.

After obtaining the feature extraction results from TF-IDF and Word2Vec, the next step is to combine the results of both by multiplying the TF-IDF and Word2Vec feature extraction result vectors with the formula in (3)

$$Combine_{(t,d)} = TF - IDF_{(t,d)} * Vdoc_{Word2Vec}. \tag{3}$$

The use of a combination of TF-IDF and Word2Vec feature extraction is done to combine the advantages of both feature extractions, so as to produce accurate classification. This combination can improve the semantic representation of documents or words by combining Word2Vec's ability to find contextual meaning with TF-IDF's advantage in understanding word uniqueness and providing word weights. TF-IDF will provide word weights, which will then be used to adjust the Word2Vec vector weights. This combination aims to improve the capabilities of the model and can increase the accuracy in recommending traditional medicine.

4. 7. Cosine similarity

Cosine similarity is a calculation method commonly used to measure the extent of similarity between items. This technique involves calculating the cosine value of the angle between two vectors and is generally used to assess the extent to which two documents are similar [16]

$$\cos\left(q,d_j\right) = \frac{\sum_{i=1}^{n} W_{i,q} W_{i,j}}{\sqrt{\sum_{i=1}^{n}\left(W_{i,q}\right)^2 \cdot \sum_{i=1}^{i}\left(W_{i,j}\right)^2}}, \tag{4}$$

$\cos(q, d_j)$ – cosine similarity between query dan document; $W_{i,q}$ – weight of document (i) query; $W_{i,j}$ – document (i) weight.

5. Results of traditional medicine recommendation system based on content-based filtering

5. 1. Preprocessing of traditional medicine dataset result

Before preprocessing the data, the herbal medicine name, efficiency and content attributes in the data are combined into 1 attribute named "metadata" then just do the data cleaning process, data cleaning involves several steps, including case folding, tokenization, removal of stopwords, and stemming. In the case folding process, the text is converted from capital letters to lowercase letters. Tokenization involves breaking a document or sentence into a number of words with spaces as separators, while removing punctuation and special characters. The result of preprocessing is shown in Table 2.

Table 2 shows the result of preprocessing data. Preprocessing data is the initial stage in text analysis. Text preprocessing involves a series of steps to clean and transform raw text into a form that is more suitable for further analysis.

Table 2

Preprocessing result

Traditional medicine	Metadata	Preprocessed metadata
Bom	Bom for vitality, and soreness, spice	Bom strong vitality soreness spice
sehat lelaki	Man health, increase stamina and health, strengthens and warms the body, soreness, muscle strain, stiff muscles, and languid	Man health increase stamina health strong warm body soreness muscle strain stiff muscle languid
pegal linu	Soreness reduce fatigue, joint pain and improve blood circulation lempuyang, laos, chili jamu red ginger, kencur, pepper, coriander, sentok peok	Soreness reduce fatigue joint pain improve circulation blood lempuyang laos chili jamu red ginger kencur pepper coriander sentok peok
Macho	Macho increase stamina and erection, increase libido, treating soreness, tired, and warm the body habbatus sauda', pinang muda, garlic lanang, temu kunci etc	Macho increase stamina erection level libido medicine soreness tired warm body habbatus sauda pinang muda garlic lanang temu kunci etc
Jamu Pegel Linu	Jamu Pegel Linu Overcoming itching and unpleasant odor in women	Jamu gel rheumatic upper itching unpleasant odor women
jamu galian montok	jamu galian montok to increase appetite curcuma, ginger	Jamu gali montok increase appetite curcuma ginger

5. 2. Term frequency-inverse document frequency feature extraction of traditional medicine dataset result

TF-IDF is one of the techniques used in natural language processing to measure the weight/importance of words in text. It is a method used to evaluate how important a word is in a document or a collection of documents (corpus). The result of TF-IDF shown in Table 3.

To calculate the weight of words in a document, TF-IDF feature extraction is required. The manual calculation process of TF-IDF can be seen in Table 3.

