

EVALUATION OF HYBRID MOVIE RECOMMENDATION SYSTEM BASED ON NEURAL NETWORKS

William Widjaja^{1*}, Robet¹, Johanes Terang Kita Perangin-Angin²

¹Informatics Engineering, STMIK TIME

²Information System, STMIK TIME

Email: *williamwidjaja12@gmail.com

Abstract: Recommendation systems are becoming increasingly important with the growth of streaming platforms. The purpose of this study is to compare the performance of Content-Based Filtering, Neural Collaborative Filtering, and a combination of both in a movie recommendation system. The method used in this study involves retrieving movie details from the TMDB API and ratings from the MovieLens 32M Dataset (2010-2023). Each model's performance is evaluated using evaluation metrics such as RMSE and MAE. The results of this study indicate that Neural Collaborative Filtering achieves the best prediction performance (RMSE = 0.785423, MAE = 0.581262), followed by the hybrid model (RMSE = 0.800863, MAE = 0.660872), while Content-Based Filtering produces low performance and limits the capabilities of the hybrid model. In conclusion, these findings highlight the superiority of latent feature-based models such as NCF that learn directly from user interaction patterns over content-based approaches in the context of modern recommendation systems.

Keywords: content-based filtering; hybrid filtering; movie recommendation; neural collaborative filtering.

Abstrak: Sistem rekomendasi menjadi semakin penting seiring berkembangnya platform streaming. Tujuan dari penelitian ini adalah membandingkan kinerja Content-Based Filtering, Neural Collaborative Filtering dan kombinasi keduanya dalam sistem rekomendasi film. Metode yang digunakan dalam penelitian ini melibatkan pengambilan detail film dari TMDB API dan rating dari dataset MovieLens 32M Dataset (2010-2023). Setiap performa model dievaluasi dengan menggunakan metrik evaluasi seperti RMSE dan MAE. Hasil dari penelitian ini menunjukkan bahwa Neural Collaborative Filtering mencapai kinerja prediksi terbaik (RMSE = 0.785423, MAE = 0.581262), diikuti oleh model hybrid (RMSE = 0.800863, MAE = 0.660872), sementara Content-Based Filtering menghasilkan performa yang rendah dan membatasi kemampuan model hybrid. Kesimpulannya, penelitian ini menyoroti superioritas model berbasis latent feature seperti NCF yang belajar langsung dari pola interaksi pengguna dibandingkan pendekatan berbasis konten dalam konteks sistem rekomendasi modern.

Kata kunci: content-based filtering; hybrid filtering; neural collaborative filtering; rekomendasi film.

INTRODUCTION

Technology is rapidly advancing, transforming how people access communication, information, and entertainment. Movies, in particular, are now diverse and readily available on various streaming platforms [1].

In 2023, IMDb recorded 30.443 new titles, making it almost impossible for audiences to choose what to watch and often leading to wasted time searching. To address this, recommender systems are needed. Recommender systems is an algorithm that suggest content based on user preference[2] [3].

In recommendation systems, Collaborative Filtering (CF) and Content-Based Filtering (CBF) are often used. CF stems from leveraging the behaviors and ratings of similar users to predict the behavior of a target user [4]. In contrast, CBF utilizes item descriptions, such as genres, crew, actors, or synopsis, to recommend similar items [5][6].

However, these approaches, especially CF, still face challenges such as data sparsity, cold start, and reliance on user feedback. To address this issue, hybrid methods combine both CF and CBF techniques [7].

Most traditional recommendation systems struggle with large-scale data analysis [8]. Researchers continuously develop methods to overcome challenges such as cold start problems and data sparsity, with neural network models offering innovative solutions [9][10].

Similarly, research [11] introduces a hybrid group recommender framework, where the state-of-the-art approach is the use of Fuzzy C-Means (FCM) clustering. Unlike traditional methods that assign a user to a single group, FCM allows them to belong to multiple groups, providing a better representation where user prefer-

ences overlap. Building on this, research [12] proposes a hybrid system where the state-of-the-art approach involves integrating a Deep Neural Network (DNN) to classify new users.

Meanwhile, research [13] proposes a multiview vision model that utilizes Transformers to integrate multiple data sources, including textual data from reviews. Furthermore, research [14] presents an ensemble system, highlighting the use of an Artificial Neural Network (ANN) as the final classifier

Collectively, these methods are consistently evaluated using standard metrics, with research [15] reaffirming RMSE and MAE as benchmark performance indicators.

