Implementation of Content-Based Cosine Similarity Algorithm With TF-IDF and SBERT for Movie Recommendation

Eliata Zefanya Irabela
Information System
Amikom University Yogyakarta
Yogyakarta, Indonesia
eliatazefanya@students.amikom.ac.id

Norhikmah
Information System
Amikom University *Yogyakarta*
Yogyakarta, Indonesia
hikmah@amikom.ac.id

Abstract— The number of films on streaming platforms continues to increase, often leaving users confused about which film to watch. To overcome this, this research develops a content-based movie recommendation system. Representation of the film information obtained through the application of TF-IDF and SBERT to genre and synopsis data. Cosine similarity is used to calculate the closeness between representations. The performance system is then evaluated through the Precision@K, MAP@K, and Recall@K metrics. According to the test results, the hybrid approach performs better and is more stable than the single method, with a MAP value of 0.95, a Recall of 0.95, and a Precision of 0.71. In the future, the development system will still be possible by utilising other types of data, including user interaction data.

Keywords—Content-Based Filtering, Recommender System, TF-IDF, Cosine Similarity, SBERT

Article info: Date Submitted: 2026-02-18 | Date Revised: 2026-04-23 | Date Accepted: 2026-04-25

I. INTRODUCTION

The proliferation of movie streaming platforms has increased the number of films available, making it difficult for users to find suitable content. To address this issue, recommendation systems are widely used to help users find the content they need. These recommendations can help viewers save time in choosing movies to watch [1].

Recommendation systems can be grouped into content-based filtering, collaborative filtering, and hybrid methods that combine both filters. Collaborative filtering utilises user activity, while content-based filtering looks at item similarities. In this study, the content-based method was chosen because movie recommendations are made by comparing the content of movies that users like, without requiring user rating history data [2].

In content-based recommendation systems, text processing is used to measure the similarity between films. One method often used is Term Frequency–Inverse Document Frequency (TF-IDF), which represents text based on the frequency of word occurrences [3]. Nevertheless, this approach still has limitations in comprehending the context and meaning of statements [4]. Because they can better comprehend sentence context, semantic representation-based techniques such as SBERT can enhance the semantic capabilities of recommendation systems [5]. This method is comparable to academic document semantic search systems that employ SBERT embeddings to capture the semantic link between queries and documents, thereby producing more pertinent search results[6].

Since the hybrid filtering technique has been demonstrated to be effective in providing more accurate, this study combines TF-IDF and SBERT content-based representations to improve the quality of movie recommendations[7][8] . Although TF-IDF and SBERT have

DOI: https://doi.org/10.37396/jsc.v9i1.633

been widely used in text-based recommendation systems, their separate application still reveals a research gap, as TF-IDF only captures word frequency and distribution without understanding context, while SBERT excels in semantic representation but overlooks explicit word distribution. This limitation leads to sub optimal recommendation performance, especially in cases with high lexical variation and complex sentence structures, thereby highlighting the need for a hybrid approach that integrates the strengths of both methods to produce more comprehensive text representations and improve recommendation accuracy.

II. METHOD / MATERIAL

The system combines TF-IDF and SBERT as hybrid features in content-based movie recommendations. The process starts from data preprocessing, followed by feature extraction using TF-IDF and SBERT. These two features are then combined and their similarity calculated using cosine similarity. Based on the similarity values, the recommendation system generates Top-K, as shown in Figure 1.