Table 3

Term frequency-inverse document frequency result

Term	TF*IDF						
	Q	D1	D2	D3	D4	D5	D6
1	2	3	4	5	6	7	8
Upper	0	0	0	0	0	0.48	0
Body	0	0	0.30	0	0.30	0	0
Odor	0	0	0	0	0	0.48	0
Garlic	0	0	0	0	0.48	0	0
Bom	0	0.48	0	0	0	0	0
Chili	0	0	0	0.48	0	0	0
Fatigue	0	0	0	0.30	0.60	0	0
Blood	0	0	0	0.48	0	0	0
Etc	0	0	0	0	0.48	0	0
Circulation	0	0	0	0.48	0	0	0
Erection	0	0	0	0	0.48	0	0
Gali	0	0	0	0	0	0	0.48
Itching	0	0	0	0	0	0.48	0
Gel	0	0	0	0	0	0.48	0
Habbatus	0	0	0	0	0.48	0	0
Warm	0	0	0.30	0	0.30	0	0
Vitality	0	0.48	0	0	0	0	0
Ginger	0	0	0	0.30	0	0	0.30
Jamu	0	0	0	0.18	0	0.18	0.18
Stiff	0	0	0.48	0	0	0	0
Kencur	0	0	0	0.48	0	0	0
Coriander	0	0	0	0.48	0	0	0
Strong	0	0.30	0.30	0	0	0	0
Kunci	0	0	0	0	0.48	0	0
Reduce	0	0	0	0.48	0	0	0
Lanang	0	0	0	0	0.48	0	0
Lancer	0	0	0	0.48	0	0	0
Laos	0	0	0	0.48	0	0	0
Man	0	0	0.48	0	0	0	0

Continuation of Table 3

1	2	3	4	5	6	7	8
Lempuyang	0	0	0	0.48	0	0	0
Fatigue	0	0	0.48	0	0	0	0
Libido	0	0	0	0	0.48	0	0
Rheumatic	0.08	0	0.08	0.08	0.08	0.08	0
Macho	0	0	0	0	0.48	0	0
Appetite	0	0	0	0	0	0	0.48
Red	0	0	0	0.48	0	0	0
Pepper	0	0	0	0.48	0	0	0
Montok	0	0	0	0	0	0	0.48
Young	0	0	0	0	0.48	0	0
Lust	0	0	0	0	0	0	0.48
Pain	0	0	0	0.48	0	0	0
medicine	0	0	0	0	0.48	0	0
Muscle	0	0	0.48	0	0	0	0
Soreness	0.08	0.16	0.08	0.08	0.08	0	0
Peok	0	0	0	0.48	0	0	0
Pinang	0	0	0	0	0.48	0	0
White	0	0	0	0	0.48	0	0
Spice	0	0.48	0	0	0	0	0
Sauda	0	0	0	0	0.48	0	0
Pleasant	0	0	0	0	0	0.48	0
Healthy	0	0	0.95	0	0	0	0
Vigor	0	0	0.48	0	0	0	0
Joints	0	0	0	0.48	0	0	0
Sentok	0	0	0	0.48	0	0	0
Stamina	0	0	0.30	0	0.30	0	0
Increase	0	0	0.18	0	0.18	0	0.18
Temu	0	0	0	0	0.48	0	0
temulawak	0	0	0	0	0	0	0.48
Level	0	0	0	0	0.48	0	0
Women	0	0	0	0	0	0.48	0
Is	0	0	0	0	0	0.48	0

Note: D – traditional medicine dataset; Q – query.

5. 3. Word2Vec feature extraction of traditional medicine dataset result

Word2vec utilizes a large set of text, called a corpus, as a training data source to form a vocabulary and create vectors that can have hundreds of dimensions. Each unique word in the corpus is represented as a vector and the vector formation process applies the Skip-gram model and the CBOW (continuous bag of words) model. The result of word2Vec is shown in Table 4.

Table 4

Word2VecResult

Weight update							
–0.70425	0.677003	–0.11995	0.66379	–0.08172	0.255043	–0.05486	0.11154
0.390087	0.351404	0.307852	0.04516	–0.95237	0.453549	–0.42915	0.592599
0.158364	0.717897	0.257353	0.309388	–0.37306	0.914384	0.33947	0.574183
0.542102	0.723593	0.897544	–0.58222	0.645778	0.444238	0.642808	0.903358
–0.745	–0.16429	–0.15225	–0.34383	0.589192	0.327859	0.185335	0.714099
0.094203	0.029825	0.419215	0.349836	–0.33641	–0.04477	0.754609	–0.43863
0.240624	0.334567	–0.54188	–0.3444	0.853223	0.052817	0.519159	–0.75858
0.458636	–0.26403	–0.65469	0.109977	0.754845	–0.53406	–0.50257	0.772816
0.325109	0.397538	–0.10736	0.276183	0.268481	–0.66999	–0.85531	0.401301
0.141069	0.215519	0.462433	0.45397	–0.44373	0.520173	0.100307	0.118228

In Table 4, after getting the W1 and W2 update values, the next step is to perform as many repetitions or epochs as necessary. Thus, the final value of the above training process is an $m \times d$ value matrix. Using the process, the word "jamu" will have a value vector of 0.613644, 0.24746, –0.45318, –0.40623, 0.508709, 0.400391, 0.158851, 0.403582, –0.23319, –0.12693.

