Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12876
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Social Media Based Film Recommender System (Twitter) on Disney+ with Hybrid Filtering Using Support Vector Machine

Helmi Sunjaya Ramadhan[1)*], Erwin Budi Setiawan[2)]
[1,2)]School of Computing, Study Program of Informatics, Telkom University, Bandung, Indonesia
[1)]helmisunjaya@students.telkomuniversity.ac.id,[2)]erwinbudisetiawan@telkomuniversity.ac.id

Submitted : Aug 12, 2023 | **Accepted** : Aug 14, 2023 | **Published** : Oct 1, 2023

Abstract: In the current era, the culture of watching TV shows and movies has been made easier by the presence of the internet. Now, watching movies on platforms can be done from anywhere, one of which is Disney+. At times, people find it challenging to decide which film to watch given the multitude of genres and film titles available on these platforms. One solution to this issue is a recommendation system that can suggest films based on ratings. The recommendation system to be utilized involves Collaborative Filtering, Content-Based Filtering, and Hybrid Filtering. This is because Collaborative Filtering and Content-Based Filtering encounter issues like cold start, sparsity, and overspecialization. Thus, the objective of this study is to develop a recommendation system using Hybrid Filtering combined with Support Vector Machine (SVM). In this research, classification will be carried out using poly, linear, and RBF kernels with varying parameters. Techniques such as TF-IDF, RMSE, tuning, and data balancing with SMOTEN will be implemented to enhance accuracy during the classification process. The evaluation employed in this study utilizes the confusion matrix. Support Vector Machine, when tuned and combined with SMOTEN, achieves noteworthy results, particularly with the RBF kernel which attains a Precision score of 0.94. Recall produces a value of 0.93 with the Poly kernel, while the highest Accuracy, at 0.93, is achieved with the RBF kernel. Furthermore, the RBF kernel also demonstrates the highest F1-Score of 0.93. These findings illustrate elevated precision, recall, accuracy, and F1-Score within the context of hybrid filtering, achieved through the application of Support Vector Machine for classification and the implementation of the SMOTEN technique.

Keywords: Disney+, Hybrid Filtering, Recommender System, SMOTEN, Support Vector Machine

INTRODUCTION

Currently, social media is one of the frequently utilized platforms by individuals, and activities are often shared by users on social media to share both the small things they do and to express their viewpoints. (Das, Chidananda, & Sahoo, 2018). The culture of watching TV shows and movies has now been made more convenient with the presence of the internet. Streaming platforms like Netflix, HBO Max, and Disney+ offer users greater flexibility to watch their favorite TV shows and movies whenever and on whatever device they prefer. The number of TV shows and movies provided by each of these platforms can be considered quite substantial. (Arfisko & Wibowo, 2022). Before watching a movie, an individual often seeks reviews about the film they are about to watch. Rotten Tomatoes is one of the platforms where users can share their movie reviews.. The movie review data from Rotten Tomatoes,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12876 p-ISSN : 2541-044X

utilized in this research, has been sourced from the Kaggle website, and this information will be processed using a recommendation system.

A recommendation system is a way of providing information or suggestions based on user preferences, using information obtained from the user. When creating a recommendation system, there are two recommendation methods that can be employed Collaborative Filtering and Content-Based Filtering. (Ricci, Shapira, & Rokach, 2015).

Collaborative Filtering represents a recommendation technique relying on the collective preferences of various users toward a specific product. Conversely, Content-Based Filtering involves suggesting items akin to those previously favored by a user. In contrast, Hybrid Filtering combines both Content-Based and Collaborative Filtering methods, capitalizing on the merits of each approach. (Eli Lavindi & Rohmani, 2019).

Several studies have been conducted regarding recommendation systems employing the Hybrid Filtering method to recommend movies to users. Based on prior research, (Imelda Lubis, Josua Napitupulu, and Satia Dharma 2020). Utilizing the Hybrid Filtering method proves to be sufficiently accurate and superior in providing recommendations, as indicated by experimental results with the lowest Mean Absolute Error (MAE) recorded at 0.3741 for a k value of 25%. This conclusion is drawn from research conducted in prior studies. (Wardani, Sawaluddin, & Sihombing, 2020). It has been demonstrated that employing the Hybrid Filtering method, coupled with research findings from combining SVM-KNN, yields an accuracy score of 94.67%.

