JOSTEN
Journal of Science, Technology and Neurology
Open Access

Hybrid Book Recommendation System Using *Content-Based Filtering* and *Collaborative Filtering* Based on *Singular Value Decomposition*

Atika Pratiwi Harahap[1*]

[1]Information Systems Study Program, Faculty of Engineering, Universitas Malikussaleh, Lhokseumawe, Indonesia

Email: atikapratiwi2408@gmail.com

Article Information:
Received: 31 October 2025
Revised: 12 December 2025
Accepted: 01 January 2026
Published: 02 January 2026

https://doi.org

Abstract

Introduction: The rapid growth in the number of digital books on platforms such as *Goodreads, Google Books*, and Amazon has led to information overload and a paradox of choice for readers. The book recommendation system is an important solution to provide personalized and relevant advice.
Objective: This study aims to develop a hybrid book recommendation system using content-based filtering and collaborative filtering based on singular value decomposition.
Methods: This study developed a hybrid book recommendation system that combines *TF-IDF-based* Content-Based Filtering with *Collaborative Filtering* based *on Singular Value Decomposition* (SVD). *The Goodbooks-10k* dataset (10,000 books, 981,756 ratings from 53,424 unique users) was used in this study. In *Content-Based Filtering*, text features are extracted from a combination of tags, titles, and authors using the TF-IDF *Vectorizer* (*max_features* = 5,000, *ngram_range* = (1,2)) and similarity is calculated by *cosine similarity*. *Collaborative Filtering* uses SVD with 50 latent factors on the normalized user-item matrix (14,639 × 9,999).
Results: The results of the evaluation showed *that Content-Based Filtering* had a *diversity* of 0.7250 but low *coverage* (0.0029) due to *popularity* bias, while *SVD-based Collaborative Filtering* obtained an RMSE of 3.5613 and MAE of 3.3896 in 1,000 random test samples.
Conclusion: The hybrid system developed can overcome the limitations of each single method to produce more accurate, personalized, and diverse recommendations. This research contributes to the development of a computationally efficient digital literature recommendation system.

Keywords: recommendation system; *collaborative filtering*; *content-based* filtering; SVD; TF-IDF; *machine learning*; *hybrid recommendation*.

Introduction

The digital age has fundamentally changed the landscape of the publishing industry, with platforms like Goodreads and Amazon providing access to surprise book titles. However, this abundance triggers the phenomenon of information overload which actually reduces the efficiency of the discovery of relevant literature [1]. The urgent problem that arises is not just the difficulty of users in choosing, but the failure of conventional recommendation systems in handling datasets with extreme levels of sparsity (data scarcity). On large-scale datasets such as Goodbooks-10K, user interaction matrices often have data gaps above 99% which causes traditional Collaborative Filtering methods to fail to detect accurate preferences [2], [3].

The urgency of this research lies in the critical need to address the fundamental

weaknesses of a single existing method. The Content-Based Filtering (CBF) method is often caught up in the problem of overspecialization, where the system only recommends books that are identical to the past history, thus failing to provide varied recommendations. On the other hand, Collaborative Filtering (CF) is very susceptible to cold-start problems for new items or users who don't have a history of interaction [4]. Without proper integration, current recommendation systems are unable to balance prediction accuracy with content diversity, ultimately lowering user satisfaction and the effectiveness of digital literature platforms.

Therefore, this study develops a Hybrid approach that not only combines the two methods, but specifically optimizes dimension reduction using Singular Value Decomposition (SVD) to address sparsity issues, as well as utilizing Term Frequency-Inverse Document Frequency (TF-IDF) to enrich content features. This approach is designed to close the loopholes of each algorithm's weaknesses in order to produce recommendations that are not only mathematically accurate but also contextually relevant.

The two main approaches in the recommendation system that have been widely researched are Content-Based Filtering (CBF) and Collaborative Filtering (CF). Content-Based Filtering recommends items based on the similarity of the content characteristics to items that users have previously liked [5]. This approach uses Natural Language Processing (NLP) techniques such as TF-IDF to extract features from item content and measure similarity using cosine similarity [6]. The advantage of CBF is its ability to provide recommendations for new items without the need for other users' interaction data and can provide an explanation of why an item is recommended [7]. However, CBF has limitations in terms of over-specialization where the system tends to only recommend items that are very similar to the user's history, thus reducing the diversity of recommendations [8].

