scientific reports

Check for updates

OPEN

Prediction techniques of movie box office using neural networks and emotional mining

Zhuqing Zhang[1]✉, Yutong Meng[2] & Daibai Xiao[3]

Box office prediction is of great significance for understanding investment risks, class construction, promotion and distribution, and theater scheduling. However, due to the insufficient selection of influencing factors of movie box office, the currently existing prediction model restricts the prediction accuracy. A total of 34 influencing factors in 11 categories, such as heat index, movie types, release date, creators, first-day box office, were selected to study the prediction technology of movie box office. The Word2vec algorithm is used to construct a feature thesaurus for nouns in movie domain; adjectives and verbs with emotional coloring are used to construct an emotional dictionary based on the movie domain; and the TF-IDF algorithm is integrated to calculate the emotional scores of movie comments. A prediction method based on comments and Multivariate Linear Regression (MLR) is designed to analyze the relationship between the influencing factors and the movie box office, which provides an important basis for the prediction of the total box office, and also provides a decision-making reference for the movie industry and the related management departments. Incorporating comments as feature values to improve the accuracy, a prediction model based on comments and Convolutional Neural Network (CNN) is constructed. The results show that the average prediction accuracy of the MLR without comments, Back-Propagation Neural Network (BPNN), and CNN is 63.4%, 68.3%, and 71.9%, respectively, and after integrating the comments, the average prediction accuracy of the MLR and CNN is improved by 16.1% and 11.8%, respectively, and the prediction accuracy is significantly improved.

Keywords Box office prediction, Movie comments, Multiple linear regression, Neural networks, Emotional dictionary

Culture has become a representative of a country's soft power. As an important cultural export medium of America, Hollywood, the center of the world's movie industry, is watched by many people to recognize American culture. Nowadays, with the rapid development of economy and technology, China's cultural industry has also stepped into the fast lane[1]. As an important carrier of cultural output, the Chinese movie industry has been developing and progressing together with the market economy for decades, and the box office has been setting records year by year[2]. Chinese movies have become the focus of the global industry, and the box office has maintained high growth every year from 2010 to 2019, with total box office revenue shown in Fig. 1.

In 2020, under the impact of COVID-19, the box office revenue of Chinese movies is only about 30% of that of 2019. In 2021, as the COVID-19 in China is effectively controlled and the mainland movie market gradually recovers, China's movie box office eventually recovers to 75% of its 2019 level in 2021. Overseas, on the other hand, was significantly affected by COVID-19, and compared with the box office revenue in 2019 ($11.4 billion), the North American box office in 2021 only recovered to 37% of the pre-epidemic level, which is about $2.9 billion lower than that of China. China's box office has been leading overseas box office significantly and has become the world's top box office[3,4]. With the effective control of COVID-19, this gap will further widen, and the importance of the Chinese movie market becomes more and more obvious. The global movie box office in the past two years is shown in Fig. 2.

The goal of a movie company is to maximize the profit of the box office. Whether it is the pre-production, mid-shooting and post-production steps in movie production, or the length of the cycle and the amount of investment, they will all affect the final box office of the movies[5]. As a result, there are a large number of high-investment, low-return movies in the current movie market, and there are also excellent movies with insufficient

[1]School of Journalism and Communication, Nanjing University, Nanjing 210093, Jiangsu, China. [2]Movie and Philosophy at Humanities, University of Southampton, Southampton SO17 1BJ, UK. [3]Faculty of Humanities and Social Sciences, City University of Macau, Macau, China. ✉email: jessiezhang@smail.nju.edu.cn

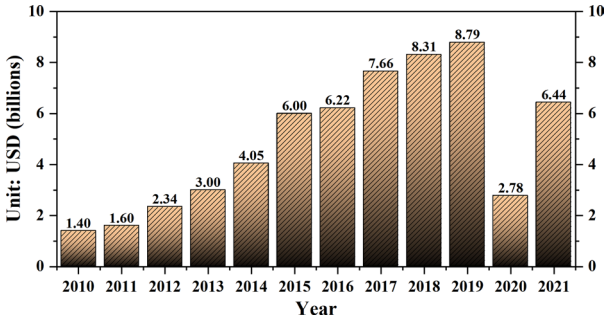

Fig. 1. Annual box office of Chinese movies in 2010–2021.

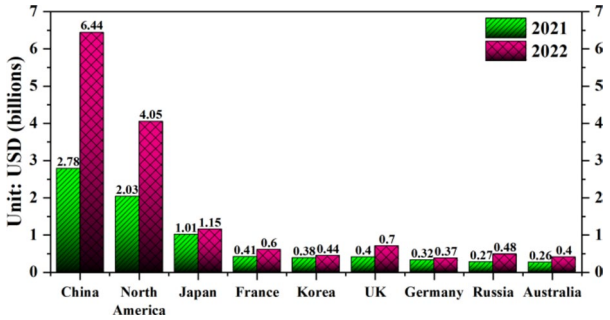

Fig. 2. Annual box office of movies in the world, 2020–2021.

budgets in the early stages, leading to a lack of momentum in the final box office. The reason for this is that in the early stages of a movie, it is difficult for investors and producers to control the final box office, leading to a mismatch of resources[6,7]. Therefore, in the early stage, it is of great significance to the movie industry to predict the box office according to the basic information and special nature of the project to be developed.

At the early stage, the investors and producers can evaluate and predict the box office in advance based on the script, the team, the actors or other influencing factors. By predicting the box office in advance, the investors will have some reference on whether to invest, the investment ratio, and the investment risk; and the producers will choose the best production team according to the predicted return on investment. In the late stage, the investors can reasonably allocate the expenses for publicity and distribution according to the predicted box office volume, and equationte a marketing plan in line with the characteristics of the movie. For movies that have already been released, based on the daily real-time box office data, they can plan the number of theaters to be scheduled in a timely manner, equationte the attendance strategy, and maximize the benefits of the box office. Therefore, predicting the box office and finding the factors affecting the box office are of great practical significance in understanding the investment risk, guiding the promotion and development, and equationting the release strategy.

Research state

Zhang et al.[8] used neural networks for the first time to build a classification model to predict box office by distinguishing movie types. The method has poor prediction effect and can only predict the range of box office. Zhang and Skiena[9] built a box office prediction model based on three factors: directors, actors, and exposure. Barman et al.[10] built a prediction model based on Back-Propagation Neural Network (BPNN) by using movie types as an influencing factor, but the experimental results were not as expected. Ru[11] used the release schedule as a reference and added a BPNN to build a model with a higher accuracy rate. Lu[12] used the amount of internet searches before and after the release of a movie as a factor, and predicted the box office, with significantly higher prediction accuracy. Zhao[3] used Sina Weibo as a data source of word-of-mouth, and mined comments from movie-related Weibo for opinion analysis. Dai[13] first selected as many as 22 relevant variables and constructed a prediction model using regression equations and BPNN. Liu[14] analyzed foreign movies together with domestic movies and used the regression model as the basis for prediction, which improved the prediction accuracy of movie box office. Kim[15] used social service network to predict movie box office. Song[16] also constructed a Sina Weibo-based emotion analysis dictionary using Support Vector Machine (SVM) to predict the box office in the first week through comments. Shen[17] selected nine types of data as influencing factors and built a box office prediction model using CNN. Jiang[18] selected 34 movies with ten features such as director awards, actor awards, and established a box office prediction model using SVM. Zhou et al.[19] obtained data from movie posters to establish a box office prediction system based on Convolutional Neural Network (CNN), which plays a key role in subsequent feature extraction. He[20] fused Multivariate Linear Regression (MLR) and BPNN to establish a prediction system of first-day box office. Ni et al.[21] used Long Short-Term Memory (LSTM) to predict box office and reduced the average prediction error. Mahmud et al.[22] extracted the emotion features of movie comments

and built a multilayer neural network model, but the accuracy was less than 50%. Ahmad et al.[23] found that the actors of a movie plays a crucial role in box office. Liao[24] established a movie prediction model based on various methods such as different models. Al Fahoum[25] introduced emotional features of movie comments and proposed a prediction model based on Stacking, which in turn improved the prediction accuracy of box office.