Although previous studies have made significant progress, they lack a comprehensive analysis of existing algorithms. This research aims to compare the performance of a hybrid neural network-based recommendation system with standalone CBF and Neural Collaborative Filtering (NCF) approaches, using RMSE and MAE as evaluation metrics.

METHOD

The system will be developed by integrating three recommendation approaches. Content-Based Filtering, Neural Collaborative Filtering and a hybrid method that combines CBF and NCF. The first approach design is illustrated in Image 1.

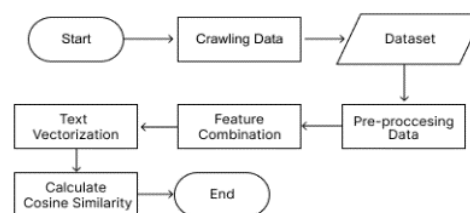


Image 1. Content-Based Flow

The second approach is the standalone NCF model in Image 2.

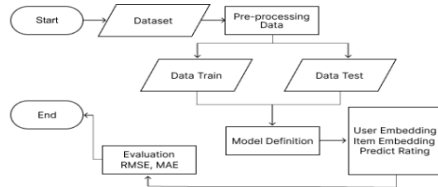


Image 2. NCF Flow

The third and final approach is Hybrid, which combines Content-Based and NCF as illustrated in Image 3.

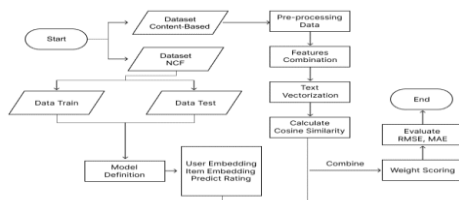


Image 3. Hybrid Flow

Dataset & Preprocessing Data

Movie attributes (title, cast, keywords, crew) were extracted from the TMDB API. At the same time, ratings were obtained from the MovieLens 32M dataset, which were then processed.

For efficiency and in consideration of the available resources, only movies that exist in both datasets were included. Table 1 shows the data before and after preprocessing (filtering movies from 2010 to 2023).

Table 1. Data Quantity

Source	Quantity	Final
Movies	87.585	35.509
Users	200.948	89.026
Rating	32.000.000	4.821.916

After preprocessing, the data will be split into a training and a testing set. Ratings in the training set will be normalized to the [0,1] range to match the sigmoid function used in the model's output.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

Description:

x is the input

e is Euler's number.

For ranking evaluation, each positive interaction in the training data is paired with four randomly chosen negative samples from items the user has not interacted with. Positive interactions are labeled as 1, while negative ones are labeled as 0. In a recommendation system, negative samples help reveal the user preferences [16].

Content-Based Filtering (CBF)

CBF methods quantify item similarity using metrics like cosine similarity. For example, if a user frequently watches action thrillers with a specific actor, the system will recommend similar movies featuring the same actors, director, or plot elements.

This approach assumes that past user preferences could predict future interest in items with similar characteristics.

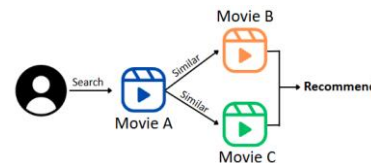


Image 4. Content-Based Illustration

Content-based systems often use cosine similarity to measure similarity between items. Cosine similarity quantifies the cosine of the angle between vectors.

$$\cos(\theta) = (A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (2)$$

Description:

$A \cdot B$ is the dot product of vectors A and B
 $\|A\| \|B\|$ are their respective magnitudes.

In CBF, vectors A and B represent the items being compared, which are created from features such as actors, genre, keywords, and directors, and then trans-

formed into a numerical representation using TF-IDF. This method weights terms based on their frequency within the item and across all items [17].

$$\text{TFIDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \quad (3)$$

Description:

t is the term of the word being analyzed
 d is to the document in which the term appears.

TF-IDF weight is calculated using two main components: Term Frequency (TF) and Inverse Document Frequency (IDF), each determining a term's importance within a document relative to the entire corpus.

$$\text{TF}(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (4)$$

Description:

$f_{t,d}$ is the number of times a term appears in a document

$\sum_k f_{k,d}$ is the total number of terms in the document.

$$\text{IDF}(t) = \log \left(\frac{N}{1 + n_t} \right) \quad (5)$$

Description:

N is the total number of documents

n_t is the number of documents containing the term.