5. 4. Cosine similarity of traditional medicine dataset result

Cosine similarity is a commonly used calculation method to measure the extent of similarity between items. This technique involves calculating the cosine value of the angle between two vectors and is commonly used to assess the extent to which two documents are similar. Cosine similarity result is shown in Table 5.

Table 5

Cosine similarity result

Cosine similarity			
Document	SUM(D^2)	SUM(Q*D)	SUM(Q*D) / SQRT(SUM(Q^2))*SQRT(SUM(D^2))
Q	0.91		
D1	3.81	0	0
D2	2.60	0	0
D3	2.55	0	0
D4	2.42	0.18	0.07
D5	1.33	0.91	0.51
D6	3.44	0	0
D7	3.44	0	0
D8	4.23	0	0

Similarities between documents will be identified through feature similarity searches using cosine similarity. If the cosine similarity value is high or close to 1, then the two vectors are identical. Conversely, if the cosine similarity value is close to 0, then the two vectors have no relationship and are negative (close to –1), then the two vectors are completely opposite. The results of the cosine similarity calculation process can be seen in Table 5. D1–D8 was a document that contain of efficiency and ingredients of traditional herbal medicine. In this example, the cosine similarity calculated between the documents/corpus and the user query "pegal linu".

5. 5. Test scenario result of traditional medicine recommendation system

Scenario 1.

The results of scenario 1 can be seen in Table 6 with the precision value of the TF-IDF method is 80%.

Table 6

TF-IDF precision result

Method	Parameter	Precision result
TF-IDF	–	80%

Scenario 2.

The results of scenario 2 can be seen in Table 7 with the precision value of the Word2Vec method between 4% and 14%.

Table 7

Word2Vec precision result

Method	Parameter	Precision result
Word2Vec	D = 50, C = 5	8%
	D = 50, C = 9	4%
	D = 50, C = 15	6%
	D = 70, C = 5	14%
	D = 70, C = 9	4%
	D = 70, C = 15	14%
	D = 100, C = 5	4%
	D = 100, C = 9	10%
	D = 100, C = 15	10%
	D = 120, C = 5	8%
	D = 120, C = 9	8%
	D = 120, C = 15	10%
	D = 150, C = 5	6%
	D = 150, C = 9	6%
	D = 150, C = 15	10%
	D = 200, C = 5	8%
	D = 200, C = 9	6%
	D = 200, C = 15	8%
	D = 250, C = 5	8%
	D = 250, C = 9	6%
	D = 250, C = 15	8%
	D = 300, C = 5	8%
	D = 300, C = 9	6%
	D = 300, C = 15	6%

The Word2Vec approach generally gives much lower precision results compared to TF-IDF. It shows that Word2Vec in these configurations is less effective for the Madura herbal medicine recommendation system. In addition, the variation of dimension and window size in the Word2Vec model also does not provide a significant improvement like shown in Fig. 2.