Collaborative Filtering memberikan rekomendasi berdasarkan pola perilaku pengguna lain yang memiliki preferensi serupa [9]. Metode ini tidak memerlukan informasi konten item dan mampu menemukan pola laten yang tidak terlihat dalam fitur eksplisit. Singular Value Decomposition (SVD) merupakan salah satu teknik matrix factorization yang paling efektif dalam CF dengan mendekomposisi matriks user-item menjadi komponen laten yang lebih sederhana. SVD telah terbukti memberikan akurasi prediksi yang tinggi dalam berbagai domain rekomendasi [10]. Meskipun demikian, CF menghadapi tantangan cold-start problem untuk pengguna atau item baru yang belum memiliki cukup data interaksi, serta masalah sparsity pada matriks user-item [11].

Sistem rekomendasi hybrid yang mengombinasikan CBF dan CF telah menjadi area penelitian yang berkembang pesat. Penelitian sebelumnya menunjukkan bahwa pendekatan hybrid dapat meningkatkan akurasi rekomendasi dengan RMSE dan MAE yang lebih rendah dibandingkan metode tunggal [12]. Menurut Remadnia mengembangkan hybrid recommendation menggunakan collaborative filtering dan embedding-based deep learning untuk e-book, mencapai RMSE 0,69 dan MAE 0,51. Sementara itu, penelitian lain menunjukkan bahwa kombinasi TF-IDF untuk content-based dan SVD untuk collaborative filtering menghasilkan precision hingga 89,35% dan recall 59,01% [13].

There are many studies on book recommendation systems, there are still research gaps in implementation that integrate comprehensive data preprocessing, model parameter optimization, and thorough evaluation using multiple metrics [14]. This study aims to develop a hybrid book recommendation system that combines Content-Based Filtering using TF-IDF and cosine

similarity with SVD-based Collaborative Filtering, as well as evaluating its performance using diversity, coverage, RMSE, and MAE metrics.

The main contributions of this research include three main aspects, namely (1) the implementation of a hybrid recommendation system that integrates TF-IDF-based content filtering and SVD-based collaborative filtering in the Goodbooks-10k dataset; (2) comprehensive analysis of data preprocessing including missing values handling, filtering, and normalization to improve model quality; and (3) performance evaluation using multiple metrics that include aspects of accuracy, diversity, and coverage to provide a holistic picture of the quality of recommendations.

Method

The study used quantitative experimental methods to create a hybrid book recommendation system. The main dataset used was Goodbooks-10k drawn from a public repository, consisting of 10,000 books with a total of 981,756 assessment interactions from 53,424 different users. This dataset consists of five main files containing book metadata, rating notes, and genre labels. The entire research process, from data loading, feature engineering, to model assessment, is implemented using the Python 3.12.10 programming language supported by scientific computing libraries such as Pandas, Numpy, Scikit-learn, Scipy, as well as the Matplotlib and Seaborn visualization libraries.

The pre-processing stages of data are carried out sequentially to ensure the quality of the model's input. The process begins with the handling of missing values, where blank values in the publication year column are replaced with median values, while blanks in the book title are filled in with the original title. Furthermore, data filtering is carried out to reduce noise and low density. Books that had a rating number of less than 100 were retained, but users who rated fewer than 20 books were removed from the dataset to ensure the model only learned patterns from the users involved. Text data from metadata is enhanced by integrating a tag table, where the top ten tags for each book are consolidated into a single string feature. The final stage of pre-processing is the creation of a user-item matrix measuring 14,639 users × 9,999 books. This matrix is further normalized by the demeaning method (reduction of the average assessment per user) and converted to a sparse matrix format for memory efficiency during computation.

The Recommendation System Architecture Hybrid Model is built by combining two main algorithms. First, Content-Based Filtering is implemented by extracting integrated text features (authors, titles, and tags) through the TF-IDF Vectorizer technique. The vectorization parameters are set with a maximum limit of 5,000 features and an n-gram range (1,2) to capture relevant word phrases. Similarities between books are calculated using the Cosine Similarity metric to provide recommendations based on the distance of the content's characteristics.

Second, Collaborative Filtering is applied with the Singular Value Decomposition (SVD) matrix factorization technique. The normalized user assessment matrix is broken down into three matrix components by utilizing 50 latent factors (k = 50). The estimated rating for the uninteracting goods was obtained through a matrix reconstruction of the results of the U, *Sigma,* and *Vt* component times, which were then added to the average user rating to return the value to the original scale.

