JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

MACHINE LEARNING CONTENT-BASED FILTERING WOMEN EMPOWERING RECOMMENDATIONS ON YOUTUBE

Yuliana[1*], Mira[1], Aloysius Hari Kristianto[2]
[1]Information Technology, Shanti Bhuana Institute
[2]Management, Shanti Bhuana Institute
*email:** yuliana@shantibhuana.ac.id

Abstract: YouTube is one of the most popular video streaming platforms, but it has constraints that can cause problems when clients have difficulty finding content according to their wishes. The main objective of this study is to increase user capacity in viewing content specifically in the field of women's empowerment. By using content-based filtering techniques, the system will analyze user preferences and interests through recommendations for women's empowerment content. The data source is via the YouTube API and is analyzed using PHP programming content-based filtering techniques. The system's recommendations provide a list of women's empowerment content with a user request display. The results of the research evaluation obtained a precision value of 62%, meaning that the recommendations match the topic being searched for, namely women's empowerment. The recall value of 84% indicates that the system has succeeded in finding relations from the database. The f1-score value of 72% indicates that there is a balance between precision and recall, meaning that a system is needed that is not only accurate but also complete. While the cosine value shows a score of 0.7071 approaching the maximum value (1.0). The recommendation of the content-based filtering method produces quite effective women's empowerment content.

Keywords: content-based filtering, recommendations, women Empowerment, youtube

Abstrak: Youtube merupakan platform streaming video yang sangat populer, namun memiliki masalah ini bisa menyebabkan permasalahan ketika klien kesulitan menemukan konten sesuai keinginan. Agar mempermudah user mencari konten yang sesuai maka dengan keinginan klien, maka tujuan utama dari penelitian ini harapannya ingin meningkatkan kapasitas pengguna untuk melihat konten khusus dalam bidang pemberdayaan perempuan dengan menggunakan teknik content-based filtering, sistem akan menganalisis preferensi dan minat penonton melalui rekomendasi konten pemberdayaan perempuan. Sumber data konten adalah youtube API dan dianalisis menggunakan bahasa pemrograman PHP dengan teknik content-based filtering. Rekomendasi tersebut memberikan daftar konten pemberdayaan perempuan yang ditayangkan untuk kepentingan pengguna tertentu. Hasil evaluasi mendapatkan nilai presisi 61.76% artinya rekomendasi sesuai dengan topik yang dicari yaitu tentang pemberdayaan perempuan. Nilai recall 84% menunjukkan sistem berhasil menemukan keterkaitan dari database. Nilai f1-score 71.19% menyatakan keseimbangan antara presisi dan recall artinya membutuhkan sistem yang tidak hanya akurat namun data juga harus lengkap. Sedangkan cosine score mendapatkan score 0.7071 mendekati nilai maksimum (1.0). Sistem rekomendasi metode content based filtering menghasilkan rekomendasi konten pemberdayaan perempuan cukup efektif.

Kata kunci: content-based filtering; pemberdayaan perempuan; rekomendasi; youtube

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

In its development, technology understands the influence of consumers in searching for streaming-based content. Various types of podcasts and music and interesting information. According to research conducted by the GMI Team, more than 2.70 billion people worldwide use YouTube every month. [1] We Are Social report informs that there are 139 million Indonesians surfing the Youtube Platform. Indonesia is the fourth country with the most users in the world. YouTube uploads 720,000 hours of new videos every day. Music, entertainment, and education are the most watched genres on YouTube. [2]

Youtube is a website that functions for users to upload videos, share videos, and select and watch videos. This platform is also a digital media that is useful in presenting women's empowerment podcast content, thus there is very good potential for women to form and increase insight related to empowered women's innovations. However, it takes an understanding of listener preferences and behavior regarding the podcast content that will be selected and sorted. for that the main challenge faced by researchers is when presenting women's empowerment content through streaming platforms such as YouTube which of course have a dominance of high-quality content. One thing that needs to be prepared is to integrate with the YouTube API to create new opportunities, distribute and evaluate in selecting content according to what users want. Machine Learning is part of Artificial Intelligence Technology or under the umbrella of artificial intelligence by focusing on the development of computer algorithms to improve in recognizing patterns in data, studying data and making decisions by predicting data into knowledge and information. And the task of Machine Learning is to find a way to complete the task based on the examples given. [3] The Recommendation System has the task of handling the problem of excessive information, with the basic goal of recommending accurate information to users according to the available options and providing experience and satisfaction as well as loyalty and active visitors so that the results obtained are an increase in the level of popularity of the application created. [4] [5][6]