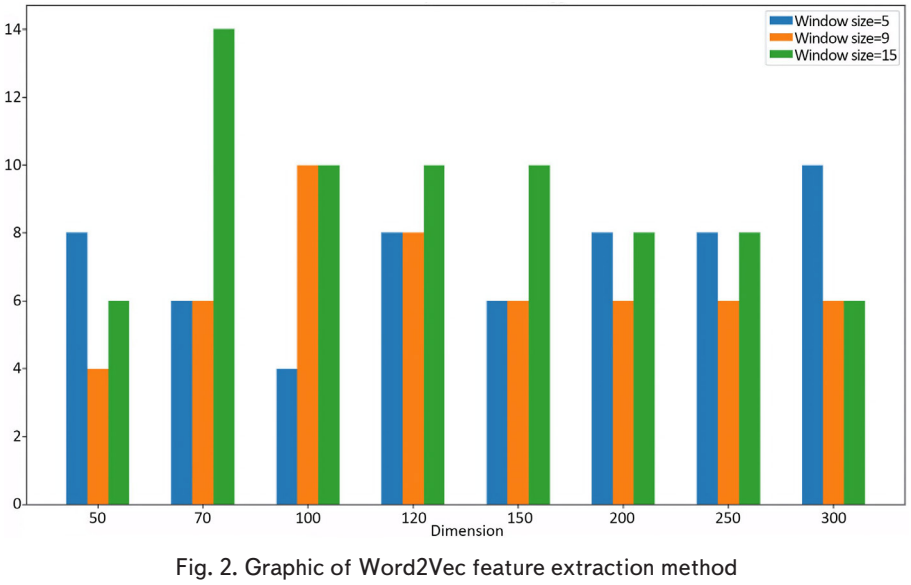

Fig. 2. Graphic of Word2Vec feature extraction method

The graph of the test results using the Word2Vec feature extraction method can be seen in Fig. 2.

Scenario 3.

The results of scenario 3 can be seen in Table 8 with the precision value of the combination of TF-IDF and Word2Vec methods between 4–82%.

Table 8

TF-IDF And Word2Vec precision result

Method	Parameter	Precision result
TF-IDF Word-2Vec	D = 50, C = 5	78%
	D = 50, C = 9	82%
	D = 50, C = 15	82%
	D = 70, C = 5	80%
	D = 70, C = 9	80%
	D = 70, C = 15	82%
	D = 100, C = 5	4%
	D = 100, C = 9	80%
	D = 100, C = 15	82%
	D = 120, C = 5	72%
	D = 120, C = 9	76%
	D = 120, C = 15	80%
	D = 150, C = 5	80%
	D = 150, C = 9	80%
	D = 150, C = 15	82%
	D = 200, C = 5	68%
	D = 200, C = 9	80%
	D = 200, C = 15	80%
	D = 250, C = 5	82%
	D = 250, C = 9	80%
	D = 250, C = 15	78%
	D = 300, C = 5	78%
	D = 300, C = 9	80%
	D = 300, C = 15	80%

The graph of the test results using the TF-IDF and Word2Vec feature extraction method can be seen in Fig. 3.

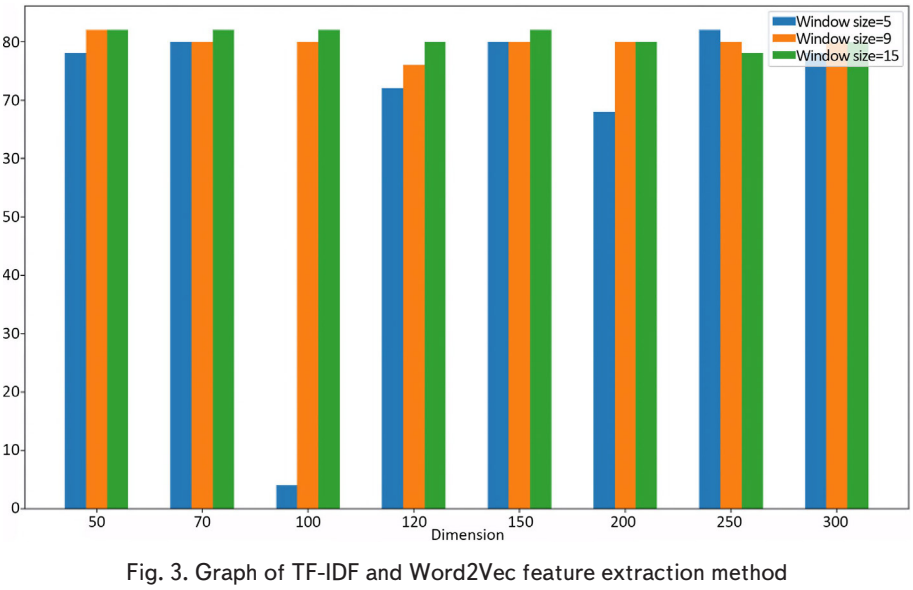

Fig. 3. Graph of TF-IDF and Word2Vec feature extraction method

Combining TF-IDF and Word2Vec significantly improves the relevance of the Madura herbal medicine recommendation system. This combination provides high and stable precision across a wide variety of dimensions and window sizes, with optimal performance at dimensions 50–200 and window sizes 9–15.