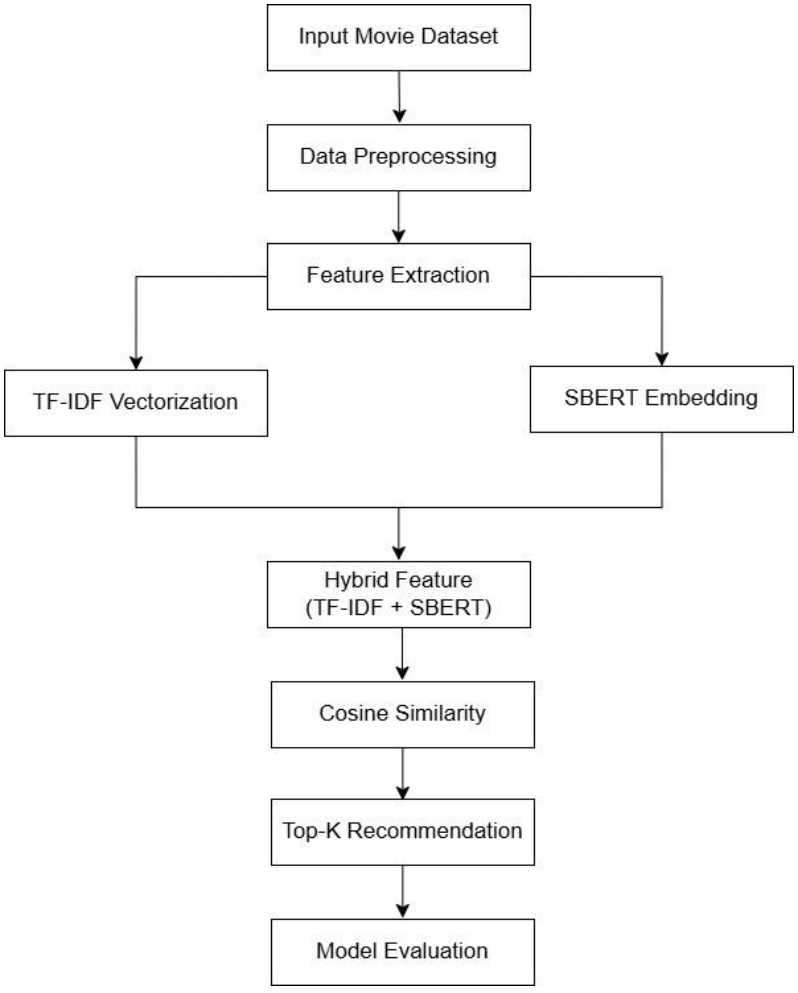

Figure 1. Flowchart of recommendation system with combination of TF-IDF and SBERT.

A. Input Movie Dataset

The study's dataset, movie_metadata.csv, is accessible to the general public via the Kaggle platform. The movie's title, genre, and summary (overview) are all included in this

DOI: https://doi.org/10.37396/jsc.v9i1.633

dataset. Because it provides a narrative account of the movie's plot and setting, the film synopsis serves as the main source for creating features.

Table 1. Sample Dataset

title	genre	overview
Toy Story	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]	Led by Woody, Andy's toys live happily in his room until Andy's birthday brings Buzz Lightyear onto the scene. Afraid of losing his place in Andy's heart, Woody plots against Buzz. But when circumstances separate Buzz and Woody from their owner, the duo eventually learns to put aside their differences.
Grumpier Old Men	[{'id': 10749, 'name': 'Romance'}, {'id': 35, 'name': 'Comedy'}]	A family wedding reignites the ancient feud between next-door neighbors and fishing buddies John and Max. Meanwhile, a sultry Italian divorcÃ©e opens a restaurant at the local bait shop, alarming the locals who worry she'll scare the fish away. But she's less interested in seafood than she is in cooking up a hot time with Max.

B. Data Preprocessing

The data preprocessing stage is performed before the feature extraction procedure to make the summary text more consistent, noise-free, and free of unnecessary or missing data. The goal of this step is to enhance the text data's quality so that the recommendation system can handle it as efficiently as possible [9].

Preprocessing includes tokenisation to divide the text into word units, case folding by changing all text to lowercase, and eliminating punctuation, symbols, and numbers [10]. In the preprocessing stage, stopwords are removed to reduce words that do not convey meaningful information [11]. Next, lemmatisation and stemming are used to simplify English word forms, enabling the synopsis text to be used more effectively in the feature extraction stage [12].