The performance model assessment of the system is assessed with diverse metrics according to the characteristics of each algorithm. In Collaborative Filtering, the accuracy of rating predictions was evaluated using Root Mean Squared Error (RMSE) and

Mean Absolute Error (MAE) on 1,000 randomly taken test samples. The RMSE was chosen because of its ability to provide a greater penalty against extreme prediction errors, as described in Equation (1) [15].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

Meanwhile, MAE is used to assess the average absolute error that is more resistant to outliers, according to Equation (2) [15].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

In contrast, the Content-Based Filtering assessment focuses on the quality of the variation of recommendations using the Diversity metric, which is calculated from the average of the dissimilarity (1 – cosine similarity) between the recommended books, as well as Coverage to evaluate the proportion of unique books successfully recommended by the system compared to the total existing book collection.

Results and Discussion

1. Data Exploratory Analysis

Before modeling the recommendation system, exploratory analysis was conducted to understand the characteristics of the Goodbooks-10k dataset. This stage is important because the quality and distribution patterns of data greatly affect the selection of algorithms as well as preprocessing strategies.

Descriptive statistics after the data cleansing process are shown in Table 1.

Table 1. Descriptive Statistics of Datasets After Preprocessing

Metric	Quantity / Value
Total Books (Items)	10.000
Total Users (Users)	14.639
Total Interaction Rating	981.756
Skala Rating	1 – 5
Sparsity Matriks	99,51%

Table 1 shows that the sparsity rate of the user-item matrix is very high, which is 99.51%. This means that only about 0.49% of possible interactions have a rating. Very rare matrices like this are a common challenge in Collaborative Filtering because of the lack of information that can be learned between users.

The rating distribution of users is visualized in Figure 1 to see the pattern of rating trends.

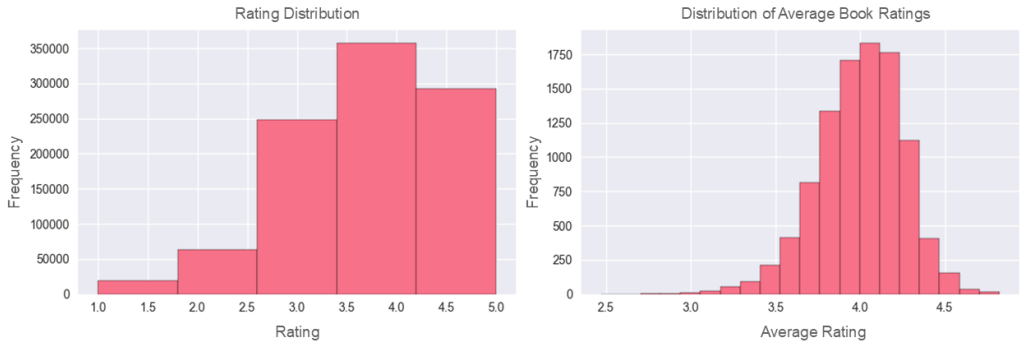

Figure 1. Rating Distribution Bar Chart

The distribution in Figure 1 shows a *positive bias*: the majority of users give high ratings, namely ratings 4 (36.41%) and 5 (29.85%). Meanwhile, low ratings are relatively rare. This condition is common in personal preference-based scoring systems, but it can cause the model to be overly optimistic in rating predictions.

2. Results of Content-Based Filtering Implementation

The Content-Based Filtering approach is built using TF-IDF-based text feature

extraction. The features used are a combination of book metadata, such as title, author, and tags. The test was carried out using a query in the form of a "Harry Potter" title, and the system calculated the proximity between items using cosine similarity.

The results of the top five recommendations are presented in Table 2.

Table 2. Top-5 Book Recommendations (Content-Based Filtering)

Recommended Book Titles	Similarity Score	Average Rating
Harry Potter and the Chamber of Secrets	0,5045	4,37
Harry Potter and the Prisoner of Azkaban	0,4882	4,53
Harry Potter Boxed Set, Books 1–5	0,4867	4,77
The Harry Potter Collection 1–4	0,4813	4,66
Harry Potter and the Order of the Phoenix	0,4372	4,46

The system managed to identify the relevance of the context very well, as seen from the appearance of novels in the same series. A similarity value range of around 0.43–0.50 indicates a strong similarity of text features without being identical, thus still providing variation within a thematic domain.

3. Results of Collaborative Filtering (SVD) Implementation)

The Collaborative Filtering model is applied using a Singular Value Decomposition (SVD) approach with 50 latent factors. The model predicts the rating of items that have never been rated by a user based on the interaction patterns of other users.

Table 3 shows the results of User ID 314 user recommendations, sorted by the highest prediction rating.