6. Discussion of traditional medicine recommendation system based on content-based filtering

This study investigated the efficiency of term frequency-inverse document frequency (TF-IDF) and word to vector (Word2Vec) in extracting response behavior features and compared the predictive, analytical, and clustering effects of classical machine learning methods (supervised and unsupervised) on response behavior.

Based on the precision results of the traditional medicine recommendation system based on content-based filtering with various feature extraction experiments used, the performance of three different approaches can be analyzed: TF-IDF, Word2Vec, and a combination of TF-IDF and Word2Vec:

– in Table 2, traditional medicine dataset is cleaned and converted from raw text into a form more suitable for further analysis. The weight/importance of words in a text (metadata) was measured in Table 3. It is a method used to evaluate how important a word is in a document or document collection (corpus). As another feature extraction, word2vec also apply in metadata in Table 4. Finally, it was measured the extent of similarity between items by using cosine similarity in Table 5;

– based on research result in Table 6, TF-IDF approach yields an excellent precision of 80%. Precision is a basic test that is commonly used in testing recommendation systems to evaluate the extent to which the results of system recommendations are relevant or in accordance with user needs. The results show that the use of TF-IDF method for feature extraction is very effective and give advantages in recommending relevant traditional medicine;

– the Word2Vec approach generally gives much lower precision results compared to TF-IDF, as shown in Table 7. It is shown that Word2Vec in these configurations is less effective for traditional medicine recommendation system. In addition, the variation of dimension and window size in the Word2Vec model also does not provide a significant increase in precision, it was illustrated in Fig. 2. In the relevance of traditional medicine recommendation system, precision tends to remain low across configurations, suggesting that Word2Vec is less suitable in this case;

– the combination of TF-IDF and Word2Vec in Table 8 gives excellent results with precision reaching up to 82%. Based on Fig. 3, this combination consistently yielded high precision across various dimension and window size configurations, suggesting that combining these two feature extraction methods can capture more relevant information for the recommendation system.

In contrast to [17, 19–21] where the similarity result using TF-IDF method is smaller than using Word2Vec, this is because TF-IDF cannot detect paraphrases. This research got that TF-IDF is higher than Word2Vec. This is made possible by this research because Word2Vec in these configurations is less effective for traditional medicine recommendation system. In addition, the variation of dimension and window size in the Word2Vec model also does not provide a significant increase in precision. in the relevance of traditional medicine recommendation system. Word2Vec becomes higher than TF-IDF, when TF-IDF is combine with Word2Vec in feature extraction.

Combining TF-IDF and Word2Vec significantly improves the relevance of the traditional medicine recommendation system. This combination provides high and stable precision across a wide variety of dimensions and window sizes, with optimal performance at dimensions 50–200 and window sizes 9–15. This combination is more effective than using TF-IDF or Word2Vec separately.

The limitations of this study include the dataset, which if using a larger and more varied dataset can improve the model's ability to understand and recommend herbs that are more relevant to user needs. In addition, it is necessary to explore other hybrid methods such as a combination of content-based filtering and collaborative filtering to provide recommendations that are more personalized and targeted.

For further research, using a larger and more varied dataset can improve the model's ability to understand and recommend herbs that are more relevant to the user's needs. Although the combination of TF-IDF and Word2Vec has provided satisfactory results, exploration of other hybrid methods such as content-based filtering and collaborative filtering could provide more personalized and targeted recommendations. These methods can capture user preferences based on broader behavioral patterns. Further optimization of model parameters such as vector dimension and window size in Word2Vec to find the most optimal configuration can improve the accuracy of the recommendation system.

7. Conclusion

1. Text processing involves cleaning and preparing raw text data for further analysis or model training. Proper text preprocessing can significantly impact the performance and accuracy of natural language preprocessing models.

2. The use of a combination of TF-IDF and Word2Vec is proven to be able to improve the relevance of traditional medicine recommendation system. The combination of TF-IDF and Word2Vec is proven to be able to improve the relevance of traditional medicine recommendation system by providing the best precision of up to 82% in various dimension and window size variations, with optimal performance in dimensions 50–200 and window size 9–15.

3. The TF-IDF method alone achieves 80% precision, while Word2Vec shows lower performance (4–14% lower than the combination) in various parameter experiments.

4. The Word2Vec approach generally gives much lower precision results compared to TF-IDF. This suggests that Word2Vec in these configurations is less effective for the Madura herbal medicine recommendation system.

