Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/18/20230998

# Personalized movie rating prediction based on the data of short video users and videos' information

**Yizhe Wang**

Chinese University of HongKong ShenZhen, Shenzhen, 0755, China

wangyizhe@link.cuhk.edu.cn

**Abstract.** This study presents a sophisticated personalized movie rating prediction model that incorporates user-generated content from Douyin, a prominent video-based social media platform. The proposed model effectively leverages user demographic data, movie metadata, historical rating records, and sentiment analysis of user comments to precisely predict a user's rating score for films. This study employs lexicon-integrated two-channel CNN-LSTM family models for conducting sentiment analysis. Furthermore, the study explores a variety of regression models, such as linear regression, ridge regression, LASSO regression, and Elastic Net regression, to determine the most suitable model for the dataset being analyzed. This analysis aims to enhance the accuracy and effectiveness of the sentiment analysis performed on the given dataset. Furthermore, we contemplate the application of collaborative filtering algorithms, such as Alternating Least Squares (ALS), in developing our personalized movie rating prediction model. By incorporating user-generated content from Douyin, our methodology enhances the accuracy of movie rating predictions and underscores the significance of personalized recommendations in the era of social media.

**Keywords:** movie rating prediction, short video, Douyin.

## 1. Introduction

Movie rating prediction is an essential research topic with significant implications for various stakeholders in the movie industry, including producers, distributors, and streaming platforms. Accurate movie rating prediction can inform the development of marketing strategies, improve recommendations, and optimize revenue generation. However, traditional movie rating prediction methods rely solely on users' historical rating records, which may not provide a complete picture of user preferences. For instance, users are only asked to rate movies from 0-5 stars, which does not capture the nuances of their opinions accurately. Movie audiences are more likely to pay 20% to 99% more for five-star ratings than for four-star ratings, according to certain studies [1]. Furthermore, users may rate movies differently based on their individual personalities, preferences, and experiences, which can impact the accuracy of the prediction model. For example, even if two people feel the same way about a movie, they may give it different rating stars. This suggests that if the model uses this inaccurate data for prediction, the results may not reflect the real conditions. Ignoring these individual characteristics can lead to inaccurate predictions, and ultimately, poor decision-making.

Several studies have investigated various methods for predicting movie ratings. Collaborative filtering (CF) has found extensive application in e-commerce and online services to provide

DOI: 10.54254/2755-2721/18/20230998

personalized recommendations to users based on their historical rating records [1]. Herlocker et al. evaluated CF recommender systems and discussed different approaches, such as user-based, item-based, and model-based methods. They also proposed a framework for evaluating CF systems based on accuracy, diversity, serendipity, and novelty [2]. However, traditional CF methods may not capture individual user preferences and personalities, leading to inaccurate predictions. To address this issue, some studies have incorporated user-generated content and other metadata into movie rating prediction models.

Social media platforms have been increasingly used to enhance personalized movie rating prediction. For instance, Oghina et al. predicted IMDB movie ratings using social network analysis and user behavior data, including tweets and blog posts on Twitter, and then used machine learning algorithms to combine these data to build a model that predicts movie ratings. They used machine learning algorithms to combine these data to build a model that predicts movie ratings [3].

Recently, some studies have proposed hybrid models that combine CF with content-based methods or sentiment analysis of user-generated content. For example, The authors, Mehdi Elahi et al., propose a novel hybrid recommender system, which improves the accuracy and effectiveness of the recommender system by analyzing the sentiment information of user reviews. Compared with traditional rating-based recommender systems, this method utilizes reviews and sentiment information to reflect user preferences more comprehensively, and experiments on two real datasets demonstrate its superiority. This method has great research value and application prospect [4]. But it is not currently used in the movie rating prediction model.

However, these studies may not be directly applicable to the Douyin platform as it has unique characteristics. Douyin is a video-based platform that differs significantly from text-based platforms like Twitter,which were the primary data sources used in previous studies. Additionally, the user demographic on Douyin is different from that on other social media platforms, as it is primarily young and mobile-first. Additionally, Douyin has a massive market and user base in China, and analyzing its data will play a significant role in many predictive models.Therefore, previous methods may not be optimal for capturing user preferences and behaviors on Douyin, and there is a need for research specific to this platform [5].

