Indonesian Journal of Electrical Engineering and Computer Science
Vol. 29, No. 3, March 2023, pp. 1750~1757
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v29.i3.pp1750-1757

Prediction of movie success based on machine learning and twitter sentiment analysis using internet movie database data

Jyoti Tripathi, Sunita Tiwari, Anu Saini, Sunita Kumari
Department of Computer Science and Engineering, G.B. Pant DSEU Okhla I Campus, New Delhi, India

Article Info

Article history:

Received Jun 25, 2022
Revised Oct 27, 2022
Accepted Oct 30, 2022

Keywords:

Decision tree
Entertainment industry
Naïve Bayes
Random forest
Regression
Supervised learning
Support vector machines

ABSTRACT

Nowadays, predicting the success of a new movie is a crucial task. In this work, the hybrid approach considers the movie features as well as sentiment expressed in the movie review to predict the success rate of a movie. Multiple movie features such as title, director, star cast, and writer. Are considered for prediction. The related raw data is collected from the internet movie database (IMDb) website and after pre-processing, the collected data is used to generate the supervised machine learning model. Different supervised learning models are compared and the one with the best results is used further. The mean squared error, root mean squared error and r2 score of the models generated are comparable with existing models. Further, sentiment analysis of the movie-related tweets is performed. The accuracy of best sentiment analysis model is 88.47%. Finally, the two models are combined to give the success prediction rating of new movies and the results of the hybrid model are encouraging. The proposed model may be used to find the top-rated movies of a particular calendar year.

Corresponding Author:

Anu Saini
Department of Computer Science and Engineering, G.B. Pant DSEU Okhla I Campus
New Delhi-110020, India
Email: anuanu16@gmail.com

1. INTRODUCTION

Movies, online videos and television are most popular source of entertainment across the globe especially in India [1]. Movie industry involve huge sum of investment in terms of money, time and effort [2]. Movie industry is producing hundreds of movies every year. Therefore, it is crucial to predict success of a movie in early stages. Success or failure of a movie is based on multiple factors. A huge amount of information related to movies such as actors, directors, critic review, user reviews, ratings, writer, budget, genre, Facebook likes, number of views on YouTube for movie trailer, and fan following on twitter. are available on web.

Success of movie in this era depends on the revenue generated in first few weeks [3]. The revenue generation in initial weeks is greatly influenced by online reviews and ratings of the movie. Since first few weeks are very crucial for the success of movie, and the movie production people put in lot of efforts on the publicity and building people's opinion. In this work, we aim to use the available information to predict the success rate of a movie in early stages. The internet movie database (IMDb) is a rich source of information which contains the data about almost all the movies.

To predict the success of a movie, the supervise machine learning algorithms are used. Different machine learning algorithms are used to build the prediction model and the results obtained from each mode are compared over root mean squared error (RMSE), mean square error (MAE) and R2 score. Further, social media such as twitter, Instagram and Facebook has become a great source of influence on people's opinion. A huge amount of data is generated through such sources and they are important means of gathering the movie

reviews. Opinions about movies are mostly expressed in form of reviews and comments. However, this huge amount of data presents a challenge of information overload. Therefore, the need to automate the process of extracting the sentiments from the information available has been posed. In this work the sentiment analysis model is developed using machine learning and natural language processing to analyse the reviews/comments and predict the sentiment of a tweet and reviews. The overall sentiments of a reviews/tweets are classified in one of the two classes positive and negative. Further, different machine learning algorithms have been applied for sentiment analysis and the results are compared to find the best suited model for the movie success rate prediction. Once the above two model (prediction and sentiment analysis models) are built, the hybrid success rating is predicted using these two models together.