Collaborative Filtering (CF)

A standalone CBF system has several limitations. It can only recommend similar movies. Meanwhile, the CF system operates differently, providing personalized recommendations based on user-item interactions rather than prior knowledge or history [18].

For example, if user 1 likes movies A, B, and C, and user 2 likes movie B, the system may predict that user two will also enjoy movies A and C.

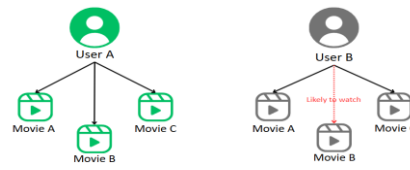


Image 5. CF Illustration

CF can be categorized into item-based and user-based approaches. Item-based CF recommends new items to a user based on similarities between items, while user-based CF recommends items by identifying users with similar preferences.

While effective, traditional CF often relies on simple matrix factorization, which may miss complex, non-linear user-item relationships. Neural Collaborative Filtering addresses this issue using neural networks to provide more accurate and personalized recommendations.

Neural Collaborative Filtering

Neural Collaborative Filtering (NCF) was introduced in 2017. NCF combines deep learning with collaborative filtering. It uses an embedding layer to capture latent features of users and movies. NCF consists of two components, generalized Matrix Factorization (GMF and Multilayer Perceptron (MLP). The NCF architecture is presented in Image 6.

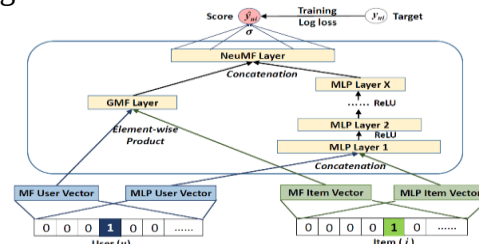


Image 6. NCF Architecture

Root Mean Square Error (RMSE)

Root Mean Squared Error (RMSE) is a metric used to evaluate the average difference between predicted values and actual values, expressed in squared form. The lower the RMSE value, the more ac-

curate the algorithm model is in making predictions [19].

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

Description:

n is the number of data points

y_i is the actual i -th value

$\hat{y}_i$ is the i -th predicted value.

The term $(y_i - \hat{y}_i)^2$ is the squared difference between the actual and predicted values.

RMSE is more sensitive to outliers because all errors are squared, making it suitable for use when errors truly need significant attention.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values. Like RMSE, the smaller the value, the more accurate the algorithm model is in making a prediction.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

Description:

N is the number of data points

y_i is the i -th predicted value.

The term $|y_i - \hat{y}_i|$ is the absolute value of the difference between the actual and predicted values.

MAE is not as sensitive to outliers. The primary reason is that it only calculates the absolute value of errors, making it more suitable for a stable interpretation of average errors.

RESULT AND DISCUSSION

The first step of this research involves building a content-based system by computing item similarities using cosine similarity and a TF-IDF vectorizer.

The second step is developing a neural collaborative filtering system using existing user ratings. Once both systems are established, they are combined into a hybrid system through weighted scoring.

Content-Based Filtering Result

The CBF approach in recommendation systems analyzes items based on their attribute similarities, but typically cannot predict numeric ratings. An illustration of how CBF works is shown in Table 2.

Table 2. Content-Based Examples

	The Avengers	Captain America: Civil War
The Avengers	1.0000	0.7179
Black Panther	0.6363	0.6370

Movies with the highest similarity score will form the top N recommendations, regardless of user preferences, so not all similar films will appeal to every user.

Since this research aims to compare methods using RMSE and MAE quantitatively, CBF is unable to predict ratings directly; therefore, it needs to be combined with a regression method to estimate user ratings.

Table 3. RMSE and MAE Evaluation

RMSE	MAE	Accuracy
0.995792	0.852235	0.405239

The results show a relatively high RMSE (0.995792) and MAE (0.852235), indicating that lower values would be preferable for better prediction accuracy. This shows that its rating predictions deviate more from actual ratings, supporting the claim that CBF struggles due to reliance on item attributes rather than user-to-user behavior.

Therefore, RMSE and MAE should be interpreted cautiously for CBF, with top N recommendation evaluation being a more suitable approach. The top 5 recommendations are shown in the Image 7.