The aim of this study is to obtain the outcomes of Hybrid Filtering through the fusion of Collaborative Filtering and Content-Based Filtering methods. Additionally, the study seeks to ascertain the performance results of the amalgamation of Hybrid Filtering and Support Vector Machine.

LITERATURE REVIEW

The recommendation system approach is used for information filtering based on user preferences and ratings, greatly assisting each user in making purchases according to their existing needs. Recommendation systems directly offer users a way to seek information tailored to their interests, enabling them to discover products they require. Recommendation systems utilize content-based filtering, collaborative filtering, and association rule mining techniques to suggest items according to users' behaviors and inclinations. This strategy is advantageous for individuals who may lack expertise in navigating a multitude of choices presented by websites. (Patel & Patidar, 2018).

The Support Vector Machine (SVM) is a classification algorithm recognized for its effectiveness and precision in classifying data, surpassing the performance of other classification techniques. This superiority is attributed to SVM's incorporation of the Structural Risk Minimization (SRM) principle, which guarantees minimal classification errors. (Samaiya, Raghuwanshi, & Pateriya, 2018). SVM operates by creating a hyperplane or n-hyperplane that is useful for classifying a set of data points into their respective classes. (Tripathi & Sharma, 2020). The optimal hyperplane is the one with the maximum distance from the points representing each class. More specifically, to minimize errors globally, it's important to maintain a higher margin between the groups of data.

SVM is also widely used in recommendation systems. (Samaiya et al. 2018). On research (Wardani et al., 2020) Conducting a study involving the Hybrid of Support Vector Machine (SVM) Algorithm and K-Nearest Neighbor (KNN) Algorithm to improve the diagnosis of eye diseases produced results that demonstrate the enhanced accuracy of the hybrid SVM-KNN approach over using SVM alone for classifying eye diseases. The combined SVM-KNN algorithm achieved an accuracy rate of 94.67%.

While on (Imelda Lubis et al., 2020) The research results indicate that hybrid filtering has proven to be quite accurate and superior in providing recommendations, as evidenced by experimental outcomes. The lowest Mean Absolute Error (MAE) recorded was 0.3741 with a k value of 25%. This is in comparison to the content-based filtering method, which yielded an MAE of 1.174201, and the collaborative filtering method, which resulted in an MAE of 0.3768 with a k value of 25%.

METHOD

The process commences by gathering data from the Rotten Tomatoes website, accessible through Kaggle. The data sourced from Rotten Tomatoes comes in the form of a raw dataset. The operational

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12876
e-ISSN : 2541-2019
p-ISSN : 2541-044X

sequence of the forthcoming system is depicted in Figure 1. In the initial phase, the data undergoes preprocessing, wherein TF-IDF is employed in content-based filtering to shape profiles for individual items. Following this, SMOTEN is employed to tackle the uneven distribution of data. RMSE is utilized to populate any missing rating values. The hybrid filtering technique is utilized to merge the weighting from collaborative filtering and content-based filtering methods. Subsequently, the Support Vector Machine (SVM) model is employed for data manipulation. Ultimately, the system's performance is assessed through the utilization of a confusion matrix.

