JOURNAL LA MULTIAPP

VOL. 05, ISSUE 05 (595-608), 2024
DOI: 10.37899/journallamultiapp.v5i5.1417

Movie Success Prediction Based on Feature and Trailer Comments Using Ensemble+LSTM Model

Nadya Sikana[1], Ronsen Purba[1]

[1]*Master of Information Technology, Informatics Faculty, Mikroskil University, Indonesia*

**Corresponding Author: Nadya Sikana*
E-mail: : nadya.sikana@students.mikroskil.ac.id

Article Info

Article history:
Received 14 August 2024
Received in revised form 07 September 2024
Accepted 28 September 2024

Keywords:
Ensemble Method
Movie Success Prediction
Sentiment Analysis

Abstract

Predicting the success of a movie is a very important aspect due to the high risks involved in movie production. The challenge lies in the uncertainty within the movie industry and selecting the appropriate machine learning model. We can combine movie features and sentiment analysis from social media using machine learning techniques to achieve movie success prediction. The methods used for predicting based on movie features are Ensemble models (Random Forest + Gradient Boosting). Meanwhile, the methods used for sentiment analysis of trailer comments is LSTM. The evaluation of the models used is based on RMSE and accuracy calculation. The final prediction of success obtains an RMSE of 0,8807 and an accuracy of 91,19%. This represents an improvement from previous research. Further research is recommended to implement the model in the movie industry.

Introduction

Predicting a movie's success is a crucial aspect due to the high risks and significant investments involved in movie production (Sindhu. & Shamsi, 2023). Filmmakers and investors need to make informed decisions early in the production process to enhance the chances of a movie's success. However, there are two main challenges in this regard. The first challenge is the uncertainty in the movie industry, with many parameters involved, such as the movie's genre, country of production, cast, director, trailer comments, social media impact, and audience preferences (Ruwantha & Kumara, 2020). The second challenge is choosing the right machine learning model (Kansari & Vijayant Verma, 2021). We can combine movie features and sentiment analysis from social media using machine learning techniques to predict a movie's success (Tripathi et al., 2023).

Agarwal et al. (2022) used an extensive IMDb dataset from Kaggle and applied various algorithms, including Simple Linear Regression, Multiple Linear Regression, and Artificial Neural Network. The features used included rating, genre, top-rated voters, total votes, duration, and release date. They found that Artificial Neural Network provided the best results with an accuracy of 86%. However, the study did not use several categorical features.

Sindhu & Shamsi (2023) conducted a study using an IMDb dataset as well as scraped data from Facebook for 210 movies. They compared two algorithms, Linear Regression and SVM. The features used were related to Facebook, including director FB likes, actor FB likes, actress FB likes, movie FB likes, movie budget, likes on FB movie page, and sentiment score of FB movie page. The results showed that SVM achieved an accuracy of 84% when the sentiment score from Facebook was added as a feature. This indicates that sentiment scores are also

ISSN: 2716-3865 (Print), 2721-1290 (Online)

important in predicting a movie's success. However, the limitation of this study was that the dataset used was too small, covering only 210 movies.

Ruwantha & Kumara (2020) classified Twitter posts to predict a movie's success. The dataset used consisted of scraped Twitter posts about 500 movies. They employed the LSTM algorithm. The attributes used were tweets and labels. The results showed that LSTM achieved an accuracy of 83.97%. In the conclusion, they suggested that future research should use ensemble methods to predict a movie's success.

Tripathi et al. (2023) used an IMDb dataset and a movie review dataset. They applied several algorithms, including Simple Regression Tree, Random Forest, Linear Regression, and several classification algorithms like Linear SVC and Naive Bayes Classifier. The features used for prediction included genre, duration, budget, crew popularity, and aspect ratio. The results showed that Linear Regression provided the best performance for predictions using movie features, while Linear SVC was the best algorithm for sentiment analysis with an accuracy of 88.47%. However, the study was incomplete in terms of feature usage. Features such as release year, cast, and country of production were not used.

Gandasari et al. (2023) used a Netflix dataset obtained through data scraping with Python. They applied several algorithms, including Random Forest, Naive Bayes, and K-Nearest Neighbor. The features used included movie type, release year, age certification, duration, genre, and several metrics such as IMDb score and TMDB popularity. The results showed that Random Forest had the highest accuracy of 81.59%. However, the study also did not use features like cast and country of production.

Ensemble algorithms, such as Random Forest and Gradient Boosting, combine the strengths of multiple models to enhance prediction accuracy. Ensemble methods build predictive models by using several different predictors, known as base learners. Base learners are simple models that typically do not perform better than random guessing, thus they are called weak learners. By aggregating the results of base learners, ensemble learning creates a strong learner capable of producing more accurate predictions (Vanneschi & Silva, 2023). Research across various sectors, including the movie industry, has demonstrated the effectiveness of ensemble machine learning approaches in improving prediction accuracy (Mhowwala et al., 2020). Zhang et al. (2022) evaluated the Gradient Boosting Random Forest model, which optimizes decision trees in Random Forest using Gradient Boosting. The evaluation results showed that Gradient Boosting Random Forest achieved 5% higher accuracy compared to Random Forest alone.

