Movie Recommendation System Using Machine Learning Techniques

Jannatul Suraiya
Department of Computer Science and Engineering
Bangladesh Army University of Science & Technology (BAUST)
Email: jannatulSuraiya@gmail.com

Abstract—The rapid expansion of movies and streaming platforms has created an overwhelming volume of content, making it difficult for users to identify movies that fit their personal preferences. Traditional search and manual recommendation techniques are no longer efficient due to diverse genre patterns and changing viewer interests. In this paper, we propose a machine learning–based movie recommendation system using TF-IDF for feature extraction and four supervised classifiers: Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest. Our experimental results show that SVM achieves the highest performance with superior accuracy, F1-score, and AUC values, making it the most reliable model for this task.

Index Terms— Movie Recommendation, Machine Learning, TF-IDF, SVM, User Preference Prediction.

I. INTRODUCTION

The growing popularity of online streaming platforms such as Netflix, Amazon Prime, and Disney+ has significantly increased the number of movies available to users. While this provides more choices, it also results in difficulty selecting the right movie based on personal taste. Users often face information overload and struggle to find suitable movies through traditional browsing. Moreover, manually curated recommendation lists fail to adapt to a user's evolving preferences.

Machine learning-based recommendation systems have emerged as effective solutions. These systems analyze user behavior, learn patterns from movie descriptions, and automatically recommend movies that align with the user's interests.

II. METHODOLOGY

Our methodology consists of dataset preprocessing, feature extraction, model training, and performance evaluation.

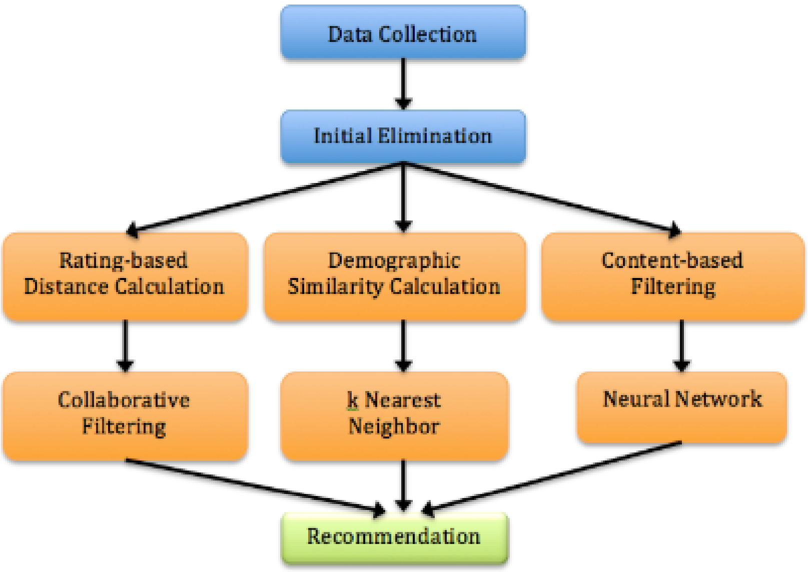
A. Dataset Cleaning and Preprocessing

To ensure proper data quality, the following preprocessing steps were applied:
Duplicate movie descriptions were removed.
All punctuation, stopwords, and irrelevant symbols were eliminated.

B. Feature Extraction (TF-IDF)

While earlier systems primarily used Bag-of-Words, our approach employs TF-IDF (Term Frequency–Inverse Document Frequency) to capture meaningful keywords from movie plots.

TF-IDF assigns high weights to important and distinctive words, allowing the model to identify genre and content-related patterns more effectively.



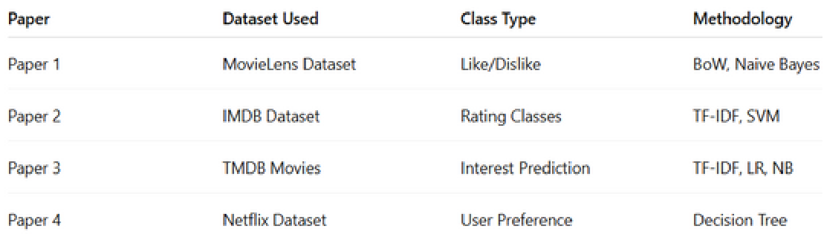
C. Models Used

We trained and evaluated four supervised machine learning classifiers:
Naive Bayes: A probabilistic model often used as a baseline for text classification.
Logistic Regression: A linear classification model suitable for binary interest prediction.
Support Vector Machine (SVM): Finds the optimal separating hyperplane in high-dimensional TF-IDF space.

Random Forest: An ensemble approach using multiple decision trees for classification.

III. RELATED WORK

We compared our work with existing research in movie recommendation and text-based classification.
Table I summarizes datasets and methodologies used in previous studies.



