IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 14, 2022, accepted April 7, 2022, date of publication April 18, 2022, date of current version April 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3168161

Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System

SANDIPAN SAHU[1], RAGHVENDRA KUMAR[1], MOHD SHAFI PATHAN[2], JANA SHAFI[3], YOGESH KUMAR[4], AND MUHAMMAD FAZAL IJAZ[5], (Member, IEEE)

[1]Department of Computer Science and Engineering, GIET University, Gunupur 765022, India
[2]Department of Computer Science and Engineering, MIT ADT University, Loni Kalbhor 412201, India
[3]Department of Computer Science, College of Arts and Science, Prince Sattam Bin Abdulaziz University, Wadi Ad-Dawasir 11991, Saudi Arabia
[4]Indus Institute of Technology and Engineering, Indus University, Ahmedabad 382115, India
[5]Department of Intelligent Mechatronics Engineering, Sejong University, Seoul 05006, South Korea

Corresponding author: Jana Shafi (j.jana@psau.edu.sa)

This work was supported by Sejong University.

ABSTRACT The movie is one of the integral components of our everyday entertainment. The worldwide movie industry is one of the most growing and significant industries and seizing the attention of people of all ages. It has been observed in the recent study that only a few of the movies achieve success. Uncertainty in the sector has created immense pressure on the film production stakeholder. Moviemakers and researchers continuously feel it necessary to have some expert systems predicting the movie success probability preceding its production with reasonable accuracy. A maximum of the research work has been conducted to predict the movie popularity in the post-production stage. To help the movie maker estimate the upcoming film and make necessary changes, we need to conduct the prediction at the early stage of movie production and provide specific observations about the upcoming movie. This study has proposed a content-based (CB) movie recommendation system (RS) using preliminary movie features like genre, cast, director, keywords, and movie description. Using RS output and movie rating and voting information of similar movies, we created a new feature set and proposed a CNN deep learning (DL) model to build a multiclass movie popularity prediction system. We also proposed a system to predict the popularity of the upcoming movie among different audience groups. We have divided the audience group into four age groups junior, teenage, mid-age and senior. This study has used publicly available Internet Movie Database (IMDb) data and The Movie Database (TMDb) data. We had implemented a multiclass classification model and achieved 96.8% accuracy, which outperforms all the benchmark models. This study highlights the potential of predictive and prescriptive data analytics in information systems to support industry decisions.

INDEX TERMS Expert systems, content-based, recommendation system, deep learning, audience groups.

I. INTRODUCTION

The worldwide movie industry is a fast-moving revenue generating industry, and the multi-billion dollar has been involved in this industry. A large number of people are associated with this industry, and massive investment is required as qualitative and quantitative. In 2019 the total box office revenue of the United States and Canada was $11.32 billion [59]. However, in-ground reality, few numbers movies has been achieved success. Film producers and researchers constantly feel it essential to have some expert systems that predict the movie's success chance leading its production with appropriate accuracy. The movie industry is massive and diversified. A significant number of parameters from different dimensions are involved in creating a movie. Representing an upcoming movie's success or degree of success is a highly complex task. Research works [32], [33], [60] have been conducted to predict movie popularity. Earlier, several works have been conducted on post-production or post-release forecast. However, it is not beneficial as the investor has already contributed their funds to the film production. The early production stage and pre-production prediction with satisfying accuracy have been beneficial to secure investment. A forecast made soon after the cast, director, and

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

storyline have been finalized would assist the investor in making a financial decision.

After a rigorous study, we have seen significant researchon movie hit prediction before the official release. Predictions performed shortly before [32], [56] or following [33], [60], [34] the official release (the last stage in film production) may have additional data to use and produce a more precise prediction [66]. Still, they are considerably delayed for investors to estimate any critical decision. Early-stage (production) forecast [39], [40] of movie success is the most beneficial. Very little work has been performed to forecast movie success at an early stage of movie production. The early-stage forecast of previous works' accuracy is not significantly good. Maximum of the works are performed only to focus success probability of the upcoming movie. Some of them classify the problem into a binary problem (hit/ flop), and in some work, they classify the problem into a multiclass problem. Movie Makers start creating a new movie while targeting a specific audience group or groups most of the time. Audience age is one of the essential criteria for the target audience [61]. Some movies are created by targeting the junior audience group. Some movies target teenage audiences, sometimes target the mid-age and senior audience group, and some movies are for all. Suppose we could predict whether the upcoming movie would be famous among the target audience group or not. Movie makers would be benefited if we could measure the influence of the upcoming movie among all the age groups at the early stage of the movie production. Then, movie Maker could make necessary changes if needed. The movie hit forecasting and target audience prediction of the upcoming movie at the early stage of the movie production are interrelated and meaningful. The outcome of this work could reduce the risk involved in the movie industry.

Our research problem proposed a system to predict movie success at an early stage of movie production and performs movie Target audience prediction. Both the above works have been done using the CB movie recommendation system. Our research work can be folded into three significant parts. In our framework, the first module is a movie recommendation system [2]. We have considered only five essential features of the movie like genre, cast, director, keywords and movie description as the feature sets to build the recommendation system. All these basic features are available at the early stage of the movie production. The proposed recommendation system provides a set of similar movies of a given upcoming movie. The second module accepts similar movies from the first module, uses movie rating and voting information of similar movies, and creates a novel feature set. Next, we have proposed a CNN model and use the newly created data set to predict the movie's popularity. We have divided the popularity of a movie into six classes super-duper hit (SDH), super hit (SH), hit (H), above average (AA), average (A) and flop (F). Next, in the third and final phase, we build a module to estimate the target audience. We have divided the audience group into four age groups junior, teenage, mid-age and senior. We used a similar movie set from the first module and created a new feature set from each age group considering movie rating and voting information. Using the new data set, we built a model using fuzzy c means and cosine similarity to estimate the popularity of the upcoming movie among all age groups.

The primary contributions of this study are as follows.

1. This research work is among the foremost in the previous studies to use a recommendation system to predict upcoming movie popularity at its early production stage.
2. Proposed a model to estimate the popularity of an upcoming movie among different age groups using fuzzy c mean and cosine distance.

The rest of this paper is arranged as follows. Section II summarizes the related work to RS and film forecasting. Section III outlines our proposed framework in detail, illustrates all the features and introduces the movie hit success criterion. In Section IV proposed model is described elaborately. Section V presents the experimental results simultaneously with a comparative study of other statistical models shown and explained—finally, research contributions and their limitations and further research directions in section VI.

II. RELATED WORKS

In this section, we have presented a detailed survey on the past related works. Our work is divided into three interrelated parts. Highlight some of the previously proposed models of recommendation system and then discuss movie popularity prediction. In the proposed work, movie recommendation and movie popularity are not interrelated. Our proposed work has used recommended movies to predict the upcoming movie's popularity and predict the movie's target audience. The recommendation system is primarily divided into three parts [1]–[3], [18], collaborative filtering (CF) [4]–[8], [17], content-based filtering(CBF) [8], [9], [11]–[13], [31] and hybrid filtering [14]–[16]. CF is a procedure that can refine things that a user might prefer based on responses by similar users. It searches a broad group of people and gets a smaller circle of users with tastes comparable to a particular user. It looks at the things they like and connects them to form a ranked list of suggestions.

A. CONTENT-BASED RECOMMENDATION

In some application situations, the recommended items must be content-wise comparable to a reference item, e.g., for similar item recommendations [19]. Also, content information allows the period of better descriptions [20], which is becoming frequently crucial in fair and open recommender systems. Content-based recommender systems utilize metadata information of items or textual items [21]. Linked Open Data (LOD) initiation suggests new ideas to extend item information with outside knowledge sources [22], [23]. Movie recommendation using CBF is one of the widely used research paradigms. A content-based movie recommender has been proposed where users with and movie features

are used [24]. Proposed movie rating using movie feature set. Content-based movie recommended system considered different movie attributes like movie genre, name of the actors, name of the directors, and other attributes to build a recommender system. The movie genre that users prefer to watch has been used to build a recommender system using Movie Lens dataset [25]. Correlations between content or attributes are measured to find out the similarity between items. A multi-attribute network has been proposed to calculate the correlations to recommend items to users [26]. The similarity between directly or indirectly correlated items is calculated using network analysis. They have proposed a hybrid model where genomic tags of the movie have been used with CBF to recommend movies with similar tastes [27]. The proposed model reduces the computational complexity by using principal component analysis (PCA) and Pearson correlation procedures to reduce redundant tags and dispense a low variance [64]. In the following work, authors have used and leveraged the gap between high-level and low-level features [28]. They have used low-level feature colors, motion, exceeds and lighting from film to make a hybrid recommendation system. A new movie recommendation system has been proposed and addresses the cold start problem for the new item [29]. They have offered audio and visual descriptions extracted from movie videos and developed a video genome. A hybrid movie recommendation system has been proposed to incorporate sentiment analysis with collaborative filtering (CF) [30]. Movie tweets have been used from micro blogging sites to understand the public sentiment, current trends, and user response. The sparsity of data is one of the significant challenges for recommendation system algorithms. In the following work, a generative adversarial network (GAN) has dealt with the sparsity of review data and rating. They have proposed Rating and Review Generative Adversarial Networks (RRGAN), an innovative framework for the recommendation [67]. GAN also been used to rank the movie according to the preference of the users. In the following work,LambdaGAN has been used for recommending top-N movies [68].

B. MOVIE HIT PREDICTION

Movie feat broadcasting is a well-known problem of research. The problem is broadly divided into two primary groups based on forecasting time. Significant work has been proposed where predictions were made very late at the production stage before the movie's release or just after the movie's official release [32]–[38], [60]. Limited works have been carried out where movie hit forecasting has been executed at the initial stage or the early stage of the production [39]–[43]. The late prediction may be facilitated by more movie attributes to increase the forecasting accuracy. On the other hand, for the early prediction, only a few attributes or features are available for making movie predictions which makes the problem much more difficult.

