Sentiment Analysis of Movies Based on Natural Language Processing

Hanfei Zhu(✉)

School Of Management And Economics, The Chinese University of Hong Kong, Shenzhen 518172, China
hanfeizhu@link.cuhk.edu.cn

Abstract. Movies are an important spiritual food for people to satisfy their spiritual needs in contemporary society. Each person's evaluation of one movie is different. The data mining of movie-related data can make a certain accuracy prediction of the final rating of a movie. This paper shows a new movie rating prediction model is proposed by using nlp technology. It analyzes the data recorded in Movielens for nearly 45,000 movies. Some of the selected data include final rating, movie duration, title, budget, story overview, genre, and so on. This includes a large amount of textual data that needs to be converted into vector data using nlp techniques. In this paper, random forest and neural network are chosen as the main models and a lot of tuning is done to obtain a relatively accurate model, i.e., a random forest movie rating prediction model using natural language processing. This research can be used for upcoming or newly released movies that lack ratings, which is a good reference for audiences' viewing choices and movie producers' promotion and investment decisions.

Keywords: Natural Language Processing · Movie Emotions · Random Forests · Neural Networks

1 Introduction

Cinema has been around for more than 100 years, from the original black-and-white silent movies to today's IMAX and Dolby Atmos technology, etc. The film industry has developed very rapidly, and in 2019, the worldwide box office reached $42.2 billion. As a comprehensive audio-visual art that integrates music art, photography art, fine art and theater art, film has enriched people's spiritual world and is an indispensable and important part of modern life.

Thousands of movies are produced every year, but a movie is often at least 90 mi.

nutes long. And movies of more than 2 h in length are gradually becoming mainstream, and some movies are even three to four hours long. Today's society is very fast-paced, people can not spare much time to watch movies, it is impossible to watch all the movies released, so there must be a trade-off. So movie evaluation platforms were created, where people who have seen a certain movie evaluate the movie according to their own feelings and understanding of the movie, often scoring, commenting on or

X. Yuan et al. (Eds.): ICEKIM 2023, AHCS 13, pp. 1232–1240, 2023.
https://doi.org/10.2991/978-94-6463-172-2_130

simply choosing good or bad movies. In this way, a large amount of data is stored in the movie rating platform, and people can then choose whether they want to watch the movie based on the final average rating, the number of people who rated it and the reviews, thus being able to save a lot of time.

Nowadays, there are mainly two kinds of rating methods in the mainstream movie review websites, one is from 1 to 10, and finally the final score is presented by some algorithm, such as IMDB, Metacritic, movielens, etc., and the other is only good or bad evaluation, such as rotten tomatoes, movielens. Most of these sites accept ratings from all people on the web, while others combine general audience ratings with expert and media ratings (e.g. Metacritic, rotten tomatoes) to give a more detailed and professional reference. This wealth of movie rating data has made the study of movie ratings feasible and has inspired many people to participate in movie rating research.

Although everyone's subjective evaluation of the film is different, and the elements concerned about a good film are also different, the film with higher score is more likely to be a high-quality film, and In the absence of other better channels for consumers to understand the quality of a movie, choosing whether to watch a movie based on its rating is really a best choice. By studying the influencing factors and prediction models of ratings, consumers can understand the inner mechanism of ratings, so that they can look at all aspects of a movie, not just the number of ratings, then judge the movie better.

For movie suppliers in the movie industry, the prediction model of movie ratings becomes more important. Although there are some movies with high ratings but low box office in the movie market, in general, there is a certain positive relationship between movie ratings and box office. Since it takes a lot of money, staff resources and time cost to produce a movie, if there is a model that can predict the approximate rating of a movie, then it can roughly predict the upper limit of box office that the movie can reach, which can provide an important reference for the investors of the movie whether to invest or not and the scheduling of the cinema. Therefore, the research on movie rating prediction is bound to have a significant impact on the whole movie industry.

Section 2 summarizes the existing research on movie scoring using traditional or current advanced algorithms, Sect. 3 data research provides a detailed description of the dataset used in this study, Sect. 4 models provides a detailed introduction and explanation of the algorithmic model used in this study, and Sect. 5 result provides a rich interpretation of the results of this study, and Sect. 6 conclusion summarizes the strengths and weaknesses of this study and the future directions for improvement.