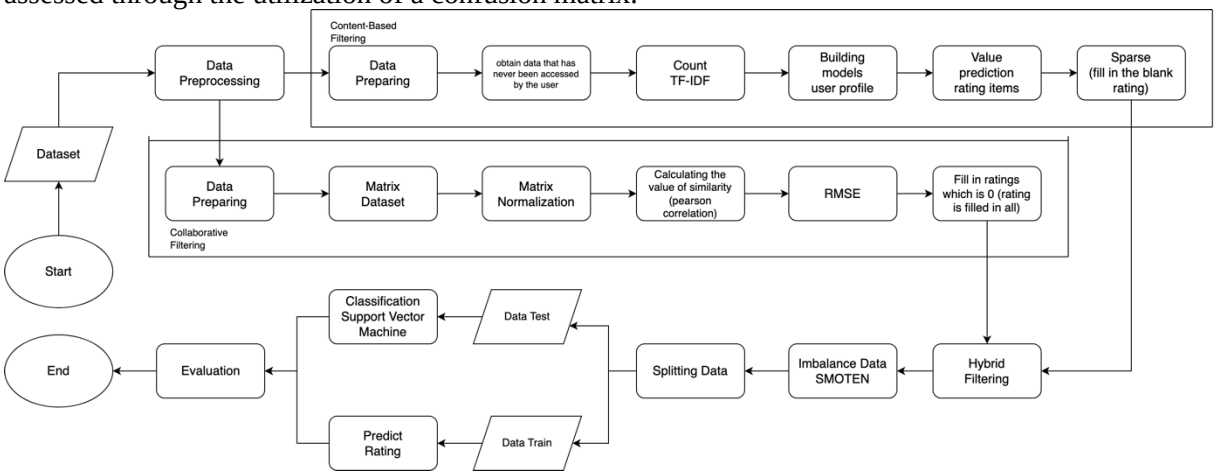

Figure 1. Flowchart of The Implementation of Classification and Evaluation

Dataset

The data utilized for this study is sourced from the Rotten Tomatoes website, accessible through Kaggle, and is presented in the CSV file format on the Kaggle platform. (Stefano Leone, 2020). The total number of data files is 1,129,887 rows. The complete dataset used comprises movie review data produced by Disney company from the year 2015 to 2020, containing 20,397 rows of data.

There are two datasets available on the Rotten Tomatoes website on Kaggle. In the "critic_reviews" dataset, there are several columns, and the columns used are only "critic_name" and "review_content." This can be observed in Table 1, where the "critic_name" column represents users who reviewed the movies, while the "review_content" column contains the users' reviews of the films they have watched.

Table 1. Sample Data Critic_reviews

Critic_name	Review_content
Jeremy Jahns	If you are officially done with Halloween, and you're ready to start feeling like Christmas is right around the corner, then this movie is a must.
Giles Hardie	While for some there is something disconcerting about the lifeless eyes of the otherwise life like characters, it certainly lends itself to large 3D cinemas, with rubber faced characters and grand sweeping movement.
Eileen Jonas	This, I told myself, is going to suck. And in most ways, it does suck as fully as expected. But there are, surprisingly, a few decent points, too.

Meanwhile, in the "movies" dataset, there are several columns as well, and the columns used are only "movie_title," "genres," "production_company," and "movie_info." This is evident from Table 2, where the "movie_title" column represents the movie title, the "genres" column indicates the movie's genre, the "production_company" column specifies the production company, and the "movie_info" column provides a brief description of the film.

Table 2. Sample Data Movies

Movie_title	Genres	Production_company	Movie_info

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12876
e-ISSN : 2541-2019
p-ISSN : 2541-044X

disney's a chrismas carol	Animation | Drama | Kids | Family | Science Fiction | Fantasy	Walt Disney Studios	though london awaits the joyful arrival of christmas, miserly ebenezer scrooge (jim carrey) thinks it's all humbug, berating his faithful clerk and cheerful nephew for their view. later, scrooge encounters the ghost of his late business partner, who warns that three spirits will visit him this night. the ghosts take scrooge on a journey through his past, present and future in the hope of transforming his bitterness.
beverly hills chihuahua	comedy	Walt Disney Studios	chloe (drew barrymore), a pampered chihuahua from beverly hills, gets an unwelcome taste of the real world when she gets lost in a tough part of mexico. with no rodeo drive boutiques in sight, she is out of her element, until scrappy street dogs delgado and papi lend her a paw, helping her find her way home.