One of the most significant parts of our problem is defining the success of a movie. No benchmark models exist which define the success of a movie. Few works have focused on whole box office revenue [40], [43]–[48]. At the same time, some have adopted the number of admissions [34], [49]. The underlying assumptions to make revenue or the number of admissions as the parameter of success. Some of the earlier works measured success as profitability. It may be a numeric value of revenue [50] or the return on investment (ROI) [39], [51], [52]. Several works distributed movies into two classes (success or not) and selected binary classifications; some considered the forecast a multiclass classification problem and tried to classify films into multiple discrete classes [47]. Predictions are also made on continuous integral values of profit metrics [32], [39], [53], with values of these metrics containing logarithmic in some works [48], [50], [54].

The movie hit forecast trusted machine learning models considering these learning techniques have developed prediction models with reasonable levels of accuracy [55], [52], [56], [65]. For instance, [56] has presented some machine learning models such as discriminate analysis, Logistic regression (LR), Decision tree (DT), and Neural Network (ANN) and measured the performance to predict a movie's success. Authors of [57] have proposed the multi-layer back propagation architecture and a more quality increased neural network model proposed by [56]. The authors [58] obtained movie data from websites like and rotten tomatoes, IMDb and executed machine learning strategies like support vector machine (SVM) and linear and Logistic regression. Authors [38] Introduced Cinema Ensemble Model (CEM) to enhance forecast accuracy, comprise seven machine learning models, and focus on selecting attributes. The research [40] proposed few new features to predict the box-office success of a movie. They have adopted a Voting system to foretell by averaging the output from various machine learning classifiers.

In most of the works, similar movies are computed using an RS from the existing movie. The recommendation system for the upcoming movie is sporadic. Content-based movie RS could be used to find out similar movies for an upcoming movie. Box-office information of all these similar movies could be used to analyze the upcoming movie. Our research work has used Content-based movie RS for an upcoming movie and find out similar movies. We analyze the output of the RS and successfully build a model to forecast movie popularity. Again, we move one step ahead and also predict the target audience from the information of the RS. Most of the research works focused on the movie hit prediction problem as a binary classification problem. Very few of the works [40] resolve the problem as a multiclass problem, but they have sacrificed the accuracy in the process. Our research work has classified the movie popularity problem into six different classes and achieved high accuracy. Maximum of the works only target the movie popularity prediction problem; they have only predicted the popularity of the upcoming movie. Research work regarding the forecast of the target audience of the upcoming movie is sporadic.

III. MATERIALS AND METHODS

This research study aimed to develop a model that will predict movie popularity and its age-wise preference using movie recommendations. Our objective is to classify the movie popularity among the six classes {SDH, SH, H, AA, A, F} at the early movie production stage. Next, our objective is to find the movie's target audience and determine its influence on audience groups. Regroup the audience into four age demography {Junior, Teenage, Mid-Age, Senior}. Our final output of the system will be age-wise movie popularity prediction.

In this study, we used a content-based movie recommendation system to find out a similar movie. In the next step, we use the voting information and rating of each recommended movies. All these data are used to train the 1-D CNN deep learning model. The output of the CNN model is the classification of the film among six classes. We predict the scale of popularity of the movie.

Our third module of the system takes recommended movie information and age-wise voting information. We have grouped the voting information into four age groups. We use Fuzzy C-Mean to calculate the movie preference for each age group.

The framework of our job has three significant steps, which are listed below Fig. 1

1. Acquire movie data and movie intrinsic features from TMDb dataset and computes similar movie using a content-based movie recommendation system.
2. Use similar movie information and voting data from the IMDb beta set. Predict the movie popularity using the Deep learning approach.
3. Compute target audience prediction using fuzzy c means.

- We have introduced a new data set containing voting and rating information of the recommended movies, used to predict the movie popularity class.
- We have proposed new parameters called Global centroid for each age group.
- We have also presented a new approach for estimating the interest or popularity of an upcoming movie among distinct age groups.

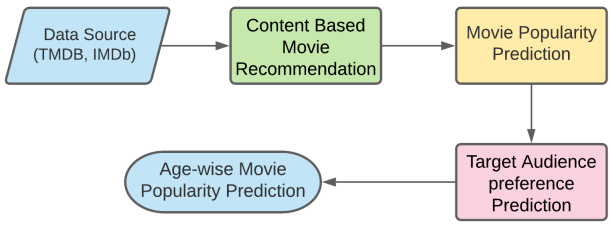

FIGURE 1. Framework workflow.

A. DATASET DESCRIPTION

The proposed system has three modules. The first module is a content-based (CB) recommendation system (RS), the first to use the TMDb database. The second and third modules make use of the IMDb database.

Multiple public databases are available in the market, and all these databases are used for movie recommendation and movie popularity prediction system. The proposed content-based movie recommendation system used tmdb_5000_movies and tmdb_5000_credits datasets, which are publicly available [62]. The movie hit prediction and target audience prediction module make use of the IMDb rating dataset. Use of two different databases (TMDb and IMDb) creates synchronization problems since they use two separate movies ID. The proposed system uses the link small dataset to merge two databases.

The tmdb_5000_movie data set is consistent with 4803 movie data. It contains 20 movie attributes. That isits accountants the movie with published year from 1916 to 2017. From the 20 attributes, we have chosen only 4 attributes. Selected attributes are keywords, overview, tagline and genre. Attributes like budget, revenue, and release year are very much dependent on time, and since we are considering movie of more than a hundred years of span, these attributes could not be selected. Other attributes like runtime, spoken language, original title, homepage are also irrelevant to the proposed system.

The tmdb_5000_credits data set consisted of 4813 movie data. The data set has 4 attributes movie_id, title, cast, and crew. We have extracted the name of three primary cast members and each movie's director name from the data set. The director is one of the most influential characters of a movie. Similarly, the first 3 cast members are also critical. The movie attributes selected for the proposed content-based movie system are shown in table 1.

TABLE 1. Example of a movie data used in CB recommendation system.

Attribute	Value
Movie_id	68721
Genres	Action, Adventure, Science Fiction
Keywords	Terrorist, war on terror, Tennessee, superhero, tony stark, war machine, extremis etc.
Overview	When Tony Stark's world is torn apart by a formidable terrorist called the Mandarin, he starts an odyssey of rebuilding and retribution.
Tagline	Unleash the power behind the armor.
Casts	Robert Downey Jr, Gwyneth Paltrow, Don Cheadle
Director	Shane Black

IMDb database is publicly available for research work. The imdb_rating dataset has been used in the proposed work, which consisted of 85856 movie rating data [63]. The data set content rating and voting details of each movie. The data set has a total of 49 attributes, which includes gender-wise and age demography wise voting and rating details. TMDb data set uses tmdb_ID, and IMDb dataset uses imdb_id; they use two different movies ID. Links for all data set creates a link between two movie IDs. Otherwise, using two different

movie databases would create a synchronization problem. Since multiple databases are used in the proposed work and each data set has a different size. After combining all the data sets and synchronizing the IMDB_ID with the tmdb_ID, the number of movies used in the work is 3100.

B. DATASET PREPROCESSING

The attributes used in the proposed CB movie recommendation system are mentioned in the table 1. The value of the attributes like genre, cast, director and keywords are present in JSON data. Convert the JSON data into the string, which removes all the metadata and contains only attribute values. Next, for each attribute, one list has been generated with all the unique values of the attributes mentioned in the dataset. Next, performs the one-hot encoding to all attitudes for each movie.

The tagline of a movie is appended with the overview attribute. Combining these two attributes needs to go through several steps before finding out the similarity between them.

Step-1: Clean the data correctly. Moreover, this will allow us to reduce sentences, paragraphs, and ultimately docs to a set of single words.

Step-2: Go through the stemming process of overcoming inflected words to their word stem, root or origin form.

Step-3: Remove stop words from the set of words. English word dictionary has been used to eliminate stop words.

Step-4: After making the set of clean and filter data, next is implementing the core functionality. Term frequency is calculated by the number of times the term t repeat in the document d.

$$TF\ (t, d) = count\ of\ tind\ /\ number\ of\ wordsind$$

Step-5: Inverse document frequency (IDF) is used to pull down the weight of frequent terms. In contrast, size up the rare ones, by calculating IDF, the log of the total number of documents N in the corpus D divided by the number of documents df_t including the term t.

$$IDF\ (t) = log\frac{N}{df_t}$$

Step-6: Finally, for a term t the weight in a particular document d is determined as the outcome of the two preceding calculations:

$$TF - IDF\ (t, d) = TF\ (t, d)\ .IDF\ (t)$$

Step-7: Compute the similarity between two movie descriptions using cosine similarity.

C. MOVIE POPULARITY LABELLING (HIT TO FLOP)

The motion picture is one of the branches of art. Several parameters are associated with the movie industry. Movie interest is a complicated and extensive industry lot of elements are associated with the industry. Different perspectives are there to consider that several parameters are there to assess a movie's success. Box office revenue could be one of the parameters. The budget and the movie's revenue are changeable and depend on the movie industry. For some industries, the budget may be generally higher than the other small movie industry or varies from film to film. Defining revenue range [40] to classify the movie success does not apply to all movies. Also, fixing the profit margin [39] will not solve the problem of classification. The scale of profit margin is undoubtedly lower for low budget and higher for high budget films.

Considering all these things, the IMDb rating is one of the significant criteria for movie success prediction, and also IMDb rating is a globally accepted rating. Figure 2 presents the histogram of the IMDb rating of all the films considered in our database. It shows that the rating is near a normal distribution, contributing to the model prediction's robustness. In our work, we have used IMDb rating as the primary parameter to determine the movie success.

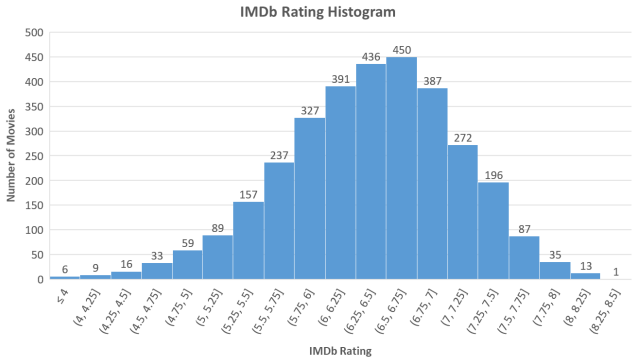

FIGURE 2. Represents the distribution of the movie number at IMDb rating.

In this work movie, popularity is classified into six classes. In our early production stage movie classification problem, the prediction module would predict the IMDb rating for a range of IMDb ratings. According to the predicted IMDb rating, the upcoming movie e is classified into six classes.

