MicroOhm
JOURNAL OF
ELECTRICAL
ANDCOMPUTER
ENGINEERING

MicroOhm Journal of Electrical and Computer Engineering
Vol. 1, No. 1; July 2025
E-ISSN: XXXX-XXXX

Movie Rating Prediction Analysis on Netflix Using Support Vector Machine Algorithm

Nanda Syamia[1], Ayu Pratiwi[2]

[1]Universitas Islam Negeri Sumatera Utara; nandasyamia@gmail.com
[2]Politeknik Negeri Medan; ayup9363@gmail.com

ABSTRACT

The rapid rise of digital streaming platforms such as Netflix has created a demand for accurate movie rating prediction systems to enhance user experience and content personalization. This study investigates the performance of the Support Vector Machine (SVM) algorithm in classifying movie ratings using historical metadata, including genre, duration, release year, and country of production. A quantitative approach is employed using RapidMiner for data mining and model evaluation. The dataset, sourced from publicly available open-access repositories, underwent preprocessing steps such as normalization, transformation of categorical attributes, and data splitting. Experimental results show that the SVM model achieved an accuracy of 83.10%. Furthermore, positive precision reached 82.98% and recall reached 100%, yielding a positive F1-score of 90.61%. However, the model exhibited limitations in detecting negative ratings, as indicated by a negative recall of only 3.98% despite achieving 100% negative precision, resulting in a low negative F1-score of 7.65%. These findings suggest that while SVM performs well in identifying positive instances, further improvement is needed for balanced classification in real-world movie rating prediction systems.

Keywords : *Data Mining, Netflix, Rating Prediction, RapidMiner, Support Vector Machine;*

Corresponding Author:
Author Name: Nanda Syamia
Affiliation: Universitas Islam Negeri Sumatera Utara
Email: nandasyamia@gmail.com

1. INTRODUCTION

The advancement of information technology has significantly impacted the digital entertainment industry, particularly through the emergence of streaming platforms such as Netflix. As these services offer vast content libraries, the demand for intelligent recommendation and rating prediction systems is growing rapidly. Predicting movie ratings based on content-related features is a critical task in recommendation systems. In particular, narrative descriptions such as synopses serve as primary information sources that influence users viewing decisions, making them valuable input for predictive modeling using natural language processing (NLP) techniques. However, utilizing movie descriptions for predictive tasks presents challenges due to the unstructured nature of text data. These descriptions often vary in style, vocabulary, semantic nuance, and sentence structure, complicating the task of feature extraction and classification. A further challenge lies in converting narrative text into numerical representations suitable for machine learning, while retaining semantic context. Variations in writing styles across genres, cultural influences, and marketing strategies further complicate this transformation. These challenges underscore the need for robust natural language processing techniques for feature extraction, alongside reliable classification algorithms. Support Vector Machines (SVM), known for their effectiveness in high-dimensional and sparse text data, offer a promising approach for this task. Accordingly, this study aims to evaluate the performance of the SVM algorithm in predicting movie ratings using only textual movie descriptions, thereby contributing to the development of more accurate and personalized content recommendation systems.

2. LITERATURE REVIEW

Netflix is a globally accessible digital streaming platform, providing on-demand movies and series content. Its popularity has driven interest in intelligent systems for enhancing user experience, including automated movie rating prediction. Movie ratings are often used as indicators of a film's reception and commercial success. Therefore, predictive modeling of ratings based on content features has become an important research area in the field of data mining and recommendation systems (Ilmi et al., 2023). Predictive analysis involves using historical data to model and forecast future outcomes. In the context of this research, it enables the prediction of movie ratings based on descriptive content attributes (Mujilahwati et al., 2020).

RapidMiner is a visual data science platform that supports various machine learning workflows, including preprocessing, modeling, and evaluation. It is particularly suitable for researchers seeking to implement predictive models without extensive programming (Rafi Nahjan et al., 2023). Support Vector Machine (SVM) is a supervised learning algorithm widely recognized for its high performance in text classification tasks. Its ability to handle high-dimensional and sparse data makes it well-suited for processing textual information such as movie descriptions. Several studies have demonstrated the effectiveness of SVM in sentiment analysis and document categorization (Handayani et al., 2023; Mahendro et al., 2022).