Through the former results, it is found that the current domestic and foreign prediction models of movie box office still have the following problems: (1) Selection of influencing factors: when selecting the factors of movie box office, at present, the basic information of the movie is mainly used as a reference, such as the movie types, the release date, and is not considered according to the different stages of the whole life cycle. (2) Prediction models: Among the algorithms and models for movie box office prediction, the single classical mathematical algorithm (e.g., multivariate linear regression) or traditional algorithms (e.g., decision tree) are usually chosen, which have their own limitations and have a greater impact on the prediction accuracy of the final box office.

Data selection

Influencing factors

The entire life cycle of a movie from preparation to release is studied, and based on the relationship between various influencing factors at different stages and the box office, the data dimensions of the dataset in the study of movie box office prediction are ultimately determined. Based on mainstream domestic film data platforms, data is obtained through writing web crawlers to establish accurate and complete datasets.

The preparation stage mainly considers factors such as script content, team building, schedule selection, and film types; The production stage is a stage jointly created by professional teams such as directors, actors, photography, art, and setting; The promotion stage mainly considers the dissemination, the audience's understanding of the movie, and the schedule; After the official release of the movie, consideration should be given to issues such as scheduling and word-of-mouth. Among them, the 1st to 5th indicators are the influencing factors in the preparation period, the 6th to 9th are the influencing factors in the promotion and distribution period, and the last two are the influencing factors in the release period. The *Cat's Eye*, and *Doubon* mentioned in the following text are all movie data platforms. The data is obtained based on these platforms.

(1) Movie Types

The movie classification is referenced to the North American Internet Movie Database (IMDB), which is divided into 31 types. They are categorized into 14 types according to past works: drama, comedy, romance, action, adventure, animation, suspense, thriller, fantasy, crime, sci-fi, family, war and others. The ratio of the number of corresponding types to the statistics is set as a dummy variable, otherwise it is 0.

(2) Production Countries

For co-produced movies, this kind of movie can integrate the hoped-for advanced technology and China's unique oriental culture into the work. Therefore, the sources are divided into 4 categories: (1) domestic movies, (2) Hong Kong, Macao and Taiwan, (3) foreign imports and (4) domestic and foreign co-productions. According to the production countries of the works in the past ten years, the ratio of the number of movies of different types to the total number of movies is used as a parameter value with the total number of movies as the denominator, and the value of the movies that are not in the 4 categories is assigned as 0.

(3) Directors

The director plays an unparalleled role in the presentation of the script and the quality of the movie in terms of the final effect of the movie. The director's favorite rating on the *Mtime* website is used instead.

(4) Actors

The influence of the actors on the movie box office has been more controversial. Instead of using the favoritism of *Mtime* website directly, the top four actors announced for each movie are selected and averaged.

(5) IP Adaptation

Intellectual Property (IP) movies are movies produced through adaptations of novels, comics, other movies. Analyzing the existing IP adaptations in the market, whether a movie is an IP adaptation or not will have a great impact on the box office. First, the original IP has a certain audience base, i.e. the original loyal fans of the IP. In addition, the original IP story is already well known to the audience, and the director will make targeted adaptations to the story based on audience feedback on the plot, which will have a positive impact on the movie. So taking whether IP adaptation or not as a factor, IP adaptation as 1 and non-IP adaptation as 0.

(6) Intended Audience

The intended audience reflects the popularity of the movie in the market and the attention. The number of people interested in the movie is also likely to watch the movie after its release, which becomes an important part of contributing to the box office in the early stage and establishing the word-of-mouth in the early stage. Therefore, the number of people who want to watch the movie on the ticketing platform is an important basis for the first-day box office. The number of intended audience on the *Cat's Eye* was used for the study.

(7) Heat Index

The heat involved in this paper includes Baidu index and 360 trend. Most audience will search for related words to understand the information of the movie at the first time after seeing the movie information, so the heat index of the movie also reflects the degree of the movie being understood by the audience.

The selection of heat data includes the average value of Baidu Index and 360 Index of the 90 days prior to the movie's release.

(8) Release Schedule

It is crucial for a movie to choose what schedule to be released. Combined with the actual situation of China's various schedules, based on the length of the schedules and the historical data and quality of movies released in various schedules, the schedules are finally categorized into: Winter Vacation, Labor Day, Summer Vacation, National Day and general schedules.

(9) Word-of-mouth

There is an "information gap" between the audience and the movie, and the movie uses preliminary marketing and publicity to "trick" the audience into the theater during the publicity period. However, with the development of the Internet platform, the audience can decide whether to buy a ticket or not through word-of-mouth. The evaluation of each movie mainly consists of two parts: one is the rating, which reflects the majority of the audience's good and bad judgment of the movie; the other part is the comments, the number of comments reflects the hotness of the movie and the authority of the final rating. Therefore, the rating of *Cat's Eye*, *Douban* and the number of comments were used as variables.

(10) Emotional Tendency of Comments

This paper analyzes the comments within 30 days before the release of the movie as data. According to the 20–30 days before the release, 10–20 days before the release, 5–10 days before the release and 1–5 days before the release, the comments are divided into four levels: I, II, III, IV, the higher the level means the heavier the impact. Based on previous research experience, the weight values are: 0.15, 0.20, 0.25 and 0.40 respectively[26].

$$T_V = \frac{1}{n}\sum_{j=1}^{k} t_{ij}(V = I, II, III, IV)$$

where n represents the number of days included in the comments of the level, i represents the comments of the i-th days, k represents the total number of comments of the day, j represents the j-th comment of the day, and t_{ij} represents the emotion score of the j-th comment of the i-th day.

The total emotion score of the movie is calculated by the equation:

$$F = \sum \left(T_V \times weight_V\right)$$

where T_V stands for the emotion score included in that level of comment and $weight_V$ stands for the weight value of that level of comment.

(11) First-day box office

The first-day box office of a movie plays a crucial role in the final total box office. Good market data on pre-marketing will attract more viewers to buy tickets and have a positive impact on theater scheduling. For movies with bad word-of-mouth and opinion, there is a high probability that the first-day box office will be close to the final box office. Therefore, the first day box office is an indicator of the preliminary publicity and a key factor in determining the total box office.

Data acquisition

At present, the mainstream movie data platforms in China are *Cat's Eye* and *Douban*. The data source is divided into two parts. The first part is the basic data of the movie, which comes from the public data of *Douban*, *Wikipedia* and other websites, including the name, directors, actors and release date of the movie. The second part includes box office, first-day box office, intended audience, which are from *Cat's Eye*. Meanwhile, this paper builds a data crawling system based on Scrapy framework to obtain data from *Cat's Eye* and *Douban*, as shown in Fig. 3.

(1) Douban

Douban records the basic information of the movie, including the name, directors, screenwriters, actors, types and other twelve dimensions of data, covering all the basic information. The platform gathers a large number of movie lovers and critics, with high overall quality, and accumulating excellent data for word-of-mouth. Therefore, the basic information and word-of-mouth data of the movie will be obtained through *Douban*.

(2) Cat's Eye

Cat's Eye is a web-based information platform dedicated to providing real-time and accurate movie data for movie practitioners, and is China's first platform that updates movie-related business data in real time. Through this platform, information on all released movies can be obtained, such as release date, box office, schedule of each movie, average attendance, attendance rate and other information. It is compiled by the Office of the State Administration Committee for Special Funds for movie Development (SASFD), and Cat's Eye synchronizes with the SASFD data on the same day, so the data is authentic and current. The total box office, the first-day box office, intended audience and other data are selected as the influencing factors of the prediction.