The LSTM algorithm has been widely used in sentiment analysis of movie reviews and has shown good performance. Bilen & Horasan (2022) compared the LSTM algorithm with other machine learning techniques for sentiment analysis on an IMDb dataset and found that LSTM performed superiorly.

This research aims to predict movie success using movie features and trailer comments uploaded on YouTube. It refers to the research by Tripathi et al. (2023), which suggested using other machine learning algorithms and sentiment analysis from different objects. Additionally, the study by Gandasari et al. (2023) is also referenced, using the same object, Netflix, but with ensemble models. Ensemble models are superior to single models in predicting movie success because they can combine the strengths of multiple models. Rating prediction based on features will be conducted using ensemble algorithms, namely Random Forest and Gradient Boosting. Meanwhile, sentiment analysis of trailer comments will be performed using the LSTM algorithm.

The structure of this article consists of 4 sections. Section 1, Introduction, covers the background of the problem, previous research, and the scope of the study. Section 2,

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Methodology, explains the steps taken to solve the problem. Section 3, Results and Discussion, presents the results and tests conducted also compares the findings of this study with similar research. Section 4, Conclusion and Suggestion, summarizes the conclusions of the study and provides suggestions for future research.

Methods

In Figure 1, the research stages begin with the collection of the dataset for training data from IMDb. Next, preprocessing is performed to clean and prepare the data before further analysis. The next stage involves building prediction models. The methods used for predicting based on movie features are Random Forest and Gradient Boosting. Meanwhile, the method used for sentiment analysis of movie trailer comments is LSTM. The results of the movie review sentiment analysis are then combined with movie features and re-predicted using the same model. The next stage involves model evaluation, where RMSE is calculated to measure model accuracy. After that, this model is used to predict movie ratings from new data not previously in the dataset.

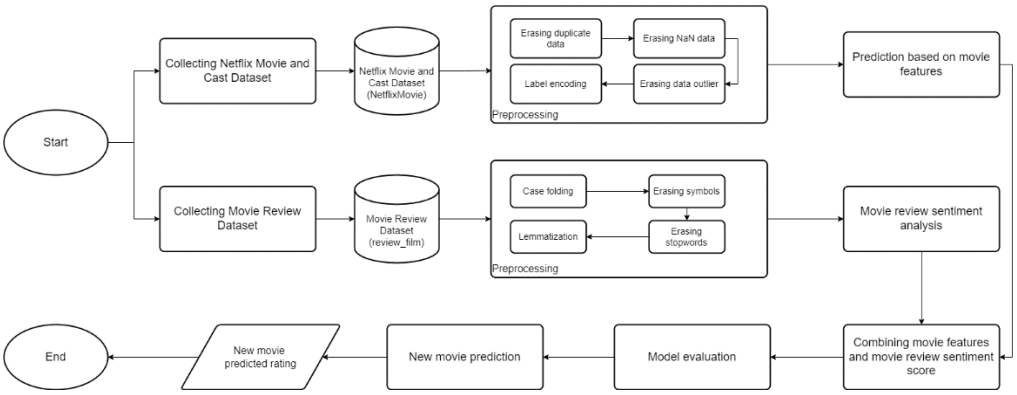

Figure 1. Research Methodology

In Figure 1, the research stages begin with the collection of the dataset for training data from IMDb. Next, preprocessing is performed to clean and prepare the data before further analysis. The next stage involves building prediction models. The methods used for predicting based on movie features are Random Forest and Gradient Boosting. Meanwhile, the method used for sentiment analysis of movie trailer comments is LSTM. The results of the movie review sentiment analysis are then combined with movie features and re-predicted using the same model. The next stage involves model evaluation, where RMSE is calculated to measure model accuracy. After that, this model is used to predict movie ratings from new data not previously in the dataset.

Data Collection

The dataset used for predicting movie success is obtained by scraping the IMDb website. Movie and cast data (NetflixMovie) are collected using Apify (accessed from https://console.apify.com/), while movie review data (review_film) is gathered using Beautiful Soup in Python.

Data Pre-Processing

For the movie and cast dataset, the preprocessing steps include removing duplicate data, empty values (NaN), outliers, and applying label encoding. They also preprocess the movie review dataset by performing case folding, removing symbols, eliminating stopwords, and applying lemmatization.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Random Forest

The Random Forest algorithm is used to predict movie success by analyzing various factors before the movie is released (e Souza et al., 2023). By applying this algorithm, the model can predict movie ratings based on movie characteristics and audience preferences from previous data (Shankhdhar et al., 2021). This algorithm is effective in overcoming feature limitations during the production stage, thereby assisting producers in making better decisions before the movie is released (Swami et al., 2021).

Random Forest consists of multiple decision trees that are combined. Preferences and influences among users are learned, and the overall model is used for prediction. Samples for each decision tree are randomly chosen, as are the variables for each subset of features. The regression results from each tree are different and averaged to obtain the final result of the Random Forest (Preprint et al., 2018). The workflow of Random Forest can be seen in Figure 2.