Classes are specified as a super-duper hit (SDH), super hit (SH), hit (H), above average (AA), average (A) and flop (F). We have prepared the movie data set and labeled the movie in different classes according to their IMDb rating. Table 2 represents how movies are classified into different classes according to their IMDb rating.

TABLE 2. Classification of the movie according to the IMDb rating.

IMDb Rating	0 – 4.9	5 – 5.9	6 – 6.9	7 – 7.9	8 – 8.9	9 - 10
Movie Class	F	A	AA	H	SH	SDH

IV. PROPOSED WORK

The proposed system has three major interrelated modules. The first module is a content-based movie recommendation system model, which produces 10 most similar movies of

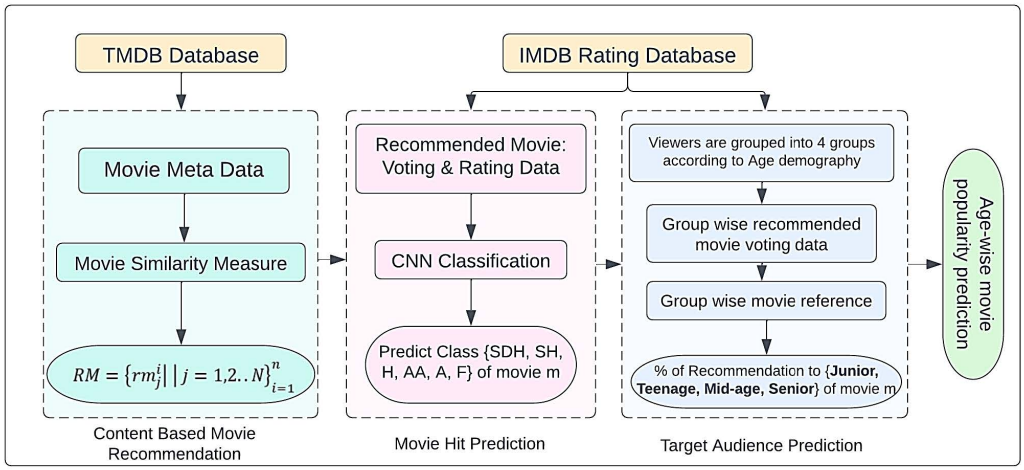

FIGURE 3. Overall process flow diagram.

all the movies listed in the data set. The second module is a movie hit prediction module. The output of the first module is the input of the second module. Next, the third module groups the audience according to their age and predict the most suitable target audience group for each movie. Finally, the whole system provides an age-wise movie popularity prediction of the upcoming movie. The overall process flow diagram has shown in figure 3.

A. CONTENT-BASED MOVIE RECOMMENDATION

Content-based filter used for finding a similar movie. Which uses movie attributes to find out the similarity between the two movies. Let feature set $F = (F_1, F_2, \ldots, F_m)$. Compute the similarity between any two movies m_i&m_j concerning the feature F_k is:

$$dist\_F_{k_{i,j}} = similarity(F_{k_i}, F_{k_j}) \tag{1}$$

$$dist_{i,j} = \cup_{k=1}^{m} \{dist_{F k_{i,j}}\} \tag{2}$$

In (1), the $dist_{i,j}$ is the distance vector between the two movies m_i & m_j. The objective of this module is to compute N most similar movies of movie m_i. The similarity between any two movies has been measured using m different features. Distance between any two movies is an m dimensional vector. The nearest neighbour algorithm has been used to find out N most similar movies. The overall similarity measure between the two movies is computed by using similarity measures like cosine similarity:

$$c\_dist_{ij} = cosine\_sim(dist_{ii}, dist_{ij}), \tag{3}$$

$$c\_dist_i = \{c\_dist_{ij}\}_{j=1}^{n} \tag{4}$$

In (4), We have computed the similarity measure between any movie m_i with all the other n movies present in the data set. The m dimensional distance vector is reduced to one-dimensional distance using the cosine similarity measure. Figure 4 presents the block diagram of the CB movie recommendation module.

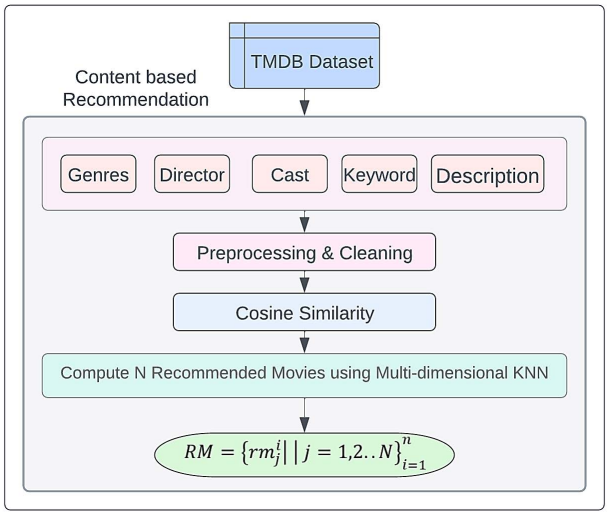

FIGURE 4. Movie recommendation module.

B. MOVIE HIT PREDICTION

The second module's purpose in this research problem is to build a movie hit prediction system. This module accepts the earlier module's output. From the previous module set of the recommended movie of each movie m_i is used as input, i.e. $RM = \cup_{i=1}^{n} \{rm_j^i \mid j = 1, 2, \ldots, N\}$. With that IMDb rating data set has been used as an input. The movie hit prediction is a multiclass classification problem. Input data are processed and fit into a deep learning model, and output is a multiclass classification. In this research problem, movies are classified into six classes, super-duper hit (SDH), super hit (SH), hit (H), above average (AA), average (A) and flop (F). Figure 4 presents the framework of the movie hit prediction module.

In this module, we have used the rating and the voting details of N recommended movie rm_j^i of the movie m_i. Let $v_{r,k}^i$ represents the number of the vote with a rating r for kth

Algorithm 1 Movie recommendation using multi-dimensional KNN.

Input:
Movie Feature set F = $\{F_1, F_2, \ldots F_N\} = \{D, K, G, DT, C\}$
Movie Description & Tag line $D = \{d_i\}_{i=1}^{n}$
Movie Keywords $K = \{k_i\}_{i=1}^{n}$,
Movie Genres $G = \{g_i\}_{i=1}^{n}$
Movie Director $DT = \{dt_i\}_{i=1}^{n}$,
Movie Cast $C = \{c_i\}_{i=1}^{n}$
Output:
N most similar movie of movie $m_i, \forall m_i\{rm_j^i | j = 1, 2, \ldots, N\}_{i=1}^{n}$

1. $D_{TF-IDF} = \{d\_ti\}_{i=1}^{n}$ calculate $TF - IDF \forall t_{ji} \in d_i$, *where* $d_i \in D$
2. $K_{BIN} = \{w_{bin}\}_{i=1}^{n}$
   where k_i transfer to binary word binw_bin_i
3. $G_{BIN} = \{g_{bin}\}_{i=1}^{n}$
   where g_i transfer to binary genrebing_bin_i
4. $DT_{BIN} = \{dt_{bin}\}_{i=1}^{n}$
   where dt_i transfer to binary director bindt_bin_i
5. $C_{BIN} = \{c_{bin}\}_{i=1}^{n}$
   where c_i transfer to binary cast binc_bin_i
6. **for** (*Each movie* $m_i \in M$)
7. **for** $\left(\text{Each movie } m_j \in M \& m_j \neq m_i\right)$
8. $dist\_d_{ij} = cosine\_sim\left(d\_ti_i, d\_ti_j\right)$
9. $dist\_w_{ij} = cosine\_sim\left(w\_bin_i, w\_bin_j\right)$
10. $dist\_g_{ij} = cosine\_sim\left(g\_bin_i, g\_bin_j\right)$
11. $dist\_dt_{ij} = cosine\_sim\left(dt\_bin_i, dt\_bin_j\right)$
12. $dist\_c_{ij} = cosine\_sim\left(c\_bin_i, c\_bin_j\right)$
13. $dist_{ij} = \{dist\_d_{ij}, dist\_w_{ij}, dist\_g_{ij}, dist\_dt_{ij}, dist\_c_{ij}$
14. $dist_i = \{dist_{ij}\}_{j=1}^{n}$
15. **for**$(\text{Each movie } m_j \in M \& m_j \neq m_i)$
16. $c\_dist_{ij} = cosine\_sim(dist_{ii}, dist_{ij})$
17. $c\_dist_i = \{c\_dist_{ij}\}_{j=1}^{n}$
18. $c\_dist_i = sort(\{c\_dist_{ij}\}_{j=1}^{n})$
19. $RM^i = \{rm_j^i | j = 1, 2, \ldots, N = firstNmoviesof\, c\_dist_i$
20. $RM = \{rm_j^i | j = 1, 2, \ldots, N\}_{i=1}^{n}$
21. **Return** RM

recommended movie of m_i. V_r^i represents the total number of the vote with a rating r for all recommended movies of m_i.

$$V_r^i = \sum_{k=1}^{N} v_{r,k}^i \tag{5}$$

$$V^i = \{V_r^i | r = 1, 2, \ldots, 10\} \tag{6}$$

In (6), the V^i represents the voting details of the movie m_i. In (7) the R_j^i rating of each recommended movie, also considered as:

$$R^i = \{R_j^i \,|\, j = 1, 2, \ldots, N\}$$
$$V = \{V^i\}_{i=1}^{n} \quad and\ R = \{R^i\}_{i=1}^{n}$$
$$\text{creates input dataset} \qquad X = V \cup R \tag{7}$$

TABLE 3. Example of voting details.

$\boldsymbol{V^i} = \{V_r^i \| r = 1, 2, .., 10\}$	$m_i =$ Avatar	$m_i =$ Spider-Man 3
V_1^i	822428	210605
V_2^i	800110	206442
V_3^i	806286	482927
V_4^i	446746	561242
V_5^i	190478	324569
V_6^i	101340	154969
V_7^i	51821	67734
V_8^i	29602	34332
V_9^i	19301	19318
V_{10}^i	38270	26599

TABLE 4. Example of rating details.

