JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 4 No. 2 (2023)
https://doi.org/10.35877/454RI.jinav1954

Social Media (Twitter) Based Movie Recommendation System On Disney+ With Hybrid Filtering Using Neighboor's K-Nearest Method

Azrina Fazira Ansshory* & Erwin Budi Setiawan

Faculty of Informatics, Telkom University, Bandung, Indonesia

Abstract

The research aims on the development of a film recommendation system that combines the Hybrid Filtering and K-Nearest Neighbors method. (KNN). Hybrid Filtering combines a variety of recommendation techniques, including collaborative filtering and content-based filtering, while KNN is a simple method for classifying data based on similarities with close neighbors. The data set used is Kaggle's Rotten Tomatoes, which includes information such as critics' names, genres, movie titles, and review content. The aim of the study was to build an accurate system of recommendations based on user ratings on Disney+ Hotstar and measure its performance using MAE (Mean Absolute Error) and Confusion Matrix assessments.The results showed that the combination of Hybrid Filtering and KNN methods resulted in better accuracy values in giving film recommendations compared to using only the Collaborative Filtering method. Graphics and performance analysis show that the developed models are able to provide film recommendations with increasing accuracy over time. In conclusion, the combination of Hybrid Filtering and K-Nearest Neighbors methods is effective in improving the accuracy of the movie recommendation system, helping users choose films that match their preferences on the Disney+ Hotstar platform. This research contributes to the development of better and more accurate recommendation systems in the film industry.

Keywords: film recommendation systems, hybrid filtering, information technology, telecommunications, K-Nearest Neighbors (KNN)

Received: 15 August 2023 Revised: 23 October 2023 Accepted: 3 November 2023

1. Introduction

Information and telecommunications technologies continue to evolve rapidly, and this has a major impact on various aspects of human life, including in the field of film making. Movies have become one of the most popular forms of entertainment in society (Ali, 2020; Kastouni & Ait Lahcen, 2022). In the past, watching movies could only be done via television, but with today's technological advances, we can watch movies via the internet with a variety of film options according to individual preferences. One of the most popular movie streaming platforms is Disney+ Hotstar, launched by the Walt Disney Company in 2019 in the United States (Adrien et al., 2017). The platform provides a digital streaming service thatins Disney's reputation as a family-friendly platform by not displaying rated content (Adrien et al., 2017; Xu, 2022).

The development of diverse film genres today often confuses the audience in choosing films that are being popular (De Groote, 2011). Therefore, it is necessary to have a system of film recommendations that can help users in choosing films that match their preferences. A recommendation system is a system that can detect user needs and preferences based on recommendations that are tailored to previous user profiles and behaviors (Ellyzabeth Sukmawati et al., 2022; Lilleberg et al., 2015; Wang et al., 2016).

One of the interactions frequently carried out by users is through social media, and one of the most popular social media platforms today is Twitter (Chen et al., 2022; Mohan Kumar et al., 2022). Twitter is one of the largest social networking sites used by many people to spread news, share articles, and socialize with others globally (Gupta et al., 2020). In giving recommendations on film recommendation systems, there are several methods that can be used, one of them is

* Corresponding author.
E-mail address: azrinafazira@student.telkomuniversity.ac.id

content-based methods (content-based), collaborative filtering, and hybrid based methods (Prasetya, 2017; Wang et al., 2021).

In this study, the authors attempted to combine between Hybrid Filtering and K-Nearest Neighbors KNN, hybrid filtering is a system that combines several recommendation techniques to produce recommendations that fit the needs of the user. In Hybrid Filtering the author combines two methods of collaborative filtering and content-based filtering, Collaborative Filltering is a method that filters and evaluates an item based on the opinions of other users, while content based filtration is a recommendation system based on user preference profiles and item descriptions (Findawati et al., 2019; Lilleberg et al., 2015; Pisarenko & Pisarenko, 2022; Shakya & Dave, 2022; Wang et al., 2016). With the results of research combining these two methods, it is expected to produce a recommendation system that meets the needs of the user.

The limitation of the problem in this research is to use the Dataset that Rotten Tomatoes available on the Kaggle website that contains Critic_name, genre, movie_title,danreview_content that has been minimized by 20,000 data and does not consider user reviews from other websites taking into account user reviews. The aim of this research is to know the process of applying the Hybrid Filtering and K-Nearest Neighbors methods in providing recommendations based on user ratings on Disney+ Hostar for each data used and obtaining results from the combination of the hybrid filtering method with K-nearest neighbors.