In this study, the author proposes a solution, namely a recommendation system. The task of this system is to help recommend content based on user observations to be more relevant according to the topic of women's empowerment podcast content. By utilizing content-based filtering techniques through the application of Machine Learning without requiring data from users as recommendation data, but utilizing parameters as items in determining recommendations that match user desires. Referring to previous research on food crop recommendation systems, the research results, namely for marketing traders in particular, will be more effective and efficient, meaning the system is able to provide recommendations for suitable land for users.

The researchers then conducted a Music Recommendation Determination using the Content-Based Filtering Method with the aim of creating a system that can provide music recommendations according to user preferences so that the user's comfort level will increase. The results of this method are finding an average song title similarity of up to 0.6684, a precision of 0.125 and a recall

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

of 0.200. The system runs well through the test results with an average response time of 3.5 seconds.[7] Research on the neural network algorithm model on the data used. The training data was 83% accurate, while the validation data achieved 78% accuracy. In the selection of elective courses, the data consisted of 70 students and 35 subject numbers. The neural network algorithm model was able to learn the data with an accuracy of 83% on the training data and 78% on the validation data. The loss on the training data was reduced to 0.41 and the validation loss to 0.62. Furthermore, the test data accuracy was 0.79 or 79%. The recommendation results provided the 10 product data with the highest scores. [8] Further research using content-based filtering with TF-IDF and Cosine Similarity bias offers podcast recommendations based on the genre and episode name that users like; the higher the similarity value, the more users like the podcast. Precision ranges between 0.52 and 0.74 on Confusion Matrix Classification Report Testing data. We found that the average weight is 0.79 and the recall is low at 0.51. Since these values are very close to 1, the information aligns with recommendations for podcasts in the education genre.[9]

METHOD

Content-based filtering is a method used in recommendation systems and analyzes data by focusing on the characteristics or content of the items to be recommended or analyzed. The approach used is the attributes or features of the items to determine the similarity between existing items and user preferences. This method attempts to provide recommendations for items that are similar to items that have been liked by the user based on content characteristics. [10] [11]

This CBF technique is customer-free, not dependent on the situation, whether the content is new content. If the user has determined content A in a certain category, the system will try to provide recommendations for content A choices with similar categories that have been provided in the YouTube search engine which are likely to match the choices sought by the user based on the preferences of other users.[12] The logic of how the CBF method works is starting from user input which can be in the form of keywords or certain descriptions such as educated, education, educators, etc. Then the content representation or through feature extraction that each video is converted into a vector representation based on content such as title, description, url. The steps that must be taken are preprocessing by performing case folding, stopword removal and tokenization, then word weighting using term frequency (TF) or TF-IDF to produce a word vector per video content. [11] Term frequency (TF) is used to measure words that often appear in a document. For example; TF(t,d) = Number of words appearing t in document d / total words in document d.
Example; if the word "girl" appears 3 times in the description totaling 14, the calculation is; TF('girl') = 3/14 =0.2
IDF (Inverse Document Frequency); reduces the weight of common words that often appear in the document.
IDF(t) = log(N/dFt) [11]

Description:
N : total number of documents
dFt: number of documents containing word t

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

example; the word 'girl' appears in 10 out of 50 videos;
IDF('girl') = log(50/20) = log(5) ~ 0.699
Combining TD and IDF to provide weighting
TF-IDF(t,d) = TF(t,d) x IDF(t) [11]

Then in the matching stage (Similarity Computation) the system calculates the similarity between the user's query vector and each video vector in the following way: Cosine Similarity: a value between 0-1 (the closer to 1, the more similar the results) with the formula: Cosine Similarity; measures the similarity between two vectors (query and document/video/content);[13][11]
Cosine_similarity(A,B) =

$$(\vec{A}, \vec{B}) = \frac{\Sigma\,(Ai \times Bi)}{\sqrt{\Sigma(Ai^2)} \times \sqrt{\Sigma(Bi^2)}} \quad (1)$$

A: Feature vector from query
B: Feature vector from video description
Cosine value: 1: very similar, 0: not similar, <0 not relevant

If the query on educational/women's empowerment content is similar to the video description, the score can reach 0.7 or more. Then enter filtering and ranking; only videos with a similarity score higher than the threshold (min_score) are displayed. Then sorted from the highest score >> lowest >> results.