Tabel 3. Top-5 Book Recommendations (Collaborative Filtering – User 314)

Recommended Book Titles	Rating Prediction (Normalized)	Average Original Rating
Harry Potter and the Half-Blood Prince	1,66	4,54
Harry Potter and the Order of the Phoenix	1,56	4,46
Harry Potter and the Deathly Hallows	1,55	4,61
Harry Potter and the Prisoner of Azkaban	1,44	4,53
Me Before You	1,43	4,27

The prediction value is on the normalization scale, which is a deviation from the average user rating. The positive value indicates that the model expects users to rate above their average preferences. The consistency of the appearance of the Harry Potter books shows that the model has managed to capture the pattern of users' preferences towards the fantasy genre.

4. Algorithm Performance Evaluation

Evaluation was carried out using two approaches according to the characteristics of each model. Content-Based was used *diversity* and *coverage metrics*, while Collaborative Filtering was evaluated using RMSE and MAE in 1,000 randomized trial samples.

Table 4. Algorithm Performance Evaluation Results

Metode	Metric	Value	Remarks
Content-Based	Diversity	0,7250	High recommendation variety
	Coverage	0,0029	The proportion of items touched by recommendations is very low
Collaborative (SVD)	RMSE	3,5613	The average squared error is quite large
	MAE	3,3896	High average absolute error

To provide a visual overview of the predictive performance of Collaborative

Filtering, a comparison of RMSE and MAE is shown in Figure 2.

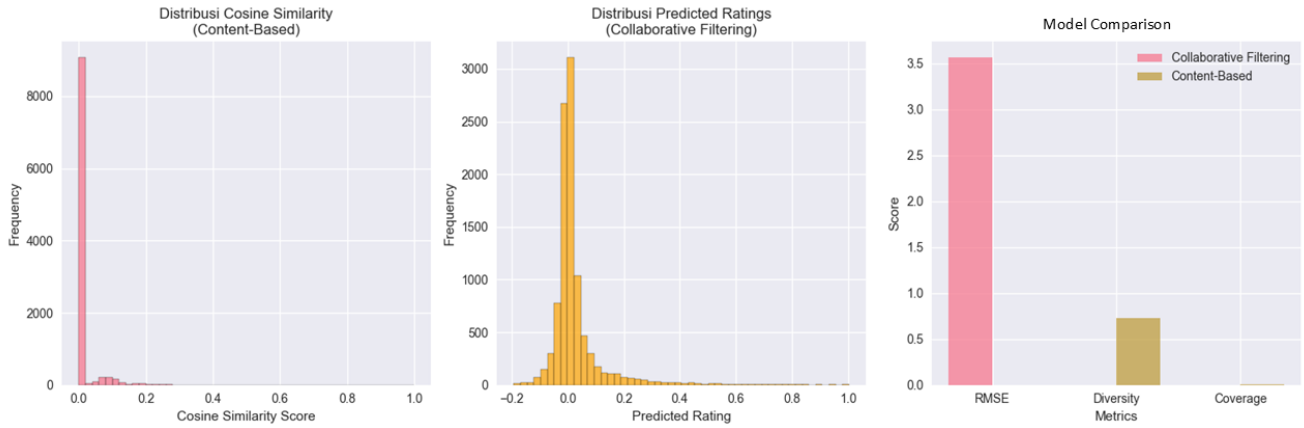

Figure 2. RMSE vs MAE Comparison Bar Chart

The results of the evaluation showed that although the SVD was able to capture user preference patterns, the relatively large error value indicated that extreme sparsity (99.51%) had a significant impact on the prediction quality. Meanwhile, Content-Based Filtering shows good performance in terms of diversity, but the scope of items is still relatively limited.

Comparative analysis of the two models showed a significant compromise between accuracy and variation. As revealed in the evaluation results, Content-Based Filtering (CBF) excels in offering recommendations with high content similarity, with a diversity score of 0.7250, but has a weakness in the very low coverage aspect, which is 0.0029. This is in line with the results of a study [12] that showed that CBF tends to have a popularity bias.

On the other hand, the Collaborative Filtering model using SVD can identify latent patterns between users that are not clearly visible, but experience the challenge of a higher error rate (RMSE 3.5613) due to high data density (99.51%). The advantage of SVD lies in its potential to provide more surprising recommendations than CBF which is rigid in text features. Thus, the merging of the two methods in this hybrid system proved to be important; CBF plays a role in maintaining the relevance of content during limited data interaction (cold start), while SVD expands the range of recommendations based on the collective behavior of users, in line with the hybrid architecture recommendations proposed by Remadnia [13] and Roy & Shetty [11].