C. Feature Extraction

One of the key stages in a recommendation system is feature extraction, in which the movie synopsis text is converted to numerical form for similarity calculation. In this study, two techniques were used: SBERT and TF-IDF. Features from both methods were combined into a hybrid representation, and the similarity between movies was then measured.

1) TF-IDF Vectoritazion

To form a text representation, TF-IDF was used by giving weight to words [13]. TF-IDF determines a word's relevance in a document; words with higher TF-IDF values are more significant [14].

The following is how the TF-IDF weights are calculated:

$$\text{TF-IDF}(t,d)=\text{TF}(t,d)\times\text{IDF}(t) \quad (1)$$

In this calculation, the value of df(t) indicates the number of documents containing the term t, while N is the total number of documents used [15]. The weight of each word in a movie summary is calculated using equation (1). Words that are commonly used in one movie but rarely in others tend to have higher TF-IDF values [16]. TF-IDF produces an n × m matrix that shows the number of films and the word features used.

2) SBERT Embedding

Sentence embeddings generated by SBERT are used to represent the concise meaning of text, which is then used to compare similarities between films. Differences in sentence structure or vocabulary in the synopsis are not an obstacle, since the film is presented as an

DOI: https://doi.org/10.37396/jsc.v9i1.633

embedded work. The cosine similarity method is applied to assess how closely the meanings of the synopses of the films match [17].

3) Hybrid Feature (TF-IDF + SBERT)

TF-IDF and BERT are not used separately; they are combined to form a hybrid representation of features, so that word information and meaning can mutually complement each other, building on previous research [18]. Compared to the single-method approach, this approach produces more suitable movie recommendations.

D. Cosine Similarity

Based on the hybrid features obtained, cosine similarity [19] is used to calculate similarities. This method measures the similarity between two feature vectors representing movies [20].

Which is used for calculate the cosine value :

$$\text{CosineSimilarity}(A,B) = \frac{A.B}{||A||\,||B||} \quad (2)$$

In equation (2), vectors A and B represent the feature vectors of two movies, where A·B denotes the dot product of the vectors, and ||A|| and ||B|| represent the magnitudes of the vectors. A higher cosine similarity value indicates greater similarity between the movie content.

E. Top-K Recommendation

Once the process is complete, the film is sorted by cosine similarity; the system selects the K films with the highest values.

F. Model Evaluation

To see the results of the evaluation used Recall@K, MAP@K, as well as Precision@K.

1) Precision@K

$$\text{Precision@K} = \frac{Jumlah\ item\ relevan\ yang\ direkomendasikan}{K} \quad (3)$$

Through equation (3), Precision@K receives assistance accuracy of movie recommendations which is included in the Top-K list [21].

2) Recall@K

$$\text{Recall@K} = \frac{Rekomendasi\ relevan\ di\ Top{-}K}{Total\ item\ relevan} \quad (4)$$

To see the system's capabilities in determining relevant films use Recall@K (4) [22]. To assess performance evaluation Precision@K and Recall@K were carried out [23].

3) Mean Average Precision (MAP)

$$\text{MAP@K} = \frac{1}{N}\sum_{i=1}^{N} \text{AP@Ki} \quad (5)$$

MAP@K is used to turn on the quality of recommendations based on the ranking order of movies in the equation (5) [24] [25].

III. RESULT AND DISCUSSION

A. Data Preprocessing

The pre-processing stage helps tidy up the synopsis text by removing redundant words and equalising word forms, so that the important parts are more visible. The pre-processed Table 2 results are ready for feature extraction.

DOI: https://doi.org/10.37396/jsc.v9i1.633

Table 2. Film Synopsis Preprocessing Results

title	overview	overview_clean
Toy Story	Led by Woody, Andy's toys live happily in hi..	led woodi andi toy live happili room andi birt..
Grumpier Old Men	A family wedding reignites the ancient feud be..	famili wed reignit ancient feud next door neig..