Preprocessing

Preprocessing is the process of refining raw data into clean and informative data that can be utilized for further analysis. The goal of preprocessing is to reduce the volume of vocabulary and standardize the form of words, making the data more structured. (Altrabsheh, Cocea, & Fallahkhair, 2014).

Data Cleaning

The initial stage of the preprocessing phase involves data cleansing, where irrelevant characters are eliminated as part of the classification process.

Lowercasing

The process involves converting the entire text into lowercase (Sari & Ruldeviyani, 2020). This can be observed in Table 3, where this lowercasing aims to change all letters into lowercase.

Table 3. Example of Lowercasing Process

Before	After

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12876 p-ISSN : 2541-044X

This, I told myself, is going to suck. And in most ways, it does suck as fully as expected. But there are, surprisingly, a few decent points, too. While for some there is something disconcerting about the lifeless eyes of the otherwise life like characters, it certainly lends itself to large 3D cinemas, with rubber faced characters and grand sweeping movement.	this, i told myself, is going to suck. and in most ways, it does suck as fully as expected. but there are, surprisingly, a few decent points, too. while for some there is something disconcerting about the lifeless eyes of the otherwise life like characters, it certainly lends itself to large 3D cinemas, with rubber faced characters and grand sweeping movement.

TextBlob

This process aims to convert a column in "review_content," which initially contained user reviews of the films they have watched, into ratings within the range of 1 to 10 based on the quality of the user's review about the film being discussed. Table 4 provides an example of the TextBlob process.

Table 4. Example of TextBlob

Before	After
This, I told myself, is going to suck. And in most ways, it does suck as fully as expected. But there are, surprisingly, a few decent points, too.	4
irritatingly unfunny.	2
While for some there is something disconcerting about the lifeless eyes of the otherwise life like characters, it certainly lends itself to large 3D cinemas, with rubber faced characters and grand sweeping movement.	6

Content-Based Filtering

TF – IDF

TF-IDF is widely used in content-based filtering. In this study, TF-IDF is employed to construct item profiles in content-based filtering (Jain et al., 2016). TF (Term Frequency) is used to determine high word frequencies, and it can be inferred that such words are important and can be used to build item profiles. (Jurafsky & Martin, 2023).

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \quad (1)$$

TF_{ij} : Term Frequency of the word "I" in document j
f_{ij} : Frequency of occurrence of "I" in document j
$max_k f_{kj}$: Total number of words in document j

However, frequency sometimes does not depict whether a word is significant; for instance, the word "The" can have a very high frequency in a document but lacks any meaningful content. IDF represents the cumulative document frequency across the entire corpus of documents.

$$IDF_i = \log \frac{N}{n_i} \quad (2)$$

IDF_i : Frequency of the word "I" in the corpus

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12876 p-ISSN : 2541-044X

N : Total number of documents in the corpus
n_i : Number of documents containing the feature (word) i

Hence, TF-IDF weighting can be concluded to nullify the effect of words with high frequency counts in determining the significance of a feature. The profile of a document is a collection of features computed from the highest TF-IDF scores for all features within the document.

$$TF - IDF\ score:\ w_{ij} = TF_{ij}\ x\ IDF_i \quad (3)$$

Collaborative Filtering

Normalisasi Data

Data normalization is a data manipulation technique. This strategy can improve the precision and effectiveness of algorithms such as k-nearest neighbors, neural networks, clustering, and classification (Al Shalabi & Shaaban, 2006). Furthermore, this method can serve to avert data duplication or repetition within the dataset. The formula employed is outlined below:

$$nr_{i,u} = r_{i,u} - \overline{r_u} \quad (4)$$

Menghitung Similarity

The procedure entails computing the similarity metric for each user, which gauges the degree of likeness between individual items. The resulting scores will span from -1 to 1. A score nearing 1 implies a high degree of similarity between the items. Various techniques exist for determining similarity scores. In this research, the chosen method is Pearson correlation, which is widely utilized for calculating similarity measures. (Fakhri, Baizal, & Setiawan, 2019). The highest n similarities will be chosen employing the Top N technique. Top N is an approach utilized to identify the most significant n similarities. The equation applied is:

$$corr\ (m,n) = \frac{\sum u\in U(R_{u,m} - \bar{R}_m)(R_{u,m} - \bar{R}_m)}{\sqrt{u\in U+(R_{u,m} - \bar{R}_m)^2}\sqrt{u\in U+(R_{u,n} - \bar{R}_m)^2}} \quad (5)$$