$\boldsymbol{R^i} = \{R_j^i \| j = 1, 2, .., N\}$	$m_i =$ Avatar	$m_i =$ Spider-Man 3
R_1^i	7.1	6.7
R_2^i	7.3	6.8
R_3^i	7.7	5.7
R_4^i	7.7	6.3
R_5^i	6.8	7.6
R_6^i	6.3	5.3
R_7^i	7.5	5.8
R_8^i	5.4	5.8
R_9^i	6.5	6.9
R_{10}^i	7.5	4.8

Table 3 and 4 presents the voting V^i and rating R^i features values of two movies (Avatar & Spider-Man 3) as an example.Prediction of the class of the upcoming movie, the data set X has been used as the input. The input data set contains 3,100 movie details. The data set has been labelled with six different classes. The input data set is again divided into the training and test part. The training part contains 2,325 movie data, and the test contains 775 movie data. The convolutional neural network classification model has been used to classify movies into 6 different classes according to popularity. Figure 5 represents the block diagram of the movie hit prediction module.

The proposed deep learning model is 1D-CNN architecture. It consists of three convolutional layers and one dense layer. The structure of the 1D-CNN is experimentally selected by a trial and error approach. The first Layer of CNN has taken an input of 22×1 array comprising all features. We have

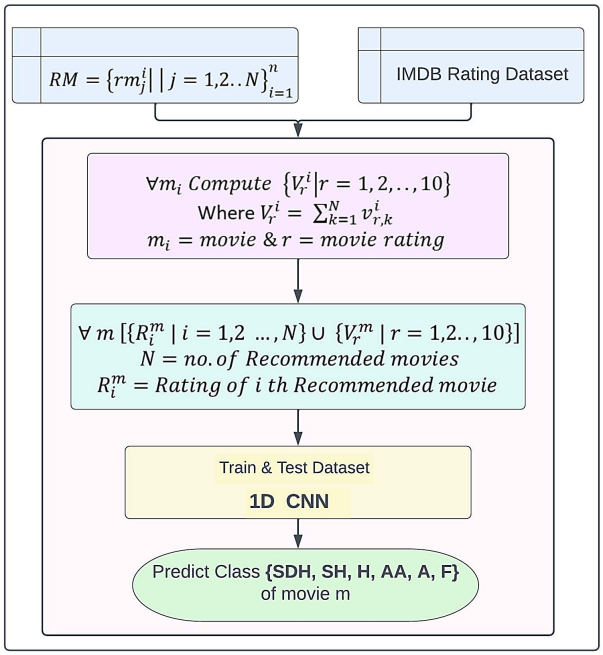

FIGURE 5. Movie hit prediction model.

used 128 filters in the first layer with kernel size 5. We have adopted the activation function Relu and dropout 0.1. With that, max-pooling is estimated as 2. The next layer maintains 128 kernel size and the same activation function Relu and dropout 0.1 and repeats this two times. Finally, a flattering layer has been used. The Last Layer is a dense layer with six over here with 6 predictive classes. Multiple Keras Optimizers have been experimented like Adam, SGD, RMSprop, and finally, we have selected RMSprop optimizer due to better accuracy. Figure 6 depicts the topology of the proposed 1D-CNN.

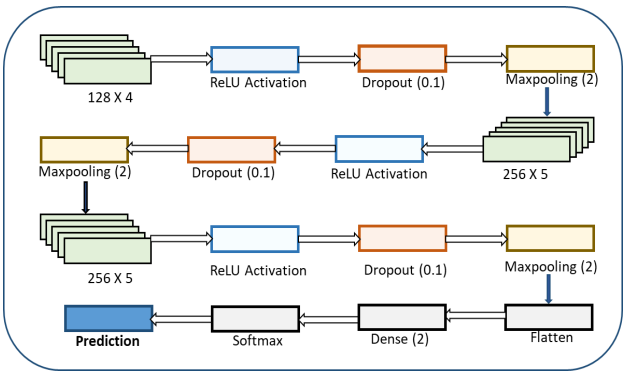

FIGURE 6. Topology of the proposed 1D-CNN.

C. TARGET AUDIENCE PREDICTION

The third and final module forecasts the target audience's preference according to the age of demography. To predict the upcoming movie's target audience, we have considered all the recommended movies delivered by our first module.

Algorithm 2 Movie popularity prediction using deep learning.

Input:
Recommended Movie details $RM = \{rm_j^i | j = 1, 2, \ldots, N\}_{i=1}^{n}$
Movie Rating Database M_r
Output:
Prediction of Popularity Class (0, 1, 2, 3, 4, 5) for *movie* m_i

1. **for**(*Each movie* $m_i \in M$)
2. **for**(*Rating* $r = 1, 2, \ldots, 10$)
3. $V_r^i = \sum_{k=1}^{N} v_{r,k}^i$,
4. *where* $v_{r,k}^i$ = no. of vote with rating r of *kth* recommended movie of m_i
5. $V^i = \{V_r^i | r = 1, 2, \ldots, 10$ voting details of each movie m_i
6. $V = \{V^i\}_{i=1}^{n}$ voting details of all movie M
7. $R = \{R_j^i | j = 1, 2, \ldots, N\}_{i=1}^{n}$, where R_j^i = rating of *jth* recommended movie of m_i
8. *Create Dataset* $X = V \cup R$
9. $Y = \{y_i | y_i \in (0, 1, 2, 3, 4, 5)\}_{i=1}^{n}$
10. $X = X_{train} \cup X_{test}$
11. Y = Y_train ∪ Y_test
12. Y ^_test = CNN(X_train,Y_train,X_test)
13. **Return** Y ^_test

The system would use user rating by each age group and take the number of vote details from each group to each recommended movie—system analysis of all the voting and rating information from each group of all the recommended movies. Ultimately, the module would predict the popularity of the upcoming movie for each age group. The target audience prediction module takes input from the first module output, and that also takes the IMDb rating data set as input. Movie recommendation module produces similar movies ($rm_j^i | j = 1, 2, \ldots, N$) of a given movie *mi*. We have used a set of recommended movies of all movie $RM = \cup_{i=1}^{n} \{rm_j^i | j = 1, 2, \ldots, N\}$ present in the data set. We have also taken the IMDb rating data set, including voting and rating information of all movies present in the data set. The proposed system divides the audience into four groups according to age demography. The proposed system forecast how much preferable the movie would be for each group for an upcoming movie. Which group would prefer the film most and which group would not like the movie. Table 5 presents the viewer's age groups.

TABLE 5. Viewers age group.

Age	Group	Group name
0-17	Gr-1	Junior
18-29	Gr-2	Teenage
30-44	Gr-3	mid-age
45 +	Gr-4	Senior

We have separated each group Gr_j. Moreover, create separate data set using the recommended movie data set and IMDb rating data set. Each data set consists of rating and voting information of all recommended movies of a movie m_i.

$$Gr_j^i = \{R^i_{Gr_j,k}, Vr^i_{Gr_j,k} | k = 1, \ldots, N\} \tag{8}$$

In (8), Gr_j^i represents the group j for a movie m_i, where $R^i_{Gr_j,k}$ describes an average rating of *kth* recommended movie of the movie m_i of group j. $Vr^i_{Gr_j,k}$ represents the voting ratio of *kth* recommended movie of movie m_i of group j. Each group Gr_j^i creates a separate cluster. For a cluster, Gr_j^i describes the preference of the age group J for a given movie m_i. For each cluster, consider all the recommended movies ($k = 1, \ldots, N$) and their average rating and voting details for the age group Gr_j^i. If the overall rating and the voting ratio of a group ($Gr_j^i | j = 1, 2, 3, 4$) are comparatively higher than the other group ($Gr_l^i | l = 1, 2, 3, 4 \,\&\, l \neq j$), then it represents the movie m_i is relatively preferable by the group Gr_j^i then Gr_l^i. Tables 6 and 7 present rating $R^i_{Gr_3}$ and voting $V^i_{Gr_3}$ features values of group-3 (mid-age) for two example movies (Avatar & Spider-Man 3).

TABLE 6. Example of rating details of group-3.

$R^i_{Gr_3} = \{R^i_{Gr_3,k} \mid k = 1,2,..,N\}$	$m_i =$ Avatar	$m_i =$ Spider-Man 3
$R^i_{Gr_3,1}$	7.6	7.2
$R^i_{Gr_3,2}$	8.6	7.2
$R^i_{Gr_3,3}$	8.4	6.2
$R^i_{Gr_3,4}$	4.6	6.4
$R^i_{Gr_3,5}$	4.8	7.7
$R^i_{Gr_3,6}$	5.1	7.5
$R^i_{Gr_3,7}$	5.2	5.4
$R^i_{Gr_3,8}$	6	6.1
$R^i_{Gr_3,9}$	7.1	6
$R^i_{Gr_3,10}$	8.6	7.4

To compare two groups, we need to have an overall preference for each group. The centroid of the cluster measures the performance of a cluster. Each group consists of several data points. Fuzzy c mean is used to determine the cluster centroid of each group.

$$(C\_R^i_{Gr_j}, C\_V^i_{Gr_j})_{j=1}^{4} = FCM(\{R^i_{Gr_j,k}, Vr^i_{Gr_j,k}| \times k = 1, \ldots, N \,\&\, j = 1, \ldots, 4\}) \tag{9}$$

$$C\_centroid^i_{Gr_j} = (C\_R^i_{Gr_j}, C\_Vr^i_{Gr_j}) \tag{10}$$

Cluster centroid is the parameter to measure the audience group's performance or likings to the movie m. Experimentally we have set a global centroid for all movies. The global centroid is unique for each group. The distance of the cluster centroid from the global centroid is measured. Figure 7 represents the block diagram of the Target Audience Prediction Model.

TABLE 7. Example of voting details of group-3.

