ANALYSIS ON IMDB MOVIE REVIEWS

A PROJECT REPORT

for

DATA MINING TECHNIQUES (SWE2009)

in

M.Tech (Integrated) Software Engineering

by

ABINESH K (16MIS0451)

Under the Guidance of

Prof. PRABHAVATHY P

School of Information Technology and Engineering

November, 2019

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled **"ANALYSIS ON IMDB MOVIE REVIEW"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (SWE2009)** is a record of bonafide project work carried out by us under the guidance of **Prof . PRABHAVATHY.P** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore   Signature

Date : 6-NOV-2019

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **"ANALYSIS ON IMDB MOVIE REVIEWS"** submitted by **VINITHA JANANI S (16MIS0410), ABINESH K (16MIS0451), SHANTHINI L(16MIS0495)**to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (SWE2009)** is a record of bonafide work carried out by them under my guidance.

Prof. PRABHAVATHY P
ASSOCIATE PROFESSOR
SITE

ANALYSIS ON IMDB MOVIE REVIEW

Abstract :

This paper presents different approaches to predict the success rate of the movies and also predicted the sentimental analysis of the moviegoers As Movie ratings and reviews at sites such as IMDb are commonly used by moviegoers to decide which movie to watch or buy next. Currently, moviegoers base their decisions as to which movie to watch by looking at the ratings of movies as well as reading some of the reviews at IMDb.In this project, we developed a mathematical model to predict the success and failure of the upcoming movies based on several attributes. Some of the criteria in calculating movie success included budget, actors, director, producer, set locations, story writer, movie release day, competing movie releases at the same time, music, release location and target audience. The success prediction of a movie plays a vital role in movie industry because it involves huge investments. This paper presents a detailed study of Adaboost, SVM, Naïve Bayes Classifier and K-Nearest Neighbors on IMDb data to predict movie box office. And also to predict the sentimental analysis by the classifiers (SVM, Naïve bayes).

Keywords – Movie Reviews, IMdb ,SVM ,naïve bayes ,KNN, Adaboost

I. INTRODUCTION

   Data mining process was used to extract patterns and trends which can be beneficial in predicting movies success. The data mining techniques were applied to movie database, but before the mining techniques could be used, the data went through the cleaning and integration process. We have provided a useful model in this study which can lower chance of failure and can provide the stakeholders with confidence and a visible prediction of success. The goal of this mathematical model is to provide a precise prediction of success, hence providing confidence to stakeholders in their investments .The success prediction of a movie plays a vital role in movie industry because it involves huge investments. This paper presents a detailed study of Adaboost, SVM Logistic Regression, Naïve Bayes Classifier and K-Nearest Neighbors on IMDb data to predict movie box office. And also to predict the sentimental analysis by the classifiers (SVM, Naïve bayes).

III. DATASET DESCRIPTION & SAMPLE DATA

- The dataset we utilized to train and test our model came from kaggle.com. The dataset includes several movies on IMDb, including set of 1,000 most popular movies on IMDB in the last 10 years. The data points included are:

- Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascrore Feel free to tinker with it and derive interesting insights.

- This information was provided in a variety of formats, including strings, integers, and floating point data.

- In order to implement the machine learning algorithms effectively and avoid underutilization of certain aspects of the movies provided in the data set, the data was converted to numerical values using the Scikit learn reprocessing library to scale the features.

	A	B	C	D	E	F	G	H	I	J	K	L
	Rank	Title	Genre	Descriptio	Director	Actors	Year	Runtime (	Rating	Votes	Revenue (	Metascore A
	1	Guardians	Action,Ad	A group of	James Gui	Chris Prat	2014	121	8.1	757074	333.13	76
	2	Promethe	Adventure	Following	Ridley Sco	Noomi Ra	2012	124	7	485820	126.46	65
	3	Split	Horror,Th	Three girl:	M. Night S	James Mc.	2016	117	7.3	157606	138.12	62
	4	Sing	Animation	In a city of	Christoph	Matthew	2016	108	7.2	60545	270.32	59
	5	Suicide Sc	Action,Ad	A secret g	David Aye	Will Smith	2016	123	6.2	393727	325.02	40
	6	The Great	Action,Ad	European	Yimou Zha	Matt Dam	2016	103	6.1	56036	45.13	42
	7	La La Land	Comedy,D	A jazz piar	Damien Cl	Ryan Gosl	2016	128	8.3	258682	151.06	93
	9	The Lost C	Action,Ad	A true-life	James Gra	Charlie Hu	2016	141	7.1	7188	8.01	78
	10	Passenger	Adventure	A spacecra	Morten Ty	Jennifer L	2016	116	7	192177	100.01	41
	11	Fantastic I	Adventure	The adver	David Yate	Eddie Red	2016	133	7.5	232072	234.02	66
	12	Hidden Fi	Biography	The story	Theodore	Taraji P. H	2016	127	7.8	93103	169.27	74
	13	Rogue On	Action,Ad	The Rebel	Gareth Ed	Felicity Jo	2016	133	7.9	323118	532.17	65
	14	Moana	Animation	In Ancient	Ron Clem	Auli'i Crav	2016	107	7.7	118151	248.75	81