$R_{u,m}$: Rating given by user m to user u
$\bar{R}_m$: Average rating of item u given by user m

RMSE (Root Mean Squared Error)

The RMSE conducted in this study is aimed at filling in ratings for the collaborative filtering method where there are still some empty ratings.

Hybrid Filtering

According to (Burke, 2007) The Hybrid recommendation system is a method that combines two or more recommendation techniques to enhance the performance of recommendations. It is typically employed to address issues present in each individual method used. (Burke, 2007). In this study, the hybrid filtering method employs a combination of collaborative filtering and content-based filtering techniques, aiming to address the issue of sparsity. The content-based filtering method is employed to predict ratings and to fill in sparse rating predictions. Subsequently, the filled sparse rating data is reused in the collaborative filtering method.

Data Balancing

SMOTE (Synthetic Minority Over-sampling Technique) is a synthetic technique used to systematically perform oversampling on a given dataset. It involves augmenting the data of the minority class to increase its representation. SMOTEN is another form of the SMOTE technique. Along with oversampling, instances or data points close to the majority class are removed. This is done before oversampling the minority class, so as not to include the most abundant data. (Alabrah, 2023).

SVM Classification

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12876 p-ISSN : 2541-044X

SVM stands out as a prominent method within the realm of pattern recognition. This algorithm is relatively recent, having been introduced in 1992. The popularity of the SVM algorithm can be attributed to its adherence to the SRM principle (Structural Risk Minimization). This concept becomes apparent in Figure 2, where the primary goal is to determine the optimal hyperplane for effectively distinguishing between two classes within a single space. This quality renders the algorithm especially well-suited for application as a classifier. Unlike neural networks that seek out hyperplane separators amidst classes, SVM endeavors to identify the most suitable hyperplane within the input space. The underlying principle that guides the SVM algorithm is a linear classifier, which is subsequently extended to address non-linear complexities through the employment of the kernel trick within high-dimensional feature space. (Wardani et al., 2020).

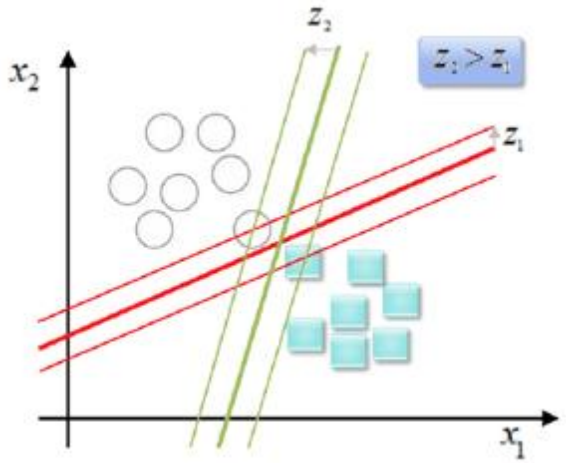

Figure 2. SVM Classification (Wardani et al., 2020).

Evaluation

After performing classification using the Support Vector Machine, an evaluation is conducted on both the training and test data using a confusion matrix. The evaluation results from this confusion matrix provide information about True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP). These values are then utilized to calculate accuracy, F1-Score, Precision, and Recall. This information is presented in Table 5, which illustrates the confusion matrix table.

Table 5. Confusion Matrix

		Actual Value	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

The validation of the model is determined by analyzing the confusion matrix derived from the experimental outcomes. The assessment metrics encompass recall (Rec), precision (Pre), accuracy (Acc), and F-score (F-Scr). The computations for Rec, Pre, Acc, and F1-Scr are elucidated in Table 6 based on the provided confusion matrix found in Table 6 itself. This matrix incorporates the actual positive and actual negative values, alongside the predicted positive and negative values. This process yields values such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which result from combining the actual and predicted values.