B. Feature Extraction

Pre-processing is complete, the film synopsis is extracted into a numeric vector. Because the results are quite large, they are not displayed and are directly used to calculate film equations. The resulting vector values show that every film has unique characteristics.

C. Cosine Similarity

The similarity between films. The hybrid similarity score, the ultimate metric for gauging content similarity across movies, serves as the foundation for generating suggestions. The results of the hybrid similarity calculation for one input film and several additional films with the highest degree of similarity are displayed in Table 3.

Table 3. Results of Hybrid Similarity Calculation

	Film	Hybrid Similarity
0	Toy Story 3	0.73
1	Toy Story 2	0.71
2	Welcome to Happiness	0.47
3	Child's Play 3	0.43

Toy Story 2 and Toy Story 3 have the highest similarity ratings to the input movies, according to Table 3. This shows that the system can identify substantial similarities in story content, particularly between movies in the same series or with similar locales. Lower similarity ratings for the other movies indicate greater variations in the story context.

D. Top-K Recommendation

Table 4.Spiderman Movie Recommendation Results (Top 10)

	title	genre_clean	overview
1	The Amazing Spider-Man	[action, adventure, fantasy]	Peter Parker is an outcast high schooler aband...
2	The Amazing Spider-Man 2	[action, adventure, fantasy]	For Peter Parker, life is busy. Between taking...
3	Spider-Man 2	[action, adventure, fantasy]	Peter Parker is going through a major identity...
4	Spider-Man 3	[fantasy, action, adventure]	The seemingly invincible Spider-Man goes up ag...
5	Spiderman: The Ultimate Villain Showdown	[action, animation, family, science fiction	Spider-Man meets some of his greatest foes inc...
6	Earth vs. the Spider	[horror, science fiction]	Teenagers from a small town and their high sch..
7	Earth vs. the Spider	[horror, science fiction]	A shy comic book fan is injected with an exper..
8	Superman	[action, crime, science fiction]	Superman comes to Earth as a child and grows u...
9	Chronicle	[science fiction, drama, thriller]	Three high school students make an incredible ...
10	Spider-Plant Man	[comedy]	Spider-Plant Man is a parody of Spider-Man, ma...

Table 4 demonstrates that the genres and narrative themes of the recommended films are comparable to those of the input film. This suggests that the system can use descriptive context and keyword occurrences to determine content similarities.

DOI: https://doi.org/10.37396/jsc.v9i1.633

E. Model Evaluation

Metrics used to assess the effectiveness of the system's recommendations include Precision@K, Recall@K, and Mean Average Precision (MAP@K). No novel approach is used in this assessment.

Table 5. Evaluation Results

Metric	Value
Precision@K	0.71
Recall@K	0.95
MAP@K	0.95

The evaluation findings for the recommendation system are shown in Table 5. Relevant films are consistently listed higher on the suggestion list, with a MAP@K value of 0.95. While the Recall@K value of 0.95 demonstrates that the system successfully retrieves the majority of relevant movies, the Precision@K value of 0.71 suggests that a significant fraction of the suggested movies are relevant. Overall, these findings show that the system offers precise suggestions that span a wide range of pertinent topics.

F. Comparison of Recommendation Methods

To evaluate the effectiveness of TF-IDF, SBERT, and hybrid approaches for producing relevant movie suggestions, the approaches are assessed. The metrics utilised for the assessment are MAP@K, Recall@K, and Precision@K.

Table 6. Result Method Comparison

Method	Precision@K	Recall@K	MAP@K
TF-IDF	0.10	0.05	0.06
SBERT	0.80	0.85	0.94
Hybrid (TF-IDF + SBERT)	0.71	0.95	0.95

Table 6 shows that the TF-IDF approach yields the lowest assessment results, with Precision@K of 0.10, Recall@K of 0.05, and MAP@K of 0.06. This suggests that the contextual similarity of movie material is not as well captured by TF-IDF, which depends on word frequency.