$V^i_{Gr_3} = \{V^i_{Gr_3,k} \mid k = 1,2,..,N\}$	$m_i =$ Avatar	$m_i =$ Spider-Man 3
$V^i_{Gr_3,1}$	76438	214612
$V^i_{Gr_3,2}$	431187	272023
$V^i_{Gr_3,3}$	275633	86073
$V^i_{Gr_3,4}$	9803	41716
$V^i_{Gr_3,5}$	19141	69057
$V^i_{Gr_3,6}$	24309	79096
$V^i_{Gr_3,7}$	19183	59705
$V^i_{Gr_3,8}$	38607	12752
$V^i_{Gr_3,9}$	55223	38607
$V^i_{Gr_3,10}$	458626	27175

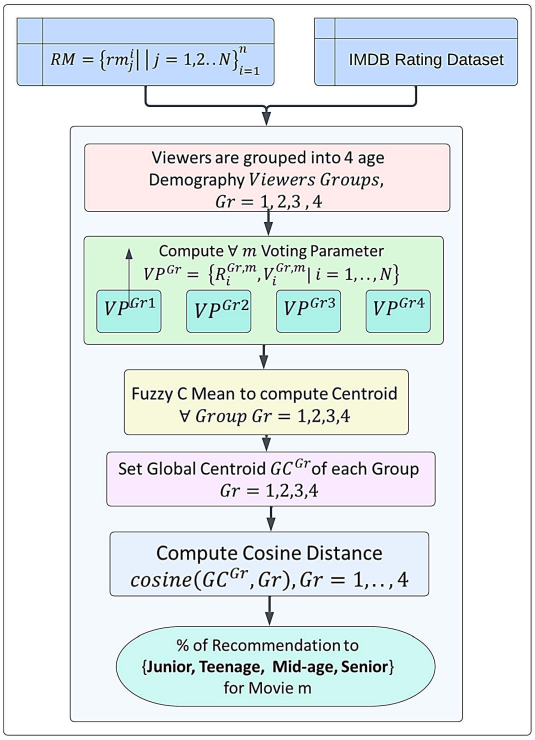

FIGURE 7. Target audience prediction model.

$$G\_centroid = (G\_R, G\_Vr) \tag{11}$$

In the earlier section, we define the popularity of a movie depending on the movie rating. Where we specify a popular movie if the movie rating $mr_i > 7$. In our global centroid parameter, we also set a global rating $G\_R = 7$. We have considered the supported number of voting must be greater than

or equal to the median value. The median value is different for each group. We have considered the voting ratio, and we have fixed the Global voting ratio $G\_Vr = 1$. Next, consolidating the Global parameters, we are set to measure the similarity among the Global centroid and movie centroid for a group. If $C\_R^i_{Gr_j} > G\_R$ we set the value $C\_R^i_{Gr_j} = G\_R$, similarly if $C\_Vr^i_{Gr_j} > G\_Vr$ we set the value of $C\_Vr^i_{Gr_j} = G\_Vr$

$$similarity^i_{Gr_j} = cosine\_sim(G\_centroid, Movie\_centroid^i_{Gr_j}) \tag{12}$$

Distance between the movie centroid $C\_centroid^i_{Gr_j}$ of the group and global centroid, $G\_centroid$ determine the movie's popularity within the group. If the distance is low, then the popularity is high. If the distance is high, the movie's popularity is low within the group. The distance measure is converted to the popularity percentage. If a group centroid has $C\_R^i_{Gr_j} \geq G\_R$ and $C\_Vr^i_{Gr_j} \geq G\_Vr$, then the popularity measure would be 100%. Finally, the module predicts the popularity of an upcoming movie among each age group.

Algorithm 3 Target audience preference prediction using fuzzy c mean.

Input:
Recommended Movie details $RM = \{rm^i_j | j = 1, 2, \ldots, N\}^n_{i=1}$
Movie Rating Database M_r
Output:
Preference percentage to each group $\{Gr_j | j = 1, 2, 3, 4\}$

1. Audience $Au = Gr_1 \cup Gr_2 \cup Gr_3 \cup Gr_4$
2. **for** (*Each movie* $m_i \in M$)
3. **for**(*Each group* $\{Gr_j | j = 1, 2, 3, 4\}$)
4. $VP^i_{Gr_j} = \{R^i_{Gr_j,k}\}^N_{k=1} \cup \{V^i_{Gr_j,k}\}^N_{k=1}$
5. $\{VP_{Gr_j} | j = 1, 2, 3, 4\}$ voting parameters for each group Gr_i
6. $VP^i_{Gr_j}$ for each movie m_i formed a cluster $Gr^i_j | j = 1, 2, 3, 4$
7. **for** (*Each movie* $m_i \in M$)
8. **for**(*Each group*$\{Gr_j | j = 1, 2, 3, 4\}$)
9. Elements in cluster $Gr^i_j = \left\{ R^i_{Gr_j,k}, V^i_{Gr_j,k} \middle/ median(\sum_{k=1}^N V^i_{Gr_j,k}) \,|\, k{=}1, \ldots, N \right\}$
10. $Gr^i_j = \{R^i_{Gr_j,k}, Vr^i_{Gr_j,k} | k = 1, \ldots, N$
11. $(C\_R^i_{Gr_j}, C\_V^i_{Gr_j})^4_{j=1} = FCM(\{R^i_{Gr_j,k}, Vr^i_{Gr_j,k} | k = 1, \ldots, N \& j = 1, \ldots, 4\})$
Centroid of each cluster using Fuzzy c mean
12. $Movie\_centroid^i_{Gr_j} = (C\_R^i_{Gr_j}, C\_Vr^i_{Gr_j})$
13. Set global centroid $= G\_centroid = (G\_R, G\_Vr)$
14. $similarity^i_{Gr_j} = cosine\_sim(G\_centroid, Movie\_centroid^i_{Gr_j})$
15. Calculate percentage similarity $Per\_similarity^i_{Gr_j}$ of each group Gr_j
16. **Return** *Per_similarity*

TABLE 8. Recommended movies of the movie The Terminator.

Recommended Movie	Similarity distance	Genres
Terminator 2: Judgment Day	1.2698	Action, ScienceFiction, Thriller
True Lies	1.3677	Action, Thriller
The Abyss	1.3725	Action, Adventure, ScienceFiction, Thriller
Aliens	1.3978	Action, Horror, ScienceFiction, Thriller
Terminator 3: Rise of the Machines	1.4491	Action, ScienceFiction, Thriller
Terminator Salvation	1.4787	Action, ScienceFiction, Thriller
Surrogates	1.5352	Action, ScienceFiction, Thriller
Terminator Genisys	1.5820	Action, Adventure, ScienceFiction, Thriller
Fortress	1.6107	Action, ScienceFiction, Thriller
Doomsday	1.6377	Action, ScienceFiction, Thriller

V. EXPERIMENTAL RESULTS AND ANALYSIS

In the research problem, three different modules perform an individual responsibility. The first model is a recommendation system module to find similar movies from the data set. The second module uses the first module's input and classifies the upcoming movie into six different classes according to the popularity prediction. Furthermore, that third module is the target audience prediction module. Each module's experimental study is critical—the recommendation system computes similar movies for a given movie from the data set. The movie hit prediction is a multiclass problem. The accuracy of the movie hit prediction module is highly dependent on the efficiency of the first model. It has been imperative to find and recommend the most similar movies to predict the upcoming movie's class. The target audience prediction module categorically estimates the likings of the movie to each age group. We have used the rating and voting date of all the recommended movies.

A. MOVIE RECOMMENDATION

The proposed content-based movie recommendation system uses features like genre, cast, director names, keywords, and movie description to measure the similarity between two movies. The similarity between the two movies is computed after calculating the similarity distance from each parameter. Table 8 shows the recommended film of the movie "The

TABLE 9. Recommended movies of the movie The Avengers.

Recommended Movie	Similarity distance	Genres
Avengers: Age of Ultron	0.5031	Action, Adventure, ScienceFiction
Ant-Man	1.0835	Action, Adventure, ScienceFiction
Captain America: The Winter Soldier	1.0908	Action, Adventure, ScienceFiction
Captain America: Civil War	1.0930	Action, Adventure, ScienceFiction
Iron Man 2	1.1864	Action, Adventure, ScienceFiction
Captain America: The First Avenger	1.2637	Action, Adventure, ScienceFiction
Iron Man	1.2887	Action, Adventure, ScienceFiction
Iron Man 3	1.3764	Action, Adventure, ScienceFiction
X-Men Origins: Wolverine	1.3784	Action, Adventure, ScienceFiction, Thriller
The Incredible Hulk	1.3816	Action, Adventure, ScienceFiction

Terminator'' and Table 9 shows the movie ''the Avenger''. The tables presented the top 10 most similar movies of each selected movie and calculated the overall similarity distance from the selected movie to each recommended movie. Also, we have presented the genre of each of the recommended movies. The recommendation system computed the similarity distance from the selected movie. The similarity distance between two movies defines the similarity between them. As distance decreases, the similarity between two movies increases and vice versa. According to the computation, ''Terminator 2'' is the most similar movie to ''The Terminator''. The measured overall similarity distance is 1.2698, the minimum among all the distances. Similarly, ''Avengers: Age of Ultron'' is the most similar movie ''The Avenger'' compared to other movies. The overall similarity distance measured by the recommendation system is 0.5031. The movie ''The Terminator'' genres are action, thriller, and science fiction. All the selected recommended movies are also having almost the same genres. The movie ''The Avengers'' has genres science fiction, action and adventure. All the selected similar movies are also having almost the same genres. Table 10 shows the recommended movies for the film American beauty. The genre of the movie is drama. Hear all recommended movies also having the same genre. Overall calculated dishes are also shown in the table. According to the system, Revolutionary Road is the most similar movie to the selected movie, and the overall similarity distance is 1.5076.

B. MOVIE HIT PREDICTION

This subsection will discuss the movie hit prediction. The movie hit prediction is a multiclass classification problem.

TABLE 10. Recommended movies of the movie American Beauty.

Recommended Movie	Similarity distance	Genres
Revolutionary Road	1.5076	Drama, Romance
Jarhead	1.6959	Drama, War
Whip It	1.7166	Drama
Philomena	1.7473	Drama
The Brown Bunny	1.7473	Drama
Life as a House	1.7631	Drama
The Color Purple	1.7664	Drama
My Week with Marilyn	1.7751	Drama
The Ice Storm	1.7849	Drama
Draft Day	1.7888	Drama

We classify the problem into six popularity classes {SDH, SH, H, AA, A, F}. The initial recommendation system module provides N similar types of movies of the upcoming movie. The system uses the voting and rating parameters of all the recommended movies. Finally, it predicts the upcoming movie's popularity—the accuracy of the prediction model is principally related to the recommendation system model's accuracy.