Table 6. Model Validation Matrix

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12876
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Metrik	Deskripsi	Persamaan
Pre	Nilai presisi	$\frac{TP}{TN + FP}$
Rec	Nilai recall	$\frac{TP}{TP + FN}$
Acc	Nilai akurasi	$\frac{TP + TN}{TP + TN + FP + FN}$
F1 - Scr	Nilai F1-Score	$\frac{2 * Pre * Rec}{Pre + Rec}$

RESULT

In the evaluation of this research, a comparison was made between three kernels for data both before and after oversampling. Several testing scenarios were conducted. In the research conducted, a 70:30 ratio was employed for each of the poly, linear, and rbf kernels. The first scenario with the untuned and non-SMOTEN kernel aimed to produce optimal results that would be carried forward to the subsequent scenarios. In the second scenario, testing was performed on the kernel with tuning to attain more optimal outcomes. Scenario three centered on mitigating data imbalance through the application of SMOTEN, with the objective of achieving equilibrium between positive and negative facets, thus elevating the overall efficiency of each kernel compared to the SMOTEN-enabled baseline. Moving to scenario four, an examination was carried out on each kernel, employing both tuning and SMOTEN, with the intention of attaining outcomes that surpass the levels of optimization achieved in the preceding scenarios.

Scenario 1

In the evaluation of this research, a comparison was made among the best-performing kernels from the test results of Scenario 1 in Table 7. Each kernel without tuning and SMOTEN resulted in the highest precision value for the poly kernel, the highest recall value for the linear kernel, the highest accuracy values for the poly and rbf kernels, and the highest F1-Score value for the linear kernel.

Table 7. Results of SVM

Kernel	Presicion	Recall	Accuracy	F1-Score
Poly	**0,45**	0,08	**0,91**	0,13
Linear	0,27	**0,1**	0,9	**0,15**
RBF	0,03	0,01	**0,91**	0,01

Scenario 2

In the evaluation of testing Scenario 2, tuning was applied to each kernel. For the poly kernel, the parameters used were C, Gamma, and degree. For the linear kernel, only the parameter C was used. Meanwhile, for the rbf kernel, the parameters employed were C and gamma. As seen in Table 8, the parameters for each kernel are different.

Table 8. Results of SVM Tuning

Kernel	Parameter	Presicion	Recall	Accuracy	F1-Score
Poly	C, Gamma & Degree	**0,41**	0,19	**0,91**	0,1
Linear	C	0,04	0,1	0,9	**0,15**
RBF	C & Gamma	0,04	**0,22**	**0,91**	0,01

As shown in Table 8, the results of testing using tuning reveal the highest precision value in the poly kernel with a score of 0.41. The highest recall value is found in the rbf kernel, with a value of 0.22. The highest accuracy values are present in both the poly and rbf kernels, each achieving a value of 0.91. The highest F1-score is observed in the linear kernel, which is 0.15. Further evaluation and model enhancement may be necessary to improve the overall performance of the movie recommendation system.

Scenario 3

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12876
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In Scenario 3, the evaluation focused on how SMOTEN handles imbalanced data. As seen in Table 9, the values of precision, recall, accuracy, and F1-Score significantly increased. The conducted tests indicate that by utilizing SMOTEN, the results for precision, recall, accuracy, and F1-Score become optimal.