A total of more than 63,418 data related to 3912 movies released in China from 01/01/2012 to 31/12/2020 were crawled. The specific descriptions of the specific indicators of the dataset are shown in Table 1:

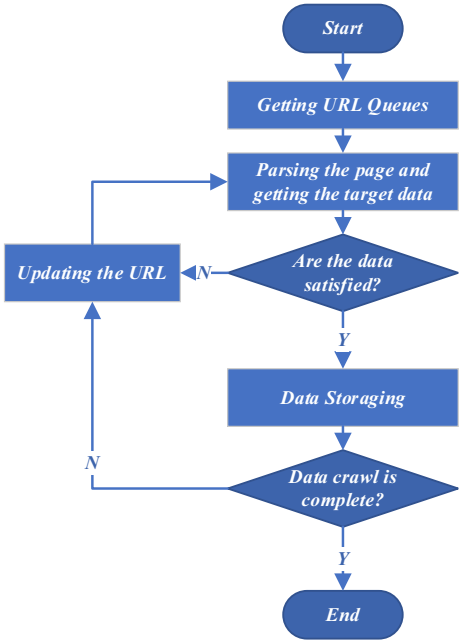

Fig. 3. Data crawling process.

Primary indicators	Secondary indicators	Description	Definition
Basic indicators	movie_name	Name of the movie	Name of the movie
	Director	Director	Director's favoritism of the movie
	Actor	Actors	Actors' favoritism of the movie
	release_date	Release date	Weighting of schedule distribution
	Type	Movie type	Weight of the type in 14 types
	Country	Movie countries	Weight of the country in 4 countries
	IP	IP adaptation	IP adaptation is defined as 1; otherwise 0
Network data indicators	intended_index	Intended audience index	The number of intended audience
	heat_index	Heat index	Heat value in the first 90 days of release
	Score	Movie rating	Movie rating
	first_box	First-day box office	The first-day box office of the movie
	comment_num	Number of comments	Total number of movie comments
	comment_emotion	Comment emotion	Emotional tendency of movie comments
	pos_com_num	Number of positive comments	Number of positive comments of movie
	neu_com_num	Number of neutral comments	Number of neutral comments of movie
	neg_com_num	Number of negative comments	Number of negative comments of movie

Table 1. Specific descriptions of indicators.

Emotional mining in movie comments

Movie-based emotional dictionary construction

The audience's evaluation for movie contains not only the movie itself (e.g., theme, narrative style, soundtrack, art effects.), but also the performance of the movie's related creators (e.g., actors, directors, screenwriters)[27]. Therefore, this paper constructs a movie-based emotional dictionary based on the characteristics of movie comments, with the following process (Fig. 4):

(1) Composing a corpus of movie comments crawled by Scrapy, and pre-processing the text data in the corpus with word splitting, lexical annotation, format conversion and so on.
(2) Obtaining the basic emotion dictionary, to construct the auxiliary emotional dictionary[28].
(3) For the processed sentiment disambiguation, Word2vec and SO-PMI are used to expand the emotion words with synonyms or near-synonyms, and build a movie-based feature thesaurus and emotional dictionary.
(4) Using Python's Snow NLP library, the TF-IDF algorithm is integrated to calculate emotional scores for movie comments.

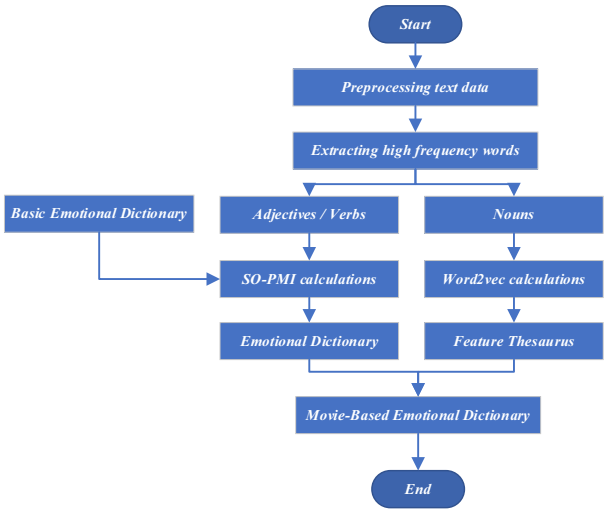

Fig. 4. Construction flow movie-based emotional dictionary.

Movie-based feature thesaurus

Feature words in movie domain

Feature words include movie elements (e.g., plot, music, art) or movie actors (e.g., director, actor, character name). The content of the current feature thesaurus in movie domain is divided into four major categories: plot, music, visual effects, and characters, as shown in Table 2. In movie comments, actors, directors, and character names are important elements mentioned by the audience, therefore a library containing characters and cast members is established, such as "Actors" in Table 2. Because the audience has a high level of attention to certain characters related to the movie (such as the director, actors), with continuous analysis and learning of comments, the more complete vocabularies of actors can be obtained. Each category of feature thesaurus contains close phrases that describe the same thing[29]. For example, "dubbing", "soundtrack", and "background music (BGM)" are all in the category of "music"; "actor", and "director" are all the categories of characters.

Feature extraction for movie comments

The Word2vec algorithm is used to extract words related to movie feature words from the corpus, and to extend and improve the feature thesaurus in the movie domain. Parameter settings: size = 50, window = 5, min-count = 3, other parameters use default values[30]. For example, the results of finding the top 5 semantically similar words with "plot" and "music" and their similarity are shown in Table 3:

Emotional dictionary construction

Basic emotional dictionary

The basic emotion dictionary, also known as a seed emotion dictionary, typically contains emotion vocabulary that is domain independent, meaning it can be applied in different fields and basic emotional vocabulary can be obtained from ordinary emotion words. There are currently a large number of mature sentiment dictionaries, as shown in Table 4. The three dictionaries in Table 4 are merged, and the repeated words are deleted as auxiliary emotional dictionaries. It should be noted that for emotional words with controversial polarity, the polarity of word selection should be determined according to the priority order of Hownet, NTUSD, and THUOCL. Then, based on the movie's comment text, 10% of the comments were extracted for testing corpus, and the remaining 90% of the corpus was subjected to unified text preprocessing, removing punctuation, segmentation, stop words, etc. The processed data is fused and reprocessed with an emotional dictionary to obtain a basic dictionary based on movie reviews.

Featured categories	Words
Music	Dubbing, soundtrack, background music(BGM), interlude, theme song, track, tone, OST
Visual Effects	Picture quality, 2 K, 4 K, cinematography, art, night scene, scenery, special effects, visual
Plots	Story structure, montage, script, frame, pacing, main line, branch line, story
Actors	Andy Lau, Teng Shen, Bo Huang, Xiaofei Zhang, Tao Guo, Jiahui Zhang, Snapdragon Wang
Characters	Cheung Kong 7, 007, The Great Sage, The Monkey King, Shiloh, Sherlock Holmes, Watson

Table 2. Feature words in movie domain.

Target words	Words	Similarity
Plots	Drama	0.974481552
	Story	0.943604046
	Thread	0.915268266
	Storyline	0.914246586
	Content	0.886183353
Music	Music	0.943604046
	Interlude	0.933781959
	Theme Song	0.930954321
	Background Music	0.921722076
	OST	0.900935343

Table 3. Synonyms and similarity.

Dictionary name	Main functions
Hownet	Containing 836 positive emotional words and 1254 negative emotional words
NTUSD	Containing 2812 positive emotional words and 8278 negative emotional words
THUOCL	Containing 5568 positive emotional words and 4470 negative emotional words

Table 4. Emotional dictionary.