Text mining involves extracting meaningful patterns from unstructured textual data. Key preprocessing steps include tokenization, stopword removal, and stemming, which prepare the data for feature extraction. In this study, these steps are applied to movie descriptions to enable vectorization and subsequent classification using the SVM algorithm (Widiari et al., 2020).

3. METHOD

This research uses data obtained from public datasets that can be accessed openly, one of which is through the Kaggle platform. The dataset contains various information about movies, such as title, genre, year of release, duration, country of production, and user ratings which are used as target variables in the prediction process. The dataset used is data from Netflix. In data analysis and modeling, this research applies the Support Vector Machine (SVM) algorithm with the help of the RapidMiner application as the main data processing tool.

The following figure presents the flowchart of the Support Vector Machine (SVM) method used in this research. The flowchart illustrates the main stages involved in building a prediction model, starting from the data collection stage, data pre-processing (such as data cleaning, text transformation, and feature extraction), data separation into training data and test data, to the model training process using the SVM algorithm. After the model is trained, an evaluation of the model's performance is conducted by measuring metrics such as accuracy, precision, recall, and F1-score to assess the extent to which the model is able to accurately predict movie ratings.

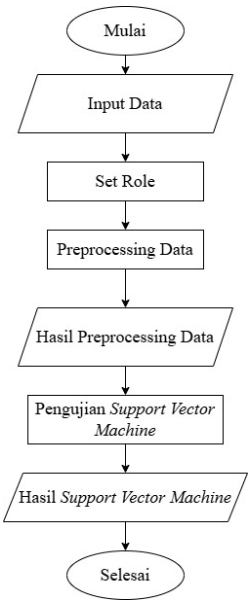

Figure 1. SVM Method Flowchard

The concept of Support Vector Machine is to find the most optimal hyperplane or dividing line that serves to separate two classes. Support Vector Machine has the basic principle of a linear classifier, namely classification cases that can be linearly separated, but SVM has been developed to work on non-linear problems by incorporating the concept of kernels in high-dimensional workspaces (Octaviani et al., 2020).

The initial steps in text data processing begin with data cleaning, which is the process of cleaning data from irrelevant characters such as punctuation marks, numbers, and special symbols that have no important meaning in the analysis. The next step is case folding, which aims to convert all letters in the text into lowercase letters for consistency in the analysis. After that, filtering is done to remove unnecessary elements, such as HTML tags or certain special characters. The process continues with tokenizing, which breaks sentences into words (tokens) so that they can be analyzed separately. Then stopword removal is performed to remove common words that do not have high information value, such as "and", "which", or "from". The last stage is stemming, which converts words to their basic form so that variations of words that have the same root can be recognized as one entity. All of these stages aim to improve the quality of text data before it is used in the classification process using the SVM algorithm. (Widiari et al., 2020).

After the data is complete, the data is divided into two parts, namely the training set is part of the dataset used to train the machine learning model. This data is used by the algorithm to learn patterns, relationships between variables, and data characteristics in order to form a predictive model. and the testing set is part of the dataset that is not used in the training process, but is used to measure the performance or accuracy of the model that has been trained. 5 Data division in this study was carried out by utilizing the Split Data operator available in RapidMiner, with a proportion of 80% for training data and 20% for test data. From a total of 1,000 comments used in the experiment, 800 comments were allocated as training data used to build and train the classification model, while the remaining 200 comments were used as test data to evaluate the performance of the model. This division aims to ensure that the model has enough data to learn while providing representative data to measure the generalization ability of the model to new data that has never been seen before (Hanafi, 2023).

The next stage is the construction of a prediction model using the Support Vector Machine (SVM) algorithm through the SVM operator in RapidMiner. In this process, several parameters such as kernel type (linear, polynomial, or RBF), C value (regularization), and gamma can be adjusted to get the best performance. The model is trained on the training data, then tested on the pre-separated test data. Model evaluation is performed using the Performance (Classification) operator, which is defined as an evaluation tool used to measure the performance of classification models (Rahmadiani & Kusrini, 2023) . This operator calculates various evaluation metrics used to assess how well the model performs classification which will produce evaluation metrics such as Accuracy is a measure that shows the percentage of correct predictions of all predictions made by the model, Precision it measures the accuracy of the model when predicting positive classes, Recall measures the ability of the model to find all positive cases that actually exist in the data. That is, how much positive data is successfully recognized by the model., and F1-score is the harmonic mean between precision and recall, depending on how the rating is classified (Romadloni et al., 2022).