5. This system helps users in choosing herbs that suit their needs and has the potential to re-popularize traditional Madurese herbs. This study was determined that TF-IDF outperforms Word2Vec. This conclusion is drawn from the research indicating that Word2Vec, under these specific configurations, is less effective for a traditional medicine recommendation system. Furthermore, variations in dimension and window size within the Word2Vec model do not yield a significant enhancement in precision regarding the relevance of the traditional medicine recommendation system. Word2Vec surpasses TF-IDF only when TF-IDF is integrated with Word2Vec during feature extraction. The integration of TF-IDF and Word2Vec greatly enhances the relevance of the traditional medicine recommendation system. This synergy delivers high and consistent precision across a broad range of dimensions and window sizes, achieving optimal performance at dimensions between 50 and 200 and window sizes from 9 to 15. This approach proves to be more effective than utilizing TF-IDF or Word2Vec independently. Additionally, it is essential to investigate other hybrid methodologies, such as combining content-based filtering with collaborative filtering, to offer recommendations that are more personalized and targeted.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The research was funding from LPPM Trunojoyo University, Indonesia.

Data availability

Manuscript has associated data in a data repository.

Acknowledgments

Thanks to the Informatics Engineering Department and LPPM Trunojoyo University Indonesia for giving us the opportunity to contribute the development of computer science and Indonesian traditional herbal medicine.

References

1. Che, C.-T., George, V., Ijinu, T. P., Pushpangadan, P., Andrae-Marobela, K. (2024). Traditional medicine. Pharmacognosy. Academic Press, 11–28. https://doi.org/10.1016/b978-0-443-18657-8.00037-2
2. WHO traditional medicine strategy: 2014–2023 (2013). World Health Organization. Available at: http://apps.who.int/iris/bitstream/10665/92455/1/9789241506090_eng.pdf Last accessed: 20.05.2025