Data from Douyin may be more useful for predicting movie ratings than data from Twitter for a number of reasons. First, compared to Twitter's text-based material, Douyin's video-based platform offers more in-depth and varied content. Users now frequently express their thoughts and experiences on social media sites like Douyin on a variety of subjects, including movies. The model can learn about people' opinions on movies, their tastes, and their personalities by examining users attitudes towards different kinds of videos. Second, a key demographic for the film industry is Douyin's user base, which is primarily young and mobile-first. Moreover, compared with Twitter, Twitter is not available in China, and the characteristics of their user groups are completely different.This demographic is a key target for movie marketers because it is thought to have more influence over cultural trends. Therefore, for a movie to succeed, it can be essential to understand their preferences and attitudes toward cinema.

This study aims to address this gap by proposing a personalized movie rating prediction model that incorporates informations from Douyin to enhance the accuracy. There are 3 contributions in this study: 1) This study fills the gaps in the existing literature and further expands and enriches the existing research in this field. 2) This study proposes a novel personalized movie rating prediction model that utilizes data from the Douyin platform, capturing the unique preferences and behaviors of its users to improve the accuracy of predictions. 3) The proposed model has the potential to greatly benefit various stakeholders in the movie industry, such as producers, distributors, and streaming platforms, enabling them to make more informed decisions, optimize marketing strategies, and ultimately, enhance revenue generation.

DOI: 10.54254/2755-2721/18/20230998

## 2. Research framework

As shown in figure 1, the research framework of this paper consists of two steps: (1) predicting the user's movie rating based on their historical ratings, and (2) enhancing the prediction using personalized data such as the user's short video viewing history, likes, comments, and other relevant information.

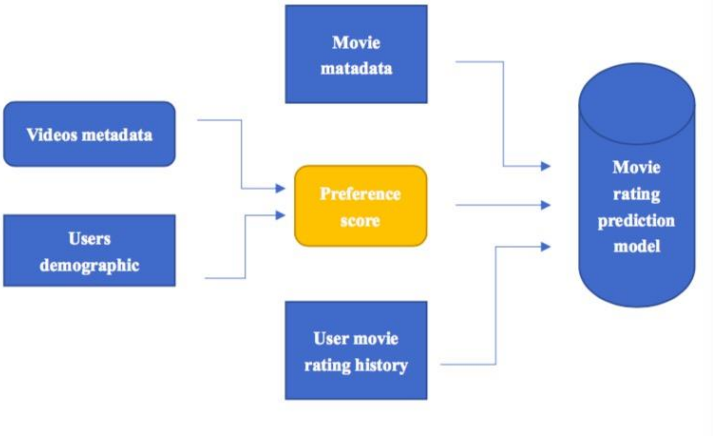


**Figure 1.** The overall framework of model.

## 3. Data set

The data set will be collected from Douyin and will include user-generated content related to movies, such as video clips, comments, and likes. Additionally, we will collect user demographic information, including age, gender, and location, as well as movie metadata such as short video genre, director, actors, and release date. We will also supplement the movie metadata using external sources, such as IMDb and Rotten Tomatoes.

The data set will be further divided into the following categories shown in table 1.

**Table 1.** The main content of the data.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| User demographic data | age | gender | location | occupation | education | income | User id |
| Movie metadata | genre | director | actors | release date | duration | Movie id | whether to share |
| User movie rating history | Movie id | Rating score | | | | | |
| User short video metadata | Video genre | Like or not | Content of comments | Viewing time | Video id | | |

## 4. Data preprocessing

### *4.1. Data cleaning*

When performing data cleaning, it is important to note that systematic data errors can significantly affect the reliability of model training. Dirty data can significantly impact the performance of machine learning (ML) models, and the high-dimensional nature of these models may produce unexpected outcomes when trained on data that has undergone certain types of cleaning procedures. As such, it is

DOI: 10.54254/2755-2721/18/20230998

crucial to be mindful of data cleanliness when working with ML problems to avoid misleading results. To avoid these issues, it is crucial to incorporate error detection and correction techniques as a part of the data cleaning process [6].

1) Duplicate content removal: We may encounter duplicate content in the dataset, which may skew the analysis results. These duplicates must be removed.

2) Handling missing data: Some records, such as incomplete user profiles, may have missing information. We may need to delete these records or use data imputation methods to fill in the gaps.