The formula used to find the value of accuracy, precision, recall and F1-score is as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\%$$

Description:
TP (True Positive) = Number of positive data successfully predicted correctly by the model.
TN (True Negative) = The amount of negative data that is successfully predicted correctly by the model.
FP (False Positive) = The amount of negative data that is incorrectly predicted as positive by the model.
FN (False Negative) = The amount of positive data that is incorrectly predicted as negative by the model.

4. RESULTS AND DISCUSSION

At this stage, the model evaluation process is carried out using the Support Vector Machine (SVM) algorithm that has been previously built in RapidMiner. This test aims to measure the performance of the model in classifying data based on patterns that have been learned during the training process. The following figure displays the workflow in RapidMiner designed for data analysis, starting with the "Read Excel" operator to import data, followed by "Set Role" to determine the role of data attributes. Next, "Nominal to Text" is used to convert nominal data to text, and "Process Documents from Data" serves for text pre-processing. The whole process is then evaluated using the "Cross Validation" operator to measure the performance and generalization of the model to be built.

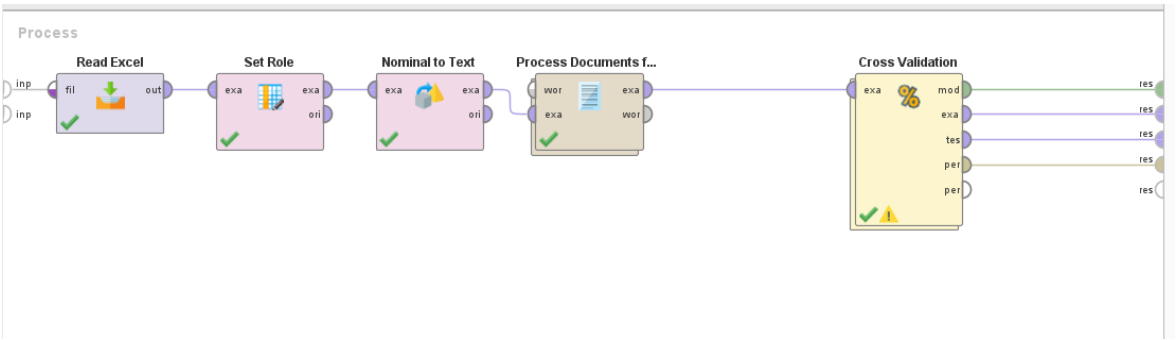

Figure 2. SVM Workflow

The following figure illustrates the stages of the data preprocessing process which includes several important steps in text cleaning and standardization, namely tokenizing, case folding, filtering, and stopword removal. Tokenizing is the process of breaking sentences into word units (tokens), while case folding aims to equalize the shape of letters by converting the entire text into lowercase letters. Filtering is used to remove unnecessary characters or symbols, and stopword removal is done to remove common words that do not have high information value. This whole process is very important to ensure that the text data to be analyzed is in a clean, structured format, and ready to be used in the classification process using algorithms such as SVM. This preprocessing stage is an important foundation in producing accurate and reliable models.

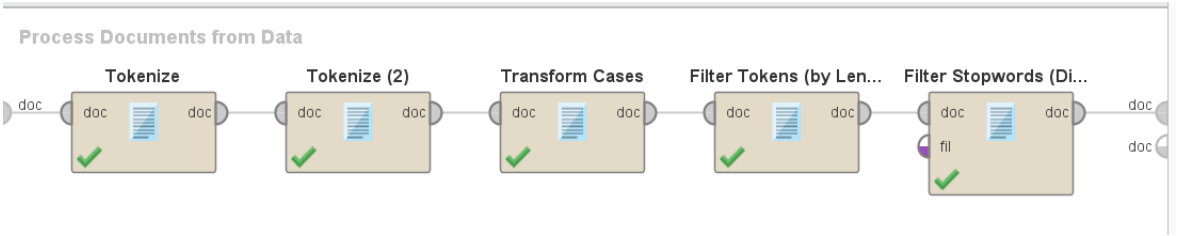

Figure 3. Data Preprocessing

The following figure shows the testing process using the Support Vector Machine (SVM) method, a machine learning technique used for classification and regression. This test aims to evaluate the ability of the SVM model to separate data based on existing features so that it can

produce accurate predictions. From this test, it can be seen how SVM works in determining the optimal hyperplane that maximizes the margin between data classes, so that the model is able to classify new data with good performance. The results of this test are important as an indicator of the effectiveness of the SVM method in solving classification problems on the dataset used.