Emotional dictionary in movie domain

A set of positive words (*PosWords*) and negative words (*NegWords*) with prominent emotional colors are selected as seed words from the basic emotional dictionary. For each target word $word_i$, the SO-PMI value with respect to the seed word is calculated according to the equation, and the emotional color of the target word $word_i$ is determined by the positive or negative of the value. If the SO-PMI value is positive then the target word is a positive word and vice versa for negative words.

$$SO-PMI(word_i) = \sum_{Pos_i \in PosWords} PMI(word_i - Pos_i) - \sum_{Neg_i \in NegWords} PMI\left(word_i - Neg_i\right)$$

where $PMI(word_i - Pos_i)$ represents the pointwise mutual information with positive words and $PMI\left(word_i - Neg_i\right)$ represents the pointwise mutual information with negative words. Taking the positive point mutual information as an example, is shown:

$$PMI(word_i - Pos_i) = \log_2 P(word_i - Pos_i)/P(word_i) - P(Pos_i)$$

where $P(word_i - Pos_i)$ represents the probability of $word_i$ and Pos_i appearing together in the corpus, and $P(word_i)$ and $P(Pos_i)$ represent the probability of the two words appearing separately. If the probability of appearing simultaneously in a single set is greater, it means that the two words are more similar; on the contrary, it means that the two words are less similar. Through the above operations, 2671 positive and 1534 negative emotion words in movie domain were extracted from the movie comments[31]. By integrating the basic emotional dictionary as well as the emotional dictionary in movie domain, the emotional dictionary in movie domain contains 7621 positive emotional words and 12,094 negative emotional words.

Emotional score calculation for movie comments

(1) Emotional Score Calculation based on Snow NLP

Snow NLP is a Python library specifically designed for processing Chinese text, with excellent performance in processing Chinese text. By analyzing the core code of Snow NLP, it was found that the sentiment classifier of Snow NLP is based on a naive Bayesian model and integrates basic training corpus[32]. In the process of natural language processing, Snow NLP can complete various tasks such as Chinese text segmentation, keyword extraction, text similarity calculation, font conversion, etc., and has a custom training interface to facilitate model training on the target dataset in the future[33].

We analyze the text emotion of the corpus based on Snow NLP, and derive the overall emotional score W_i for each comment text, W_i representing the i-th comment. The range of W_i is between 0 and 1, where the value of [0, 0.5) indicates that the emotional tendency of the text is negative, and the closer to 0 the more negative the emotional tendency is; the value of (0.5, 1] indicates that the emotional tendency of the comment text is positive, and the closer to 1 the more positive the emotional tendency is; when $W_i = 0.5$, the emotional tendency is neutral.

(2) Emotional Weight Calculation based on TF-IDF

Each short comment is analyzed in the movie comment corpus. If a comment has one or more emotional words that can express the emotional tendency and intensity of the entire comment, then the emotional weight of an emotional word can be measured by its importance. The higher the importance of emotional words, the more obvious the emotional tendency in this sentence. The TF-IDF algorithm is used to determine the importance of all words in the movie comment corpus. Words without practical meaning and nouns without emotional tendencies are removed, while words with emotional tendencies are retained. The importance score of emotional words in the entire corpus is calculated as the emotional weight score of the word in the corpus.

The emotion weight T_{ij} of the emotional words in the base emotional dictionary is calculated, where i represents the i-th comment and j represents the j-th emotional word. T_{ij} ranges between [0, 1], with larger values representing higher importance of words[34]. The calculation results are output in reverse order of word importance, from highest to lowest.

(3) Emotional Score Calculation for Movie Comments

Through the above two methods, the emotional tendency of each comment and the value of the emotional weight can be calculated, using the following equation to directly take the product of the two parameters of the base emotional score W_i and the emotional weight T_{ij} as the emotional score of the comment.

$$F_i = W_i \times \sum_{0}^{j} T_{ij}$$

where i represents the i-th comment and j represents the j-th emotional word.

It is found that the value of emotional weight obtained by the TF-IDF algorithm is less than 1. This is mainly due to the fact that the proportion of emotional words in the comments in the corpus is larger than that in the standard corpus, which causes these emotional words to have lower weights in the corpus. So, the calculation equation was improved as:

$$F_i = W_i \times \sum_{0}^{j} \left(T_{ij} + 1\right)$$

In addition, in order to be able to express the emotional tendency more clearly with the plus and minus signs, this paper takes the base emotional score and changes the value range of W_i from [0, 1] to [−0.5,0.5]:

$$F_i = (W_i - 0.5) \times \left(T_{ij} + 1\right)$$

After processing, F_i is a numerical value with a positive and negative sign, a positive number indicates positive emotion, a negative number indicates negative emotion, 0 indicates neutrality, and the larger the absolute value of the value the higher the intensity of emotion.

Methods of predicting movie box office

The MLR utilizes the correlation between independent variables and dependent variables to establish a prediction model to explain the relationship between each independent variable and dependent variable and the degree of their influence. In this study, the MLR models of first-day box office and total box office are constructed separately, through which the model can intuitively explain the relationship of each influencing factor on the box office of the movie, and judge the degree of influence of each factor on the prediction results according to the models.

Prediction model of first-day box office based on multiple linear regression

To establish a model for first-day box office prediction:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$

where Y represents the box office of the first day of the movie, X_1 represents the number of intended audience, X_2 represents the heat index, β_0, β_1, β_2 represent the regression coefficients of the model, and μ is the random error. Least squares regression is performed on the organized data, as shown in Table 5.

Through the above modeling, the first-day box office prediction model of the movie is:

$$Y = -1.0246 + 1.281158X_1 + 1.125741X_2$$

	Estimate	Std. error	t value	Pr(>|t|)
(Intercept)	−1.0246	0.5614	−2.9	2.15e−16***
Intended Audience	1.281158	0.008001	21.84	< 2e−16***
Heat Index	1.125741	0.002357	−1.24	2.1e−10**

Table 5. Multiple regression coefficients and significance test results.

From the model, it can be seen that the number of intended audience and the heat index play a contributing role in the improvement of the first-day box office, and the higher the data of these two items indicate that the more attention the movie receives, the higher the data of the first-day box office will be accordingly.

Prediction model of total box office based on multiple linear regression

The prediction model of total box office based on multiple linear regression is established, in which the model with comment text and the model without comment text are shown:

$$Y_{Comments} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{34} X_{34} + \mu$$
$$Y_{No\ comments} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{30} X_{30} + \mu$$

where Y is the total box office, X_1 represents the first-day box office, X_2 represents the number of intended audience, X_3 represents the heat index, X_4 represents the Drama movie, X_5 represents the Comedy movie, and X_6 represents the Romance movie. X_7 stands for Action movie, X_8 stands for Adventure movie, X_9 stands for Animation movie, X_{10} stands for Suspense movie, X_{11} stands for Thriller movie, X_{12} stands for Fantasy movie, X_{13} stands for Crime movie, X_{14} stands for Sci-Fi movie, X_{15} stands for Family movie, X_{16} stands for War movie, X_{17} stands for other types, X_{18} stands for domestic movie, X_{19} stands for Hong Kong, Macao and Taiwan movie, X_{20} stands for imported movie, and X_{21} stands for co-production movie, X_{22} stands for director, X_{23} stands for actors, X_{24} stands for whether the movie is based on IP, X_{25} stands for winter vacation, X_{26} stands for Labor Day, X_{27} stands for summer vacation. X_{28} stands for National Day, X_{29} stands for general schedule, X_{30} stands for movie ratings, X_{31} stands for the number of positive movie comments, X_{32} stands for the number of neutral movie comments, X_{33} stands for the number of negative movie comments, and X_{34} stands for the emotional tendency of movie comments. $\beta_1, \beta_2, \ldots, \beta_{34}$, are the estimated regression coefficients and μ is the random error.