IV. LITERATURE SURVEY :

Title of paper	Algorithms used	Dataset being used	Performance measures	Gap identified	Scope for future work
Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B. (2017, July). Movie success prediction using data mining. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE.	Chi square analysis between all attributes and find correlation between them	Simulation data was used for this analysis and hundreds of records were cleaned, integrated and transformed. A random subset of this data was utilized for each set of analysis.	Mathematical model based on ratings and statistical figure	Here they focused only on Bollywood films	Expand their model to include Hollywood movies
Dhir, R., & Raj, A. (2018, December). Movie Success Prediction using Machine Learning Algorithms and their Comparison. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 385390). IEEE.	Here they have used 4 classifiers models such as SVM, ADA boost , KNN, random forest	The dataset from kaggle.com. The dataset includes information about several movies on IMdb.	Accuracy comparative analysis	In order to implement the machine learning algorithms effectively they have avoided underutilization of certain aspects of the movies provided in the dataset.	They would like to increase both the number of movies and features in the dataset. They would also like to include other social media sources such as Twitter.

Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015, January). Role of different factors in predicting movie	Here they have discussed what are the steps to be followed to achieve prediction , like ways of data	they gather the data through various APIs and file formats available	Accuracy and parameter like Gross box office collection.	Their predictive models are usually based on classical factors such as	The interrelation among the classical factors and Integration
success. In *2015 International Conference on Pervasive Computing (ICPC)* (pp. 1-4). IEEE.	acquisition, data pre processing , feature extraction, classification and result		And Critics rating	Cast, Producers, Directors, Genre, movie revenue, movie production budget	of classical factors and social media
Kanitkar, A. (2018, October). Bollywood Movie Success Prediction using Machine Learning Algorithms. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)* (pp. 1-4). IEEE.	Here they have implemented 5 machine learning algorithms, they are *Linear Regression, Polynomial Regression, Logistic Regression, Artificial Neural Network (ANN), K Nearest Neighbors (KNN).*	Their dataset contains 250 bollywood movies released between 2014 and 2017.their data sources are Wikipedia[5], RadioMirchi[6] and BoxOfficeIndia[7].	accuracy, precision and f1 score	Bollywood movies they have focused only on	sentiment of peo ple while watching the trailer (both in case of theatre audience and ho me viewing audience) could be a pivotal feature to consider as well. By using several IOT and CNN algorithms

Verma, G., & Verma, H. (2019, February). Predicting Bollywood Movies Success Using Machine Learning Technique. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* (pp. 102105). IEEE.	Logistics regression (LR) is one of the most popular algorithm that can be used to predict a binary outcome and one or multiple continuous or categorical predictor variables	data for their study had been collected from variety of online sources such as imdb.com, bollymoviereviewz.com, planetbollywood.com, boxofficeindia.com, and bollywoodhungama.com	the Receiver Operating Characteristics (ROC) curve, using three predictors like No. of screens, IMDb ratings, MusicRating	sample size being small as not enough data is	In future they may use other predictors
Quader, N., Gani, M. O., & Chaki, D. (2017, December). Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In *2017 3rd Technology (EICT)* (pp. 1-6). IEEE.	They have used seven machine learning methods like *Logistic Regression, Neural Network, Stochastic Gradient Descent, AdaBoost, Gaussian Naive Bayes, Random Forest*	their dataset contains 755 movies released in between 2012 to 2015. their data sources are IMDb, Rotten Tomatoes, Metacritic and Box Office Mojo.	performance comparison for all features and accuracy comparison	Most of the features are missing for most of the movies. In many cases, movies budgets are unavailable. So they removed those movies	They also suggest to analyze and include the number of audience for analysis. They can get number of annual audience by using total ticket sold in a particular year