3) Error correction: The data may contain errors such as misspellings or incorrect information. These errors must be corrected before proceeding with the analysis.

4) Data format standardization: The data may be in a variety of formats, such as different date formats or inconsistent naming conventions. Standardizing the format will help ensure consistency and ease of analysis.

5) Removing stop words: Stop words are a type of common words that carry little or no significant information in the text, and thus, they should be eliminated to enhance the accuracy of sentiment analysis. This removal of stop words plays a vital role in reducing the noise in the data and producing more reliable results.

6) Lemmatization or stemming: To improve the precision of sentiment analysis and detect patterns in the data, performing techniques like lemmatization or stemming are necessary. These techniques aim to normalize words and group their inflected forms, which can lead to more accurate analysis and better understanding of the sentiment conveyed in the text.

### *4.2. Comments analysis*

Before comments analyzing, acquiring a firm grasp of sentiment analysis is imperative in the field of natural language processing. Sentiment analysis refers to the process of extracting subjective information from textual data, such as product reviews or social media posts, and is also referred to as opinion mining. Understanding this process is essential for accurately interpreting the sentiment conveyed in textual data and making informed decisions based on the outcomes. It entails analyzing the language used in the text to determine the writer's overall sentiment or opinion [7].

Sentiment analysis will be used in our project to analyze user-generated content on Douyin, such as comments and like, to determine user preferences and attitudes toward movies. We can gain insights into how users feel about certain movies and what factors may be influencing their opinions by extracting sentiment information from user comments.

Table 2 and table 3 are some examples of judging attitudes towards videos in sentiment analysis by analyzing words and emoji expressions in comments:

**Table 2.** Examples of words in comments to do sentiment analysis.

| Positive words | Negative words |
|---|---|
| Like | Dislike |
| Interesting | Boring |
| Amazing | Terrible |
| Incredible | Awful |
| Impressive | Disapointing |
| Awesome | Disgusting |
| Creative | Dull |

DOI: 10.54254/2755-2721/18/20230998

**Table 3.** Examples of emoji expressions in comments to do sentiment analysis.

| Positive emojis | Nagative emojis |
|---|---|
| 😊 (Smiling) | 😔 (Sad) |
| 😂 (Joyful) | 😡 (Angry) |
| 🔥 (Fire) | 😞 (Disappointed) |
| 👍 (Thumbs up) | 💔 (Broken heart) |
| ❤️ (Love) | 👎 (Thumbs down) |
| 😍 (Heart eyes) | 😒 (Unamused) |
| 🎉 (Celebration) | 😩 (Tired) |

In the model proposed, the sentiment analysis of user-generated content on Douyin is performed using lexicon-integrated two-channel CNN-LSTM family models, specifically focusing on comments and likes. This approach has been demonstrated to be effective in previous studies, such as the work conducted by Li [8] which applied a similar methodology to analyze sentiment in textual reviews. Notably, the proposed method features a novel padding technique called sentiment padding, ensuring consistent input data size and improving the proportion of sentiment information in each review. Experimental results on various datasets, including the Stanford Sentiment Treebank, have demonstrated the superiority of this approach over many baseline methods [9].

By employing this approach for sentiment analysis, the accuracy and efficiency of the prediction model can be enhanced. The fusion of CNN and LSTM/BiLSTM branches provides an effective means of capturing sentiment information in user comments while considering the sequential nature of words and their interrelationships. Consequently, this methodology enables the extraction of insights into user attitudes and preferences toward different types of short videos based on their comments on the Douyin platform. Leveraging this analysis, user characteristics can be accurately discerned and combined with movie metadata, such as genre, director, and actors, to refine the prediction of the user's movie rating. Relevant features are extracted from the preprocessed data to train the personalized movie rating prediction model. These features include user demographic information, such as gender, age, and location, historical rating records, and sentiment scores. Additionally, features from the movie metadata, such as genre, director, actors, and release date, are extracted. Identifying patterns and trends in users' behavior and preferences is facilitated through the analysis of these features.

## 5. Data combination

To enhance the accuracy of the movie rating prediction model, this study analyzes user behavior on Douyin, such as the types of short videos they watch, their likes, comments, and viewing time. By utilizing sentiment analysis techniques, it is possible to classify the sentiment of user comments to gain insights into their preferences and attitudes towards different types of short videos.