The Precision@K, Recall@K, and MAP@K values of 0.80, 0.85, and 0.94, respectively, indicate that SBERT is quite effective at recognising film similarities from their synopses. The hybrid approach (TF-IDF + SBERT) produces pretty good recommendations, with a Precision@K of 0.71 and the highest Recall@K and MAP@K values of 0.95.

DOI: https://doi.org/10.37396/jsc.v9i1.633

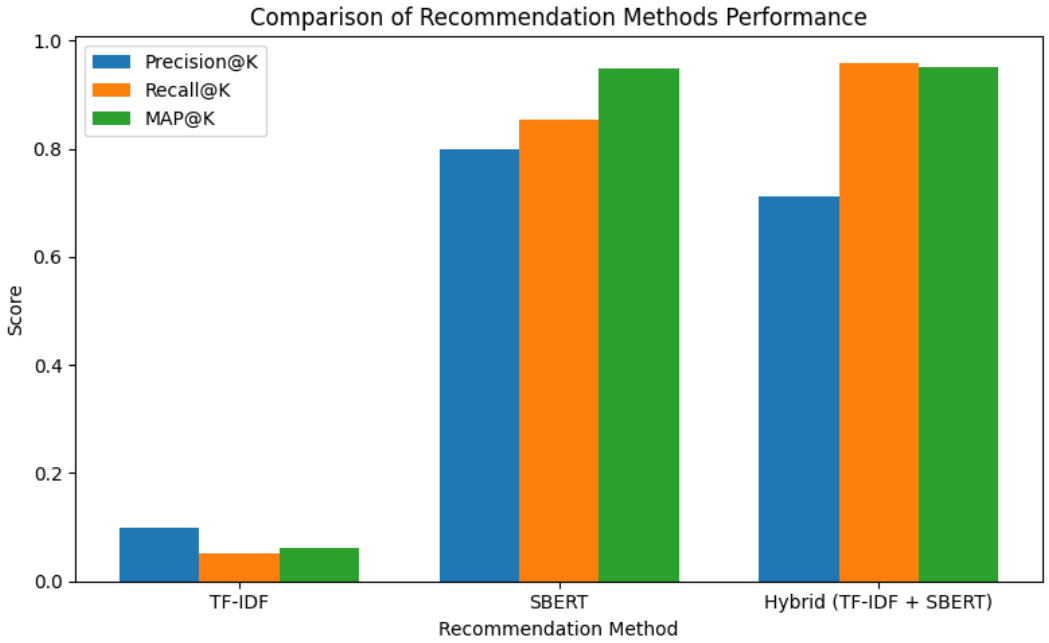

Figure 2. Comparison Chart of Recommendation Methods

Figure 2 compares the use of TF-IDF, SBERT, and hybrid methods in recommendation systems. TF-IDF performs the worst, according to the graph. The hybrid approach has the highest Recall@K and MAP@K, but Precision@K is marginally lower than SBERT. Based on the evaluation results, the hybrid approach shows more balanced performance compared to single methods, including SBERT.

IV. CONCLUSION

Based on the experimental results, the proposed content-based recommendation system successfully generates relevant movie recommendations by leveraging both synopsis and genre features. Furthermore, the hybrid approach demonstrates its effectiveness in text-based recommendation scenarios by achieving a balanced performance, with a Precision@K of 0.73, Recall@K of 0.75, and MAP@K of 0.60.

In the future, this system can be developed by adding user interaction data, such as ratings or viewing history, to make recommendations more personalized. In addition, other, more complex methods can be explored to improve the quality of recommendations.

REFERENCES

[1] V. Sandrya, W. Wasino, and D. Arisandi, "Sistem Rekomendasi Film Menggunakan Metode Multiple Attribute Utility Theory," *Computatio : Journal of Computer Science and Information Systems*, vol. 6, no. 1, p. 19, Jun. 2022, doi: 10.24912/computatio.v6i1.17081.