A lot of work related to movie rating prediction and sentiment analysis of reviews can be found in literature, specifically in movie domain. Most of the sentiment analysis and rating prediction work use machine learning and natural language processing to some extents. The work focusing on predicting the polarity of movie reviews includes the following. Turney [4] used unsupervised learning to classify the movie based on their average semantic orientation. Pang *et al.* [5] Compared the performance of machine learning algorithms on movie reviews and according to their analysis support vector machines (SVM) classifier has shown the best performance. Mishne and Glance [6] used blogger sentiments to predict movie sales. Used combination of machine learning and simple technique based on counting of positive and negative words in review [7]. Proposed movie gross prediction trough new analysis [8]. Claimed that the performance of baseline machine learning algorithm used for text classification varies greatly based on the selection of model variant and features used [9], [10]. García-Cumbreras *et al.* [10] proposed an approach to improve collaborative filtering using sentiment analysis. Combined generative and discriminative model together for sentiment prediction [11]. Presented an improved sentiment analysis of online movie reviews based on clustering [12]. Proposed hybrid convolutional neural network (CNN)-long short-term memory (LSTM) based approach for sentiment analysis [13], [14]. Proposed an attention-based long short-term memory, 'Senti_ALSTM,' model for sentiment analysis [15]. Has proposed the predictive model for Bollywood movies [16]. Proposed lexicon and neural network based approach for sentiment analysis [17]. Proposed movie success rate prediction model using random forest method [18]. Proposed the decision tree based approach for rating prediction [19]. Proposed LSTM based sentiment analysis for movie review [20]. Authors proposed news based recommendation system using multinomial Naïve Bayes (MNB) classifier [21]. In contrast to the existing work, the aim of proposed work utilizes a hybrid model which combines the results of prediction model and twitter sentiment analysis model to predict the success of movie. The rest of the paper is organized as follows. Section 2 discusses the methods. The results of proposed work are deliberated in section 3. Finally, section 4 concludes our work.

2. METHOD

The entire work is divided in two components namely rating prediction model and sentiment analysis model. Figure 1 shows the system design. Initially the supervised machine learning model is generated for rating prediction based on multiple features described in dataset section. The model is evaluated over RMSE, MAE, and R2 score. The prediction model contains multiple independent features (genre, runtime, budget, crew's popularity, and aspect ratio) and one output variable i.e., rating. Once this model is generated the rating of a new movies may be predicted based on the input features of new movie.

Figure 1. Overall system design

The dataset of the text reviews is used to generate the supervised machine learning model for sentiment analysis [22]. The sentiment analysis model can be used to identify the positive and negative sentiments of the movie reviews and a sentiment score is computed. Further, these two models are used to predict rating and sentiment score of the new movie concurrently. Weighted sum of the two generated values is computed. The experiments are conducted for different weight combinations and the results shown that when the 0.6 weight is assigned to rating prediction and 0.4 weight is assigned to the sentiment score the results obtained are optimal. This hybrid rating is the final output (predicted movie success rating) of the system. The forthcoming section describes the system components in detail. For rating prediction and sentiment analysis both the supervised learning is used. Figure 1 shows the overall supervised learning architecture used for both rating prediction and sentiment analysis. Both of these components are discussed in detail in subsequent sections.

2.1. Rating prediction

Each stage shown in Figure 2 in context of rating prediction model is deliberated as follows:

- Dataset: Part of IMDb data is available for non-commercial use. The dataset is downloaded from official IMDb website [23]. The downloaded dataset contains more than one million entries. The 'title.akas.tsv' and 'title.basics.tsv' contains all the information about movie titles. The 'title.crew.tsv' and 'title.principals.tsv' contains all the information about people (directors, actors, and actresses) involved. All the data in these files are mostly nominal in nature. The 'titles.ratings.tsv' has the most important information that is rating and number of votes the title has received.
- Pre-processing and transformation: The prediction of rating is based on the movie features such as genre, runtime, budget, crew's popularity, and aspect ratio. The data in IMDb is in raw form and it needs further pre-processing. Data were spread across multiple .tsv files and therefore this data is converted to csv format and merged into a single file with python script. For experimental purpose, only movie released in English language are filtered out and used for model generation and evaluation. The data was thoroughly analyzed to check for any illegal characters. Missing values are handled by replacing them by mean values in case of numerical data and replacing by most frequent values in case of categorical data. Since models can handle numerical data types therefore the categorial and ordinal features are converted to numerical features using data transformation methods. Figure 3 shows different distribution figures such as movie runtime, rating, budget, and vote count distributions.
- Splitting training and testing data: The processed data is now divided into two subsets, training and testing dataset. Training dataset is used to fit the model and testing dataset is used to evaluate the prediction accuracy. The data is divided into a ratio of 80:20 for training and testing. The data division is done in this ratio intuitively based on Pareto Principle (80/20 rule). Model training: The aim of training is to fit the model using the training data. The data contains multiple features and one target variable (rating). Features and target variables are separated and the various supervised learning models are applied. The model is trained using random forest, simple regression tree and linear regression algorithms.
- Simple regression tree is a non parametric supervised learning which predicts target by applying simple decision rules. It partitions the dataset into subsets and fits the simple model for each subset. However, a single tree model is mostly unstable. Random forest is a supervised learning algorithm and estimator fits number of decision tress on various sub sets of datasets. To improve prediction accuracy and control overfitting, random forest use averaging. Linear regression is also a supervised learning algorithm which predicts the dependent variable value based on the given dependent features. The equation for regression used is shown in (1).

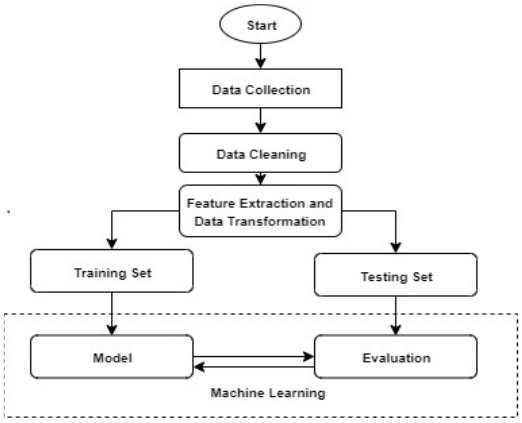

Figure 2. Machine learning model

$$y\_i = \beta\_0 + \beta\_1\, x\_i1 + \beta\_2\, x\_i2 + \cdots .. + \beta\_p\, x\_ip + \epsilon \quad (1)$$

Where, for i=n observations, y_i is dependent variable, x_i is the explanatory variable, β_0 is y-intercept (constant), β_p are the slop coefficients for each explanatory variables and ϵ the error term for model (it is also known as residuals).