The data set has 3310 movie data. We have separated the data set into an 80:20 ratio. 80% means 2684 movie data for training purposes and the remaining 626 data for testing purposes. We have this newly created novel data set and experimented with different machine learning models for better accuracy. Table 11 compare different machine learning model using several evolution parameters. The proposed CNN model performs significantly better than all machine learning models, comparing all the parameters. Fig. 8 shows the accuracy of the machine learning model.

TABLE 11. Compare different machine learning model.

Machine Learning Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.811	0.81	0.81	0.79
KNN	0.592	0.60	0.59	0.59
Random Forest	0.946	0.95	0.95	0.94
CNN	0.968	0.96	0.96	0.96

The experimental analysis of the data set indicates that the proposed convolutional neural network model's overall performance outperforms all the baseline conventional machine learning models. Figure 9 and 10 shows the proposed CNN model accuracy curve and loss curve respectively.

The proposed movie hit prediction module also outperforms the predicted system presented in the past. We have

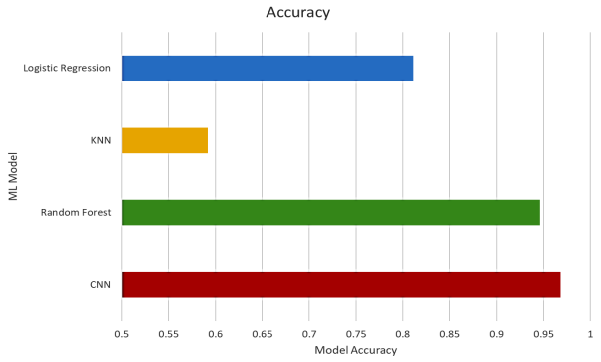

FIGURE 8. Compare different machine learning model accuracy.

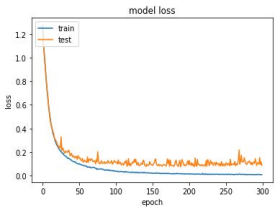

FIGURE 9. CNN model accuracy curve.

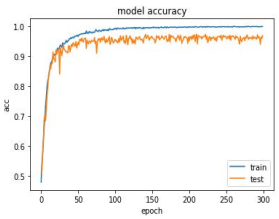

FIGURE 10. CNN model loss curve.

presented a comparative study of our work with some relevant research work done in the past. Ahmed *et al.* (2019) [22] used a hybrid voting system to obtain the early production stage forecast. In this study, they introduced new features to improve prediction efficiency. They classified the movie into eight different success classes and gained 85% accuracy. Abidi *et al.* (2020) [27] examined each movie's attributes and selected all the features relevant to movie prediction's early stage. They have Executed five various machine learning models with binary classification and gained a height of 76.6% accuracy with the Generalized linear model (GLM). Michael and Kang (2016) [10] introduced novel features and predicted movie success at an early movie production stage. They evaluated Machine learning models multiple times with distinct success measures. Michael and Kang (2016) achieved a maximum of 90.4% accuracy with the Random Forest model using binary classification. For multiclass classification, they have achieved 84.7% accuracy. Verma and Garima (2019) [24] proposed "music rating" as one of the significant features to forecast movie hits. They attained 87.0% accuracy with the Random Forest model using binary classification. Our proposed model outperforms all the previous models and achieved 96.8% accuracy. Table 12 presents Comparative Analysis.

TABLE 12. Comparative analysis of our proposed work with related works.

Author and Year	Proposed Work	Methodologies/ Parameters	Classification Type	Accuracy
Ahmed et al. 2019	Pre-production box-office success prediction	Incorporate new features and Hybrid voting classifier	Multiclass	0.85
Abidi et al. 2020	Movie's popularity prediction	Incorporate novel features and generalized linear model (GLM)	Binary class	0.766
Michael and Kang 2016	Movie profitability using what, who and when features.	Incorporate novel features and using RF machine learning model	Binary class	0.904
Michael and Kang 2016	Movie profitability using what, who and when features.	Incorporate novel features and using RF machine learning model	Multiclass	0.847
Verma and Garima 2019	Bollywood movie success prediction	Incorporate music rating and use RF machine learning model	Binary class	0.87
Our proposed system	Predicting movie success	Using content-based RS and using CNN deep learning model	Multiclass	0.968

C. PREDICT PREFERRED AUDIENCE GROUP

This subsection will discuss the result and analysis of the final module of the movie target audience prediction. Audiences

TABLE 13. Animation movie's target audience.

Movie title	Genres	Popularity			
		Junior	Teenage	Mid-age	Senior
Cars	Animation, Adventure, Comedy, Family	99.71	95.34	89.76	88.57
The Good Dinosaur	Adventure, Animation, Family	95.91	87.24	78.92	75.91
WALLÂ·E	Animation, Family	96.24	89.13	81.19	78.62
Kung Fu Panda	Adventure, Animation, Family, Comedy	98.83	89.7	81.4	78.49
Ice Age: Dawn of the Dinosaurs	Animation, Comedy, Family, Adventure	91.92	83.1	78.03	77.15

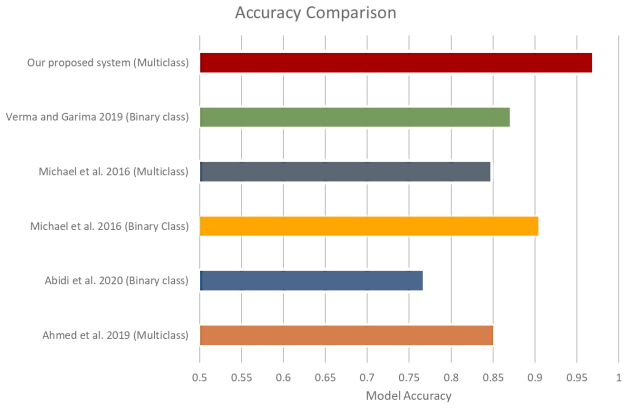

FIGURE 11. Comparative analysis of our proposed work with related works.

are grouped into four age groups {junior, teenage, mid-age, senior}. To predict the upcoming movie's target audience, we have considered all the recommended movies delivered by our first module. Then, the system uses each recommended movie's rating by each age group and takes each group's number of vote details for each recommended movie—system analysis of all the voting and rating information from each group of all the recommended movies. Ultimately, the module would predict the popularity of the upcoming movie for each age group.

Some movies are created targeting only the junior age group. Generally, animation movies are specially targeted to the junior age group. Our proposed system focused on how popular the upcoming movie would be among all age groups. The Movie Maker can estimate the popularity of the upcoming animation movie among the junior age group. Usually,

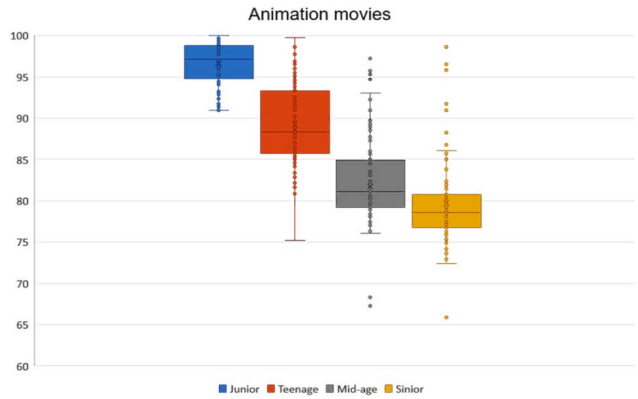

FIGURE 12. Popularity box plot of animation movies.

TABLE 14. Comedy movie's target audience.

Movie title	Genres	Popularity			
		Junior	Teenage	Mid-age	Senior
Eulogy	Comedy, Drama	54.8	69.53	70.95	73.3
Death at a Funeral	Comedy, Drama	52.3	67.8	70.68	73.22
And So It Goes	Comedy, Drama, Romance	63.45	83.27	85.33	88.06

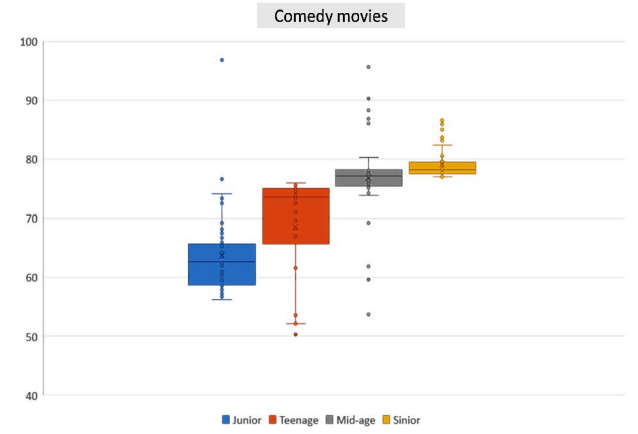

FIGURE 13. Popularity box plot of comedy movies exclude science fiction, adventure and animation movies.

hit movies always make a good impact among all age groups. Table 13 shows the popularity of the five leading animation movies among all age groups—junior groups like animated movies most among all the age groups. Figure 12 shows the popularity of animation movies within the junior group, and it also reveals the stepwise decrement in popularity as the age increased.

TABLE 15. Science fiction movie's target audience.

Movie title	Genres	Popularity			
		Junior	Teenage	Mid-age	Senior
Ant-Man	Science Fiction, Action, Adventure	91.75	99.89	94.27	92.83
Star Wars: Episode III - Revenge of the Sith	Science Fiction, Adventure, Action	94.37	99.88	95.25	95.62
The Avengers	Science Fiction, Action, Adventure	91.4	98.91	92.83	89.98
Star Trek IV: The Voyage Home	Science Fiction, Adventure	89.44	97.75	85.68	84.34
Iron Man	Action, Science Fiction, Adventure	97.3	99.8	90.7	86.73

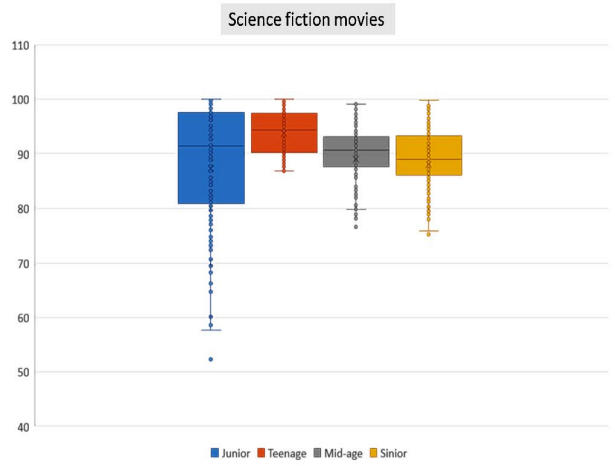

FIGURE 14. Popularity box plot of science fiction movies.