[2] A. D. Saputro and F. Amin, "Sistem Rekomendasi Content-Based Filtering Skincare Pria Di E-Commerce Shopee," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 7, no. 1, pp. 106–113, Jan. 2024, doi: 10.31539/intecoms.v7i1.8036.

[3] N. Azizah and A. F. Rozi, "Sistem Rekomendasi Produk Somethinc Menggunakan Metode Content-based Filtering," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 3, pp. 461–468, Jul. 2024, doi: 10.47233/jteksis.v6i3.1411.

[4] Hanafi *et al.*, "Improvement of E-commerce Recommender System Using Hybridization of Bert, Matrix Factorization and Attention Mechanism," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 5, pp. 725–740, 2024, doi: 10.22266/ijies2024.1031.55.
[5] H. Hartatik and A. Syafrianto, "Penerapan Model Sentence-bert Untuk Sistem Rekomendasi Buku Berbasis Konten Di Perpustakaan Digital," *Jurnal Dialektika Informatika (Detika)*, vol. 6, no. 1, pp. 12–19, Nov. 2025, doi: 10.24176/detika.v6i1.15916.
[6] M. A. Hafizh Fathuddin 1, E. Prakarsa Mandyartha 2, and A. Lina Nurlaili 3, "Penerapan Sentence-Bert dan Cosine Similarity untuk Pencarian Semantik Dokumen Skripsi dalam Format PDF," *Ranah Research : Journal of Multidisciplinary Research and Development*, vol. 8, no. 1, Oct. 2025, doi: 10.38035/rrj.v8i1.
[7] M. Abdul, H. Fathuddin, E. Prakarsa Mandyartha, and A. L. Nurlaili, "Penerapan Sentence-Bert dan Cosine Similarity untuk Pencarian Semantik Dokumen Skripsi dalam Format PDF," *R2J*, vol. 8, no. 1, 2025, doi: 10.38035/rrj.v8i1.
[8] A. A. P. Yudha, Munir, and Ani Anisyah, "Perancangan Sistem Rekomendasi Akomodasi pada Event Konser dengan Metode Hybrid Filtering," *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)*, vol. 4, no. 2, pp. 631–641, Jul. 2025, doi: 10.62712/juktisi.v4i2.493.
[9] A. Rizky Mangunsong, V. Sihombing, and I. Rasyid Munthe, "Pengembangan Sistem Rekomendasi Produk Berdasarkan Pola Pembelian dengan Pendekatan Algoritma Apriori," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 82–86, Jan. 2024, doi: 10.55338/jikomsi.v7i1.2718.
[10] T. Safitri, Y. Umaidah, and I. Maulana, "Analisis Sentimen Pengguna Twitter Terhadap Grup Musik BTS Menggunakan Algoritma Support Vector Machine," *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 28–35, Jul. 2023, doi: 10.30871/jaic.v7i1.5039.
[11] A. Rachmaniar1, S. Widayati2, and K. Rokoyah3, "Sistem Rekomendasi Produk E-commerce Menggunakan Collaborative Filtering Dan Content-based Filtering," *Journal of Information System, Informatics and Computing*, vol. 9, no. 1, pp. 1–15, Jun. 2025, doi: 10.52362/jisicom.v9i1.1904.
[12] R. M. Holis, P. E. P. Utomo, and B. F. Hutabarat, "Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 2, pp. 915–922, Oct. 2025, doi: 10.47709/brilliance.v5i2.7027.
[13] M. Y. Ridho and E. Yulianti, "From Text to Truth: Leveraging IndoBERT and Machine Learning Models for Hoax Detection in Indonesian News," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 10, no. 3, pp. 544–555, Sep. 2024, doi: 10.26555/jiteki.v10i3.29450.
[14] A. H. J. P. Juni Permana and Agung Toto Wibowo, "Movie Recommendation System Based on Synopsis Using Content-Based Filtering with TF-IDF and Cosine Similarity," *International Journal on Information and Communication Technology (IJoICT)*, vol. 9, no. 2, pp. 1–14, Dec. 2023, doi: 10.21108/ijoict.v9i2.747.
[15] A. Febrian and E. D. Permana, "Sistem Rekomendasi Film Menggunakan Metode Content Based Filtering Dengan Algoritma Tf-idf," *Al-Aqlu: Jurnal Matematika, Teknik dan Sains*, vol. 4, no. 1, pp. 19–25, Jan. 2026, doi: 10.59896/aqlu.v4i1.494.
[16] A. Serlina, A. Rahim, and Arbansyah, "Comparative Analysis of Naïve Bayes Algorithm Performance in English and Indonesian Text Sentiment Classification on Duolingo Application in Playstore," *Teknika*, vol. 14, no. 1, pp. 165–171, Mar. 2025, doi: 10.34148/teknika.v14i1.1207.
[17] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results in Engineering*, vol. 17, p. 100856, Mar. 2023, doi: 10.1016/j.rineng.2022.100856.