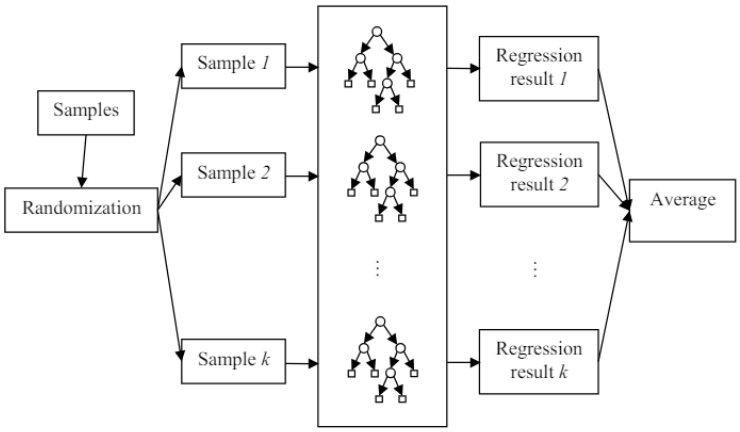

Figure 2. Random Forest Workflow

Gradient Boosting

The Gradient Boosting algorithm works by sequentially building prediction models through a series of weak prediction models, such as decision trees, to reduce errors and improve accuracy (Adetunji et al., 2020). By using Gradient Boosting, the movie industry can predict movie success before release, aiding strategic decision-making regarding marketing and release timing (Shankhdhar et al., 2021). This algorithm assigns weights to factors such as budget, cast, ratings, and social media data to predict movie success based on historical information, thereby enhancing prediction accuracy (Chakraborty et al., 2019). The pseudocode for Gradient Boosting can be seen in Figure 3.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Algorithm 1 Friedman's Gradient Boost algorithm

Inputs:

- input data $(x, y)_{i=1}^{N}$
- number of iterations M
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

1: initialize $\widehat{f}_0$ with a constant
2: **for** $t = 1$ to M **do**
3: compute the negative gradient $g_t(x)$
4: fit a new base-learner function $h(x, \theta_t)$
5: find the best gradient descent step-size ρ_t:

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi\left[y_i, \widehat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right]$$

6: update the function estimate:
$\widehat{f}_t \leftarrow \widehat{f}_{t-1} + \rho_t h(x, \theta_t)$
7: **end for**

Figure 3. Gradient Boosting Pseudocode (Natekin & Knoll, 2013)

In Figure 3, the Gradient Boosting algorithm by Friedman is presented. This algorithm is used to construct a robust prediction model by combining several weak prediction models. Its input consists of data pairs (x, y), where x is the feature vector and y is the target variable, along with the desired number of iterations, M. The algorithm also requires the selection of a loss function $\psi(y, f)$ and a base learner model $h(x, \theta)$.

The process begins by initializing the estimated function $\hat{f}_0$ with a constant. For each iteration from 1 to M, the algorithm computes the negative gradient $g_t(x)$ of the loss function with respect to the current estimated function. Then, the algorithm learns a new base learner model, $h(x, \theta_t)$, that corresponds to this negative gradient. Subsequently, the algorithm determines the optimal step size ρ_t through a gradient descent search on the loss function. Using this step size, the estimated function is updated by adding the product of ρ_t and the newly learned base learner model. This process is repeated until the specified number of iterations is reached (Natekin & Knoll, 2013).

Long-Short Term Memory (LSTM)

The Long Short-Term Memory (LSTM) algorithm is effective in analyzing sentiment in movie reviews because it can capture sequence dependencies in textual data (Dubey et al., 2023). LSTM networks excel in processing long text sequences by addressing the vanishing gradient problem through gate mechanisms such as the forgetting gate, which allows the model to retain important information over long periods (Zhang et al., 2022). LSTM was developed to overcome the issues of exploding and vanishing gradients often encountered during training of traditional RNNs. One of LSTM's advantages over traditional RNN, Hidden Markov Models, and other learning methods is its insensitivity to the length of gaps. There are several architectures of LSTM units. The typical architecture consists of a cell (the memory part of the LSTM unit) and three "gates" that control the flow of information within the LSTM unit: the input gate, output gate, and forget gate. Some variations of LSTM units may lack one or more of these gates or even introduce different types of gates. An illustration of the LSTM architecture can be seen in Figure 4.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

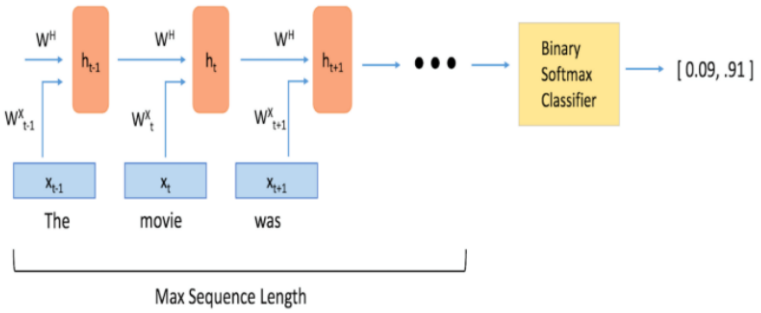

Figure 4. LSTM Architecture et al., 2020)