3. Bodeker, G., Graz, B.; Ryan, E. T., Hill, D. R., Solomon, T., Aronson, N. E., Endy, T. P. (Eds.) (2020). Traditional Medicine. Hunter's Tropical Medicine and Emerging Infectious Diseases. Elsevier, 194–199. https://doi.org/10.1016/b978-0-323-55512-8.00025-9
4. Kamboj, V. P. (2000). Herbal medicine. Current Science, 78 (1), 35–39.
5. Pal, S. K., Shukla, Y. (2003). Herbal medicine: Current status and the future. Asian Pacific Journal of Cancer Prevention, 4, 281–288. Available at: https://www.researchgate.net/profile/Sanjoy-Pal-2/publication/8914668_Herbal_medicine_Current_status_and_the_future/links/0c96051fd33d11991d000000/Herbal-medicine-Current-status-and-the-future.pdf
6. WHO global report on traditional and complementary medicine 2019 (2019). Geneva: World Health Organization, 226. Available at: https://iris.who.int/bitstream/handle/10665/312342/9789241515436-eng.pdf?sequence=1
7. Sianipar, E. A. (2021). The potential of Indonesian traditional herbal medicine as immunomodulatory agents: A review. International Journal of Pharmaceutical Sciences and Research, 12 (10), 5229–5237. https://doi.org/10.13040/IJPSR.0975-8232.12(10).5229-37
8. Pradipta, I. S., Aprilio, K., Febriyanti, R. M., Ningsih, Y. F., Pratama, M. A. A., Indradi, R. B. et al. (2023). Traditional medicine users in a treated chronic disease population: a cross-sectional study in Indonesia. BMC Complementary Medicine and Therapies, 23 (1). https://doi.org/10.1186/s12906-023-03947-4
9. Muharrami, L. K., Santoso, M., Fatmawati, S. (2024). Traditional Medicine Uses of Madurese Ethnic, Indonesia: Indigenous Knowledge "Jamu" in Relation with Medicinal Plants. Journal of Hunan University Natural Sciences, 51 (10). https://doi.org/10.55463/issn.1674-2974.51.10.2
10. Yunitarini, R., Widiaswanti, E. (2024). Analysis and Design of Indonesian Traditional Medicine (Jamu) Information System by using Prototyping Model (Case Study: Madura Island). E3S Web of Conferences, 483, 03012. https://doi.org/10.1051/e3sconf/202448303012
11. Vall, A., Dorfer, M., Eghbal-zadeh, H., Schedl, M., Burjorjee, K., Widmer, G. (2019). Feature-combination hybrid recommender systems for automated music playlist continuation. User Modeling and User-Adapted Interaction, 29 (2), 527–572. https://doi.org/10.1007/s11257-018-9215-8
12. Widayanti, R., Chakim, M., Lukita, C., Rahardja, U., Lutfiani, N. (2023). Improving Recommender Systems using Hybrid Techniques of Collaborative Filtering and Content-Based Filtering. Journal of Applied Data Sciences, 4 (3), 289–302. https://doi.org/10.47738/jads.v4i3.115
13. Van Balen, J., Goethals, B. (2021). High-dimensional Sparse Embeddings for Collaborative Filtering. Proceedings of the Web Conference 2021. Ljubljana, 575–581. https://doi.org/10.1145/3442381.3450054
14. Gunarto, S. A., Honggara, E. S., Purwanto, D. D. (2023). Website Sistem Rekomendasi dengan Content Based Filtering pada Produk Perawatan Kulit. Jurnal Sistem Dan Teknologi Informasi, 11 (3), 399. https://doi.org/10.26418/justin.v11i3.59049
15. Nastiti, P. (2019). Penerapan Metode Content Based Filtering Dalam Implementasi Sistem Rekomendasi Tanaman Pangan. Teknika, 8 (1), 1–10. https://doi.org/10.34148/teknika.v8i1.139
16. Huda, A. A., Fajarudin, R., Hadinegoro, A. (2022). Sistem Rekomendasi Content-based Filtering Menggunakan TF-IDF Vector Similarity Untuk Rekomendasi Artikel Berita. Building of Informatics, Technology and Science, 4 (3), 1679–1686. https://doi.org/10.47065/bits.v4i3.2511
17. Putri, M. W., Muchayan, A., Kamisutara, M. (2020). Sistem Rekomendasi Produk Pena Eksklusif Menggunakan Metode Content-Based Filtering dan TF-IDF. Journal of Information Technology and Computer Science, 5 (3), 229. https://doi.org/10.31328/jointecs.v5i3.1563
18. Negara, E. S., Sulaiman, Andryani, R., Saksono, P. H., Widyanti, Y. (2023). Recommendation System with Content-Based Filtering in NFT Marketplace. Journal of Advances in Information Technology, 14 (3), 518–522. https://doi.org/10.12720/jait.14.3.518-522
19. Nawangsari, R. P., Kusumaningrum, R., Wibowo, A. (2019). Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study. Procedia Computer Science, 157, 360–366. https://doi.org/10.1016/j.procs.2019.08.178
20. Khomsah, S. (2021). Sentiment Analysis on YouTube Comments Using Word2Vec and Random Forest. Telematika, 18 (1), 61–72. https://doi.org/10.31315/telematika.v18i1.4493
21. Ramadhanti, N. R., Mariyah, S. (2019). Document Similarity Detection Using Indonesian Language Word2vec Model. 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). Semarang: IEEE, 1–6. https://doi.org/10.1109/icicos48119.2019.8982432
22. Cahyani, S. N., Saraswati, G. W. (2023). Implementation of support vector machine method in classifying school library books with combination of TF-IDF and WORD2VEC. Jurnal Teknik Informatika, 4 (6), 1555–1566. https://doi.org/10.52436/1.jutif.2023.4.6.1536
23. Liang, M., Niu, T. (2022). Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs. Procedia Computer Science, 208, 460–470. https://doi.org/10.1016/j.procs.2022.10.064
24. Nurfalah, F., Asriyanik, Pambudi, A. (2022). Sistem Rekomendasi Event Online Menggunakan Metode Content Based Filtering. Elkom : Jurnal Elektronika Dan Komputer, 15 (2), 271–279. https://doi.org/10.51903/elkom.v15i2.736
25. Irvandani, A., Auliasari, K., Primaswara Prasetya, R. (2020). Sistem Rekomendasi Pemilihan Fotografer dengan Metode Haversine dan TF-IDF di Malang Raya. Jurnal Mahasiswa Teknik Informatika, 4 (1), 137–146. https://doi.org/10.36040/jati.v4i1.2330
26. Yutika, C. H., Adiwijaya, A., Faraby, S. A. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. Jurnal Media Informatika Budidarma, 5 (2), 422. https://doi.org/10.30865/mib.v5i2.2845