[18] P. Aprilio, M. Felix, P. S. Nugraha, and H. Fahmi, "Hybrid Feature Combination of TF-IDF and BERT for Enhanced Information Retrieval Accuracy," *JISA(Jurnal Informatika dan Sains)*, vol. 8, no. 1, pp. 8–15, Jun. 2025, doi: 10.31326/jisa.v8i1.2179.
[19] A. Maitaigahasse, J. L. K. Ebongue Fendji, and M. Atemkeng, "Offline Content-based Recommendation System for Wikimedia Commons Contents," *Procedia Computer Science*, vol. 257, pp. 485–494, 2025, doi: 10.1016/j.procs.2025.03.063.
[20] D. Velamentosa and E. Zuliarso, "Sistem Rekomendasi Film Menggunakan Metode Content-based Filtering," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 2918–2922, Mar. 2025, doi: 10.36040/jati.v9i2.13251.
[21] D. R. P. Noordi, H. Hasanah, and S. Sumarlinda, "Marvel Movie Recommendation System Using Hybrid Item-Based and Content-Based Filtering Methods," *TIERS Information Technology Journal*, vol. 5, no. 1, pp. 13–19, Jun. 2024, doi: 10.38043/tiers.v5i1.5209.
[22] L. Palupi, E. Ihsanto, and F. Nugroho, "Analisis Validasi dan Evaluasi Model Deteksi Objek Varian Jahe Menggunakan Algoritma Yolov5," *Journal of Information System Research (JOSH)*, vol. 5, no. 1, pp. 234–241, Oct. 2023, doi: 10.47065/josh.v5i1.4380.
[23] . Amdahl J. R. Sari, K. Sadik, A. M. Soleh, and C. Suhaeni, "Evaluasi Model Klasifikasi Dalam Deteksi Penipuan Transaksi: Studi Kasus Pada Data Tidak Seimbang," *Jurnal Gaussian*, vol. 14, no. 2, pp. 565–576, Dec. 2025, doi: 10.14710/j.gauss.14.2.565-576.
[24] D. Çelik Ertuğrul and S. Bitirim, "Job recommender systems: a systematic literature review, applications, open issues, and challenges," *Journal of Big Data*, vol. 12, no. 1, Jun. 2025, doi: 10.1186/s40537-025-01173-y.
[25] J. M. Azri Saputra, L. M. Huizen, and D. B. Arianto, "Sistem Rekomendasi Film pada Platform Streaming Menggunakan Metode Content-Based Filtering," *Jurnal Transformatika*, vol. 22, no. 1, pp. 10–21, Jul. 2024, doi: 10.26623/transformatika.v22i1.7041.