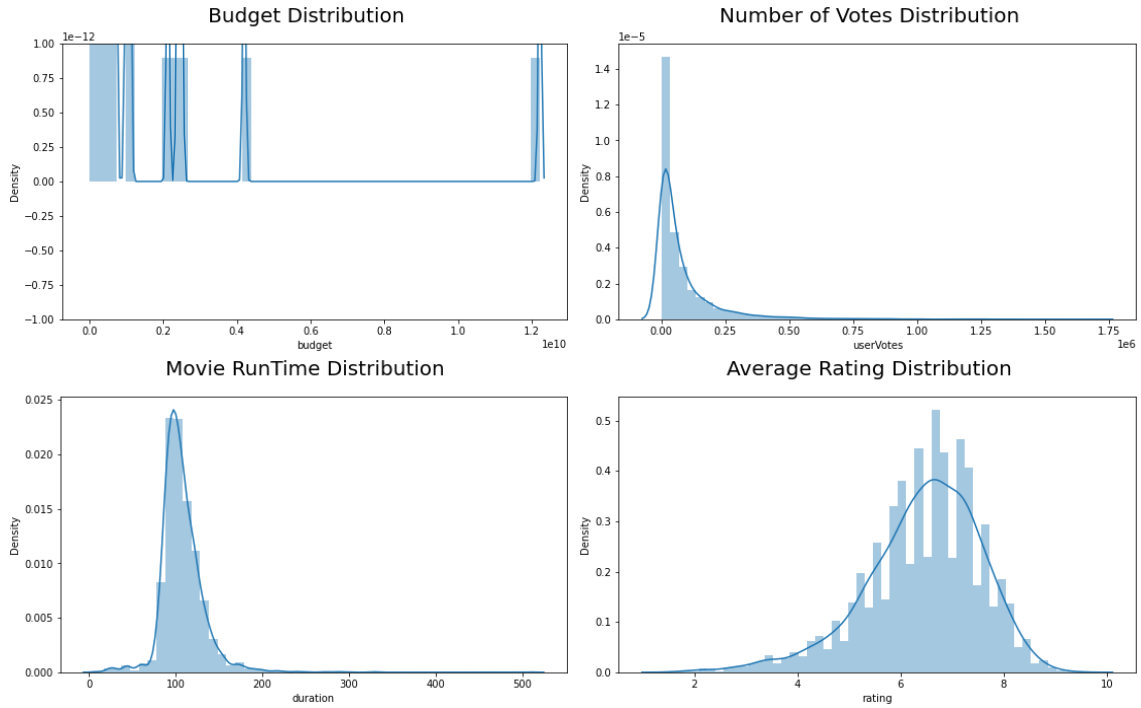

Figure 3. Density distribution

Testing and validation: After training the model, prediction model is tested on the remaining testing data and performance is evaluated. In the validation step, the accuracy of results is verified, comparing the test data with prediction.

2.2. Sentiment analysis

Each stage shown in Figure 2 in context of sentiment analysis model is discussed as follows: -

- Dataset: The dataset of sentiment analysis is downloaded from [22]. This dataset contains IMDb movie reviews along with their respective sentiment polarity labels i.e. negative or positive. The core dataset contains around 50K reviews and they are further split into 25K training set and 25K testing set. Train and test set contains disjoint set of movies.
- Data pre-processing and transformation: This is an important process in text classification and it transforms the raw data into understandable format for learning model. The pre-processing steps followed are tokenization (breaking stream of text into words or meaningful tokens), stemming (reducing the inflectional forms). The reviews text may contain non-alphabetic characters, stop words, and URLs. During pre-processing of tweets, the special characters, emails, urls, and repetitive characters. Have been removed. Table 1 shows the reviews before and after pre-processing. Unigram based bag of word (BOW) approach is used as feature selection approach.
- Splitting training and testing data: The dataset already contain the train and test split in ratio of 50:50.
- Model training: To perform sentiment analysis linear support vector classifier (SVC), Naïve Bayes Classifier and logistic regression models are used for classification of review text into one of the two categories positive and negative.
- Testing and validation: Once the best fit hyperplane is obtained, the prediction for new features can be calculated. Finally, performance evaluation of the model is done.

This model is now used for classification of tweets related to a particular movie. The Figure 4 shows the overall process. Tweets are collected from twitter application by using the public application programming

interface (API) provided by Twitter. Twitter provides the authentication keys for extraction of tweets for authentic requests. Consumer key, consumer secret key, access token and access token key are the unique keys required to fetch the tweets. The retrieved tweets contain information like user ID, date of tweet, retweet count, tweet ID and so on. All the tweets related to a particular movie and all the news and comments related to particular movie are fetched.

Next, the sentiment analysis model is applied on the collected tweets. Firstly, various parameters like keywords, language, and node entities are set. Next, the user credentials are extracted. Once the user credentials are set, tweets are analysed for identifying the polarity of tweet using the sentiment analysis model generated above. The scores of a movie are decided on the bases of the polarity, if polarity is positive then just add 1 to the score, otherwise remove 1 from the score. The score computed are used for hybrid rating model. These scores may also be used to find the most popular movies of a particular calendar year.