2 Literature Review

The author applied SVD(Singular Value Decomposition) with a Bayesian approach to alleviate overfitting, which gave significantly improved results over vanilla SVD. The author focused on the user's conformity in rating with a matrix-factorization-based conformity modeling technique. The result was a significant improvement compared to RMSE, so conformity modeling is important for movie rating studies [1].

In fear that the common movie review and ratings would be corrupted by fake news, the author developed a model using machine learning and deep learning algorithms to rate a movie based on the emotion of the spectators by initiating a real-time analysis.

The auther used data from Kaggle with some other data from Facebook, Youtube and OMDb APIs, to study models to predict movie ratings, and gave its best result using J48 decision tree algorithm with bagging. By transforming the 1–10 rating result to only three categories——"Great", "OK", and "Poor", it gave a higher accuracy [2].

Many research have been conducted using machine learning algorithms such as Random Forest. One of those researchers collected movie data from IMDB and social media data from YouTube and Wikipedia, then compared the performance of two machine learning algorithms—Random Forest and XGboost.(2) Another auther used data from Kaggle with some other data from Facebook, Youtube and OMDb APIs, to study models predicting movie ratings, and gave its best result using J48 decision tree algorithm with bagging. By transforming the 1–10 rating result to only three categories——"Great", "OK", and "Poor", it gave a higher accuracy.(1) In a research to predict movie ratings before the movie is released, the author used data from IMDB including IMDB score, director, gross, budget and so on to train the mode. It concluded that Random forest gave the best prediction accuracy and number of voted users, number of critics for review, numbers of Facebook likes, duration of the movie and gross collection of movie affect the score most strongly [9].

Using neural network framework to predict movie ratings based on the estimating parameters was explored in early years [8]. However, in recent years, neural network has been more and more popular. One of these works include a regression model based on generative convolutional neural networks for movie rating prediction. It used the pre-release and intrinsic attributes such as genres, budget, cast, director and plot information [7]. Another author used neural network Levenberg-Marquardt back propagation algorithm to predict movie ratings to train six parameters and their individual ratings. The data set contains of 150 movies and the data is collected from IMDB. The proposed system achieves an accuracy of 97.33% and a sensitivity of 98.63% [9].

There are also other models applied in this field. In fear that the common movie review and ratings would be corrupted by fake news, a research developed a model using machine learning and deep learning algorithms to rate a movie based on the emotion of the spectators by initiating a real-time analysis [3]. Traditional methods have also been implemented. SVD(Singular Value Decomposition) is one of them. One author applied SVD with a Bayesian approach to alleviate overfitting, which gave significantly improved results over vanilla SVD [4]. In another work, the author proposed a method of rating prediction based on Gaussian Mixture Model(GMM), which can avoid the influence of malicious rating because GMM is not sensitive to exception. With 4 features of the movies taken into account and data from DouBan, an improved performance of rating prediction is achieved compared with the benchmark of linear regression [5]. Besides, one author focused on the user's conformity in rating with a matrix-factorization-based conformity modeling technique. The result was a significant improvement compared to RMSE, so conformity modeling is important for movie rating studies [6].

3 Data Research

The dataset for this research is obtained from the website Kaggle. This dataset consists of information for 45,000 movies that are recorded in the MovieLens Dataset, including TMDB vote counts, vote averages, titles, languages, production companies, countries,

Table 1. 1. Statistics of features

	Mean	Std	Min	25%	50%	75%	Max
Rating	3.43	0.55	1	3.1	3.5	3.84	4.75
Adult	0	0	0	0	0	0	0
Budget	3.00E + 07	4.39E + 07	0	2.5	1.25E + 07	4.00E + 07	3.8E + 08
Popularity	1546.44	891.66	0	776.5	1543	2316.5	3092
Revenue	1.05E + 08	1.89E + 08	0	6.98E + 05	3.38E + 07	1.23E + 08	2.79E + 09
Runtime	111.29	21.98	5	97	108	122	287
Vote_Average	6.66	0.8	2	6.2	6.7	7.25	8.5
Vote_Count	917.2	1485.89	1	126	338	996	14075

overview, cast, crew, plot keywords, tagline, budget, revenue, posters, release dates, and so on. The total 26 million ratings from GroupLens website is voted by 270,000 users, ranging from 1 to 5.