Higher age groups commonly favour movies like comedy, drama, and romance. Moviemakers usually target mid-age and senior groups for comedy and drama genres for movies. Table 14 shows the popularity of the five movies from comedy and drama genres among all age groups—senior and mid-age groups like such movies most among all the age groups. Figure 13 shows the popularity of comedy movies, excluding science fiction, adventure, and Animation movies. It shows popularity among the senior group is maximum, next mid-age group. Popularity among junior is the least.

Science fiction movies are usually preferred by all age groups, depending upon the storyline of the movie—particularly teenage group like the science fiction movies most. Table 15 shows the popularity of the five leading science fiction movies among all age groups—the teenage group likes the most among all the age groups. All five movies are hit movies; hence they are popular among all age groups. Figure 14 shows the popularity of science fiction movies among all groups. It shows that popularity among the teenage group is maximum. Popularity within the junior group varies to a large extent. The average popularity of science fiction movies is usually high.

VI. CONCLUSION

A substantial amount of financing is consumed in every box-office movie. However, most movies fail to achieve success. Earlier, the most significant number of works have been done on post-production or post-release forecast. The estimate does not influence as the investor has already consumed their funds on the film production. The pre-production or early production stage forecast needs high accuracy and the best time to ensure investment. The objective of our study is to propose an expert system that could help the movie maker execute necessary changes if needed at the appropriate time. Our system can food cost the level of popularity of the upcoming movie before the production has started for the earliest stage of the production and with significant accuracy. About system focused not only on the popularity of the upcoming movie but also on the movie's popularity among all age groups. Movie Maker can estimate the target audience and assess how the different audience groups would respond to the upcoming movie. Further, our target is to build a robust system applicable to all movie industries. We have used the last hundred years (1915-2016) of movie data from TMDb and the IMDb database. Our approach to focused movie popularity and finding out the target audience of an upcoming movie is very much unique. In our approach, we have used a recommendation system to find similar movies from a given movie and use similar movies for forecasting purposes. Moreover, it has been challenging to simultaneously use to separate the database (TMDb and IMDb). The size of the TMDb data set is 4803, and the size of the IMDb rating data set is 85855. Since we are using both the dataset, the size of the merged data set comes down to 23332 only. We need to judge each of the features that can be available at the beginning of movie production. We have carefully picked only five movie attributes for our recommendation system. We have used total votes for each movie. It has been observed that the number of votes for the old movie is relatively less than the new movie. Moreover, several voting information for the junior group is significantly less relative to the other groups. Using the voting and rating information to create a new feature set is challenging to work in this scenario. The proposed system is an excellent tool for the movie industry. In future work, multimedia data like audio and video data could be incorporated and also, the poster of the upcoming movie could be used for better results. Recent train tickets could be analyzed using sentiment analysis of the social media data. Information regarding recent trained on the market expectation from the movie industry will be

beneficial for the movie makers. The audience group could be divided according to age and according to the demography or profession of the audience. That will be much easier for targeting and promoting an upcoming movie.

Compliance with Ethical Standards: None

ACKNOWLEDGMENT

The work of Jana Shafi was supported by the Deanship of Scientific Research, Prince Sattam Bin Abdulaziz University.

REFERENCES

[1] L. Sharma and A. Gera, "A survey of recommendation system: Research challenges," *Int. J. Eng. Trends Technol.*, vol. 4, no. 5, pp. 1989–1992, 2013.
[2] N. Das, S. Borra, N. Dey, and S. Borah, "Social networking in web based movie recommendation system," in *Social Networks Science: Design, Implementation, Security, and Challenges*. Cham, Switzerland: Springer, 2018, pp. 25–45.
[3] P. Nagarnaik and A. Thomas, "Survey on recommendation system methods," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 1603–1608.
[4] M. A. Hameed, O. Al Jadaan, and S. Ramachandram, "Collaborative filtering based recommendation system: A survey," *Int. J. Comput. Sci. Eng.*, vol. 4, no. 5, p. 859, 2012.
[5] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web (WWW)*, 2001, pp. 285–295.
[6] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2015, pp. 77–118.
[7] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proc. ACM Conf. Comput. supported Cooperat. Work (CSCW)*, 2000, pp. 241–250.
[8] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artif. Intell. Rev.*, vol. 13, no. 5, pp. 393–408, 1999.
[9] R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation," in *Proc. Mach. Learn. Inf. Age, MLnet/ECML2000 Workshop*, vol. 30, 2000, pp. 47–56.
[10] P. B. Thorat, R. M. Goudar, and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *Int. J. Comput. Appl.*, vol. 110, no. 4, pp. 31–36, Jan. 2015.
[11] P. Lops, M. Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2011, pp. 73–105.
[12] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web*. Berlin, Germany: Springer, 2007, pp. 325–341.
[13] D. Wang, Y. Liang, D. Xu, X. Feng, and R. Guan, "A content-based recommender system for computer science publications," *Knowl.-Based Syst.*, vol. 157, pp. 1–9, Oct. 2018.
[14] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.
[15] M. A. Ghazanfar and A. Prugel-Bennett, "A scalable, accurate hybrid recommender system," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining*, Jan. 2010, pp. 94–98.
[16] T. K. Paradarami, N. D. Bastian, and J. L. Wightman, "A hybrid recommender system using artificial neural networks," *Expert Syst. Appl.*, vol. 83, pp. 300–313, Oct. 2017.
[17] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," in *Advances in Artificial Intelligence*. London, U.K., 2009.
[18] P. Melville and V. Sindhwani, "Recommender systems," *Encyclopedia Mach. Learn.*, vol. 1, pp. 829–838, Apr. 2010.
[19] Y. Yao and F. M. Harper, "Judging similarity: A user-centric study of related item recommendations," in *Proc. 12th ACM Conf. Recommender Syst.*, Sep. 2018, pp. 288–296.
[20] F. Gedikli, D. Jannach, and M. Ge, "How should i explain? A comparison of different explanation types for recommender systems," *Int. J. Hum.-Comput. Stud.*, vol. 72, no. 4, pp. 367–382, Apr. 2014.
[21] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" *Commun. ACM*, vol. 35, no. 12, pp. 29–38, Dec. 1992.
[22] T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker, "Linked open data to support content-based recommender systems," in *Proc. 8th Int. Conf. Semantic Syst. (SEMANTICS)*, 2012, pp. 1–8.
[23] C. Musto, P. Lops, M. de Gemmis, and G. Semeraro, "Semantics-aware recommender systems exploiting linked open data and graph-based features," *Knowl.-Based Syst.*, vol. 136, pp. 1–14, Nov. 2017.
[24] M. Uluyagmur, Z. Cataltepe, and E. Tayfur, "Content-based movie recommendation using different feature sets," in *Proc. World Congr. Eng. Comput. Sci.*, vol. 1, 2012, pp. 17–24.
[25] S. R. S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, "Content-based movie recommendation system using genre correlation," in *Smart Intelligent Computing and Applications*. Singapore: Springer, 2019, pp. 391–397.
[26] J. Son and S. B. Kim, "Content-based filtering for recommendation systems using multiattribute networks," *Expert Syst. Appl.*, vol. 89, pp. 404–412, Dec. 2017.
[27] S. M. Ali, G. K. Nayak, R. K. Lenka, and R. K. Barik, "Movie recommendation system using genome tags and content-based filtering," in *Advances in Data and Information Sciences*. Singapore: Springer, 2018, pp. 85–94.
[28] M. Elahi, Y. Deldjoo, F. Bakhshandegan Moghaddam, L. Cella, S. Cereda, and P. Cremonesi, "Exploring the semantic gap for movie recommendations," in *Proc. 11th ACM Conf. Recommender Syst.*, Aug. 2017, pp. 326–330.
[29] Y. Deldjoo, M. F. Dacrema, M. G. Constantin, H. Eghbal-zadeh, S. Cereda, M. Schedl, B. Ionescu, and P. Cremonesi, "Movie genome: Alleviating new item cold start in movie recommendation," *User Model. User-Adapted Interact.*, vol. 29, no. 2, pp. 291–343, Apr. 2019.
[30] S. Kumar, K. De, and P. P. Roy, "Movie recommendation system using sentiment analysis from microblogging data," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 4, pp. 915–923, Aug. 2020.
[31] S. Philip, P. B. Shola, and A. Ovye, "Application of content-based approach in research paper recommendation system for a digital library," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 10, pp. 1–4, 2014.
[32] J. Eliashberg, J.-J. Jonker, M. S. Sawhney, and B. Wierenga, "MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures," *Marketing Sci.*, vol. 19, no. 3, pp. 226–243, Aug. 2000.
[33] P. Boccardelli, F. Brunetta, and F. Vicentini, "What is critical to success in the movie industry? A study on key success factors in the Italian motion picture industry," Tech. Rep., 2008. [Online]. Available: https://publires.unicatt.it/en/publications/what-is-critical-to-success-in-the-movie-industry-a-study-on-key–10
[34] W. Zhang and S. Skiena, "Improving movie gross prediction through news analysis," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, Sep. 2009, pp. 301–304.
[35] W. Bai, Y. Zhang, W. Huang, Y. Zhou, D. Wu, G. Liu, and L. Xiao, "DeepFusion: Predicting movie popularity via cross-platform feature fusion," *Multimedia Tools Appl.*, vol. 79, nos. 27–28, pp. 19289–19306, Jul. 2020.
[36] M. Galvão and R. Henriques, "Forecasting movie box office profitability," *J. Inf. Syst. Eng. Manage.*, vol. 3, no. 3, pp. 1–9, Jul. 2018.
[37] Y. Zhou, L. Zhang, and Z. Yi, "Predicting movie box-office revenues using deep neural networks," *Neural Comput. Appl.*, vol. 31, no. 6, pp. 1855–1865, Jun. 2019.
[38] K. Lee, J. Park, I. Kim, and Y. Choi, "Predicting movie success with machine learning techniques: Ways to improve accuracy," *Inf. Syst. Frontiers*, vol. 20, no. 3, pp. 577–588, Jun. 2018.
[39] M. T. Lash and K. Zhao, "Early predictions of movie success: The who, what, and when of profitability," *J. Manage. Inf. Syst.*, vol. 33, no. 3, pp. 874–903, Jul. 2016.
[40] U. Ahmed, H. Waqas, and M. T. Afzal, "Pre-production box-office success quotient forecasting," *Soft Comput.*, vol. 24, no. 9, pp. 6635–6653, May 2020.
[41] S. M. R. Abidi, Y. Xu, J. Ni, X. Wang, and W. Zhang, "Popularity prediction of movies: From statistical modeling to machine learning techniques," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 35583–35617, Dec. 2020.
[42] H. Verma and G. Verma, "Prediction model for bollywood movie success: A comparative analysis of performance of supervised machine learning algorithms," *Rev. Socionetwork Strategies*, vol. 14, no. 1, pp. 1–17, Apr. 2020.
[43] M. Mestyán, T. Yasseri, and J. Kertész, "Early prediction of movie box office success based on Wikipedia activity big data," *PLoS ONE*, vol. 8, no. 8, Aug. 2013, Art. no. e71226.