Table 1. Reviews before and after pre-processing

Original Reviews	Reviews after Pre-processing
When I first tuned in on this morning news, I…	when i first tuned in on this morning news i t…
Mere thoughts of "Going Overboard" (aka "Babes…	mere thoughts of going overboard aka babes aho…
Why does this movie fall WELL below standards?...	why does this movie fall well below standards…
Wow and I thought that any Steven Segal movie…	wow and i thought that any steven segal movie…
The story is seen before, but that does'n matt…	the story is seen before but that doesand matt…

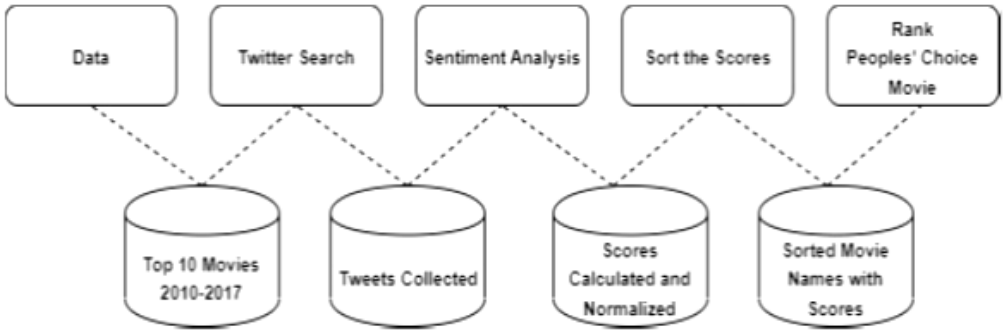

Figure 4. Sentiment analysis process

2.3. Hybrid rating

The rating obtained from the rating prediction model and the sentiment analysis score (as discussed in above two sections). Of the movie are combined to get the overall rating of the movie. This rating is now considering the movie features as well the social media sentiment for prediction the rating.

2.4. Implementation

Python APIs are used to implement the prediction model discussed above. The important libraries used are [24]-[27].

- Data pre-processing: In this work, pandas library is used to create dataframes for the data processing and transformation.
- Splitting the training and testing dataset: train_test_split() function of sklearn is used for splitting the dataset.
- Training the model: Random forest regressor, decision tree regressor, linear regression, linear SVC, Naïve Bayes, logistic regression model of sklearn are used for training.
- Testing: Testing and evaluation are done using sklearn.metrics.
- Matplotlib is used for plotting the data and results and plotly is mainly used for displaying graphs.

For sentiment analysis of collected tweets, TwitterSearch package is used. Firstly, the Twitter Search Order class available in Twitter Search package is used to set various parameters like setting keywords, language, and node entities. Next, the twitter search object is created where user credentials such as key, secret_access_token are set. Further, the algorithm iterates through the tweets and analyse them by checking the sentiment polarity using textblob for tweets.

3. RESULTS AND EVALUATION

This section discusses the results obtained. The MAE, RMSE and R2 of various models is shown in Table 2. The random forest algorithm performs the worst and Linear regression performs best in terms of MAE. It is also found that the linear regression model was very stable in predicting values. Therefore, the linear regression model is taken as the model for rating prediction. Figure 5 shows the actual and predicted ratings of some of the testing samples for linear regression model. The predicted value is very close to the real value in most of the cases.

The accuracy of the sentimental analysis model is shown in Table 3. The accuracy of linear SVC is highest among three i.e. 88.47% and therefore it is used for the hybrid model. The top ten movies of a particular year are identified using the proposed model. The accuracy of the hybrid model is further higher than the individual ones. These movies are identified considering the twitter sentiment analysis. The top movies identified for year 2017 are shown in Figure 6. The actual top movies of 2017 based on user votes are Blade Runner, Thor: Ragnarok, Call Me By Your Name, Wind River, The Greatest Showman. This shows that the results predicted are close to the actual results.