Until the evaluation stage using several metrics such as precision, recall and f1 score, each has its own task.[13] [14]. Precision does an estimate of how many results will be shown that are relevant.
Precision = number of relevant videos recommended / total number of videos recommended
Recall how much of all content is related to the results found.
Recall = number of relevant videos recommended / total number of relevant videos in the database
F1-Score does a balance between precision and recall.
F1 = 2 . (Precision x Recall) / (precision+recall).

The disadvantage of the CBF method is that it has limitations in providing recommendations for similar content so that the opportunity to find content or items is very rare when the item appears suddenly. The way CBF works is based on the similarity of new items being recommended. When the user determines the choice of content on the YouTube platform, the recommendation choices are immediately presented based on the user's choice. Items are calculated for similarity by comparing items from the content/podcast selected by the user with items from other content/podcasts after going through the preprocessing stage. The following is a flowchart of the content-based filtering technique:

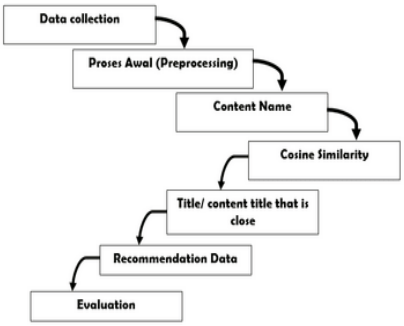

Image 1. Research Stages

RESULT AND DISCUSSION

Data collection: Collecting datasets through Youtube Data API v3 based on the keyword women's empowerment, then converted into a CSV file containing information about women's empowerment, then the data is input into the mySQL database output.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

To get the dataset, make a request to the API using the file_get_contens() command to retrieve video search results from youtube based on keywords:

Image 2. Request command to API

Part=snippet asks for title and description details, q= for search keywords, type=video to fetch videos not channels or playlists, maxResult=10 fetches a maximum of 10 videos and key= each creator's API. Then we fetch the video data display.

Image 3. Video data fetch command

Looping foreach and storing data neatly will get a list of videos from YouTube according to keywords and stored in CSV or database.

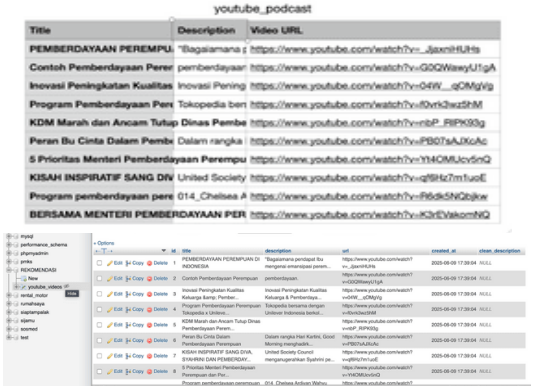

Image 4. Database .csv and.sql

In the data above, we have sorted the women's empowerment podcast data in the sql and csv databases on the attributes/fields id, title, description, url.

Preprocessing: At this stage, changes will be made to the dataset from data collection obtained from the Youtube API to be converted into the women's empowerment podcast data alphabet to be more structured and become the final data. By sorting the data, the dataset will be more neatly arranged and sequentially alphabetically.

Image 5. Data Sorting Commands

To remove punctuation and numb ers using the preg_replace command, while removing text that is too often used and considered meaningless using the stopwords command, until the text is filtered until the sentence is recombined or imploded. The Preprocessing stage or text pre-processing to clean and prepare the text so that it can be analyzed accurately. The pre-processing stage is very important so that the co-sine similarity results are valid. Implementa tion of content based filtering: the recommendation principle must be based on similar items after going through the preprocessing stage and then compared to find their similarity by calculating the cosine similarity function.

Image 6. Stage of the cosine similarity function

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

The recommendation system page uses PHP programming language based on Content-Based Filtering using TF-IDF and Cosine Simi-larity. This section will receive queries from the URL (?q=women's empowerment).