[44] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio, "Prediction of movies box office performance using social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2013, pp. 1209–1214.
[45] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Aug. 2010, pp. 492–499.
[46] S. Gopinath, P. K. Chintagunta, and S. Venkataraman, "Blogs, advertising, and local-market movie box office performance," *Manage. Sci.*, vol. 59, no. 12, pp. 2635–2654, Dec. 2013.
[47] R. Parimi and D. Caragea, "Pre-release box-office success prediction for motion pictures," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.* Berlin, Germany: Springer, 2013, pp. 571–585.
[48] J. S. Simonoff and I. R. Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," *Chance*, vol. 13, no. 3, pp. 15–24, Jun. 2000.
[49] M. Baimbridge, "Movie admissions and rental income: The case of James Bond," *Appl. Econ. Lett.*, vol. 4, no. 1, pp. 57–61, Jan. 1997.
[50] A. De Vany and W. D. Walls, "Uncertainty in the movie industry: Does star power reduce the terror of the box office?" *J. Cultural Econ.*, vol. 23, no. 4, pp. 285–318, 1999.
[51] A. Elberse, "The power of stars: Do star actors drive the success of movies?" *J. Marketing*, vol. 71, no. 4, pp. 102–120, Oct. 2007.
[52] J. Eliashberg, S. K. Hui, and Z. J. Zhang, "From story line to box office: A new approach for green-lighting movie scripts," *Manage. Sci.*, vol. 53, no. 6, pp. 881–893, Jun. 2007.
[53] W. D. Walls, "Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns," *J. Cultural Econ.*, vol. 29, no. 3, pp. 177–190, Aug. 2005.
[54] A. Zaheer and G. Soda, "Network evolution: The origins of structural holes," *Administ. Sci. Quart.*, vol. 54, no. 1, pp. 1–31, Mar. 2009.
[55] J. Du, H. Xu, and X. Huang, "Box office prediction based on microblog," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1680–1689, Mar. 2014.
[56] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Syst. Appl.*, vol. 30, no. 2, pp. 243–254, Feb. 2006.
[57] L. Zhang, J. Luo, and S. Yang, "Forecasting box office revenue of movies with BP neural network," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6580–6587, Apr. 2009.
[58] V. R. Nithin, "Predicting movie success based on imdb data," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. V, no. X, pp. 504–507, Oct. 2017.
[59] *Box Office Revenue in the United States and Canada From 1980 to 2021*. Accessed: May 5, 2021. [Online]. Available: https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since1980/
[60] B. Meiseberg and T. Ehrmann, "Diversity in teams and the success of cultural products," *J. Cultural Econ.*, vol. 37, no. 1, pp. 61–86, Feb. 2013.
[61] N. Kurniasih, E. Rizal, Y. Winoto, Sukaesih, N. Kurniawati, Sujito, A. Sudirman, A. Hasibuan, A. Daengs GS, and K. Saddhono, "Online media as a movie reference," *J. Phys., Conf.*, vol. 1114, Nov. 2018, Art. no. 012087.
[62] *Kaggle TMDB Movie Dataset*. Accessed: May 5, 2021. [Online]. Available: https://www.kaggle.com/tmdb/tmdb-movie-metadata
[63] *Kaggle IMDB Movie Dataset*. Accessed: May 5, 2021. [Online]. Available: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+ratings.csv
[64] M. A. Belarbi, S. Mahmoudi, and G. Belalem, "PCA as dimensionality reduction for large-scale image retrieval systems," *Int. J. Ambient Comput. Intell.*, vol. 8, no. 4, pp. 45–58, Oct. 2017.
[65] S. K. Das, S. P. Das, N. Dey, and A. E. Hassanien. *Machine Learning Algorithms for Industrial Applications*. Amsterdam, The Netherlands: Elsevier, 2020.
[66] S. Sivakumar and R. Rajalakshmi, "Analysis of sentiment on movie reviews using word embedding self-attentive LSTM," *Int. J. Ambient Comput. Intell.*, vol. 12, no. 2, pp. 33–52, Apr. 2021.
[67] W. Chen, H.-T. Zheng, Y. Wang, W. Wang, and R. Zhang, "Utilizing generative adversarial networks for recommendation based on ratings and reviews," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[68] Y. Wang, H.-T. Zheng, W. Chen, and R. Zhang, "LambdaGAN: Generative adversarial nets for recommendation task with lambda strategy," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

SANDIPAN SAHU received the M.Tech. degree from WBUT, West Bengal. He is currently pursuing the Ph.D. degree with the Department of Computer Science, GIET University, India. His research interests include in the areas of data mining, machine learning, and predictive analytics. Specific interests and ongoing research areas include utility-based data mining, adversarial learning, and survival analytics and learning.

RAGHVENDRA KUMAR received the B.Tech., M.Tech., and Ph.D. degrees in computer science and engineering in India. He is working as an Associate Professor with the Computer Science and Engineering Department, GIET University, India. He was a Postdoctoral Fellow with the Institute of Information Technology, Virtual Reality and Multimedia, Vietnam. He has published 13 chapters in edited book published by IGI Global, Springer, and Elsevier. He has authored and edited 23 computer science books in field of the Internet of Things, data mining, biomedical engineering, big data, and robotics; and in IGI Global Publication, USA; IOS Press, The Netherland; Springer; Elsevier; and CRC Press, USA. His research interests include computer networks, data mining, cloud computing and secure multiparty computations, theory of computer science, and design of algorithms. He has published number of research papers in international journal (SCI/SCIE/ESCI/Scopus) and conferences, including IEEE and Springer as well as served as the Organizing Chair for RICE-2019 and 2020; a Volume Editor for RICE-2018; and a keynote speaker, the session chair, the co-chair, the publicity chair, the publication chair, an Advisory Board Member, and a Technical Program Committee Member for many international and national conferences. He has served as a guest editor in many special issues from reputed journals (indexed by: Scopus, ESCI, and SCI). He serves as a Series Editor for *Internet of Everything (IOE): Security and Privacy Paradigm*, *Green Engineering and Technology: Concepts and Applications* (CRC Press, Taylor & Francis Group, USA), and *Biomedical Engineering: Techniques and Applications* (Apple Academic Press, CRC Press, Taylor & Francis Group, USA). He also serves as an Acquisition Editor for *Computer Science* (Apple Academic Press, CRC Press, Taylor & Francis Group).

MOHD SHAFI PATHAN worked as a Lecturer with the MIT Engineering College, Aurangabad, from July 1999 to July 2006, and as a Lecturer and an Associate Professor with the Smt. Kashibai Navale College of Engineering, Pune, from July 2006 to January 2017. He is currently working as a Professor with the Department of Computer Science and Engineering, MITSOE, MIT ADT University, Loni Kalbhor, Pune. He completed the university funded research project on "Public key cryptography for cross-realm authentication in Kerberos" costing two lakh rupees within a duration of two years.

JANA SHAFI is affiliated with the Department of Computer Science, Prince Sattam Bin Abdulaziz University, Saudi Arabia. She has more than eight years of teaching and research experience. She has published in numerous journals, such as *Sensors*, IEEE Access, *Diagnostics*, *Symmetry*, *Mathematics and Wireless Communications*, and *Mobile Computing*. Her research interests include online social networks, wearable technology, artificial intelligence, machine learning, deep learning, smart health, and the IoMT.

YOGESH KUMAR received the Ph.D. degree in CSE from Punjabi University, Patiala. He is currently working as an Associate Professor-CSE with Indus University, Ahmedabad. He has a total of 15 years of experience (including teaching and research) and a post Ph.D. experience of two years and 11 months. He has published 66 research articles, including 14 articles in high-impact SCI journals, 31 articles in Scopus and peer-reviewed journals, 14 papers in international conferences in India and abroad, and eight high-impact book chapters. He has also published two books and granted seven patents. His research interests include the artificial intelligence, deep learning, speech recognition systems, and data science.

MUHAMMAD FAZAL IJAZ (Member, IEEE) received the B.Eng. degree in industrial engineering and management from the University of the Punjab, Lahore, Pakistan, in 2011, and the Dr.Eng. degree in industrial and systems engineering from Dongguk University, Seoul, South Korea, in 2019. From 2019 to 2020, he worked as an Assistant Professor with the Department of Industrial and Systems Engineering, Dongguk University. Currently, he is working as an Assistant Professor with the Department of Intelligent Mechatronics Engineering, Sejong University, Seoul. He has published numerous research articles in several international peer-reviewed journals, including *Scientific Reports*, *Journal of Ambient Intelligence and Humanized Computing*, IEEE Access, *Sensors*, *Remote Sensing*, *Diagnostics*, *Journal of Food Engineering*, *Applied Sciences*, *Asia Pacific Journal of Marketing and Logistics*, and *Sustainability*. His research interests include digital health, machine learning, blockchain, healthcare engineering, the Internet of Things, and big data.