The organized data is regressed. Firstly, the data of movie box office is logarithmically calculated. Subsequently, on the basis of stepwise regression, Weighted Least Squares (WLS) was used for correction, with weights set to the reciprocal of the squared residuals and the reciprocal of the absolute residuals, respectively. After comparing the two, it was found that the accuracy of using the reciprocal of the absolute residual value for weight selection is higher. The calculation results of the data are shown in Table 6. Among them, "Estimate" is used as the estimated regression coefficient for various influencing factors, "Pr(>|t|)" is the basis for judging whether the item is significant, and "***" represents significant. For example, in Table 6, the estimated regression coefficients for "Thriller" with comments and without comments are − 1.120271 and − 1.012831, respectively. Therefore, the representations for "$Y_{comments}$" and "$Y_{no\ comments}$" are "$-1.120271X_{11}$" and "$-1.012831X_{11}$"; the estimated regression coefficient for "Number of Neutral Comments" with comments is 0.23571, and the representations for "$Y_{comments}$" is "$0.23571X_{32}$".

The regression equation obtained from the above modeling is:

(1) Multiple linear regression equation with comments:

$$\begin{aligned} Y_{comments} = {} & 1.22249X_1 + 0.96154X_2 + 0.86115X_3 + 0.01271X_4 + 0.43324X_5 \\ & + 0.52031X_6 + 0.22167X_7 - 0.32191X_8 + 0.11027X_9 + 0.02187X_{10} \\ & - 1.12027X_{11} - 0.12587X_{12} + 0.52163X_{13} + 0.09422X_{14} + 0.20167X_{15} \\ & + 0.02497X_{16} - 1.13107X_{17} + 0.033218X_{18} + 0.03157X_{19} + 0.13247X_{20} \\ & - 0.02157X_{21} + 1.02198X_{22} + 1.2157X_{23} + 0.36915X_{24} + 0.91891X_{25} \\ & + 0.81352X_{26} + 0.76215X_{27} + 0.85928X_{28} - 0.26513X_{29} + 0.69731X_{30} \\ & + 1.86171X_{31} + 0.23571X_{32} - 2.3691X_{33} + 1.0199X_{34} + 0.31357 \end{aligned}$$

(2) Multiple linear regression equation without comments:

$$\begin{aligned} Y_{noComments} = {} & 1.264591X_1 + 0.809411X_2 + 0.706306X_3 + 0.503056X_4 + 0.485566X_5 \\ & + 0.607348X_6 + 0.245951X_7 - 0.529008X_8 + 0.133367X_9 - 0.551346X_{10} \\ & - 1.012831X_{11} - 0.316774X_{12} + 0.073368X_{13} + 0.330818X_{14} + 0.260668X_{15} \\ & - 0.071289X_{16} - 2.733961X_{17} + 0.083742X_{18} + 0.62734X_{19} + 0.513196X_{20} \\ & - 0.061683X_{21} + 0.742167X_{22} + 0.865497X_{23} + 0.275468X_{24} + 0.33673X_{25} \\ & + 0.097542X_{26} + 0.03425X_{27} + 0.259297X_{28} - 0.390832X_{29} + 0.294998X_{30} + 0.294998 \end{aligned}$$

Prediction model of total box office based on BPNN

In the BPNN, the sample is input by the input layer, processed through the multi-layer hidden layer, and finally the prediction result is output by the output layer. It is judged whether the error reaches the expected error range, and when it exceeds the threshold, it enters the backward propagation process of the error. The error is transmitted layer by layer to each neuron in the hidden layer, and the weights of the neuron nodes in each layer are constantly adjusted. The prediction flow of the box office based on BPNN is shown in Fig. 5:

Prediction model of total box office based on comments and CNN

Data preprocessing

The data is mainly categorized into three types: basic data, time data, and comment data.

	$Estimate_{comments}$	$Pr(>\|t\|)_{comments}$	$Estimate_{no\ comments}$	$Pr(>\|t\|)_{no\ comments}$
(Intercept)	0.31357	< 2e−16***	0.294998	< 2e−16***
First-day box office	1.22249	< 2e−16***	1.264591	< 2e−16***
Intended audience	0.96154	1.27e−12***	0.809411	1.27e−12***
Heat index	0.86115	< 2e−16***	0.706306	< 2e−16***
Drama	0.01271	< 2e−16***	0.503056	< 2e−16***
Comedy	0.43324	< 2e−16***	0.485566	< 2e−16***
Romance	0.52031	< 2e−16***	0.607348	< 2e−16***
Action	0.22167	1.9e−12***	0.245951	0.21e−16***
Adventure	−0.32191	< 2e−16***	−0.529008	< 2e−16***
Animation	0.11027	6.1e−16***	0.133367	6.1e−16***
Suspense	0.02187	9.3e−13***	−0.551346	0.93e−16***
Thriller	−1.120271	< 2e−16***	−1.012831	< 2e−16***
Fantasy	−0.12587	< 2e−16***	−0.316774	< 2e−16***
Crime	0.52163	2.1e−13***	0.073368	0.22e−13***
Sci-Fi	0.09422	< 2e−16***	0.330818	< 2e−16***
Family	0.20167	< 2e−16***	0.260668	< 2e−16***
War	0.02497	1.9e−13***	−0.071289	0.9e−16***
Other types	−1.13107	1.2e−12***	−2.733961	1.2e−12***
Domestic	0.033218	3.3e−15***	0.083742	3.3e−16***
Hong Kong, Macau and Taiwan	0.03157	1.7e−11***	0.62734	0.7e−11***
Imported	0.13247	2.2e−11***	0.513196	0.2e−16***
Co-production	−0.02157	3.74e−14***	−0.061683	3.74e−14***
Directors	1.02198	< 2e−16***	0.742167	< 2e−16***
Actors	1.2157	< 2e−16***	0.865497	< 2e−16***
IP or not	0.36915	2.7e−13***	0.275468	1.1e−16***
Winter vacation	0.91891	1.33e−13***	0.33673	1.33e−13***
Labor day	0.81352	< 2e−16***	0.097542	< 2e−16***
Summer vacation	0.76215	2.4e−13***	0.03425	2.4e−16***
National day	0.85928	< 2e−16***	0.259297	1.9e−13***
General schedule	−0.26513	< 2e−16***	−0.390832	3.4e−16***
Movie rating	0.69731	1.6e−13***	0.294998	1.02e−14***
Number of positive comments	1.86171	1.4e−12***	–	–
Number of neutral comments	0.23571	3.3e−16***	–	–
Number of negative comments	−2.3691	2.6e−11***	–	–
Emotional tendency	1.0199	3.7e−16***	–	–

Table 6. Multiple regression coefficients and significance test results.

(1) Basic Data Processing

The irregular data are screened and unified, and all the data with part of the first-day box office less than 500,000, directors, actors, types, production countries, and movie rating missing are discarded.

One-hot coding is used to discretize the feature values, and n values corresponding to n positions are used for representation to circumvent the problem of bad processing of certain data. The feature values are as follows:

Production Countries: China, Hong Kong, Macao and Taiwan, foreign import, domestic and foreign co-production.

Release Schedule: Winter Vacation, Labor Day, Summer Vacation, National Day and general schedule.

Movie Types: Drama, Comedy, Romance, Action, Adventure, Animation, Suspense, Thriller, Fantasy, Crime, Sci-Fi, Family, War, Other.

A total of 27 features were formed after processing, and for some data of the features missing, the positional data was replaced by 0. Finally, all the data were standardized and processed.

(2) Time Data Processing

The time data includes two parts: intended audience and heat index. Each movie has the Baidu index, 360 trend index, and intended audience for the 90 days prior to release, so the time data is preprocessed into a 3×90 2D matrix, which is convenient to input into the input layer of the CNN.

(3) Comment Data Processing

The comment data contains the number of positive comments, number of neutral comments, number of negative comments, emotional tendency for the 30 days prior to release. The data is preprocessed into a 4×30 2D matrix and input to the CNN.