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017, December). Predicting movie box office success using multiple regression and SVM. In *2017 internationa on intelligent sustainable systems (ICISS)* (pp. 182186). IEEE.	*Here they have used Support Vector Machine Classification and Data Pruning*	The dataset was populated with information scraped from BoxOfficeMojo and Wikipedia for movies released in 2016. Trailer views were taken from YouTube.	Adjusted ROI and accuracy	They remove junk values and discard movies which do not have available information about them.	Their efforts at trying to predict movie box office success have focused only on one feature and its im pact on movie success
Magdum, S. S., & Megha, J. V. (2017, June).Mining reviews, tweets for predicting sales performance and success of movies. In *2017 International Conference on Intelligent Computing Systems (ICICCS)* (pp. 334339). IEEE.	*SPLSA(Sentiment-Probabilistic Latent Semantic Analysis), Autoregressive Sentiment Model*	Data from reviews and tweets, which is used for predicting sales performance. The online reviews are collected from the IMDB website and tweets from Twitter	They have plain metric called P-N ratio for predicting success of movies	If a movie receiving poor online reviews and also movie's predicted box office collection using these models declining day by day.	Hollywood movies using sentiment information mined from reviews and tweets.

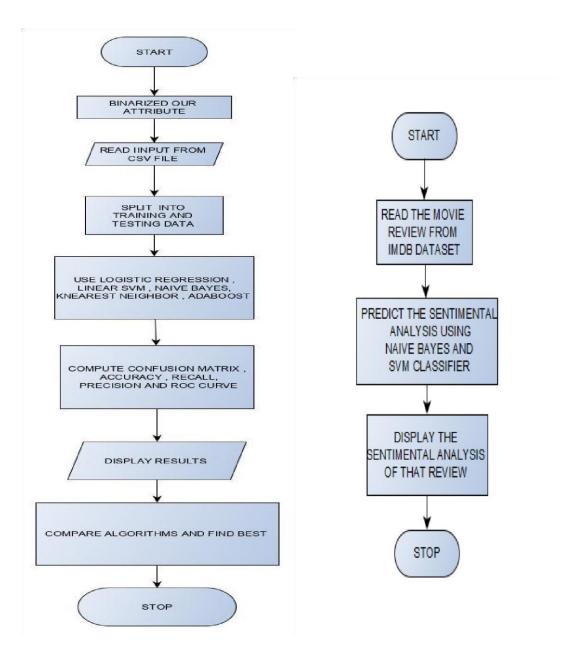|					
Quader, N., Gani, M. O., Chaki, D., & Ali, M. H. (2017, December). A machine learning approach to predict movie box-office success. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1-7). IEEE.	here they have used *Sentiment Analysis, Support Vector Machine (SVM), Neural Network Analysis*	This dataset contains 755 movies released in between 2012 to 2015.. their data sources are IMDb, Rotten Tomatoes, Metacritic and Box Office Mojo.	Accuracy for *Sentiment Analysis, Support Vector Machine (SVM), Neural Network Analysis*	Here they do not consider genre and sequel to a movie as features. Prediction of a sequel movie is terrible	A country's GDP rate can be used as a feature to know if there is financial stability , this can be consider for future purpose.
Rahim, M. S., Chowdhury, A. E., Islam, M. A., & Islam, M. R. (2017, December). Mining trailers data from youtube for predicting gross income of movies. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10HTC)* (pp. 551-554). IEEE.	regression analysis, Spearman's rho correlation technique.	Movie dataset contains two types of information. One is movie related information which is collected from Box Office Mojo [20] and another one is movie trailer related information which is collected from YouTube	Correlation Coefficients between all the attributes (R^2).	they only worked with the trailer data collected from YouTube.	they recommend performing an experiment by adding the data from different sources.

V. PROPOSED ALGORITHM WITH FLOWCHART

1. **Searched for available datasets** to support our idea and thoroughly scrutinized them, to get most suitable dataset for our idea.

2. **Shortlisted few datasets**, we picked the most suitable dataset for our project.

3. **Pruned the data** which we required, most suitable for our prediction and analysis.

4. Collected ground truth data, and **saved in the csv file** format. **We also binarized our attributes** and **used an additional success column**, based on the average revenue, rating and votes received by the movie.

5. We used **this data** as an **input** to the machine learning and data mining algorithms **for prediction of ANALYSIS OF MOVIE SUCCESS** .