Table 2. Comparison of MAE and RMSE for different models

Model Used	MAE	RMSE	R2
Random Forest	0.491	0.511	0.712
Simple Regression Tree	0.452	0.699	0.601
Linear Regression	0.392	0.501	0.718

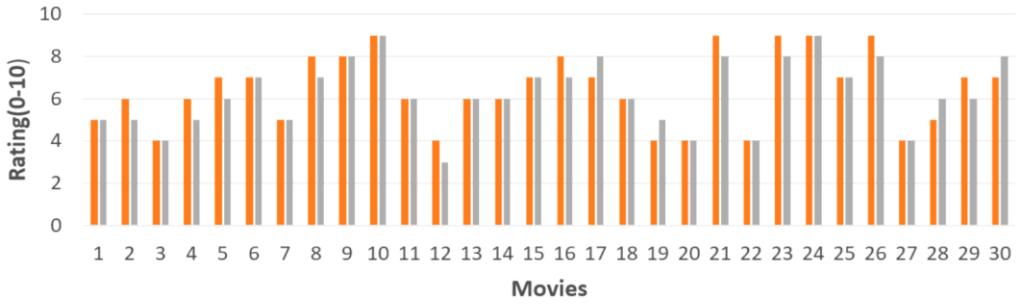

Figure 5. Actual vs predicted ratings

Table 3. Performance comparison of sentiment analysis

Model	Precision	F1 Score	Accuracy
Logistic Regression	0.81	0.84	0.8480
Linear SVC	0.84	0.88	0.8847
Naïve Bayes	0.80	0.83	0.838

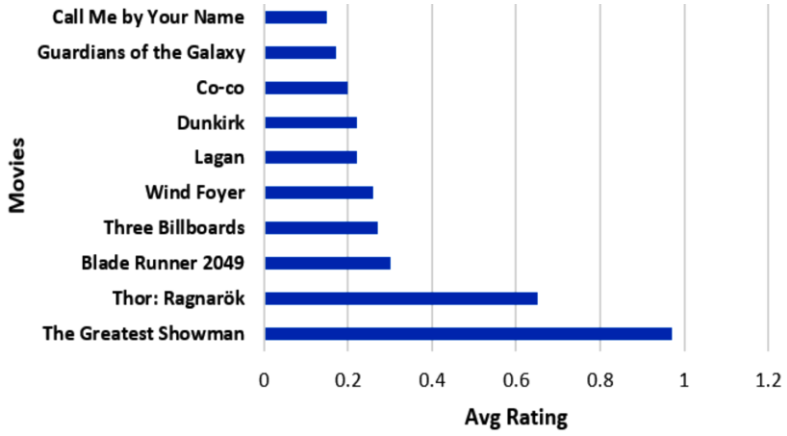

Figure 6. Top 10 movies of year 2017

4. CONCLUSION

Success of movie does not only depend on the features of movie but also influenced by social media reviews and comments. In this work the hybrid approach is used to predict the success rate of a movie. The hybrid approach considers both, the movie features as well as sentiment expressed in the movie review. For rating prediction, random forest, simple regression tree and linear regression models are generated and compared. Based on the performance of all the model over RMSE, MAE and r2 score, the linear regression model is selected for the proposed movie success prediction approach. Further, for review sentiment analysis, linear SVC, Naïve Bays and logistic regression models are compared over accuracy, precision and f1 score. The performance of linear SVC was best among the tree models and therefore, it is selected for the proposed success rate prediction approach. Finally, the success rate of a new movie is predicted by combing the two models generated and analysing the related tweets scrapped from twitter. In future, other machine learning models may be implemented and tested for movie success rate prediction.

REFERENCES

[1] "Statista Research Department, Which is your favourite source of entertainment?," https://www.statista.com/statistics/248352/most-preferred-sources-of-entertainment-in-india/ (2012).