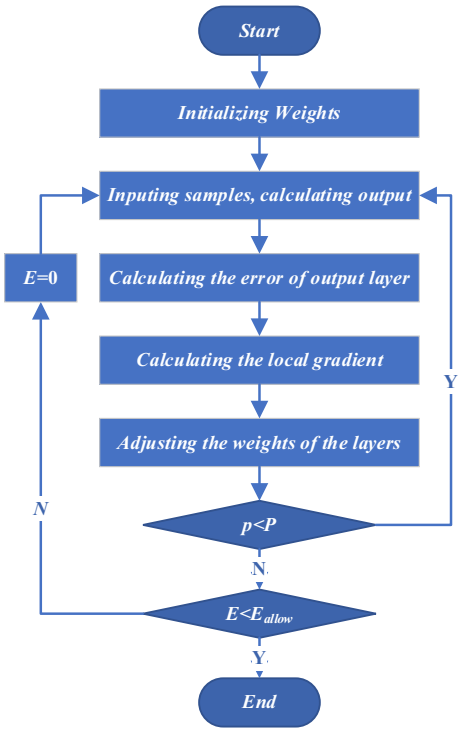

Fig. 5. Flow of BPNN.

Model construction

The prediction model based on CNN is built, including the basic data and the time data of the movie. The structure of the conventional CNN model built in this paper is shown in Fig. 6. The sub networks for processing basic data and time data both set the size of the kernel in the first convolutional layer to 1×3, and the size of the kernel in the second convolutional layer to 3×3, with the aim of fully extracting feature data from the movie for three consecutive days. Meanwhile, the number of kernels in the basic data subnetwork is 128 and 128, respectively; The number of kernels in the time data subnetwork is 64 and 128, respectively; The number of neurons in the fully connected layer is 512 and 128, respectively, with the Dropout of 75%. The Sigmoid function is used.

The prediction model based on comment and CNN, including movie basic data, time sequence data and movie comment data. The structure of the CNN model incorporating comment text is shown in Fig. 7. The sub networks for processing basic data and time data both set the size of the kernel in the first convolutional layer to 1×3, and the size of the kernel in the second convolutional layer to 3×3, with the aim of fully extracting feature data from the movie for three consecutive days. Meanwhile, the number of kernels in the basic data subnetwork is 128 and 128, respectively; The number of kernels in the time data subnetwork and the comment data subnetwork is 64 and 128, respectively; The number of neurons in the fully connected layer is 512 and 128, respectively, with the Dropout of 75%. The Sigmoid function is used.

Figure 8 shows the training and testing processes of CNN and CNN-C. As can be seen from Fig. 8, the accuracy of both CNN and CNN-C gradually increases with the increase of the number of iterations, until the accuracy basically peaks after the number of iterations reaches 1000, and the data are well fitted. Meanwhile, the prediction accuracy of the CNN-C model after fusing the comments data is significantly improved.

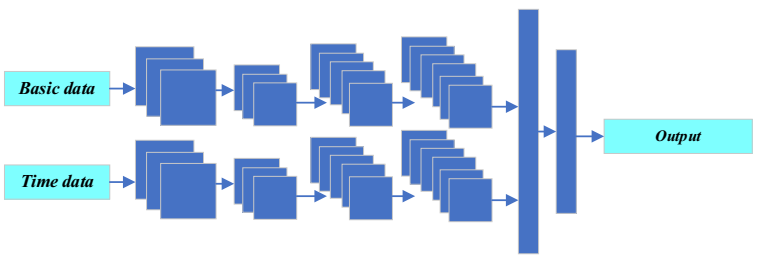

Fig. 6. Structure of total box office based on CNN.

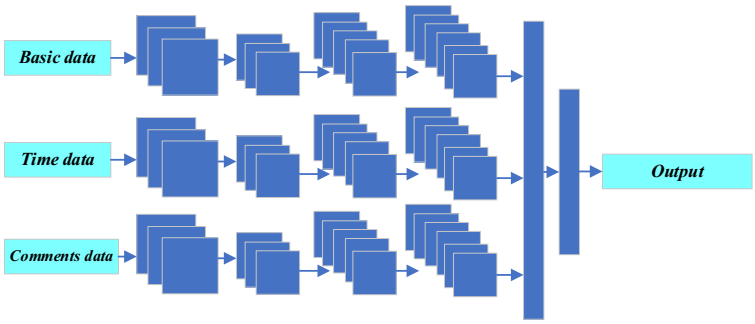

Fig. 7. Structure of total box office based on comments and CNN (CNN-C).

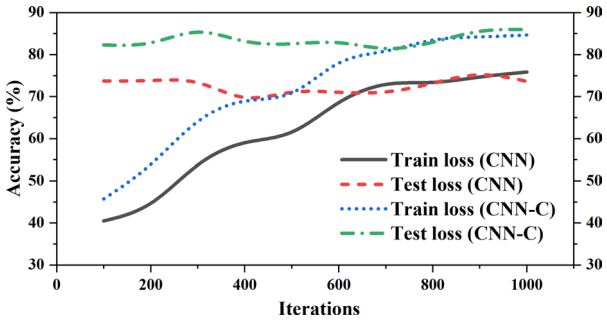

Fig. 8. Training and testing curves for CNN and CNN-C.

Prediction results of box office

Prediction of first-day box office and analysis of influencing factors

The prediction results of first-day box office are shown in Table 7. The mean prediction error is within 30%. Since the first-day box office is also affected by other factors, an error of 25% or less can be recognized by practitioners. During the movie's promotion period, we set targets for intended audience and heat index according to different box office targets for movies of different grades; we predict the first-day box office according to the current intended audience and heat index, and judge the effect of each material placement, so that we can adjust the promotion and distribution strategies at any time.

Analysis of influencing factors of total box office

Different prediction models of box office are compared, as shown in Table 8. In the prediction model of box office based on MLR, the fusion of comments can improve the prediction accuracy of the model. The average prediction accuracy of the ordinary regression prediction model is 64.8%, and the average prediction accuracy

Movie	Truth of first-day box office	Prediction of first-day box office	Error
Mojin: The Worm Valley	7119.75	5747.64	− 19.3%
Kill Mobile	1971.2	2457.864	24.7%
Mad Ebriety	2632.08	3184.83	21.0%
Long Day's Journey Into Night	28905.36	22033.83	− 23.8%
The Wandering Earth	20981.51	14477.21	− 31.0%
Crazy Alien	45121.67	34661.64	− 23.2%
Pegasus	35414.39	25852.53	− 27.0%
The New King of Comedy	30128.12	34313.22	13.9%
Boonie Bears: Blast into the Past	8147.81	9919.584	21.7%
The Knight of Shadows: Between Yin and Yang	7312.25	9261.324	26.7%
Peppa Celebrates Chinese New Year	6471.08	5012.16	− 22.5%
Integrity	6268.02	4512.96	− 28.0%
Average error			23.6%

Table 7. Prediction comparison of first-day box office (unit: US$ 10,000).

Movie	Truth of total box office	MLR-C	Error	MLR	Error
Mojin: The Worm Valley	16526.4	12800.4	− 22.5%	11455.15	− 30.7%
Kill Mobile	70552.9	58263.6	− 17.4%	49316.3	− 30.1%
Mad Ebriety	5548.4	6367.85	14.8%	7657.1	38.0%
Long Day's Journey Into Night	31040.9	22688.4	− 26.9%	18935.4	− 39.0%
The Wandering Earth	515493	399835.2	− 22.4%	310842.4	− 39.7%
Crazy Alien	243453.1	197064	− 19.1%	324280	33.2%
Pegasus	190106.4	155126.4	− 18.4%	264818.4	39.3%
The New King of Comedy	69027.2	79823.75	15.6%	89873.3	30.2%
Boonie Bears: Blast into the Past	78914	90643.3	14.9%	101799.5	29.0%
The Knight of Shadows: Between Yin and Yang	17708.9	20163.5	13.9%	24456.3	38.1%
Peppa Celebrates Chinese New Year	13751.1	15949.55	16.0%	19141.1	39.2%
Integrity	13671.9	11372.25	− 16.8%	18525.1	35.5%
Average error			19.2%		35.2%

Table 8. Prediction comparison of total box office (unit: US$ 10,000).

of the regression model with fused comments is 81.8%. When fusing the comments, the number of negative comments become the most negative factors that can affect the box office of the movie, so in movie box office prediction, fusing comments is crucial for predicting the outcome.