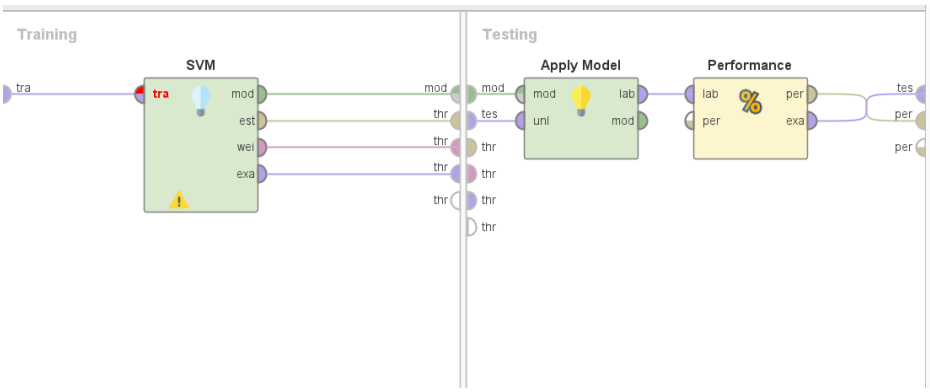

Figure 4. SVM Testing

The figure below displays the confusion matrix along with the accuracy value of the tested model. The confusion matrix provides a detailed overview of the model's performance in classifying the data, by showing the number of correct and incorrect predictions for each class. Meanwhile, the accuracy value shows the overall percentage of correct predictions by the model compared to the total data tested. By looking at these two components together, we can assess how well the model does in recognizing patterns and distinguishing classes correctly, while also identifying areas that still need improvement.

accuracy: 83.10% +/- 1.20% (micro average: 83.10%)

	true Positive	true Negative	class precision
pred. Positive	824	169	82.98%
pred. Negative	0	7	100.00%
class recall	100.00%	3.98%	

Figure 5. Confusion Matrix

After obtaining the confusion matrix, precision, recall, and F1-score are calculated to assess the accuracy, detectability, and balance of the model's performance in classification, making it easier to identify the strengths and weaknesses of the model.

Positive TP = 824
Positive FP = 169
FN Positive = 0
TN Positive = 7

$$Precision\ Positif = \frac{824}{824+169} \times 100\% = 82{,}98\%$$

$$Recall\ Positif = \frac{824}{824+0} \times 100\% = 100\%$$

$$F1 - Score\ Positif = 2 \times \frac{82{,}98\% \times 100\%}{82{,}98\% + 100\%} = 90{,}61\%$$

TP (Negative) = 7
FP (Negative) = 0
FN (Negative) = 169
TN (Negative) = 824

$$Precision\ Negatif = \frac{7}{7+0} \times 100\% = 100\%$$

$$Recall\ Negatif = \frac{7}{7+169} \times 100\% = 3{,}98\%$$

$$F1 - Score\ Negatiftif = 2 \times \frac{100\% \times 3{,}98\%}{100\% + 3{,}98\%} = 7{,}65\%$$

5. CONCLUSION

Based on the research on movie rating prediction on Netflix using the Support Vector Machine (SVM) algorithm, it can be concluded that SVM shows good and effective performance in classifying movie ratings. The model managed to achieve an accuracy of 83.10%, which indicates a fairly high level of prediction accuracy. In addition, other evaluation metrics show that the positive precision is 82.98% and the positive recall reaches 100%, resulting in a positive F1-score of 90.61%, indicating that the model is very reliable in identifying positive classes. Although the negative precision reached 100%, the low negative recall of 3.98% led to a negative F1-score of only 7.65%, indicating the limitation of the model in recognizing negative classes. Overall, these results confirm that SVM is well suited for movie rating classification tasks, but more attention needs to be paid to improving negative class detection for a more balanced model performance.

REFRENCES

Arifin, N., Enri, U., & Sulistiyowati, N. (2021). Application of Support Vector Machine (SVM) algorithm with TF-IDF n-gram for text classification. *STRING (Unit of Research Writing and Technological Innovation)*, *6*(2), 129. https://doi.org/10.30998/string.v6i2.10133

Farid Naufal, M. (2021). Comparison and analysis of SVM algorithm and CNN. *Journal of Information Technology and Computer Science*, *8*(2), 311–318. https://doi.org/10.25126/jtiik.202184553

Hanafi, H. (2023). Data cleaning in big data: Review. *[Unpublished manuscript]*, 1–5.