For instance, if a user frequently watches and engages positively with humorous short videos, it can be inferred that they have a liking for comedic content. Similarly, if a user watches and engages positively with fitness-related short videos, it can be inferred that they have an interest in fitness-related content.

Based on the degree of liking for each type of short video, a score can be assigned to indicate the level of preference and combined with the user's historical movie rating records to predict movie

DOI: 10.54254/2755-2721/18/20230998

ratings accurately. A similar approach can also be used to predict a user's movie rating based on their interest in specific movie genres, which can be determined by analyzing their engagement with relevant content on Douyin and combining it with their historical ratings for movies of those genres.

Besides, movie commentary in Douyin is an essential factor for movie rating. Thus, crawler technology is used in sort out the short video related to movie commentary. Then put a higher weight on the indicators in movie commentary video than user behavior in the movie rating prediction model, as the movie commentary has lager influence power.

Model selection and training are crucial steps in building a movie rating prediction model. In the following sections, a detailed breakdown of these steps is provided, along with explanations of the rationale behind each decision.

## 6. Methodology

### *6.1. Model selection and training*

To develop a personalized movie rating prediction model, machine learning algorithms, specifically regression models, are utilized to predict the user's rating score. Regression models are a suitable choice for this problem because they can handle continuous target variables, such as the movie rating score.

To identify the optimal model for the dataset, a variety of regression models are examined, such as linear regression, ridge regression, LASSO regression, and Elastic Net regression, and their performance is evaluated.

Cross-validation techniques are also utilized to ensure that the model is not overfitting and to estimate its performance on unseen data.

Upon selecting the best model, it is trained on the preprocessed dataset, using user demographic information, movie metadata, historical rating records, and sentiment scores as features. Feature engineering techniques, such as feature scaling and feature selection, are employed to ensure a robust and accurate model.

Additionally, hyperparameter tuning is performed to optimize the model's performance. In machine learning, hyperparameters are parameters that can be tuned to control the behavior of a model, including but not limited to the regularization strength and learning rate. Tuning these parameters can improve the model's accuracy and prevent overfitting.

Finally, the performance of the personalized movie rating prediction model is evaluated on a test set and compared with traditional movie rating prediction models that only rely on historical rating records. The model's correctness is measured using common assessment metrics like mean squared error (MSE) and root mean square error (RMSE).

In addition to regression models, collaborative filtering algorithms, such as Alternating Least Squares (ALS), can also be considered for developing a personalized movie rating prediction model. ALS is a matrix factorization technique that is commonly used in recommendation systems.

According to Figure 2, ALS divides the user-item rating matrix into two lower-dimensional matrices, one of which represents user preferences and the other item attributes. These matrices are then multiplied to reconstruct the original rating matrix. By minimizing the difference between the predicted and actual ratings using an iterative process, ALS can learn the latent features that influence user preferences and item characteristics [10].

DOI: 10.54254/2755-2721/18/20230998

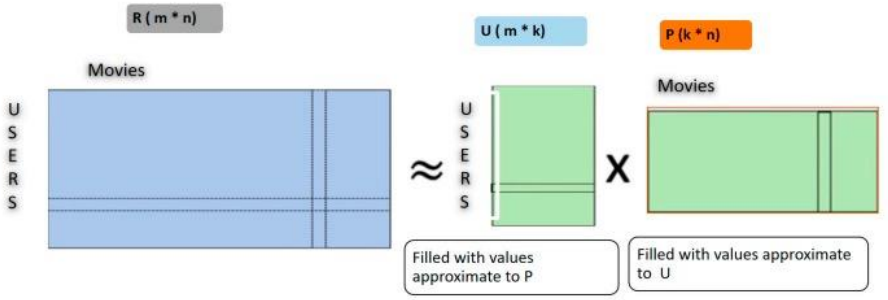


**Figure 2.** ALS matrix factorization[11].

One advantage of ALS is its ability to handle large and sparse datasets, which is common in movie recommendation systems. Additionally, ALS can capture complex user-item interactions and provide personalized recommendations even for new or unseen items.