Through the model's explanation of the movie box office, it can be seen that the audience's evaluation of the movie has a strong subjective consciousness, and different audiences of the same movie will give very different views. Through the study of movie ratings, the distribution of the number of positive, neutral and negative comments, and the emotional tendency of the comments, it is found that the audience's choice of movie is more obviously influenced by the movie comments. When the comments continue to increase, the proportion of positive comments continues to increase, the audience's willingness to purchase tickets for the movie will increase rapidly. Therefore, attracting audiences to post favorable movie comments on online platforms has a more obvious effect on the growth of movie box office.

In recent years, because of the impact of COVID-19, the audience's choice of movie types has also changed. In the first few years, visual effect action blockbusters were the types with high box office, but recently, due to the influence of the general environment, the audience is more inclined to choose warm and relaxing works, so comedy, romance and other types are favored by the audience, and adventure and thriller types are not favored by the audience. In this study, these types play a more obvious negative impact on the movie box office. With the change of the times, the audience's enthusiasm for adrenaline-pumping movies has diminished in the past, and they are more inclined to warm and light-hearted movies as their daily leisure and entertainment choices.

Prediction analysis of total box office

Different prediction models are compared and analyzed, subdivided into five prediction models: model based on Multiple Linear Regression (MLR), model based on Multiple Linear Regression with comments (MLR-C), model based on BP Neural Network (BPNN), model based on Convolutional Neural Network (CNN), model based on Convolutional Neural Network with comments (CNN-C). Meanwhile, LightGBM[35] developed by Microsoft, which is more popular, is added for comparative analysis. The predicted box office of the six methods is shown in Table 9, and the prediction accuracy is shown in Fig. 9.

Analyzing Fig. 9, it can be seen that the MLR model without comments has the lowest prediction accuracy, which is maintained between 60 and 70%, with a larger error compared to other models. The LightGBM developed by Microsoft has an accuracy of 75% for prediction, also with a large error. The accuracy of the MLR-C, BPNN model and CNN model for box office prediction is between 70 and 80%, and the errors of the three models

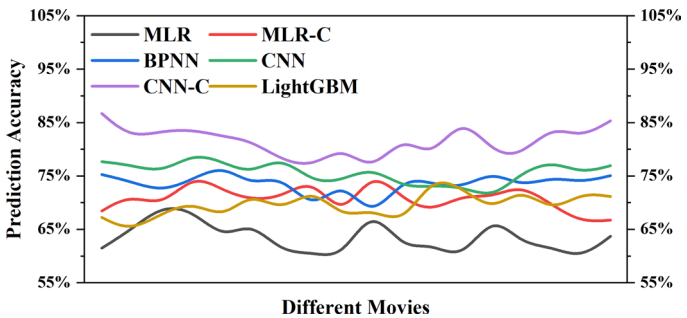

Fig. 9. Prediction accuracy of different models.

Name	Truth	Prediction					
		MLR	BPNN	CNN	LightGBM	MLR-C	CNN-C
Shadow	69187.8	95686.8	90774.2	86138.8	84477.8	78389.3	84519.7
Project Gutenberg	140115.8	188735.8	179488.1	176545.6	172342.5	165196.9	168776.4
Savages	3102.0	4023.8	4057.9	3971.0	3858.8	3610.2	3782.8
Demon Bell	39946.5	27283.3	30279.7	29959.6	31677.8	33555.5	36627.1
Operation Red Sea	401678.2	253458.7	292020.3	308891.0	313308.6	331786.4	361450.3
Monster Hunt 2	246094.2	165621.5	174234.5	179648.7	185554.6	201550.8	218876.1
Itazura na Kiss	19129.0	26627.7	24676.3	23796.3	23146.2	23280.4	23136.5
Hanson and the Beast	32146.4	19191.7	21409.3	24141.7	24013.0	25974.3	27848.3
The Bravest	187752.4	114528.7	131990.1	141565.6	137058.9	156209.9	166470.8
Captain Marvel	113867.6	156226.4	150191.8	143359.7	144383.8	139260.0	139539.3
The Face of My Gene	18683.5	11079.2	13358.4	13638.9	13620.2	16254.7	16688.9
Hello Mrs Money	66481.8	45340.9	47136.1	50792.5	47136.1	53119.0	59165.8
Fat Buddies	28717.7	17833.2	21279.5	20963.8	21825.1	22600.6	26062.6
Detective Chinatown II	373753.6	517275.0	483263.0	468312.9	455979.7	429443.3	452671.5
The Monkey King 3	80014.0	112339.7	106818.8	95456.9	99858.0	94176.5	98757.3
Boonie Bears: The Big Shrink	66606.1	90583.9	88852.5	83391.0	81925.8	76330.1	82171.9
Average error		36.6%	31.7%	28.1%	26.2%	20.5%	16.3%

Table 9. Prediction comparison of total box office under different methods (Unit: US$ 10,000).

are closer. After incorporating comments, the prediction accuracy of each model is improved. Among them, the CNN-C has the highest prediction accuracy of 83.7%.

Overall, the average prediction accuracy of various models is shown in Fig. 10. The prediction accuracy of each model is significantly improved after fusing the comments. The MLR improved the average prediction accuracy by 16.1% after fusing the comments. The CNN model improved the average accuracy by 11.8% after fusing the comments. The average prediction accuracy of LightGBM is 73.8%, which is lower than the MLR model with comments, the CNN model, and the CNN-C model with fused comments. The effectiveness of comments for box office prediction is verified through comparison, and in addition, the results show that the CNN-C model with fused comments has the highest prediction, with an average prediction accuracy of 83.7%.

Conclusion

Box office prediction research is of great significance for understanding investment risks, class construction, promotion and distribution, and theater scheduling. However, due to the insufficient selection of influencing factors of movie box office, the currently existing prediction model restricts the prediction accuracy of box office. Therefore, this paper selects multi-dimensional data to conduct research on movie box office prediction:

(1) Based on the three different stages of project, production and distribution before the release of the movie, a total of 34 influencing factors in 11 categories, such as heat index, movie types, release date, creators, first-day box office, are selected to study the prediction technology of movie box office.
(2) The nouns in the corpus are used to construct a feature thesaurus in movie domain through Word2vec algorithm; for adjectives and verbs with emotional color, three emotional dictionaries, namely, Hownet, NTUSD, and THUOCL, are used to establish an emotional dictionary based on the movie; and emotional scores of the comments are computed through Snow NLP fused with TF-IDF algorithm.

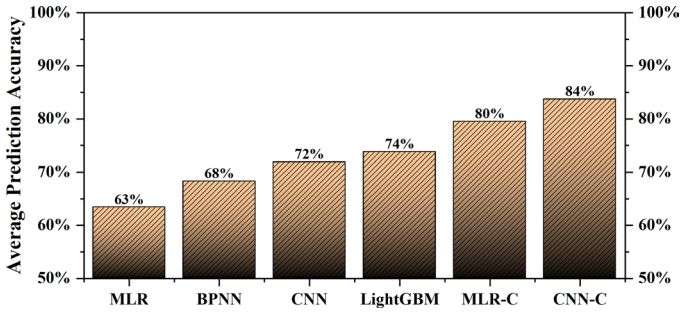

Fig. 10. Average prediction accuracy.