Table 1 presents the data in numerical form, showing the mean, standard deviation, minimum, first quarter, median, third quarter, and maximum values for each feature. Rating is the y-value that we want to predict, and the data is taken from the Official GroupLens website. It is a continuous value taken between 1 and 4.75. To facilitate prediction, the value of rating will be rounded in this study, so the problem is converted from a regression problem to a classification problem (rating 1, 2, 3, 4, 5 into 5 categories). Adult refers to whether the movie is an adult-restricted movie or not, in fact, there are very few adult-restricted movies in the dataset, and the indicator does not have a large impact on the results in terms of outcomes. Budget records the cost of the movie, and popularity is the popularity index of the movie based on the number of website hits, etc. Although this data will change after the movie is released, the popularity index before the movie is released can still make some contribution to the rating prediction model. The value of Revenue is only known after the movie is shown, so it is important for the evaluation of movie ratings after the release, but it is not available when predicting movie ratings before the release, so it cannot be a characteristic value in this study.Vote_average and vote_count are two other data on ratings from TMDB, which can be used as another reference. Belongs_to_collection, genre, original_tile/title, overview, production_companies, production_countries, spoken_language, tagline, cast, crew, keywords, these are the data composed of words, which occupy most of the data set, so we use the method of nlp (natural language processing) to transform the words into data that can be processed by the computer, i.e., in the form of word vectors, and then each eigenvalue processed by the word vector constitutes a sparse matrix, so as to enables the later data analysis.

4 Models

4.1 Random Forest

Random forest is an integrated algorithm based on decision trees. A decision tree is a classifier that classifies feature data by some specific rules of nodes, each node splits the data, and finally the final splitting result is obtained through multiple layers of splitting. An important issue to be considered in constructing the splitter is how to select the features that constitute the node rules, the core principle of which is to make the nodes maximize the information gain before and after the split, which is determined by the difference of the information entropy before and after the split.

$$\Delta i(p) = i(p) - \sum cPci(pc)$$

where $i(p)$ represents the information entropy and the formula is as follows:

$$i(p) = -\sum jP(\omega j)\log 2P(\omega j)$$

The advantages of decision trees include fast training, the ability to handle non-numerical features, and the ability to achieve non-linear classification. The disadvantages of decision trees are that they are unstable, sensitive to training samples, and easily overfitted. To address these drawbacks, researchers have proposed various decision-tree-based classifier integration methods, such as bootstrap and bagging, where the core idea of the bootstrap method is to sample with put-back, and with the bootstrap method, we derive N new bootstrap datasets from an original dataset. The Bagging method is an application of the Bootstrap idea to machine learning. We generate N Bootstrap datasets from the original dataset, train a weak classifier for each Bootstrap dataset, and finally combine them into a strong classifier by voting and averaging. The training of N weak classifiers is performed in parallel, so Bagging is a parallel method. For unstable weak classifiers (e.g., decision trees, neural networks), Bagging can significantly improve the correct prediction rate while avoiding overfitting.

A random forest is a machine learning algorithm that is a combination of many decision trees. The random forest algorithm builds each decision tree according to a two-step approach. The first step is called “row sampling”, where a Bootstrap dataset is obtained by sampling back and forth from the entire training sample. The second step, called “column sampling”, randomly selects m features (m less than M) from all M features and trains a decision tree with the m features of the Bootstrap dataset as a new training set. Finally, all N decision trees are combined in a voting process.

The advantages of random forest are firstly that the model is simple and easy to understand, and it can handle high-dimensional data without doing feature selection, which can eventually give relatively important features, its training speed is fast, and the trees are independent of each other during training, yet the interaction between features can be detected during the training process. In terms of accuracy, for unbalanced data sets, random forests can balance the error and still maintain accuracy if a large portion of features are missing.

4.2 Neural Network

The neural network algorithm simulates the structure of neurons in the human brain, as shown in Figure X. Four nodes from x1 to x4, form the input layer with input feature data, and θ1 to θ4 are the connection weights, which represent the importance of different features and need to be adjusted by training. After that is the activation function, which simulates the process of neuron activation. The commonly used activation functions include sigmoid, tanh function, etc.