Equations for the forward pass of an LSTM with a forgetting gate are listed in equations 1, 2, 3, 4, 5, and 6 below.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$\tilde{c}_t = \sigma_g(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad (5)$$

$$h_t = o_t \circ \sigma_h(c_t) \quad (6)$$

In this model, the initial values are $c_0 = 0$ and $h_0 = 0$, and the operator $\circ$ denotes the Hadamard product (element-wise product). The subscript t indexes the time step. Here, σ is the sigmoid activation function, $tanh$ is the hyperbolic tangent activation function, x_t is the input at time t. Meanwhile, $W_i, W_c, W_f, W_o, U_i, U_c, U_f, U_o$ are weight matrices for input control, and b_i, b_c, b_f, b_o are bias vectors (N Murthy et al., 2020).

Evaluation

After the prediction steps using the mentioned method have been completed, the next step will involve evaluating the prediction results using RMSE (Root Mean Square Error). RMSE has been used as a standard statistical metric to measure model performance in studies of meteorology, air quality, and climate (Hodson, 2022). The formula for RMSE can be seen in equation 7 below.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (7)$$

With $\hat{y}_i$ representing the predicted values, y_i representing the observed values, and n indicating the number of observations. The smaller the RMSE, the better the results obtained.

Next, accuracy percentation can be calculated from RMSE using the following equation 8.

$$Accuracy\ Percentation = \left(1 - \frac{RMSE}{R}\right) \times 100 \quad (8)$$

Here, R denotes the range of the target variable, which in this case is 10, as movie ratings range from 0 to 10.

Results and Discussion

Results and Discussion

This section describes the results of the conducted research. The results of this study are divided into several stages as follows.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Data Reading

This stage uses the Python programming language with Google Colab to read the dataset. Python is a versatile programming language known for its robust libraries and frameworks, making it ideal for data manipulation, cleaning and analysis. Google Colab offers an accessible and powerful cloud-based environment, allowing for seamless collaboration and the execution of complex computations without the need for local resources. These are the reasons why we use Python and Google Colab. The movie and cast dataset consists of 4,021 records across 13 attribute columns. The movies included in the dataset are those available on Netflix and released between January 2000 and June 2024. The raw movie dataset can be seen in Table 1. The dataset encompasses various important attributes, including movie title (title), original title (ori), information on whether the title is an episode (isEp), duration (run), age certification (AC), year of release (year), IMDb rating (rating), number of ratings received (rCount), movie description (desc), list of main cast (stars), director (director), genre (genre) and country of production (country). In this study, the movie and cast dataset attributes used are runtime, AC, year, rating, stars, director, genre, and country. Runtime can influence viewer engagement, while AC can affect audience demographics. The year of release is crucial for understanding trends and the impact of technological advancements in movie production. The stars and directors are often critical to a movie's success due to their influence on public interest and marketing. Genre and country can provide cultural context, impacting viewer preferences and market performance. Rating provides direct indicators of a movie's reception and popularity.

Table 1 Raw Movie and Cast Dataset

title	ori	isEp	run	AC	year	rating	rCount	desc	stars	director	genre	country
American Psycho (2000)	NaN	FALSE	102	R	2000	7.6	722436	A wealthy New York …	Christian Bale, Justin Theroux, Josh Lucas	Mary Harron	Crime, Drama, Horror	United States
Paranoid (2000)	NaN	FALSE	93	PG-13	2000	4	2406	A fashion model, living in London …	Jessica Alba, Iain Glen, Jeanne Tripplehorn	John Duigan	Crime, Horror, Thriller	Hong Kong
Gladiator (2000)	NaN	FALSE	155	R	2000	8.5	1630665	A former Roman General sets out to …	Russell Crowe, Joaquin Phoenix, Connie Nielsen	Ridley Scott	Action, Adventure, Drama	United States
…	…	…	…	…	…	…	…	…	…	…	…	…
Maharaj (2024)	NaN	FALSE	131	R	2024	NaN	NaN	Based on a real-life …	Junaid Khan, Jaideep Ahlawat, Shalini Pandey	Malhotra P. Siddharth	Biography, Crime, Drama	United States
A Family Affair (2024)	NaN	FALSE	111	PG-13	2024	NaN	NaN	An unexpected romance …	Nicole Kidman, Zac Efron, Joey King	Richard LaGravenese	Comedy, Romance	United States
Under Paris (2024)	Sous La Seine	FALSE	104	R	2024	5.2	18688	To save Paris from a bloodbath …	Bérénice Bejo, Nassim Lyes, Léa Léviant	Xavier Gens	Action, Drama, Horror	United States

ISSN: 2716-3865 (Print), 2721-1290 (Online)

The movie review dataset consists of 109,523 records across 3 attribute columns. The raw movie review dataset can be seen in Table 2. This dataset comprises attributes including movie title (title), movie review (review), and the rating given by reviewers to the movie (rate).