6. We **split the data into training and testing data**.

7. The machine learning algorithms we have used are **Logistic Regression, Linear**

SVM, KNearest Neighbor, Naïve Bayes Classifier and Adaboost.

8. We have computed the results of our algorithms by means of **confusion matrix, accuracy, recall, precision rate and ROC curve.**

9. We have also used this dataset for analysis of effect of various attributes on the success percentage of movie. These attributes **include rating, votes, actors, directors, revenue and metascore**.

10. we also **done some sentimental analysis** on the reviews from the dataset.by reading the review of the movie **and predict the sentimental analysis using naïve bayes classifier** and **SVM classifier** and display the sentimental analysis of the moviegoers.

VI. EXPERIMENTS RESULTS

To visualize the performance of an algorithm, typically a supervised learning confusion matrix is used. Also, known as error matrix, each column of the confusion matrix signifies an instance of a predicted class and each row signifies an instance of the actual class. If the prediction and ground truth are equal, then it is either True-positive or True negative based on the classification labels. If the prediction is not equal to the ground truth, then it is either False positive or FalseNegative based on the classification labels. From the table we are calculating the Accuracy, Precision and Recall. We are also determining the ROC curve to evaluate the performance of the algorithm.
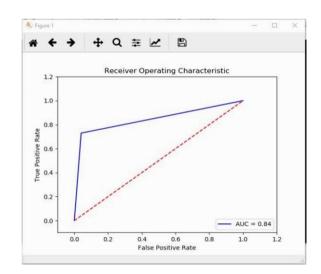
a)SVM


```
C:\Users\ABINESH LKD\Desktop\DM>SVM.py
Confusion Matrix is :
[[166   7]
 [ 10  27]]

Accuracy is :
0.919047619047619

Precision is :
0.7941176470588235

Recall is:
0.7297297297297297
```
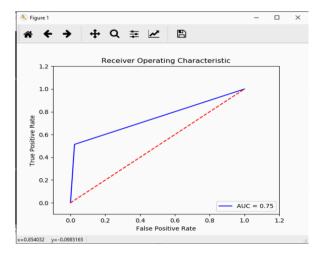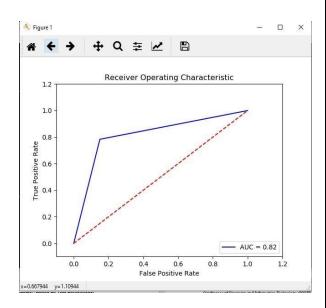
b)K NEAREST NEIGHBOUR


```
C:\Users\ABINESH LKD\Desktop\DM>Kneighbour.py
Confusion Matrix for KNN is :
[[169   4]
 [ 18  19]]

Accuracy for KNN is :
0.8952380952380953

Precision is :
0.8260869565217391

Recall is:
0.5135135135135135
```
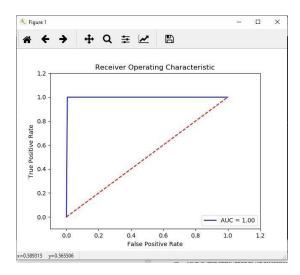
c)NAÏVE BAYES

```
C:\Users\ABINESH LKD\Desktop\DM>Kneighbour.py
Confusion Matrix for KNN is :
[[169   4]
 [ 18  19]]

Accuracy for KNN is :
0.8952380952380953

Precision is :
0.8260869565217391

Recall is:
0.5135135135135135
```
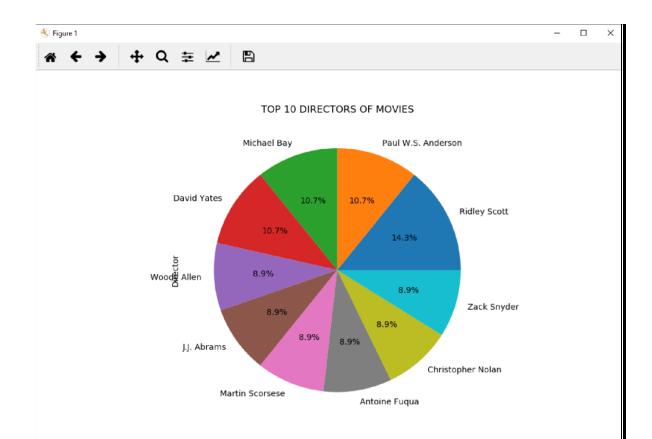

d)ADABOOST

```
C:\Users\ABINESH LKD\Desktop\DM>Adaboost.py
Confusion Matrix is :
[[172   1]
 [ 0  37]]

Accuracy is :
0.9952380952380953

Precision is :
0.9736842105263158

Recall is:
1.0
```