2. Research Methods

Where in this assessment there will be two methods (Bairagi, Vinayak; Munot et al., 2019), namely KNN and the combination of intercollaborative filtration combined with Content-based Filltering. There are several stages to complete this trial such as, data preprocessing, Hybrid recommendation, hybrid and E merger, as well as K-Nearest Neighbors classification and modeling evaluation.

2.1. Data Collection

For this assessment, use the Rotten Tomatoes dataset available on the Kaggle website. The data set used contains Critic_name, genre, movie_title, and review_content, where on review_content data that contains user opinions about the film will be converted into numbers to give rating on the film.

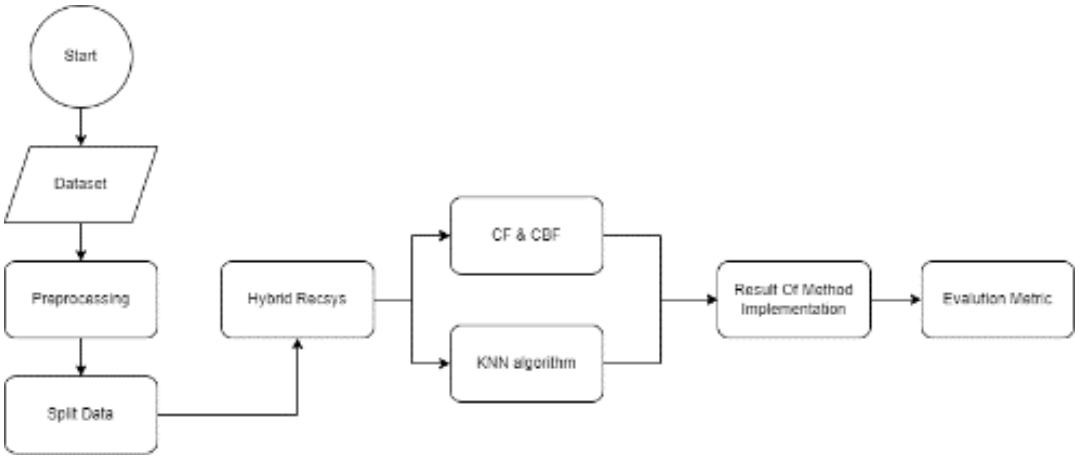

Figure 1. Planning for a Hybrid Filtering System and K-Nearest Neighbors

2.2. Preprocessing

Preprocessing data is the stage of processing raw data into more efficient data. At this stage, the data will be filtered by removing unnecessary data, from the rating given by the user to the film, so as to obtain more structured results. This rating can be used as a recommendation system.

Tabel 1 Dataset

movie_title	genres	ratings	rater_label	movie_label
Disney's A Christmas Carol	Animation, Drama, Kids & Family	6.0	1.0	1.0
Disney's A Christmas Carol	Animation, Drama, Kids & Family	6.0	2.0	1.0
Disney's A Christmas Carol	Animation, Drama, Kids & Family	4.0	3.0	1.0

2.3. Model Hybrid Filtering

Hybrid Filtering is one of the best hybrid filtering that uses content-based filtering, ccolaborative filtering and demographic suggested to address the 'cold start' problem. To predict new user rankings and therefore to find items similar to the environment, hybrid filtering uses demographic details. This proposed approach uses the advantages and overcomes the shortcomings of existing recommendation methods such as CB and CF.

This method is used at least to eliminate inefficiencies. For example, the CF method does not use product properties, and only uses user interactions that, given the fact that new users have little interaction with the system, this method doesn't work well in giving accurate recommendations to it. Thus, with the combination of CF and CBF we can more accurately understand the desires of the user and provide more effective recommendations.

2.4. K-Nearest Neighboors

The K-Nearest Neighbor algorithm (K-NN) is one of the simplest methods for solving classification problems. KNN is a classification method that trains model based in supervised learning. By defining a new class based on a common function, with the value K referring to the number of neighbors 3.5 Evaluasi Sistem

At this stage, the performance of the system will be evaluated using Hybrid Filtering using the K-Nearest Neighbor method that aims to see the results of the predicted film rating, The lower the MAE, the higher the accuracy can be produced. Whether it matches the original rating and to see which method is more effective. Using the Confusion Matrix equation to calculate the value of the classification result with the result of the minimum value MAE