Table 2. Raw Movie Review Dataset

Title	Review	Rate
American Psycho (2000)	I think best adaptation of a bret eston elli book (and maybe the most adaptable). This yuppie representation serial killer, is an unforgettable event and could be the best interpretation by Christian Bale. Christian Bale always build his body according to the role (you think Leonardo Di Caprio could look like that), you have to see "The Machinist" to confirm it, anad then he returns to his original shape with Batman Begins, etc., but this is a very introspective job and you can tell how well he was communicating with his director, the one and only Mary Harron. Excellent!. You can tell how well she knew story and how much she liked it.	10
…	…	…
Under Paris (2024)	If you want pure B movie cheer, this is it. This movie is so incredibly bad it's amazing. I would say it ranks right up there with the sharknado movies. Netflix is raising the bar extremely high for putting out garbage lately. I didn't think Atlas could be beat but Under Paris actually does it. There have been some decent movie released since the pandemic but the overall quality movies has plummeted over the last few years, the writing & directing has fallen off a cliff. It's a pity because there is a lot really good acting talent out there that is being wasted in movie like this. Hopefully it gets better.	1

Data Pre-Processing

In this stage, data preprocessing is conducted using the Python programming language with Google Colab. In the movie and cast dataset, the steps include removing duplicate titles, eliminating irrelevant columns, handling empty data (NaN), extracting primary genres from the genre column, and removing inappropriate data such as movies with only 1 or 2 cast members. The cleaned view of the movie and cast dataset can be seen in Table 3.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Table 3. Cleaned Movie and Cast Dataset

Title	Run Time	Ac	Year	Country	Cast_1	Cast_2	Cast_3	Director	Genre	Rating
American Psycho	102	R	2000	United States	Christian Bale	Justin Theroux	Josh Lucas	Mary Harron	Crime	7,6
Paranoid	93	PG-13	2000	Hong Kong	Jessica Alba	Iain Glen	Jeanne Tripplehorn	John Duigan	Crime	4,0
Gladiator	155	R	2000	United States	Russell Crowe	Joaquin Phoenix	Connie Nielsen	Ridley Scott	Action	8,5
…	…	…	…	…	…	…	…	…	…	…
How to Rob a Bank	131	R	2024	United States	Scott Scurlock	Ellen Glasser	Shawn Johnson	Stephen Robert Morse	Documentary	6,6
Baki Hanma VS Kengan Ashura	111	R	2024	United States	Nobunaga Shimazaki	Tatsuhisa Suzuki	Yutaka Aoyama	Toshiki Hirano	Animation	5,8
Under Paris	104	R	2024	United States	Bérénice Bejo	Nassim Lyes	Léa Léviant	Toshiki Hirano	Action	5,2

Then, several categorical columns such as AC, country, cast_1, cast_2, cast_3, director, and genre undergo encoding using a label encoder because prediction models can only work with numeric data. The label encoder converts each category into a unique numeric value based on alphabetical order. The total of cleaned data available for use is 3,161 records. The encoded view of the movie and cast dataset can be seen in Table 4.

Table 4. Encoded Movie and Cast Dataset

title	runtime	AC	year	country	cast_1	cast_2	cast_3	director	genre	rating
American Psycho	102	4	2000	56	1398	3629	3515	1558	5	7,6
Paranoid	93	3	2000	17	3261	2841	3177	1163	5	4,0
Gladiator	155	4	2000	56	5929	3330	1524	2010	0	8,5
…	…	…	…	…	…	…	…	…	…	…
How to Rob a Bank	131	4	2024	56	6148	2073	6245	2267	6	6,6
Baki Hanma VS Kengan Ashura	111	4	2024	56	5198	6617	7171	2405	2	5,8
Under Paris	104	4	2024	56	1148	5043	4318	2504	0	5,2

In the movie review dataset, several preprocessing steps are performed, including converting all text to lowercase (case folding), removing irrelevant symbols, eliminating stopwords, and converting words to their base form (lemmatization). Next, a label column is added based on the rate column, where if the rate is above 5, it is labeled as 1 (positive), and if the rate is 5 or less, it is labeled as 0 (negative). The total of cleaned data available for use is 109,523 records. An example of the cleaned text from the movie review dataset can be seen in Table 5.

Table 5. Cleaned Movie Review Dataset

Title	Review	Rate	Label
American Psycho (2000)	think best adaptation bret eston elli book maybe adaptable yuppie representation serial killer unforgettable event could best interpretation christian bale christian bale always build body accord role think leonardo di caprio could look like see machinist confirm return original shape batman begin etc introspective job tell well communicate director one mary harron excellent tell well knew story much like	10	1
…	…	…	…

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Under Paris (2024)	want pure movie cheer movie incredibly bad amaze would say rank right sharknado movie netflix raise bar extremely high put garbage lately think atlas could beat paris actually decent movie release since pandemic overall quality movie plummet last year write direct fall cliff pity lot really good act talent waste movie like hopefully get well	1	0

Building Predictive Model

The data split for building the feature prediction model is done with a ratio of 70% for training and 30% for testing. The movie feature prediction model utilizes an Ensemble approach, where the output from the Random Forest model is used as input for the Gradient Boosting model. The results show an RMSE of 0.9452 and an accuracy of 90.55%. Table 6 displays the comparison of RMSE and accuracy for the Linear Regression, Random Forest, Gradient Boosting, and Ensemble models.