The neurons in the neural network algorithm simulate the architecture of neurons in the real world. The BP neural network which consists of an input layer, a hidden layer and an output layer, where the hidden layers can be multiple layers in Fig. 1. x1 to x4 nodes form the input layer, which represents the input feature information and is equivalent to the dendritic part of the neuron. The nodes θ1 to θ4 are called connection weights, which represent the importance of different information and need to be adjusted by training. The information in the input layers x1 to x4 is summed by the weights and subsequently enters a nonlinear activation function $h\theta(x)$, which simulates the process of neuron activation. The commonly used activation functions include sigmoid function, tanh function, etc.

The mathematical model of the neuron is:

$$a = f\left(\sum_{i=1}^{q} w_i + b_j\right)$$

In a neural network, the activation functions of the input and output layers are determined by the different needs of data processing, and the activation function of the hidden layer is an S-shaped function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Figure 2 shows the structure of a BP neural network. It consists of an input layer, a hidden layer and an output layer. The data is input in the input layer, and there are usually multiple layers in the hidden layer, where the data is processed according to a specific activation function and then the result is obtained in the output layer.

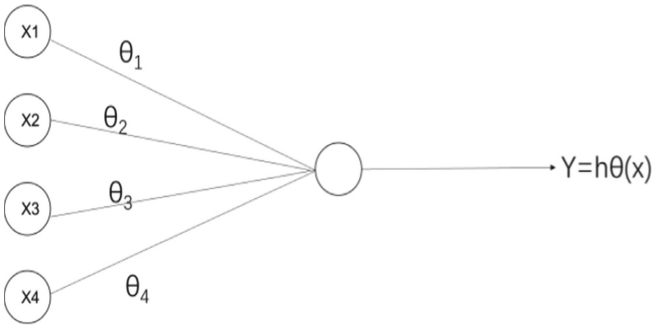

Fig. 1. Neuron structure diagram

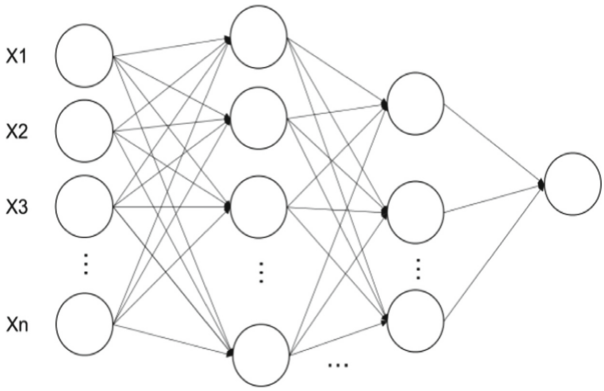

Fig. 2. BP neural network structure diagram

The advantage of Bp neural network is that it implements the nonlinear mapping function from input to output and can approximate almost any nonlinear continuous function with arbitrary accuracy, and it also has some adaptive ability to learn some reasonable parameters by itself. It also has strong adaptability, and can maintain high accuracy for new data with noise. In addition, it has some fault tolerance, and the bp neural network will not have a great impact on the global training results after its local neurons receive damage.

5 Results

There are many parameters that we can choose from sklearn RandomForestClassfier. The first one is class_weight{"balanced", "balanced_subsample"}. This parameter refers to weights associated with classes. If not given, all classes are supposed to have weight one. The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as n_samples. The "balanced_subsample" mode is the same as "balanced" except that weights are computed based on the bootstrap sample for every tree grown.

'n_estimators' represents the number of trees in the forest. Here we choose 150 estimators. Criterion{"gini", "entropy"}.Here we can choose the function to measure the quality of a split which help decide whether to split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain. 'Max_depth' represents the maximum depth of the tree. If we choose None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

In BP network model, hidden_layer_sizes represents the number of neurons in each hidden layer. 'Activation' represents Activation function for the hidden layer including four following functions:'identity', no-op activation, useful to implement linear bottleneck, returns f(x) = x; 'logistic', the logistic sigmoid function, returns f(x) = 1 / (1 + exp(-x)); 'tanh', the hyperbolic tan function, returns f(x) = tanh(x); 'relu', the rectified linear unit function, returns f(x) = max(0, x). Solver represents the solver for weight optimization which include four following solvers: 'lbfgs' is an optimizer in the family

Table 2. Prediction evalution of two models

Models	Accuracy	Precision	Recall	f1_Score
Random forest	0.6952	0.7255	0.6952	0.6043
BP network	0.1029	0.0876	0.1029	0.0892

of quasi-Newton methods; 'sgd' refers to stochastic gradient descent; 'adam' refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba. The default solver 'adam' works pretty well on relatively large datasets (with thousands of training samples or more) in terms of both training time and validation score so we choose adam. Alpha is the L2 penalty (regularization term) parameter. We take default as 0.0001.