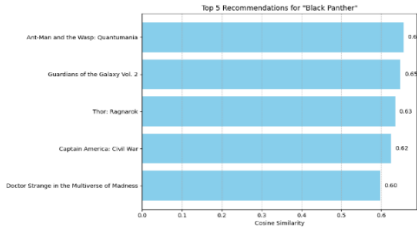


Image 7. Content-Based Results

Neural Collaborative Filtering Result

Neural Collaborative Filtering (NCF) models address complex non-linear user-item interactions by combining GMF for linear interactions and an MLP with ReLU for non-linear interactions. It is trained for rating prediction on normalized ratings [0, 1] and for ranking evaluation using a 1:4 negative sampling ratio, and tested by comparing each positive item against 99 negatives. The hit ratio is reported for reference but is not the primary evaluation metric.

Table 4. RMSE and MAE Evaluation		
RMSE	MAE	Hit Ratio
0.785423	0.581262	94.76%

The NCF model achieved an RMSE of 0.785423 and an MAE of 0.581262, reducing the error by 21% and 31.8% compared to the CBF model. It also reached an HR@10 of 94.76%, indicating that it often includes at least one relevant item in the top 10; however, this metric does not fully reflect the quality. Overall, this result demonstrates that NCF significantly outperforms CBF. For the top N recommendations, we will take user 28 as an example.

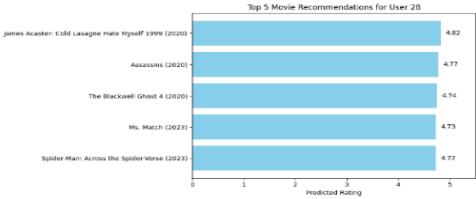


Image 8. NCF Result

Hybrid Filtering Result

The hybrid approach combines CBF and NCF using a weighted scoring of 0.7 NCF and 0.3 CBF. The CBF component primarily addresses cold start users who have not provided any ratings. Nonetheless, RMSE and MAE are still reported to maintain consistency in comparison across methods, which motivates the use of weighted scoring ability.

Table 5. RMSE and MAE Evaluation		
RMSE	MAE	Accuracy
0.800863	0.660872	0.479683

The evaluation shows that the hybrid model achieved an RMSE of 0.800863 and an MAE of 0.660872, lower than the CBF method but slightly higher than the NCF method.

The hybrid model outperforms CBF in accuracy but lags slightly behind NCF, showing that combining methods can improve performance. CBF, adapted here as a regression with weighted scoring, limits the hybrid’s effectiveness. In practice, hybrids utilize CBF to deal with sparse user ratings, while NCF is employed when sufficient data is available.

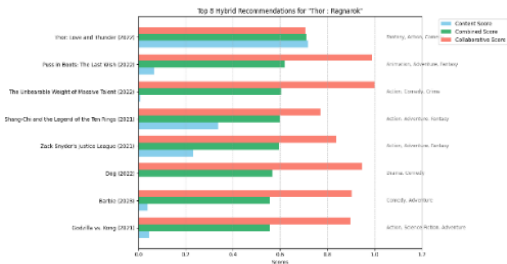


Image 9. Hybrid Result

ABLATION ANALYSIS

Ablation analysis is a method in which specific model components or features are selectively removed to evaluate their impact on overall performance, thereby helping to identify critical elements and understand the model's behavior.

Neural Collaborative Filtering

Ablation analysis was performed by evaluating the individual contributions of the GMF and MLP, which were trained and tested under identical conditions.

Table 6. RMSE and MAE Evaluation

Variant	RMSE	MAE
GMF	1.046584	0.806133
MLP	0.779635	0.575131

The results indicate that the MLP pathway performs slightly better than the complete model alone. In contrast, the GMF exhibits significantly lower performance compared to both the complete model and the MLP. This suggests that the dataset primarily exhibits non-linear patterns, which allows MLP to achieve superior results. However, the poor performance of GMF negatively affects the overall performance of the whole model. Meanwhile, the results of ablation on the negative sampling ratio:

Table 7. Negative Sampling Evals

	Easy Negative	Hard Negative
Hit Ratio	94.76%	62.24%

Easy negatives are generated through purely random negative sampling, where items are randomly selected under the assumption that they have not been interacted with by the user. In contrast, hard negatives are constructed by considering item popularity, making them difficult for the model to distinguish from actual positive.

Hybrid Filtering

The ablation analysis, we varied the weights between NCF and CBF components to assess their impact on recommendation performance.