(3) A prediction method of movie box office based on comments and multiple linear regression is designed to analyze the relationship between influencing factors and movie box office. The first-day box office prediction model is established using the number of intended audience and the heat index, which provides an important basis for the prediction of total box office in the later stage, and also provides decision-making references for the movie industry and related management departments.
(4) Incorporating comments as feature values to improve the accuracy, a prediction model of box office based on comments and convolutional neural network is constructed. The results show that the average prediction accuracy of MLR model, BPNN model and CNN model without comments is 64%, 74% and 76%, respectively, and the average prediction accuracy of MLR model and CNN model are improved by 10.9% and 7.9%, respectively, after integrating the comments, and the prediction accuracy has been significantly improved.

This research studies the movie box office prediction models, but there are still many shortcomings. 1. The selection of data features can still be further studied, especially for actors, directors, screenwriters, and other actors who require more precise quantitative evaluations. 2. With the COVID-19 raging around the world, the impact of the COVID-19 on the movie industry is more obvious. In subsequent research, the impact of public health accidents on the Chinese and even global movie market should also be considered. In future research, new prediction models based on new network structures can be designed, which will make the prediction results more precise, stable, and perform better.

Data availability

Douban data comes from https://movie.douban.com/; *Cat's eye* data comes from https://piaofang.maoyan.com/dashboard.

Received: 5 October 2023; Accepted: 5 September 2024
Published online: 11 September 2024

References

1. Delre, S. A. & Luffarelli, J. Consumer reviews and product life cycle: On the temporal dynamics of electronic word of mouth on movie box office. *J. Bus. Res.* **156**, 113329 (2023).
2. Kim, J., Lee, Y. & Song, I. From intuition to intelligence: A text mining–based approach for movies' green-lighting process. *Internet Res.* **32**(3), 1003–1022 (2022).
3. Zhao, J., Xiong, F. & Jin, P. Enhancing short-term sales prediction with microblogs: A case study of the movie box office. *Futur. Internet* **14**(5), 141 (2022).
4. Liao, L. & Huang, T. The effect of different social media marketing channels and events on movie box office: An elaboration likelihood model perspective. *Inf. Manag.* **58**(7), 103481 (2021).
5. Gao, W. *et al.* A big data analysis of the factors influencing movie box office in China. In *Intelligent analytics with advanced multi-industry applications* (ed. Sun, Z.) 232–249 (IGI Global, 2021). https://doi.org/10.4018/978-1-7998-4963-6.ch011.
6. Wang, R. *et al.* Analyzing the impact of covid-19 on the cross-correlations between financial search engine data and movie box office. *Fluct. Noise Lett.* **20**(05), 2150021 (2021).
7. Byun, J. H. *et al.* Movie box-office prediction using deep learning and feature selection: Focusing on multivariate time series. *J. Korea Soc. Comput. Inf.* **25**(6), 35–47 (2020).
8. Zhang, G. *et al.* Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *Eur. J. Oper. Res.* **116**(1), 16–32 (1999).
9. Zhang, W. & Skiena, S. Improving movie gross prediction through news analysis. in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, **1**: 301–304 (2009).
10. Barman, D., Chowdhury, N. & Singha, R. K. To predict possible profit/loss of a movie to be launched using MLP with back-propagation learning. in *2012 International Conference on Communications, Devices and Intelligent Systems (CODIS)*. IEEE, 322–325 (2012).
11. Ru, Y. *et al.* An effective daily box office prediction model based on deep neural networks. *Cogn. Syst. Res.* **52**, 182–191 (2018).
12. Lu, W. & Xing, R. Research on movie box office prediction model with conjoint analysis. *Int. J. Inf. Syst. Supply Chain Manag. (IJISSCM)* **12**(3), 72–84 (2019).
13. Dai, D. & Chen, J. Research on mathematical model of box office forecast through BP neural network and big data technology. *J. Phys. Conf. Ser.* **1952**(4), 042118 (2021).
14. Liu, L. & Zhao, Y. Research of box-office prediction based on rough set and support vector machine. *Int. J. Hybrid Inf. Technol.* **9**(2), 417–426 (2016).
15. Kim, T., Hong, J. & Kang, P. Box office forecasting using machine learning algorithms based on SNS data. *Int. J. Forecast.* **31**(2), 364–390 (2015).
16. Song, T. *et al.* Using user-and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Inf. Syst. Res.* **30**(1), 191–203 (2019).
17. Shen, D. Movie box office prediction via joint actor representations and social media sentiment. arXiv:2006.13417 (2020).
18. Jiang, L. & Wang, Z. Predicting box office and audience rating of Chinese films using machine learning. in *Proceedings of the 2018 International Conference on Education Technology Management*. 58–62 (2018).
19. Zhou, Y., Zhang, L. & Yi, Z. Predicting movie box-office revenues using deep neural networks. *Neural Comput. Appl.* **31**, 1855–1865 (2019).
20. He, Q. & Hu, B. Research on the influencing factors of film consumption and box office forecast in the digital era: Based on the perspective of machine learning and model integration. *Wirel. Commun. Mobile Comput.* **2021**, 1–10 (2021).
21. Ni, Y. *et al.* Movie box office prediction based on multi-model ensembles. *Information* **13**(6), 299 (2022).
22. Mahmud, Q. I. *et al.* A machine learning approach to predict movie revenue based on pre-released movie metadata. *J. Comput. Sci.* **16**(6), 749–767 (2020).
23. Ahmad, I. S. *et al.* A survey on machine learning techniques in movie revenue prediction. *SN Comput. Sci.* **1**, 1–14 (2020).
24. Liao, Y., Peng, Y. & Shi, S. *et al.* Early box office prediction in China's film market based on a stacking fusion model. Ann. Oper. Res. 1–18 (2022).
25. Al Fahoum, A. & Ghobon, T. A. Accurate machine learning predictions of sci-fi film performance. *J. New Media* **5**(1), 1–22 (2023).

26. Mi, C., Li, M. & Wulandari, A. F. Predicting video views of web series based on comment sentiment analysis and improved stacking ensemble model. *Electron. Commer. Res.* https://doi.org/10.1007/s10660-022-09642-9 (2022).
27. Greco, F. & Polli, A. Emotional text mining: Customer profiling in brand management. *Int. J. Inf. Manag.* **51**, 101934 (2020).
28. Cheng, Q. *et al.* Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *J. Med. Internet Res.* **19**(7), e243 (2017).
29. Li, H., Liu, H. & Zhang, Z. Online persuasion of review emotional intensity: A text mining analysis of restaurant reviews. *Int. J. Hosp. Manag.* **89**, 102558 (2020).
30. van Lissa, C. J. Mapping phenomena relevant to adolescent emotion regulation: A text-mining systematic review. *Adolesc. Res. Rev.* **7**(1), 127–139 (2022).
31. Greco, F. & Polli, A. Vaccines in Italy: The emotional text mining of social media. *Rivista Italiana di Economia Demografia e Statistica* **73**(1), 89–98 (2019).
32. Zhang, D. & Wu, C. What online review features really matter? An explainable deep learning approach for hotel demand forecasting. *J. Assoc. Inf. Sci. Technol.* **74**(9), 1100–1117 (2023).
33. Tang, T., Huang, L. & Chen, Y. Evaluation of Chinese sentiment analysis APIs based on online reviews. in *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE, 923–927 (2020).
34. Deng, Y. *et al.* Emotional discourse analysis of COVID-19 patients and their mental health: A text mining study. *Plos one* **17**(9), e0274247 (2022).
35. Al, D. E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *Int. J. Comput. Inf. Eng.* **13**(1), 6–10 (2019).

Acknowledgements
Thanks to Dr. Zehua Li from College of Communication, Northwest Normal University, for providing technical support such as software debugging and data analysis for the work of this study.

Author contributions
Conceptualization, Y.M. and D.X.; methodology, D.X.; Resources: Z.Z.; validation: Z.Z., Y.M. and D.X.; Supervision: Z.Z. and Y.M.; formal analysis: Z.Z. and Y.M.; data curation: Z.Z. and Y.M.; writing—review and editing: Z.Z. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Z.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.