2.5. Performance Measurement

Mean Absolute Error (MAE) and Confusion Matrix. MAE is a method used to measure the accuracy of a predictive model. The MAE value indicates the absolute error average between the prediction results and the real value, because the MAE is less sensitive to outliers and is a good marker for models with large weights that can make big changes in the data.

$$MAE = \frac{\sum(pi-qi)}{N} \quad (1)$$

where:

pi = predicted rating
Qi = Actual Rating
N = number of original predicted rating pairs

(MAE) is used to evaluate each algorithm and ultimately produce the top N predictions for the user. The MAE score will tell you many different models against the predicted targets. The goal of the MAE is to provide a model result with the minimum value with a better result (Mikolov et al., 2006). In addition to using MAE, the writer also uses the Confusion Matrix. The Confusion Matrix is used to determine the level of accuracy performed by the system, whether the effectiveness of the recommended results is appropriate.

In the table 2, we can see several results obtained from several equations: True positive (TP), False positive (FP), false negative (FN), and True negative.(TN). If the film recommendation is appropriate, it will get a True Positive (TP) value and if the recommended film is not suitable it will receive a True Negative value.(TN). If the film is not recommended

and not suitable, then the value is given False Negative (FN) and if the movie is not suggested and suitable then it is given the value False Positive (FP) (Purbolaksono & Suryani, n.d.).

Table 2. Confusion Matrix

Model		Actual
Predict	True	False
True	TP(True Positive)	FP(False Positive)
False	FN(False Negative)	TN(True Negative)

Several values will be obtained that will help evaluate the results of the classification among others are:

Accuracy is the result of a correct sum divided by the total amount of data, which can be calculated by the eq. 2.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is the comparison of the value of a positive category divided by the total of positive data that is true or not, can be calculated by the eq. 3.

$$precission = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Recall shows the total amount of correct results given by the system, can be calculated by the eq. 4.

$$recall = \frac{TP}{TP+FN} \quad (4)$$

The F1 score is the result of considering the value of the precision and recall ratio, which can be calculated with the eq.5.

$$F1\ Score = \frac{2x(recall\ x\ precission)}{(recall+precission)} \quad (5)$$

2.6. Precision

Precision is a test indicator that describes how accurately models predict positive events through a variety of predictive operations. The presidium value is calculated using the eq. 6.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

2.7. Loss

Loss is a test indicator that explains the calculations that cause the system's inaccuracies in the identification of objects. Loss can be calculated using the eq. 7.

$$\text{Loss} = H\ (p,\ q) = -\sum_{i=0}^{n} p(x_i) \ln q(x_i) \quad (7)$$

2.8. Recall

Recall is a test indicator that determines the success of the type of image already identified. The recall value can be calculated using the eq. 8.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

2.9. F1-score

F1 scores are calculations of evaluations in search information combining recall results and accuracy. In some cases, the recall and precision values may have different weights. The F1 score can be calculated using the eq. 9 (Kadhim, 2018).

$$F1 - Score = 2x \quad (9)$$

3. Results and Discussion

3.1. Data Result

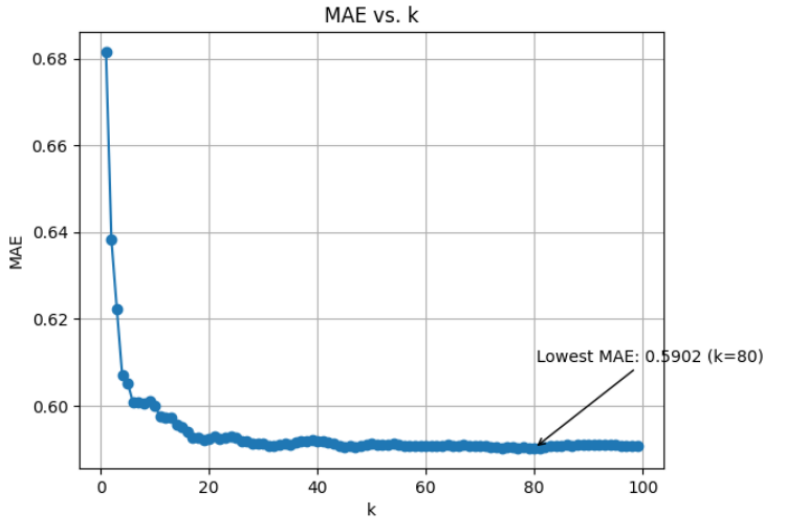

Figure 2. Collaborative filtering method

These results indicate that a recommendation system that uses the Collaborative Filtering method with KNN and the parameter K = 80 performs quite well in predicting user preferences based on patterns of collaboration between users. However, the performance of the recommendation system can vary depending on the nature of the dataset and the application context. It should be noted that comparisons with other methods and further analysis may be necessary to give a more complete picture of the effectiveness of these systems in a particular situation.