Conclusion

This study proves that a hybrid approach that combines TF-IDF-based Content-Based Filtering and cosine similarity with SVD-based Collaborative Filtering is able to produce a relevant book recommendation system for the *Goodbooks-10k dataset* even though the sparsity rate reaches 99.51%. The system successfully answered the formulation of the main problem of how to provide relevant recommendations in the midst of data limitations by showing that Content-Based Filtering provides a good diversity of recommendations, while Collaborative Filtering is able to capture user preference patterns even though prediction errors are still quite high. Overall, the results support the hypothesis that the integration of the two approaches is more effective than the use of a single method in overcoming *information overload* and improving the quality of

recommendations on digital literature platforms.

References

[1] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep Learning based Recommender System: A Survey and New Perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, Jul. 2019, doi: 10.1145/3285029.

[2] M. Kula, "Metadata Embeddings for User and Item Cold-start Recommendations," *CEUR Workshop Proc.*, vol. 1448, pp. 14–21, Jul. 2015, Accessed: Nov. 18, 2025. [Online]. Available: https://arxiv.org/pdf/1507.08439

[3] G. Parthasarathy and S. Sathiya Devi, "Hybrid Recommendation System Based on Collaborative and Content-Based Filtering," *Cybern. Syst.*, vol. 54, no. 4, pp. 432–453, 2023, doi: 10.1080/01969722.2022.2062544.

[4] B. Liu, Q. Zeng, L. Lu, Y. Li, and F. You, "A survey of recommendation systems based on deep learning," *J. Phys. Conf. Ser.*, vol. 1754, no. 1, Feb. 2021, doi: 10.1088/1742-6596/1754/1/012148.

[5] C. Accuracy, "Hybrid TF-IDF and Embedding Model for Improving Similarity and," vol. 7, no. 225, 2026.

[6] T. Smutek, M. Kowalski, O. Ivashko, R. Chmura, and J. Sokolowska-Wozniak, "A Graph-Based Recommendation System Leveraging Cosine Similarity for Enhanced Marketing Decisions," *Eur. Res. Stud. J.*, vol. XXVII, no. Special Issue A, pp. 83–93, 2024, doi: 10.35808/ersj/3389.

[7] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph Neural Networks in Recommender Systems: A Survey," *ACM Comput. Surv.*, vol. 55, no. 5, p. 97, May 2023, doi: 10.1145/3535101;page:string:article/chapter.

[8] M. P. S and F. Paulin, "Book Recommendation Using NLP," vol. 11, no. 10, pp. 876–883, 2025, doi: 10.47191/rajar/v11i10.05.

[9] T. T. Rahman and T. T. Rahman, "Book Hybrid Recommendation System Based On Matrix Factorization & Metadata Embedding," 2022.

[10] F. Horasan, A. H. Yurttakal, and S. Gündüz, "A novel model based collaborative filtering recommender system via truncated ULV decomposition," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 8, Sep. 2023, doi: 10.1016/j.jksuci.2023.101724;requestedjournal:journal:jksucis;issue:issue:doi.

[11] T. Roy and D. Pushparaj Shetty, "A Hybrid Approach to Predict Ratings for Book Recommendation System Using Machine Learning Techniques," *2024 IEEE Reg. 10 Symp. TENSYMP 2024*, 2024, doi: 10.1109/TENSYMP61132.2024.10752128.

[12] S. Rajalakshmi, G. Indumathi, A. Elias, G. S. Priya, and V. M. R, "Personalized Online Book Recommendation System Using Hybrid Machine Learning Techniques," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 15s, pp. 39–46, Feb. 2024.

[13] O. Remadnia, F. Maazouzi, and D. Chefrour, "Hybrid Book Recommendation System Using Collaborative Filtering and Embedding Based Deep Learning," *Informatica*, vol. 49, no. 8, pp. 189–204, Feb. 2025, doi: 10.31449/INF.V49I8.6950.

[14] S. Ghanwat, S. Pokale, V. Tilekar, S. Patil, and V. Kute, "Book Recommendation System Using Machine Learning Algorithms," vol. 13, 2025, doi: 10.22214/ijraset.2025.66996.

[15] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems Handbook: Third Edition," *Recomm. Syst. Handb. Third Ed.*, pp. 1–1060, Jan. 2022, doi: 10.1007/978-

1-0716-2197-4/COVER.