Handayani, A., & Zufria, I. (2023). Sentiment analysis of Indonesia 2024 presidential candidates on Twitter using SVM algorithm. *Journal of Information System Research (JOSH)*, *5*(1), 53–63. https://doi.org/10.47065/josh.v5i1.4379

Harahap, N., & Nasution, L. (2022). The effect of employee work professionalism on public service quality at the Medan City BPS Office. *Student Business Journal*, *2*(1), 43–52.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Huda, A. S., Nafsika, S. S., & Salman, S. (2023). Film as a media in changing the way of human view in humanitarian principles. *Irama: Journal of Art Design and Learning*, *5*(1), 9–14.

Husada, H. C., & Paramita, A. S. (2021). Sentiment analysis on airlines on the Twitter platform using the Support Vector Machine (SVM) algorithm. *Teknika*, *10*(1), 18–26. https://doi.org/10.34148/teknika.v10i1.311

Ilmi, R. R., Kurniawan, F., & Harini, S. (2023). IMDb film rating prediction using decision tree. *Journal of Information Technology and Computer Science*, *10*(4), 791–798. https://doi.org/10.25126/jtiik.2023106615

Mahendrajaya, R., Buntoro, G. A., & Setyawan, M. B. (2020). Sentiment analysis of GoPay users using lexicon-based and Support Vector Machine methods. *Computer*, *3*(2), 52. https://doi.org/10.24269/jkt.v3i2.270

Mahendro, I., & Abimanto, D. (2022). Analysis of student satisfaction with e-learning using Support Vector Machine algorithm. *Journal of Maritime Science and Technology*, *23*(1), 97. https://doi.org/10.33556/jstm.v23i1.333

Mujilahwati, S., Nawafilah, N. Q., & Aliyudin, M. (2020). Analysis of prediction results with the PROMETHEE method. *Mnemonic Journal*, *2*(1), 35–40. https://doi.org/10.36040/mnemonic.v2i1.49

Octaviani, P. A., Wilandari, Y., & Ispiriyanti, D. (2020). Application of SVM method on elementary school accreditation data in Magelang Regency. *Gaussian Journal*, *3*(8), 811–820.

Priskilla, R., & Suparni, S. (2024). Analysis of public sentiment towards Netflix online movie streaming application with Naïve Bayes method. *JATI (Journal of Informatics Engineering Students)*, *8*(2), 1339–1342. https://doi.org/10.36040/jati.v8i2.7520

Rafi Nahjan, M., Heryana, N., & Voutama, A. (2023). Implementation of RapidMiner with K-Means clustering method for sales analysis at Oj Cell Store. *JATI (Journal of Informatics Engineering Students)*, *7*(1), 101–104. https://doi.org/10.36040/jati.v7i1.6094

Rahmadiani, P., & Kusrini, E. (2023). Operator performance analysis using overall labor effectiveness method with root cause analysis approach. *Asian Journal of Social and Humanities*, *1*(11), 918–927. https://doi.org/10.59888/ajosh.v1i11.106

Romadloni, P., Kusuma, B. A., & Baihaqi, W. M. (2022). Comparison of machine learning methods for implementation of decision making in determining employee promotion. *JATI (Journal of Informatics Engineering Students)*, *6*(2), 622–628. https://doi.org/10.36040/jati.v6i2.5238

Sari, N. P. P. A., Suryawati, I. G. A. A. S., & Pradipta, A. D. (2021). Motives and satisfaction of Netflix users as streaming media among teenagers in Denpasar City. *Journal of Communication Science MEDIUM*, *1*(1), 1.

Siti Aisah, I., Irawan, B., & Suprapti, T. (2024). Support Vector Machine (SVM) algorithm for sentiment analysis of digital Qur'an application reviews. *JATI (Journal of Informatics Engineering Students)*, *7*(6), 3759–3765. https://doi.org/10.36040/jati.v7i6.8263

Widiari, N. P. A., Suarjaya, I. M. A. D., & Githa, D. P. (2020). Data cleaning technique using Snowflake for case study of tourism objects in Bali. *Merpati Scientific Journal (Information Technology Academics Research Tower)*, *8*(2), 137. https://doi.org/10.24843/jim.2020.v08.i02.p07