3.2. Test Results

The system testing on this study uses python and will run using Google Colab (Bisong, 2019; Paper, 2021). There are two processes used in this system: data training that serves as a data training model and testing as data testing. The aim of testing this system is to find the best performance results of the best K value Collaborative filtering and content-based goal is to obtain the optimal accuracy value.

3.3. Analysis of Test Results

Based on the image 3 parameter testing using bd with MAE 0, 2936 obtained the value of K for K = 10. This may indicate showing that the recommendation model has a good level of accuracy in predicting film rankings by users. The lower the MAE value, the less prediction error, and the recommendation becomes more consistent with user preferences.used to predict sentiment based on user reviews. The main objective of the study was to test the effectiveness of combining Hybrid Filtering and KNN methods in providing film recommendations based on user ratings on Disney+ Hotstar. The test results showed that this model has a MAE value of 0.2936 and the best K for KNN is 10. Low MAE values indicate good prediction accuracy.

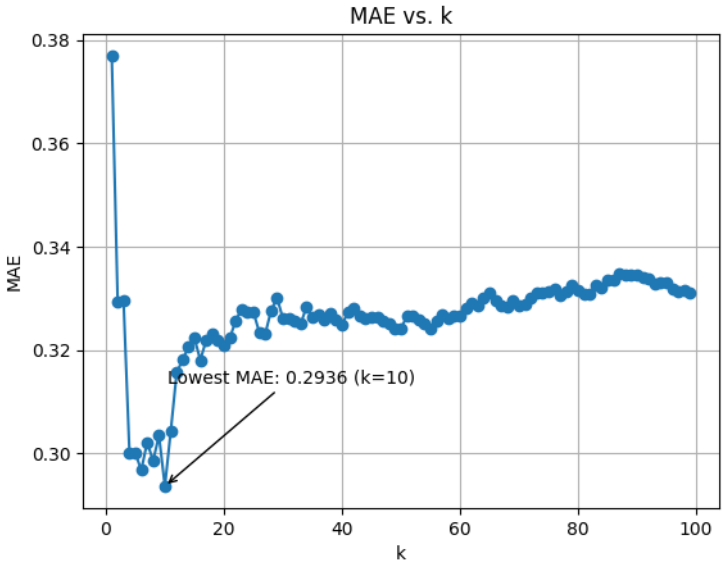

Figure 3. Content-based Filtering Method

Thus, these results indicate that recommendation systems that use the Content-Based Filtering method with KNN and the parameter K = 10 perform quite well in predicting user preferences based on the attributes or features of the item. However, it is important to keep in mind that this system of recommendations may be more suitable for situations where the characteristics or content of the item have a significant impact on user preferences, and this quality of performance needs to be further analyzed in a more specific application context.

Table 3. Hybrid recomendation result

Movie id	Title	Genre	.....	Similarity
168.	The film has the feel of ipad	Action	....	1.0
199	The scenes just dragon and seem	Action	....	1.0

A hybrid recommendation system using Collaborative Filtering (CF) and Content-Based Filtering(CBF) will give a film recommendation that has a similarity of 1.0. This means that the recommendation fits perfectly with similar user preferences (CFs) and has identical content features (CBF). However, this is rare and may indicate a problem. The main objective is to a balance between the two methods to give good and diverse recommendations.

4. Conclusion

Research shows that the Collaborative Filtering method with KNN and the K=80 parameter is effective in predicting user preferences based on collaborative patterns, although performance may vary depending on the data set and application context. Parameter testing showed that the recommended model has a good accuracy rate with an MAE of 0.2936 and the best K is 10 in predicting film ratings by users. The hybrid recommendation system that combines Hybrid Filtering and KNN also provides positive perf**ormance with a low MAE value (0.**2936) and the best K for KNN is 10. However, this hybrid recommendation system may be more suitable for situations where the characteristics or content of the item have a major influence on user preferences. The use of the Hybrid Filtering (CF+CBF) method in film recommendations may result in a similarity of 1.0, but this is a rare case and may indicate a problem. The goal is to balance and variation in recommendations so that better results can be achieved.