DICTION

10 DIRECTORS

Figure 1

TOP 10 DIRECTORS OF MOVIES

Michael Bay

Paul W.S. Anderson

David Yates

Ridley Scott

10.7%

10.7%

10.7%

14.3%

Director

Woody Allen

8.9%

8.9%

Zack Snyder

8.9%

8.9%

8.9%
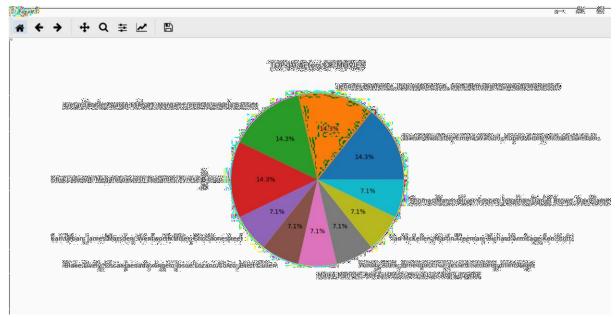
8.9%

J.J. Abrams

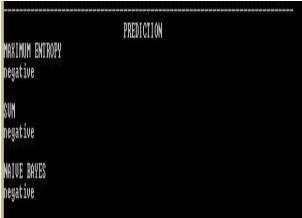Christopher Nolan

Martin Scorsese

Antoine Fuqua

TOP 10 ACTORS

a)SENTIMENTAL ANALYSIS

VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

Adaboost:

From the above results, it can be inferred that the Adaboost has a very high accuracy of 99.5%. The ROC curve gives an AUC of 1 which is a perfect score indicating a perfect test.

Naïve bayes:

acy for Naive Bayes Classifier is 83.8%. ROC is good but in comparison to other models which are scope of our study, NBC seems to an underperforming model for our dataset. AUC is 0.82 which is a ure as well.

K nearest neighbor :

he above results, it can be inferred that the K-Nearest Neighbor classifier at k = 5 has a good accuracy

and the ROC curve gives an AUC of 0.75.

Support vector machine:
s of SVM algorithm, the accuracy of 91%, and the AUC is good, even though we get good accuracy, it
to be the best suited algorithm for our prediction.

ALGORITHM	ACCURACY	PRECISION	RECALL	AUC
SVM	0.9190	0.7941	0.7297	0.8400
NAÏVE BAYES	0.5333	0.2706	0.9729	0.7100
KNN	0.8309	0.5272	0.7837	0.8200
ADABOOST	0.9952	0.9736	1.0000	1.0000

VIII. CONCLUSION AND FUTURE WORK

- After building the models we found out that the success percentage for all models were nearly the same however the Adaboost had the highest accuracy in our case for predicting the movies success.

- A larger training set is the key to improving the performance of the model.

- We need to consider additional features such as geographic location, age of viewers and voters, current trends, news analysis, movie plot analysis and social networks data analysis could be done and the information thus obtained could be added to the training set. We can also use Google trends result to improve the result.

- We also consider some other classifier for the sentimental analysis and also consider the reviews from other regional languages useful words for the prediction in future **IX. REFERENCES**

1. Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B. (2017, July). Movie success prediction
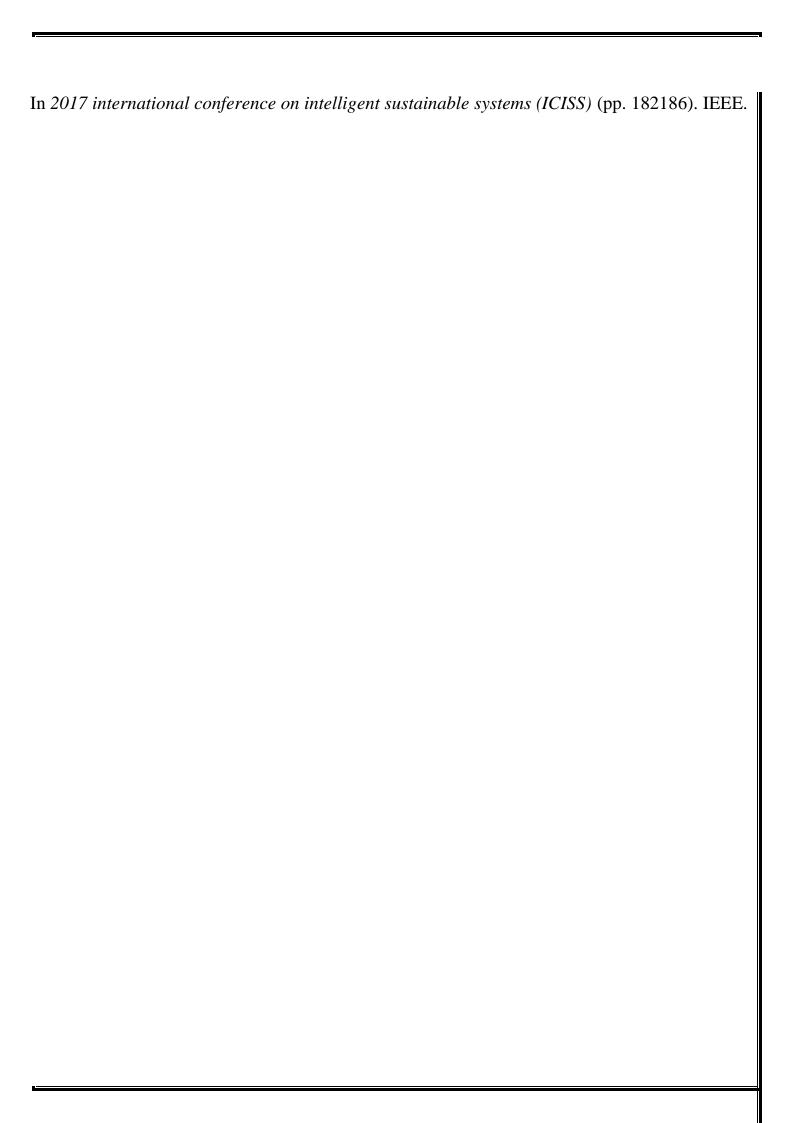
using data mining. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE.

2. Dhir, R., & Raj, A. (2018, December). Movie Success Prediction using Machine Learning Algorithms and their Comparison. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 385-390). IEEE.

3. Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015, January). Role of different factors in predicting movie success. In *2015 International Conference on Pervasive Computing (ICPC)* (pp. 1-4). IEEE.

4.Kanitkar, A. (2018, October). Bollywood Movie Success Prediction using Machine

Learning Algorithms. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)* (pp. 1-4). IEEE.

6. Verma, G., & Verma, H. (2019, February). Predicting Bollywood Movies Success Using Machine Learning Technique. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* (pp. 102-105). IEEE.

7. Quader, N., Gani, M. O., & Chaki, D. (2017, December). Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.

8. Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017,). Predicting movie box office success using multiple regression and SVM.

In *2017 international conference on intelligent sustainable systems (ICISS)* (pp. 182186). IEEE.

9. Magdum, S. S., & Megha, J. V. (2017, June). Mining online reviews and tweets for predicting sales performance and success of movies. In *2017 International*

Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 334-339). IEEE.

10. Quader, N., Gani, M. O., Chaki, D., & Ali, M. H. (2017, December). A machine learning approach to predict movie box-office success. In *2017 20th*

International Conference of Computer and Information Technology (ICCIT) (pp. 1-7). IEEE.

11. Rahim, M. S., Chowdhury, A. E., Islam, M. A., & Islam, M. R. (2017, December). Mining trailers data from youtube for predicting gross income of movies. In *2017*

IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 551-554) . IEEE

12. Wang, F., Cai, R., & Huang, M. (2010, May). Forecasting movie-going behavior based on online pre-release WOM and opening strength. In *2010 2nd International Workshop on Intelligent Systems and Applications* (pp. 1-4). IEEE.

13. Rhee, T. G., & Zulkernine, F. (2016, December). Predicting movie box office profitability: a neural network approach. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 665-670). IEEE.

13.Zhang, S., Tang, Y., Lv, X., & Dong, Z. (2018, December). Movie Short-Text Reviews Sentiment Analysis Based on Multi-Feature Fusion. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence* (p. 66). ACM.

4.Ahmed, M., Jahangir, M., Afzal, H., Majeed, A., & Siddiqi, I. (2015, December). Using Crowd

source based features from social media and Conventional features to predict the movies popularity. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)* (pp. 273-278). IEEE.

15.Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., & Siehndel, P. (2010, August). Analyzing the blogosphere for predicting the success of music and movie products. In *2010 International Conference on Advances in Social Networks Analysis and Mining* (pp. 276-280). IEEE.