Table 6. Comparison of Four Models RMSE and Accuracy for Prediction Based on Movie Features

Model	RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy
	9:1		8:2		7:3	
Linear Regression	1,0722	89,28%	1,0705	89,29%	1,0475	89,52%
Random Forest	0,9537	90,46%	0,9562	90,44%	0,9486	90,51%
Gradient Boosting	0,9583	90,42%	0,9802	90,20%	0,9924	90,08%
Ensemble	0,9514	90,49%	0,9530	90,47%	0,9452	90,55%

Building Sentiment Analysis Model

The data is split into 80% for training and 20% for testing in the sentiment analysis model for trailer comments. An LSTM model was chosen for this analysis after evaluating the model using batch sizes of 256 and 128, and epochs of 5 and 10. The results show that the configuration with batch size 128 and epochs 10 provides the most optimal performance with an accuracy of 88.42%. Table 7 displays the accuracy comparison for the LSTM model across each data split.

Table 7. Accuracy Comparison of LSTM Model for Trailer Comments Sentiment Analysis

LSTM		Accuracy		
Batch Size	Epochs	9:1	8:2	7:3
256	5	86,73%	86,46%	83,92%
	10	87,17%	87,54%	87,69%
128	5	87,06%	85,32%	81,61%
	10	87,43%	88,42%	87,22%

Training the LSTM model, especially with larger batch sizes and more epochs, resulted in long training times. To optimize and reduce the training time, we use GPU instead of CPU. Also early stopping was employed to halt training once the validation loss stopped improving, preventing unnecessary computations.

Combining Movie Features and Trailer Comments Sentiment Score

In this stage, the sentiment scores for each movie review are first analyzed using an LSTM model with a batch size of 128 and 10 epochs. After obtaining the sentiment scores for each review, these scores are averaged for each movie title to produce an average sentiment score. This average sentiment score is then combined with the movie features. Subsequently, a re-prediction of the ratings is performed using the same Ensemble model (Random Forest + Gradient Boosting). The combined dataset of features and reviews can be seen in Table 8. Upon combining them, the prediction model achieves an RMSE of 0.8807 and an accuracy of

ISSN: 2716-3865 (Print), 2721-1290 (Online)

91.19%. This represents an improvement compared to the prediction based solely on movie features, which yielded an RMSE of 0.9452 and an accuracy of 90.55%.

Table 8. Combined Dataset of Movie Features and Reviews

Title	Run Time	Ac	Year	Country	Cast_1	Cast_2	Cast_3	Director	Genre	Average Sentiment	Rating
American Psycho	102	R	2000	United States	Christian Bale	Justin Theroux	Josh Lucas	Mary Harron	Crime	0,45	7,6
Paranoid	93	PG-13	2000	Hong Kong	Jessica Alba	Iain Glen	Jeanne Tripplehorn	John Duigan	Crime	0,06	4,0
Gladiator	155	R	2000	United States	Russell Crowe	Joaquin Phoenix	Connie Nielsen	Ridley Scott	Action	0,44	8,5
…	…	…	…	…	…	…	…	…	…		…
How to Rob a Bank	131	R	2024	United States	Scott Scurlock	Ellen Glasser	Shawn Johnson	Stephen Robert Morse	Documentary	0,45	6,6
Baki Hanma VS Kengan Ashura	111	R	2024	United States	Nobunaga Shimazaki	Tatsuhisa Suzuki	Yutaka Aoyama	Toshiki Hirano	Animation	0,34	5,8
Under Paris	104	R	2024	United States	Bérénice Bejo	Nassim Lyes	Léa Léviant	Toshiki Hirano	Action	0,38	5,2

New Data Prediction

In this stage, 10 new data are provided that have never been seen by the prediction model or sentiment analysis. The movie features are obtained from the IMDb website (accessed from https://www.imdb.com/). The movie features data to be predicted can be viewed in Table 9.

Table 9. New Movie Features Data

Title	Year	Runtime	Ac	Country	Genre	Cast_1	Cast_2	Cast_3	Director
Blood & Gold	2023	98	R	Germany	Drama	Robert Maaser	Marie Hacke	Alexander Scheer	Peter Thorwarth
Chupa	2023	98	PG	United States	Adventure	Demián Bichir	Christian Slater	Evan Whitten	Jonás Cuarón
Damsel	2024	110	PG-13	United States	Action	Millie Bobby Brown	Ray Winstone	Angela Bassett	Juan Carlos Fresnadillo
Godzilla Minus One	2023	125	PG-13	Japan	Action	Minami Hamabe	Ryunosuke Kamiki	Munetaka Aoki	Takashi Yamazaki
Jung_E	2023	99	PG-13	South Korea	Sci-Fi	Kang Soo-youn	Kim Hyun-joo	Ryu Kyung-soo	Yeon Sang-ho
Lift	2024	107	PG-13	United States	Action	Kevin Hart	Gugu Mbatha-Raw	Sam Worthington	F. Gary Gray
My Oni Girl	2024	112	PG	Japan	Animation	Miyu Tomita	Kensho Ono	Shintaro Asanuma	Tomotaka Shibayama
Shehzada	2023	142	PG-13	India	Action	Kartik Aaryan	Kriti Sanon	Paresh Rawal	Rohit Dhawan
Spaceman	2024	107	R	United States	Sci-Fi	Adam Sandler	Carey Mulligan	Paul Dano	Johan Renck
The Wages of Fear	2024	106	R	France	Action	Franck Gastambide	Alban Lenoir	Sofiane Zermani	Julien Leclercq