References

Adrien et al. (2017). Walt Disney Company Case Study. *Global Strategic Management*.

Ali, E. (2020). Geographic Information System (GIS): Definition , Development , Applications & Components. *Academia*, *79*.

Bairagi, Vinayak; Munot, M. V, JE, P. C., Sultan, M. T., Selvan, C. P., Irulappasamy, S., Mustapha, F., Basri, A. A., Safri, S. N. A. N. A., Taamneh, S., Tsiamyrtzis, P., Dcosta, M., Buddharaju, P., Khatri, A., Manser, M., Ferris, T., Wunderlich, R., Pavlidis, I., Rastgoo, M. N., Wang, X. X., … Nichele, S. (2019). Research Methodology A Practical and Scientific Approach. In *Transportation Research Part F: Traffic Psychology and Behaviour* (Vol. 4, Issue March 2019).

Bisong, E. (2019). Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. https://doi.org/10.1007/978-1-4842-4470-8_7

Chen, K., Duan, Z., & Yang, S. (2022). Twitter as research data. *Politics and the Life Sciences*, *41*(1). https://doi.org/10.1017/pls.2021.19

De Groote, P. (2011). Globalisation of commercial theme parks case: The Walt Disney Company. *Applied Studies in Agribusiness and Commerce*, *5*(3–4). https://doi.org/10.19041/apstract/2011/3-4/2

Ellyzabeth Sukmawati, Iwan Adhicandra, & Nur Sucahyo. (2022). Information System Design of Online-Based Technology News Forum. *International Journal Of Artificial Intelligence Research*, *1.2*. https://doi.org/https://doi.org/10.29099/ijair.v6i1.2.593

Findawati, Y., Astutik, I. R. I., Fitroni, A. S., Indrawati, I., & Yuniasih, N. (2019). Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast. *Journal of Physics: Conference Series*, *1402*(6). https://doi.org/10.1088/1742-6596/1402/6/066046

Gupta, A., Kumar, L., Jain, R., & Nagrath, P. (2020). Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019). In *Lecture Notes in Networks and Systems* (Vol. 121, Issue Ic4s).

Kadhim, A. I. (2018). An Evaluation of Preprocessing Techniques for Text Classification. *International Journal of Computer Science and Information Security*, *16*(6), 22–32.

Kastouni, M. Z., & Ait Lahcen, A. (2022). Big data analytics in telecommunications: Governance, architecture and use cases. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 34, Issue 6). https://doi.org/10.1016/j.jksuci.2020.11.024

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2015*, 136–140. https://doi.org/10.1109/ICCI-CC.2015.7259377

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2006). Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems*, *1*, 1–9.

Mohan Kumar, A. V., Suhas, M., & Fedrich, N. (2022). Sentiment Analysis on Twitter Data. In *Cognitive Science and Technology*. https://doi.org/10.1007/978-981-19-2350-0_43

Paper, D. (2021). Build Your First Neural Network with Google Colab. In *TensorFlow 2.x in the Colaboratory Cloud*. https://doi.org/10.1007/978-1-4842-6649-6_2

Pisarenko, V. F., & Pisarenko, D. V. (2022). A Modified k-Nearest-Neighbors Method and Its Application to Estimation of Seismic Intensity. *Pure and Applied Geophysics*, *179*(11). https://doi.org/10.1007/s00024-021-02717-y

Prasetya, C. S. D. (2017). Sistem Rekomendasi Pada E-Commerce Menggunakan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *4*(3). https://doi.org/10.25126/jtiik.201743392

Purbolaksono, M. D., & Suryani, A. A. (n.d.). *Skip-Gram Negative Sample for Word Embedding in Indonesian-Translation Text Classification using Backpropagation Classifier*. *x*, 1–9.

Shakya, S., & Dave, M. (2022). *Analysis, Detection, and Classification of Android Malware using System Calls*.

Wang, X., Dai, Z., Li, H., & Yang, J. (2021). Research on Hybrid Collaborative Filtering Recommendation Algorithm Based on the Time Effect and Sentiment Analysis. *Complexity*, *2021*. https://doi.org/10.1155/2021/6635202

Wang, X., Jiang, W., & Luo, Z. (2016). Combination of convolutional and recurrent neural network for sentiment

analysis of short texts. *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2428–2437.

Xu, Y. (2022). Introduction and Development Strategy of Walt Disney Company. *Highlights in Business, Economics and Management, 1*. https://doi.org/10.54097/hbem.v1i.2529