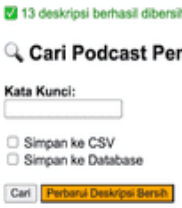

Image 7. Search engine request kontent

The following data displays information about podcasts with the topic of women's empowerment and only displays recommendation results based on the similarity of previously processed clean description content.

Image 8. Content Recommendation Page

The Evaluation Stage is carried out using the precision and recall methods. This evaluation uses test data on API v3 data based on keywords, when only using the keyword "women's empowerment" the score results only show 0.5774 and are still low.

Image 9. Recommendation Results low score

But finally adding more specific keywords "women's empowerment", "business training for women", "women's MSMEs", "maternal and child health", "women's education", "women's rights", "women in leadership" then the result obtained is 0.7071.

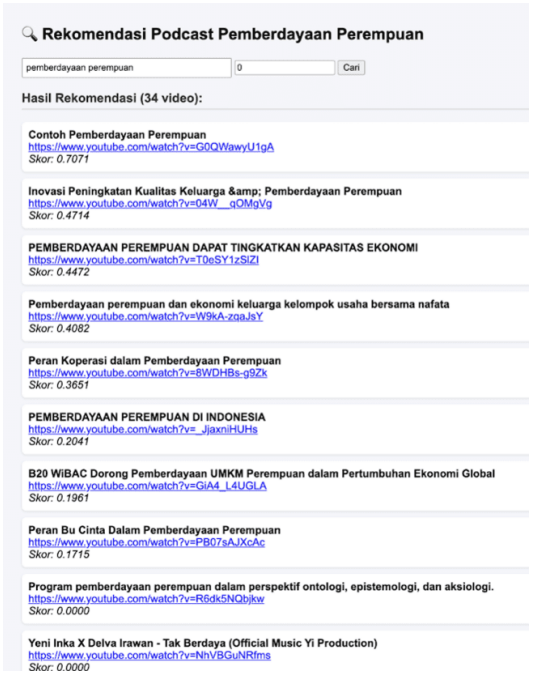

Image 10. High Score Recommendation Results

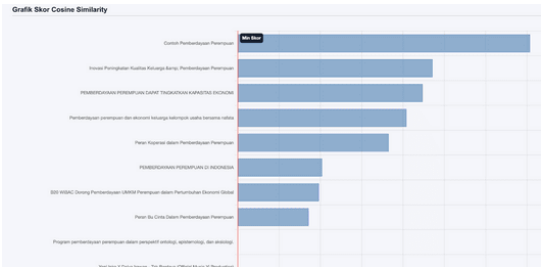

Image 11. Cosine similarity score graph

Then the comparison of the results of the diagram recommendations can be seen in Figure 12. The results show the precision or recall values and F1 Score.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 4, September 2025, hlm. 669 – 676 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v11i4.4154
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

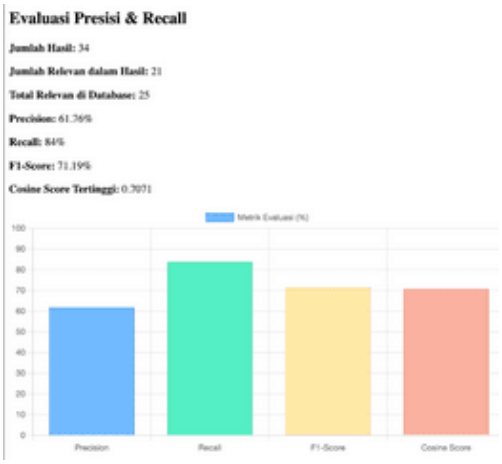

Image 12. The results of the precision or recall values and F1 Score.

In the evaluation test results with a total of 34 recommended videos based on scores above the threshold, the number of relevant videos is only 21 relevant videos labeled 1 so that the total relevant in the database is 25 videos. The precision results with a proportion of relevant recommendation results of 21/34 are 61.76% while the results on recall are 84% with a proportion of relevant item results that were successfully found, namely 21/25 and F1-Score 71.19% the average balance between precision and recall is 2 x P x / (P + R). So that the highest cosine result with a score of 0.7071 is the highest similarity score among the results.

CONCLUSION

Based on the results of the evaluation, the precision value was 61.76%, meaning that the recommendations were in accordance with the topic being searched for, namely women's empowerment. The recall value of 84% indicates that the system has succeeded in finding a relationship from the database. The f1-score value of 71.19% indicates a balance between precision and recall, meaning that a system that is not only accurate but also requires complete data is required. Meanwhile, the cosine score obtained a score of 0.7071, approaching the maximum value (1.0).