Table 8. RMSE and MAE Evaluation

Weight	RMSE	MAE
0.7+0.3	0.804104	0.664280
0.3+0.7	0.890555	0.755020

Ablation analysis with varied NCF-CBF weights reveals that a higher NCF weight yields the best performance. CBF signals alone cannot generate ratings, so emphasizing CBF too much harms prediction accuracy. NCF effectively captures collaborative patterns, while CBF can serve as auxiliary information without disrupting rating prediction.

ERROR ANALYSIS

This part, to gain deeper insights beyond aggregate metrics, we examined individual prediction cases in detail. Tables 9 and 10 present examples where NCF achieved perfect accuracy as well as cases where it produced the most significant errors, illustrating both the strengths and limitations of the model.

Table 9. NCF Excels

UID	Title	Predicted	Actual
69910	Star Trek	3.0	3.0
76079	The Martian	4.0	4.0

Table 10. NCF Fails

UID	Title	Predicted	Actual
22564	Get Out	4.9	0.5
69634	Licorice Plaza	4.8	0.5

The results show that NCF can accurately predict some ratings, but also overestimate in extreme cases. This indicates that it captures user-item patterns well but can be overly confident, reflecting issues such as outliers.

CONCLUSION

The research demonstrates that NCF, which combines GMF and MLP, outperforms both CBF and Hybrid approaches. The weak performance of CBF negatively affected the Hybrid results. Ablation analysis further revealed that NCF's high performance is primarily driven by the MLP component, which excels at capturing non-linear interactions in large datasets. Although Hit Ratio was not the primary evaluation metric, NCF still achieved a strong score of 94.76%. However, when the negative sampling strategy was changed, the Hit Ratio dropped significantly to 62.24%, showing the model's sensitivity to the choice of sampling approach.

BIBLIOGRAPHY

- [1] Anang Furkon Rifai and Erwin Budi Setiawan, "Memory-based Collaborative Filtering on Twitter Using Support Vector Machine Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 702–709, Oct. 2022, doi: 10.29207/resti.v6i5.4270.
- [2] A. Nilla and E. B. Setiawan, "Film Recommendation System Using Content-Based Filtering and the Convolutional Neural Network (CNN) Classification Methods," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 10, no. 1, p. 17, Feb. 2024, doi: 10.26555/jiteki.v9i4.28113.
- [3] P. A. Sedyo Mukti and Z. K. A. Baizal, "Enhancing Neural Collaborative Filtering with Metadata for Book Recommender System," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 19, no. 1, p. 61, Jan. 2025, doi: 10.22146/ijccs.103611.
- [4] L. Wu, X. He, X. Wang, K. Zhang, and M. Wang, "A Survey on Accuracy-oriented Neural Recommendation: From Collaborative Filtering to Information-rich Recommendation," Apr. 2021, doi: 10.1109/TKDE.2022.3145690.
- [5] M. A. Pradana and A. T. Wibowo, "MOVIE Recommendation System Using Hybrid Filtering With Word2vec And Restricted Boltzmann Machines," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 1, pp. 231–241, Feb. 2024, doi: 10.29100/jupi.v9i1.4306.
- [6] D. Velamentosa, E. Zuliarso, and J. Raya Tri Lomba Juang, "Sistem Rekomendasi Film Menggunakan Metode Content-Based Filtering," 2025.
- [7] J. Aisyiah and L. Cahyani, "Sistem Rekomendasi Program Studi Menggunakan Metode Hybrid Recommendation (Studi Kasus: MAN Sumenep)," *Jurnal Eksplora Informatika*, vol. 12, no. 1, pp. 59–72, Jan. 2024, doi: 10.30864/eksplora.v12i1.992.
- [8] B. Drammeh and H. Li, "Enhancing neural collaborative filtering using hybrid feature selection for recommendation," *PeerJ Comput Sci*, vol. 9, 2023, doi: 10.7717/peerj-cs.1456.
- [9] C. Li, I. Ishak, H. Ibrahim, M. Zolkepli, F. Sidi, and C. Li, "Deep Learning-Based Recommendation System: Systematic Review and Classification," 2023, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2023.3323353.
- [10] D. Laras and H. Hasrullah, "Analisis Kinerja Sistem Rekomendasi Film Berbasis Deep Learning Menggunakan Model Neural Network Pada Dataset MovieLens," *Jurnal Locus Penelitian dan Pengabdian*, vol. 4, no. 1, pp. 1047–1054, Jan. 2025, doi: 10.58344/locus.v4i1.3768.