Table 9. Results from SVM using SMOTEN

	Kernel	Presicion	Recall	Accuracy	F1-Score
Without SMOTEN	Poly	**0,45**	0,08	**0,91**	0,13
	Linear	0,27	**0,1**	0,9	**0,15**
	RBF	0,03	0,01	**0,91**	0,01
With SMOTEN	Poly	0,65	**0,93**	0,72	0,77
	Linear	0,77	0,88	0,81	0,82
	RBF	**0,94**	0,89	**0,92**	**0,92**

Scenario 4

In Scenario 4, the tuning tests were also applied with SMOTEN, aiming to balance the data, and the results of tuning using SMOTEN were even more optimal. As evident from Table 10, with tuning, the highest values for precision are 0.41, recall is 0.22, accuracy is 0.91, and F1-Score is 0.15. On the other hand, when using SMOTEN along with tuning, the precision value is 0.94, recall is 0.93, accuracy is 0.93, and F1-Score is 0.93.

Table 10. Results of SVM Tuning using SMOTEN

	Kernel	Parameter	Presicion	Recall	Accuracy	F1-Score
Without SMOTEN	Poly	C, Gamma & Degree	**0,41**	0,19	**0,91**	0,1
	Linear	C	0,04	0,1	0,9	**0,15**
	RBF	C & Gamma	0,04	**0,22**	**0,91**	0,01
With SMOTEN	Poly	C, Gamma & Degree	0,86	**0,93**	0,89	0,9
	Linear	C	0,75	0,88	0,79	0,81
	RBF	C & Gamma	**0,94**	0,92	**0,93**	**0,93**

DISCUSSIONS

Four scenarios were conducted by comparing the baseline, SVM tuning, baseline with SMOTEN, and SVM tuning with SMOTEN. For the precision, recall, accuracy, and F1-Score values, each scenario was evaluated using a confusion matrix.

Table 11. Results of Scenario Test

	Kernel	Presicion	Recall	Accuracy	F1-Score
Baseline	Poly	**0,45**	0,08	**0,91**	0,13
	Linear	0,27	**0,1**	0,9	**0,15**
	RBF	0,03	0,01	**0,91**	0,01
Baseline SMOTEN	Poly	0,65	**0,93**	0,72	0,77
	Linear	0,77	0,88	0,81	0,82
	RBF	**0,94**	0,89	**0,92**	**0,92**
Tuning	Poly	**0,41**	0,19	0,91	0,1
	Linear	0,04	0,1	0,9	0,15
	RBF	0,04	**0,22**	0,91	0,01
Tuning SMOTEN	Poly	0,86	**0,93**	0,89	0,9
	Linear	0,75	0,88	0,79	0,81
	RBF	**0,94**	0,92	**0,93**	**0,93**

*name of corresponding author

As observed in Table 11, multiple scenarios were employed to attain the best precision, recall, accuracy, and F1-Score values. In this study, a 70:30 ratio was utilized for each kernel poly, linear, and rbf. First, support vector machine was executed without tuning and SMOTEN technique. Second, support vector machine was performed with tuning for each kernel, using different parameters. Third, support vector machine was employed with the SMOTEN technique. Lastly, support vector machine was executed with both tuning and SMOTEN. Ultimately, the highest precision value reached 0.94, recall was 0.93, accuracy attained 0.93, and F1-Score reached 0.93, indicating significantly favorable outcomes.

CONCLUSION

This study accomplished results in hybrid filtering by merging collaborative filtering and content-based filtering techniques. Furthermore, it investigated the performance results of hybrid filtering and support vector machine in the film recommendation system. The classification of the Support Vector Machine, when utilizing tuning and the SMOTEN technique, achieved its peak precision of 0.94 in the rbf kernel, a recall of 0.93 in the poly kernel, the highest accuracy of 0.93 in the rbf kernel, and the highest F1-Score of 0.93 in the rbf kernel. The primary goal of this study was to ascertain the superior results among the three tested kernels, revealing optimal outcomes when applied to an English dataset extracted from the Rotten Tomatoes website, accessible on Kaggle in CSV format. Tuning parameters involve adapting the parameters to enhance performance. SMOTEN, a derivative of the SMOTE algorithm, amalgamates oversampling with data removal, aiming to recognize and eliminate nearest neighbors of the majority class before oversampling the minority class. In the conducted research, by employing a 70:30 ratio and employing tuning and SMOTEN, the highest precision was achieved in the rbf kernel, the highest recall in the poly kernel, the highest accuracy in the rbf kernel, and the highest F1-Score in the rbf kernel.