The content-based filtering method recommendation system produ ces quite effective women's empowerm ent content recommendations; there is still room to improve precision by filtering content results that are too general, recall is already high, meaning that the system has succeeded in retrieving most of the videos from the database. The F1 score above 70% indicates that the system is stable and quite balanced in accuracy and reach. The researcher's next suggestion is to add tags, categories and video duration features, then use a more complex TF-IDF technique and combine other methods such as collaborative filtering.

BIBLIOGRAPHY

[1] T. R. GMI, “Statistik Youtube 2025 (demografi) Pengguna penurut negara dan lainnya,”globalmediainsight.com, 2025.[Online].Available:https://www.globalmediainsight.com/blog/youtube-users-statistics/.

[2] S. KEMP, “Digital 2024 :Indonesia,” datareportal,2024.[Online].Available: https://datareportal.com/reports/digital-2024-indonesia. [Accessed: 17-Jun-2025].

[3] E. A. H. K. Yuliana, M Kom, Azriel Christian Nurcahyo, S Kom, M Kom, S Paradise, M Kom, Machine Learning Algorithms and their Use Cases. Widina Media Utama, 2025.

[4] Y. H. Alfaifi, “Recommender Systems Applications: Data Sources, Features, and Challenges,” Inf., vol. 15, no. 10, 2024, doi: 10.3390/info15100660.

[5] T. Kanwal and T. Amjad, Research paper recommendation system based on multiple features from citation network, vol. 129, no. 9. Springer International Publishing, 2024.

[6] A. Rianti, N. W. A. Majid, and A. Fauzi, "Machine Learning Journal Article Recommendation System Using Content Based Filtering," JUTI J. Ilm. Teknol. Inf., pp. 1–10, 2024, doi:10.12962/j24068535.v22i1.a1193.

[7] A. I. Putra and R. R. Santika, "Implementasi Machine Learning dalam Penentuan Rekomendasi Musik dengan Metode Content-Based Filtering," Edumatic J. Pendidik. Inform., vol. 4, no. 1, pp. 121–130, 2020, doi: 10.29408/edumatic.v4i1.2162.

[8] F. Nur Fajri, A. Tholib, and W. Yuliana, "Application of Machine Learning Algorithm for Determining Elective Courses in Informatics Study Program," J. Tek. Inform. dan Sist. Inf., vol. 8, no. 3, pp. 485–496, 2022, doi: 10.28932/jutisi.v8i3.3990.

[9] M. M. Raharjo and F. Arifin, "Machine Learning System Implementation of Education Podcast Recommendations on Spotify Applications Using Content-Based Filtering and TF-IDF," vol. 8, no. 2, pp. 2477–2399, 2023.

[10] A. A. & A. M. Shristi Shakya Khanal, P.W.C. Prasad, "A systematic review: machine learning based recommendation systems for e-learning," Educ Inf Technol, vol. Volume 25, no. July 2020, pp. 2635–2664, 2020.

[11] K. Tejaswini, V. Umadevi, S. M. Kadiwal, and S. Revanna, "Design and development of machine learning based resume ranking system," Glob. Transitions Proc., vol. 3, no. 2, pp. 371–375, 2022, doi: 10.1016/j.gltp.2021.10.002.

[12] S. Javed, U., Shaukat, K., A. Hameed, I., Iqbal, F., Mahboob Alam, T. & Luo, "A Review of Content-Based and Context-Based Recommendation Systems. International Journal of Emerging Technologies in Learning (iJET)," Int. J. Emerg. Technol. Learn., vol. 16(3), no. 1863–0383, pp. 274–306, 2021.

[13] I. S. dan D. A. Kristiyanti, MACHINE LEARNING untuk pemula. Bandung: Informatika Bandung, 2022.

[14] Y. Yuliana, P. Paradise, and M. Qulub, "Detection of Children'S Nutritional Status Using Machine Learning With Logistic Regression Algorithm," JURTEKSI (Jurnal Teknol. dan Sist. Informasi), vol. 10, no. 2, pp. 267–274, 2024, doi:10.33330/jurteksi.v10i2.2973.