[2] A. Watson. "Film industry in the United States and Canada – statistics & facts," statista.com. https://www.statista.com/topics/964/film/ (accessed Aug. 1, 2021).

[3] D. Zipin. "How Exactly Do Movies Make Money?," investopedia.com. https://www.investopedia.com/articles/investing/093015/how-exactly-do-movies-make-money.asp (accessed Aug 5, 2021).

[4] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *arXiv preprint cs/0212032*, 2002, doi: 10.48550/arXiv.cs/0212032.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*. 2002, doi: 10.3115/1118693.1118704.

[6] G. Mishne, and N. S. Glance, "Predicting movie sales from blogger sentiment," In *AAAI spring symposium: computational approaches to analyzing weblogs*, pp. 155-158, 2006.

[7] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational intelligence*, vol. 22, no. 2, pp. 110-125, 2006, doi: 10.1111/j.1467-8640.2006.00277.x.

[8] W. Zhang and S. Skiena, "Improving movie gross prediction through news analysis," In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 301-304. IEEE, 2009, doi: 10.1109/WI-IAT.2009.53.

[9] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90-94, 2012, doi: 10.1.1.727.1193.

[10] M. Á. García-Cumbreras, A. Montejo-Ráez, and M. C. Díaz-Galiano, "Pessimists and optimists: Improving collaborative filtering through sentiment analysis," *Expert Systems with Applications*, vol. 40, no. 17, 2013, pp. 6758-6765, doi: 10.1016/j.eswa.2013.06.049.

[11] G. Mesnil, T. Mikolov, M. A. Ranzato, and Y. Bengio, "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews," *arXiv preprint arXiv:1412.5335*, 2014, doi: 10.48550/arXiv.1412.5335.

[12] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office predictionm" In *International conference on computing, communication & automation*, pp. 933-937, IEEE, 2015, doi: 10.1109/CCAA.2015.7148530.

[13] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, pp. 540-546, IEEE, 2017, doi: 10.1109/UEMCON.2017.8249013.

[14] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26597-266132019, doi: 10.1007/s11042-019-07788-7.

[15] G. Charu, G. Chawla, K. Rawlley, K. Bisht, and M. Sharma, "Senti_ALSTM: sentiment analysis of movie reviews using attention-based-LSTM," In *Proceedings of 3rd International Conference on Computing Informatics and Networks*, pp. 211-219, Springer, Singapore, 2021, doi: 10.1007/978-981-15-9712-1_18.

[16] G. Verma and H. Verma, "Predicting bollywood movies success using machine learning technique," In *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 102-105, IEEE, 2019, doi: 10.1109/AICAI.2019.8701239.

[17] Z. Shaukat, A. A. Zulfiqar, C. Xiao, M. Azeem, and T. Mahmood, "Sentiment analysis on IMDB using lexicon and neural networks," *SN Applied Sciences*, vol. 2, no. 2, pp. 1-10, 2020, doi: 10.1007/s42452-019-1926-x.

[18] S. Pirunthavi, P. R. Vithusia, K. Abishankar, E. M. U. W. J. B. Ekanayake, and M. Yanusha, "Movie success and rating prediction using data mining algorithm," 2020, doi: 10.13140/RG.2.2.18052.86402.

[19] K. Pradeep, C. R. T. Rosmin, S. S. Durom, and G. S. Anisha, "Decision tree algorithms for accurate prediction of movie rating," In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 853-858. IEEE, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-000158.

[20] J. D. Bodapati, N. Veeranjaneyulu, and S. Shaik, "Sentiment analysis from movie reviews using LSTMs," *Ingenierie des Systemes d'Information*, vol. 24, no. 1, 2019, doi: 10.18280/isi.240119.

[21] S. Tiwari, S. Kumar, V. Jethwani, D. Kumar, and V. Dadhich, "PNTRS: personalized news and tweet recommendation system," *Journal of Cases on Information Technology (JCIT)*, vol. 24, no. 3, pp. 1-19, 2022, doi: 10.4018/JCIT.20220701.oa9.