As a result, Table 2 shows Random forest has 0.70 accuracy, 0.73 precision rate, 0.70 recall rate, and 0.60 f1_score. BP network 0.10 accuracy, 0.09 precision rate, 0.10 recall rate, and 0.09 f1_score.

6 Conclusions

With the advent of the era of big data, the rapid maturation of technologies such as data mining, machine learning, and deep learning, and the rapid development of the movie industry, there is ample data for scholars to study for each movie, which also provides us with the possibility to seek movie rating prediction through the data of movies. This study extracts a large amount of data that can be obtained before the release of a movie, including the movie synopsis, the movie's actors and directors, and other personnel information, and this textual information can be useful in predicting the movie rating modalities with the help of nlp technology. Movie rating is an important reference for audiences to choose whether to go to the cinema or not. However, before the release of a movie or when it is just released, the rating websites do not have a valid rating for audiences to refer to because there are not enough ratings, so a movie rating prediction model can play a great role to give reference for audiences to choose a movie and for movie producers to choose investment and promotion. On the basis of previous studies, this paper conducts nlp processing on the Movie feature information and user rating data in Movie Lens, mainly uses random forest and neural network algorithm, and constructs a rating prediction model, which has better prediction effect than the traditional prediction model.

Although the movie scoring model designed in this study has some improvement in prediction accuracy compared with the traditional movie prediction model, the study still has many shortcomings due to the limitations of technology and our own level. In fact, there are very many existing machine learning algorithms today, and only two algorithms were selected in this study, so more machine learning algorithms such as xgboost can be introduced in subsequent studies to compare the accuracy of different algorithms. Random forests and neural networks also have very many parameters that can be tuned, and due to time constraints, it is impossible to cover all parameter tuning in this study, so models with higher accuracy can definitely be found by tuning parameters in the future.

In addition, there are many feature values that affect movie ratings, and only some of the indicators were used in this study. In future studies, richer data can be collected to improve the accuracy and usefulness of the model. Movielens is an internationally influential professional movie website, and the movie rating prediction model designed in this study is mainly based on Movielens' data. In the follow-up work, more movie scoring platforms can be referred for research to improve the accuracy and applicability of the movie scoring model.

References

1. Mhowwala, Zahabiya, A. Razia Sulthana, and Sujala D. Shetty. "Movie Rating Prediction using Ensemble Learning Algorithms."
2. Shetty, Chidanand, et al. "Movie Review Prediction System by Real Time Analysis of Facial Expression." 2021 6th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2021.
3. Lim, Yew Jin, and Yee Whye Teh. "Variational Bayesian approach to movie rating prediction." Proceedings of KDD cup and workshop. Vol. 7. 2007.
4. Zhu, Jiaxin, et al. "Gaussian mixture model based prediction method of movie rating." 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2016.
5. Liu, Yiming, Xuezhi Cao, and Yong Yu. "Are You Influenced by Others When Rating? Improve Rating Prediction by Conformity Modeling." Proceedings of the 10th ACM Conference on Recommender Systems. 2016.
6. Ning, Xiaodong, et al. "Rating prediction via generative convolutional neural networks based regression." Pattern Recognition Letters 132 (2020): 12–20.
7. Augustine, Achal, and Manas Pathak. "User rating prediction for movies." Technical Report. University of Texas at Austin. (2008).
8. Basu, Somdutta. "Movie rating prediction system based on opinion mining and artificial neural networks." International Conference on Advanced Computing Networking and Informatics. Springer, Singapore, 2019.
9. Dhir, Rijul, and Anand Raj. "Movie success prediction using machine learning algorithms and their comparison." 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE, 2018.