The comments from the movie trailers are obtained from the Netflix YouTube account (accessed from https://www.youtube.com/@Netflix). 100 comments are retrieved for each movie. Examples of trailer comments for sentiment analysis can be seen in Table 10.

Table 10. New Movie Trailer Comments

Title	Comment
Blood & Gold	To those of you who are wondering what cover of Paint it Black this is, it is a 1970 cover by Karel Gott, also known as the Golden Voice of Prague. The cover title is Rot und schwarz.
Blood & Gold	Doesn't matter how many versions one has heard of Paint it Black .. every new version just adds more kick to it
Blood & Gold	Is it the German "Inglourious Basterds" ?
…	…
The Wages of Fear	Interesting ,can't wait.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

The Wages of Fear	French detected, trailer rejected.
The Wages of Fear	Sorcerer 1977

The categorical movie features are encoded using a label encoder, similar to what was done with the training dataset. Next, the trailer comments are analyzed using the previously created LSTM model. After generating sentiment scores for each comment, all scores are averaged per movie title. The movie feature data and average sentiment scores are then combined into one dataset. The combined data can be seen in Table 11.

Table 11. Encoded and Combined New Movie Data

Title	Year	Runtime	Ac	Country	Genre	Cast_1	Cast_2	Cast_3	Director	Sentiment
Blood & Gold	2023	98	4	15	7	2051	1752	118	1891	0,746018
Chupa	2023	98	2	56	1	610	548	924	1203	0,727302
Damsel	2024	110	3	56	0	1734	2232	219	1230	0,669241
Godzilla Minus One	2023	125	3	24	0	1737	2355	2004	2325	0,623551
Jung_E	2023	99	3	46	15	1289	1517	2384	2520	0,789245
Lift	2024	107	3	56	0	1358	1011	2411	720	0,69448
My Oni Girl	2024	112	2	24	2	1746	1488	2530	2400	0,697157
Shehzada	2023	142	3	19	0	1307	1552	2130	2072	0,630322
Spaceman	2024	107	4	56	15	25	461	2152	1153	0,794407
The Wages Of Fear	2024	106	4	14	0	811	76	2562	1246	0,628039

After predicting the new data using the created model, predicted ratings have been generated by the model. A comparison between the actual ratings and the predicted ratings by the model can be seen in Table 12.

Table 12. Comparison between Predicted Rating and Actual Rating

Title	Predicted Rating	Actual Rating
Blood & Gold	6,500044	6,5
Chupa	6,000009	5,5
Damsel	4,900051	6,1
Godzilla Minus One	6,601325	7,8
Jung_E	5,600093	5,5
Lift	5,400096	5,5
My Oni Girl	6,301622	6,0
Shehzada	6,799958	4,5
Spaceman	5,600093	5,7
The Wages of Fear	5,400096	4,5

From Table 12, it can be observed that movies like "Blood & Gold", "Jung_E", "Lift", "My Oni Girl", and "Spaceman" have good rating predictions because the difference between the predicted rating and the actual rating is less than or equal to 0.5. For movies like "Chupa" and "The Wages of Fear", the predictions are quite good with differences less than 1.0. However, for movies like "Damsel", "Godzilla Minus One", and "Shehzada", the predictions are not good as the differences are more than 1.0. The differences observed in movies like “Damsel”, “Godzilla Minus One” and “Shehzada” could be due to several factors. One potential reason is the complexity and uniqueness of these movies, which may not be fully captured by the features used in our models. To address these differences, future models could incorporate additional features such as historical performance data of similar movies. Additionally, exploring more advanced techniques such as deep learning models could improve prediction accuracy.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

This research refers to the works of Tripathi et al. (2023) and Gandasari et al. (2023). According to Tripathi et al. (2023), predicting movie success based on features and sentiment analysis of Twitter achieves an accuracy of 88.47%. Meanwhile, Gandasari et al. (2023) predict the success of Netflix movies and shows based on features with an accuracy of 81.59%.

In this research, a combined prediction approach for Netflix movies is conducted using both movie features and sentiment analysis of trailer comments on the Netflix YouTube account. This approach successfully increases the prediction accuracy to 91.19%, indicating an improvement over previous research. This demonstrates that the combination of movie feature analysis and public sentiment from trailer comments can provide more accurate predictions regarding the success of Netflix movies compared to previous approaches that rely solely on features.