To employ ALS for movie rating prediction, the dataset must be preprocessed to create a user-item rating matrix. The matrix would contain the historical rating records of users for movies, as well as any additional features extracted from Douyin, such as sentiment scores or interest in specific genres. Subsequently, ALS can be applied to the matrix to learn the latent features that influence movie ratings and predict the ratings for new movies.

User-item rating matrix: The initial step involves creating a user-item rating matrix that captures the historical ratings of users for movies. This matrix can also include additional features extracted from Douyin, such as sentiment scores or interest in specific genres.

ALS divides the user-item rating matrix into two lower-dimensional matrices, one expressing user preferences and the other representing item attributes. This process is known as matrix factorization. The original rating matrix is then recreated by multiplying these matrices [11].

Learning the latent features: By minimizing the difference between the predicted and actual ratings using an iterative process, ALS can learn the latent features that influence user preferences and item characteristics. These features can capture complex user-item interactions and provide personalized recommendations even for new or unseen items.

Predicting movie ratings: Once the model has learned the latent features, it can predict the ratings for new movies by multiplying the user preferences and item features matrices. The evaluation process is similar to that of the regression model mentioned earlier.

## 7. Conclusion

In this study, a personalized movie rating prediction model that incorporates user-generated content from the Douyin platform is proposed. By integrating movie-related metadata, user behavior data, and sentiment analysis of user-generated content, the approach aims to capture individual user preferences and personalities to enhance the accuracy of movie rating prediction. Various regression models, as well as collaborative filtering algorithms such as ALS, are explored for model selection and training. The proposed model has significant implications for the movie industry, offering a new and more effective approach to movie rating prediction and enhancing the user experience. The results of the experimental evaluation demonstrate the effectiveness of the approach in improving the accuracy of movie rating prediction. Future research may explore additional data sources and feature engineering techniques to further enhance the accuracy of the model.

However, a limitation of this study is the inability to access the necessary data from the Douyin platform to put the model into practice. This presents an opportunity for improvement in future research, which may focus on obtaining relevant data from the platform to validate and refine the proposed model.

DOI: 10.54254/2755-2721/18/20230998

## References

[1] Wu, J., Liu, Y., & Tan, T. (2019). Personalized movie recommendation system based on user's preference mining from Weibo. Information, 10(1), 1. https://doi.org/10.3390/info10010001

[2] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1), 5-53.

[3] Oghina, A., Breuss, M., Tsagkias, M., & De Rijke, M. (2012, April). Predicting IMDB Movie Ratings Using Social Media. In ECIR (pp. 503-507).

[4] Elahi, M., Khosh Kholgh, D., Kiarostami, M. S., Oussalah, M., & Saghari, S. (2023). Hybrid recommendation by incorporating the sentiment of product reviews. Information Sciences, 625, 738-756. https://doi.org/10.1016/j.ins.2023.01.051

[5] Chen, Z., He, Q., Mao, Z., Chung, H.-M., & Maharjan, S. (2019). A Study on the Characteristics of Douyin Short Videos and Implications for Edge Caching. arXiv preprint arXiv:1903.12399. https://arxiv.org/pdf/1903.12399.pdf

[6] Krishnan, S., Wang, J., Wu, E., Franklin, M. J., & Goldberg, K. (2015). Activeclean: Interactive data cleaning while learning convex loss models. arXiv preprint arXiv:1601.03797.

[7] K. Mouthami, K. N. Devi, & V. M. Bhaskaran. (2013). Sentiment analysis and classification based on textual reviews. In 2013 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 271-276). Chennai, India: IEEE. https://doi.org/10.1109/ICICES.2013.6508366

[8] Li, W., Zhu, L., Shi, Y., Guo, K., & Cambria, E. (2020). User reviews: Sentiment analysis using lexicon integrated two-channel CNN-LSTM family models. Applied Soft Computing, 96, 106435. doi: 10.1016/j.asoc.2020.106435.

[9] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37. https://doi.org/10.1109/MC.2009.263

[10] Netflix. (2006). Recommender Systems [Netflix]. Retrieved from https://datajobs.com/data-science-repo/Recommender-Systems-%5BNetflix%5D.pdf

[11] Awan, M. J., Khan, R. A., Nobanee, H., Yasin, A., Anwar, S. M., Naseem, U., & Singh, V. P. (2021). A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. Electronics, 10(10), 1215. https://doi.org/10.3390/electronics10101215