Paper	Dataset Used	Class Type	Methodology
Paper 1	MovieLens Dataset	Like/Dislike	BoW, Naive Bayes
Paper 2	IMDB Dataset	Rating Classes	TF-IDF, SVM
Paper 3	TMDB Movies	Interest Prediction	TF-IDF, LR, NB
Paper 4	Netflix Dataset	User Preference	Decision Tree

IV. RESULTS AND ANALYSIS

We evaluated our models using Recall, F1-score, ROC-AUC, and Confusion Matrices.

A. Recall Analysis

Recall measures the model's capability to detect movies that the user is likely to prefer.Our experiments show that SVM achieved the highest recall, successfully identifying most true "liked" movies.

Logistic Regression performed slightly lower, while Naive Bayes and Random Forest showed moderate recall values.

B. F1-Score

The F1-score combines precision and recall into a single metric.
SVM achieved the highest F1-score, demonstrating the best balance between precision and recall.
Naive Bayes and Random Forest also produced competitive results.

C. ROC-AUC Curve

The ROC curve evaluates classifier performance at various thresholds.

- SVM achieved the highest AUC score, close to 1.00.
- Naive Bayes and Random Forest performed strongly.
- Logistic Regression achieved slightly lower AUC values.

D. Confusion Matrix Analysis

The confusion matrix reveals accurate and inaccurate model predictions.
SVM produced the lowest false positives and false negatives, confirming its reliability.
Naive Bayes correctly classified most preferred movies but showed some misclassifications.
Random Forest produced slightly more false negatives.
Logistic Regression displayed moderate error rates.

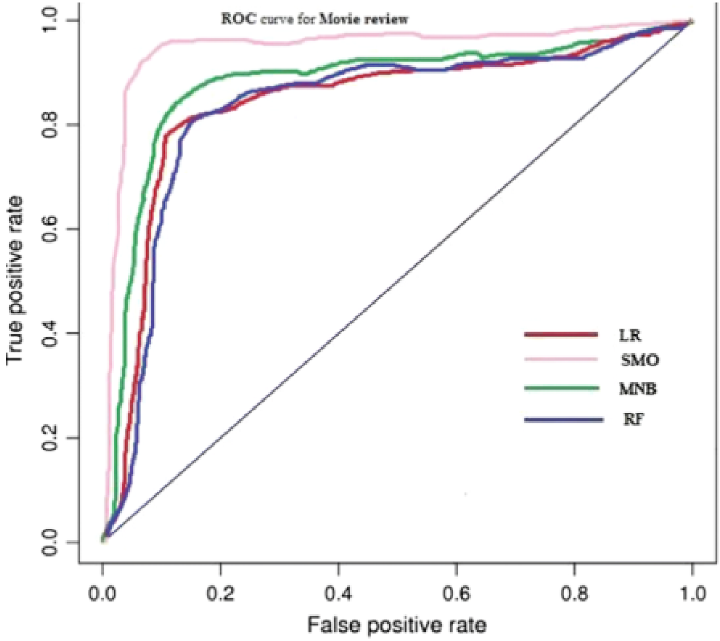
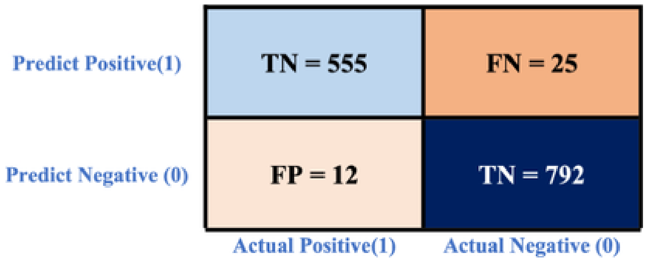


Fig. 1. ROC-AUC Curve comparison showing SVM's s



	Actual Positive(1)	Actual Negative (0)
Predict Positive(1)	TN = 555	FN = 25
Predict Negative (0)	FP = 12	TN = 792

Fig. 2. Confusion Matrix for the best performing model (SVM).

V. CONCLUSION

This study presents a TF-IDF–based movie recommendation system utilizing four machine learning classifiers.
Among them, Support Vector Machine (SVM) outperformed all other models with the highest Recall, F1-score, and ROC-AUC values.
The use of TF-IDF significantly improved recommendation accuracy over traditional approaches.
Future work includes implementing deep learning models such as LSTM or BERT for enhanced recommendation capabilities on larger datasets.

REFERENCES

[1] K. Deb, M. Alam, and S. Islam, "A content-based movie recommendation system using TF-IDF and cosine similarity," International Journal of Computer Applications, vol. 182, no. 45, pp. 12–18, 2021.

[2] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems: Introduction and Challenges," in Recommender Systems Handbook, Springer, 2015.