REFERENCES

Al Shalabi, L., & Shaaban, Z. (2006). *Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix*.

Alabrah, A. (2023). An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method. *Sensors*, *23*(9). https://doi.org/10.3390/s23094406

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014). *Sentiment analysis: towards a tool for analysing real-time students feedback*.

Arfisko, H. H., & Wibowo, A. T. (2022). *Sistem Rekomendasi Film Menggunakan Metode Hybrid Collaborative Filtering Dan Content-Based Filtering*.

Burke, R. (2007). Hybrid web recommender systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4321 LNCS*, 377–408. Springer Verlag. https://doi.org/10.1007/978-3-540-72079-9_12

Das, D., Chidananda, H. T., & Sahoo, L. (2018). Personalized movie recommendation system using twitter data. *Advances in Intelligent Systems and Computing*, *710*, 339–347. Springer Verlag. https://doi.org/10.1007/978-981-10-7871-2_33

Eli Lavindi, E., & Rohmani, A. (2019). Aplikasi Hybrid Filtering Dan Naïve Bayes Untuk Sistem Rekomendasi Pembelian Laptop Hybrid Filtering and Naïve Bayes Application for Laptop Purchase Recommendation Systems. *Journal of Information System*, *4*(1), 54–64.

Fakhri, A. A., Baizal, Z. K. A., & Setiawan, E. B. (2019). Restaurant Recommender System Using User-Based Collaborative Filtering Approach: A Case Study at Bandung Raya Region. *Journal of Physics: Conference Series*, *1192*(1). Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/1192/1/012023

Imelda Lubis, Y., Josua Napitupulu, D., & Satia Dharma, A. (2020). *Implementasi Metode Hybrid Filtering (Collaborative dan Content-based) untuk Sistem Rekomendasi Pariwisata Implementation of Hybrid Filtering (Collaborative and Content-based) Methods for the Tourism Recommendation System*.

*name of corresponding author

Jain, S., Tomar, D. S., Saxena, S., Maulana Azad National Institute of Technology, IEEE Asia-Pacific Region. MP Subsection, & Institute of Electrical and Electronics Engineers. (2016). *A Personalized Recommender System using Machine Learning based Sentiment Analysis over Social Data*.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Third Edition draft Summary of Contents*.

Patel, D., & Patidar, H. (2018). Hybrid Recommendation Solution for Online Book Portal. *International Journal for Research in Applied Science and Engineering Technology*, *6*(5), 1367–1373. https://doi.org/10.22214/ijraset.2018.5225

Ricci, F., Shapira, B., & Rokach, L. (2015). Recommender systems: Introduction and challenges. In *Recommender Systems Handbook, Second Edition* (pp. 1–34). Springer US. https://doi.org/10.1007/978-1-4899-7637-6_1

Samaiya, N., Raghuwanshi, S. K., & Pateriya, R. K. (2018). Shilling Attack Detection in Recommender System Using PCA and SVM. *Advances in Intelligent Systems and Computing*, *813*, 629–637. Springer Verlag. https://doi.org/10.1007/978-981-13-1498-8_55

Sari, I. C., & Ruldeviyani, Y. (2020). Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers. *2020 International Workshop on Big Data and Information Security, IWBIS 2020*, 23–28. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/IWBIS50925.2020.9255531

Stefano Leone. (2020). Rotten Tomatoes movies and critic reviews dataset.

Tripathi, A., & Sharma, A. K. (2020). *Recommending Restaurants: A Collaborative Filtering Approach*.

Wardani, S., Sawaluddin, & Sihombing, P. (2020). Hybrid of Support Vector Machine Algorithm and K-Nearest Neighbor Algorithm to Optimize the Diagnosis of Eye Disease. *MECnIT 2020 - International Conference on Mechanical, Electronics, Computer, and Industrial Technology*, 321–326. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MECnIT48290.2020.9166599

*name of corresponding author