[22] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Large movie review dataset," In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, 2011.

[23] IMDb dataset. imdb.com. https://www.imdb.com/interfaces/. (accessed Aug 5,2021).

[24] Pandas library. Pandas.com. https://pandas.pydata.org/ (accessed Aug 5, 2021).

[25] Scikit-learn library. scikit-learn.org. https://scikit-learn.org/stable/modules/classes.html (accessed Aug 5, 2021).

[26] Matplotlib library. Matplotlib.org. https://matplotlib.org/ (accessed Aug 5, 2021).

[27] Plotly library. Plotly.com. https://plotly.com/. (accessed Aug 5, 2021).

BIOGRAPHIES OF AUTHORS

Ms. Jyoti Tripathi has more than 18 years of teaching experience. She has research interests in Natural language processing, machine learning, data analytics and big data. She has published many research papers and book chapters in various international conferences, journals and books. She has also been awarded the best paper presenter award in a couple of conferences. She is currently serving as Head in the Department of Computer Science and Engineering at G.B. Pant DSEU Okhla-I Campus. She is also looking into the college library in the capacity of the library in charge. She has also organised and supervised several projects/seminars/sessions/workshops under the National Institute of Technical Teachers Training & Research and International Conference on Information System and Computer Networks. She has been awarded the best paper award for "Stemmer for Hindi Documents: A GA Based Approach" and "Drowsiness Detection System using PPG sensor's measured Physiological Parameter" at (ICICC-21), Delhi, India. She can be contacted at email: loginjyoti@gmail.com.

Dr. Sunita Tiwari has over 20 years of teaching experience. Her research interests include Soft Computing, Location Based Services, GPS Log Mining, Recommender Systems, and Web Service Recommendations. She has published many books and refereed papers in international journals and conferences. Her publications 'Crowdsourcing based Fuzzy Information Enrichment of Tourist Spot Recommender System' at 15th International Conference on Computational Science and Applications and 'WSS-NFP: Tool for Web Service Selection Based on Non-Functional Properties Using Soft Computing Technique' at International Conference on Machine and Web Intelligence (ICMWI) 2010 were selected as the best papers. She has accomplished various administrative roles over the years such as Dean for Student Welfare, Head of Training Placement Cell, Public Information Officer, M. Tech. Coordinator and many more. She has conducted many short term courses and workshops under the National Institute of Technical Teachers Training & Research and chaired many national and international conferences. She can be contacted at email: sutiwari@gmail.com.

Dr. Anu Saini has over 10 years of teaching experience. She has research interests in Service Oriented Architecture, Data Warehouse, Database, Software Engineering, Recommendation Systems, Big Data. She has received a Senior Research Fellowship (SRF) from Council for Scientific and Industrial Research (CSIR). She is Oracle Certified Professional (OCP) Oracle 9i Developer Suite. She has been awarded a Travel Grant from 'Council for Scientific and Industrial Research (CSIR)' for participating in the conference, Melbourne, Australia. She has also published and presented research work in several national and international conferences and journals. She has also authored book chapters. She is a lifetime member of Indian Society for Technical Education (ISTE). She can be contacted at email: anuanu16@gmail.com.

Ms. Sunita Kumari has more than 17 years of teaching experience. She has research interests in machine learning, big data, software engineering and data sciences. She has supervised 21 M. Tech. thesis/dissertation. She has published several patents on her name and has published research work in several national and international conferences and Scopus and Science Citation Indexed Journals. She has also authored book chapters. She is a lifetime member of the International Association of Computer Science and Information Technology (IACST), International Association of Engineers), ISOC (Internet Society Chapter Membership (IAENG), Universal Association of Computer and Electronics Engineers (UACEE), Science and Engineering Institute (SCIEI) and Society of Digital Information and Wireless Communications (SDIWC). She can be contacted at email: sunita2009@gmail.com.