Conclusion

Combining movie features with sentiment analysis of trailer comments for Netflix movies shows an improvement in prediction accuracy compared to using movie features alone, with an increase of approximately 0.64%. When predicting by combining movie features and trailer comments, the Ensemble approach such as Random Forest + Gradient Boosting combined with LSTM stands out with higher accuracy compared to methods like Linear Regression and Linear SVC, with a difference of approximately 2.72%. Further research is recommended to implement this model in the movie industry.

References

Adetunji, O., Hadiza, M., & Otuneme, N. (2020). Design of a movie review rating prediction (MR2P) algorithm. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 423–432. Technoscience Academy.

Agarwal, M., Venugopal, S., Kashyap, R., & Bharathi, R. (2022). Movie success prediction and performance comparison using various statistical approaches. *International Journal of Artificial Intelligence & Applications*, *13*(1), 19–36. Academy and Industry Research Collaboration Center (AIRCC).

Bilen, B., & Horasan, F. (2022). LSTM network based sentiment analysis for customer reviews. *Politeknik Dergisi*, *25*(3), 959–966. Politeknik Dergisi.

Chakraborty, P., Rahman, M. Z., & Rahman, S. (2019). Movie success prediction using historical and current data mining. *International Journal of Computer Applications*, *178*(47), 975–8887.

Dubey, G., Khera, R., Grover, A., Kaur, A., Goyal, A., Rajkuiiiar, Khatter, H., et al. (2023). A hybrid Convolutional Network and Long Short-Term Memory (HBCNLS) model for sentiment analysis on movie reviews. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(4), 341–348. Auricle Global Society of Education and Research.

e Souza, T. L. D., Nishijima, M., & Pires, R. (2023). Revisiting predictions of movie economic success: Random Forest applied to profits. *Multimedia Tools and Applications*, *82*(25), 38397–38420. Retrieved from https://doi.org/10.1007/s11042-023-15169-4

Gandasari, R. A., Wildan, M., Al-Abdillah, B. I., Nurzaman, A. F., & Anisa, N. (2023). Predicting over the top services movies and shows success using machine learning. *Proceedings of 2023 International Conference on Information Management and Technology, ICIMTech 2023* (pp. 89–94). Institute of Electrical and Electronics Engineers Inc.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

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. Retrieved from https://gmd.copernicus.org/articles/15/5481/2022/

Kansari, M., & Vijayant Verma, D. (2021). A movie success prediction based on public reviews using supervised learning. *Webology*, *18*(5), 3573. Retrieved from http://www.webology.org

Mhowwala, Z., Razia Sulthana, A., & Shetty, S. D. (2020). Movie rating prediction using ensemble learning algorithms. *IJACSA) International Journal of Advanced Computer Science and Applications*, *11*(8). Retrieved from www.ijacsa.thesai.org

N Murthy, G. S., Rao Allu, S., Andhavarapu, B., Bagadi, M., & Belusonti, M. (2020). Text based sentiment analysis using LSTM. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, *9*(5). Retrieved from www.ijert.org

Natekin, A., & Knoll, A. (2013). Gradient Boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(DEC). Frontiers Research Foundation.

Preprint, E., Ma, Y., & Gan, M. (2018). *A Random Forest Regression-based personalized recommendation method*.

Ruwantha, W. M. D. R., & Kumara, B. T. G. S. (2020). LSTM based approach for classifying Twitter posts for movie success prediction. *2020 International Conference on Decision Aid Sciences and Application (DASA)* (pp. 1160–1165).

Shankhdhar, A., Agrawal, V., & Rajpoot, V. (2021). Analysing movie success based on machine learning algorithm. *IOP Conference Series: Materials Science and Engineering*, *1119*(1), 012008. IOP Publishing. Retrieved from https://dx.doi.org/10.1088/1757-899X/1119/1/012008

Sindhu., I., & Shamsi, F. (2023). Prediction of IMDB movie score & movie success by using the Facebook. *2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT)* (Vol. I, pp. 1–5).

Swami, D., Phogat, Y., Batlaw, A., & Goyal, A. (2021). Analyzing movies to predict their commercial viability for producers. *arXiv*. Retrieved from http://arxiv.org/abs/2101.01697

Tripathi, J., Tiwari, S., Saini, A., & Kumari, S. (2023). Prediction of movie success based on machine learning and Twitter sentiment analysis using internet movie database data. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(3), 1750–1757. Institute of Advanced Engineering and Science.

Vanneschi, L., & Silva, S. (2023). Ensemble methods. *Natural Computing Series* (pp. 283–288). Springer Science and Business Media Deutschland GmbH.

Zhang, H., Xu, J., Lei, L., Jianlin, Q., & Alshalabi, R. (2022). A sentiment analysis method based on Bidirectional Long Short-Term Memory networks. *Applied Mathematics and Nonlinear Sciences*. Sciendo.

Zhang, Z., Zhu, X., & Liu, D. (2022). Model of Gradient Boosting Random Forest prediction. *2022 IEEE International Conference on Networking, Sensing and Control (ICNSC)* (pp. 1–6).

ISSN: 2716-3865 (Print), 2721-1290 (Online)