2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397206

A Machine Learning Based Movie Status Evaluation System for Bangladesh Movies

Sharmin Akter
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh
sharmin.cse051@gmail.com

MN Huda
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh
mnh@cse.uiu.ac.bd

***Abstract*— This study develops a machine learning based on movie status evaluation system for Bangladesh movies. The number of movie production rate is growing day by day worldwide and the movie maker invests highly in the movie industry. In such a scenario, this is very important to evaluate movie status. In our proposed research, we evaluate the status of Dhallywood movie (Bangladesh Cinema) based on three different types of Machine Learning (ML) based classification, Binary classifier that includes two targeted classes, Triple classifier that includes three targeted classes and, four classifier that includes four targeted classes. Here, we will give our detailed analysis of data because Bangladeshi movie data collection is the main challenge of our work and consequently, we analyze our data in different ways to set target variable to improve the accuracy of models. For the first time any research focuses on Dhallywood movie data where we have used different machine learning based models for analyzing data. We apply the same ML algorithm for each of the three different class classifications to find which classifier is performing well for our data and the problem by comparing the obtained accuracies. From the experiments, it is observed that the triple class classification accuracy is higher than binary and four class classifications. Among the five applied ML algorithms, the Random Forest shows the best accuracy around 85%. Our research provides a quite different approach to set target variable class based on Wikipedia data, news, actor- actress biography, and viewer response on YouTube for a particular movie. We go for this approach because Bangladeshi movie rating is not perfect on IMDb also the budgets and revenues are not found for all movies.**

Keywords*— *Machine Learning, Data Mining; Internet Movie Database; Random Forest; Decision Tree

I. INTRODUCTION

Films have always been the most popular medium of entertainment at all times. Film reciprocates the advancing expectations of the public which has led to an exponential rise in the risk associated with making any film. The film industry faces a challenge that is related to profit and loss. There is a risk always exist related to movie income. The risk is the film will be able to earn targeted revenue or not according to the huge budget it invests. The severity of the problem increases when we realize a multitude of factors that impact the revenue of the movie. Dhallywood movie promotion background is not similar to Hollywood or Bollywood. The common practice of Hollywood or Bollywood movies is to release a movie trailer or item song before the movie release. The movie trailer and item song have a great impact on movie success. Film promotion in a reality show is one of the best techniques followed by the film promoters. But the scenario of Dhallywood movie is different. Most of the movies released without any kind of promotion plan.

Previously investigated studies show that predicting the success of a movie using attributes such as: budget, rating, actors, movie rating, meta score, and revenue is possible [1]-[3]. Thus, the goal of this study is to further examine the possibility of using a new data set with some new those are features previously not used with machine learning. In our work we focus on Wikipedia movie information, the biography of actors-actress and sometimes YouTube viewer response to analyze Dhallywood movie status instead of directly predict success. Previous many researcher of others country focus on movie data analysis to enhance film industry [4]-[6][10]. Motivated from their outcome, we have devoted our research to find factors that are playing a very important role for making a movie excellent, good or bad. Also, we have made a dataset of Dhallywood movie for this research work.

II. LITETURE REVIEW

In the early days, many people prioritized gross box office revenue as a parameter of movie success measure [1][2]. Javaria Ahmad, Prakash Duraisamy, Amr Yousef and Bill Buckles was developed a model that can predict the success and failure of the upcoming movies based on several attributes those are closely related to a movie and some of the criteria in calculating movie success included budget, actors, director, producer, set locations, story writer, movie release day, competing for movie releases at the same time, music, release location and target audience [3]. Rijul Dhir and Anand Raj provided a method with different approach to predict IMDb score on IMDb movie dataset. They were tried to discover the significant factors those are influencing the score of IMDb Movie Data. Different algorithms were investigated in the research work for analysis. Random Forest gave the best prediction accuracy among all algorithms which is also better in comparison with the previous studies[4]. The exploratory analysis found that the number of voted users, number of critics for reviews, the number of Facebook likes, duration of the movie and gross collection of movies affect the IMDb score strongly.

Nahid Quader and Dipankar Chaki developed a system that can predict an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic.

ISBN: 978-1-7281-8529-3/21/$31.00 ©2021 IEEE

Using Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-release features. A dataset was prepared with larger number of user reviews from IMDb and Rotten Tomato to analyze user sentiment. This research use Microsoft Power BI Desktop application to calculate sentiment from collected data. With Power BI Desktop they use Microsoft Azure's cognitive service of Text Analytics API. This paper showed that the Neural Network based algorithm gives an accuracy of around 84% for pre-released features and around 90% for all features. On the other hand SVM was gain around 83% and 89% accuracy for pre-released features and all features respectively. They were identified that budget, IMDb votes, and the number of screens are the most important features. These features play an important role to predict a movie's box-office success. Authors in their paper served two parameters base movie prediction system where the two parameters were: Gross box office collection and Critics rating. Beside this they was found that, if a particular actor or actress works with a particular film production house, their films mostly perform well in the box office [5]. Another study where Anand Bhave, Himanshu Kulkarni, Vinay Biramane, and Pranali Kosamkar found that along with the classification factor, social media feedback improves accuracy (FB, Twitter, YouTube) [6]. To adjust and propose a new marketing strategy, another study combines temporal abstraction with data mining techniques to find some important rules. The framework has three main modules those are data attribute selection module, data pre-processing module, and temporal abstraction module. IMDb is a most popular movie data source. Like some other research L. Chen, C. Chi, and L. Huang was gathered most of the basic information from IMDb, such as the number of actors (stars) and comments. Other attributes data was gathered from Amazon and Box Office Mojo [7]. A paper analyzed the movie review mining using semantic orientation analysis and machine learning. The approaches semantic orientation analysis and machine learning were adapted to the movie review domain for comparison. The outcome of this research shows better findings than before the study. There also shows that comparing with many other types of review mining, the movie review mining is a more challenging application than [8]. Another research performed by P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha where they validated feature effectiveness using sentiment classification by embedding text preprocessing with text transformation and clustering. Here, researchers used fuzzy clustering for sentiment classification and also apply SVM for final sentiment classification beside this apply different data mining techniques for a better outcome. They observed that the TF-IDF based feature selection performance is better and gave an accuracy of 10% more than the 14 keywords based technique for box-office prediction [9].

A. Samad, H. Basari, B. Hussin, I. G. Pramudya, and J. Zeniarja in their research applied opinion mining with natural language processing, computational linguistics and text mining to identify or classify whether the movie is good or not based on message opinion. Support Vector Machine is a machine learning supervised method that analyzes data and recognizes the patterns that are used for classification. This research focuses on binary class classification which target was classified into two classes, those were positive and negative. The negative class contains the message of negative opinion of certain movies and the positive class contains message of positive opinion of certain movies. Two model validation techniques cross validation and confusion matrix was applied to justify the accuracy level of SVM [10].

A study was done by R. Niraj and J. Singh, they proved that professional critics review and user-generated review also play a vital role in movie success. This study contributes by implementing a new measure of balance in the movie literature from group and single psychology literature [11]. M. Lash, S. Fu, S. Wang, and K. Zhao in 2015, was discussed in their work regarding three factors 'who', 'what' and 'when'. The early stage of production it is possible to predict the profit of a movie, which can be achieved by 'who', 'what' and 'when' factor, respectively with -its actor, actress directors, and social network, genre and rating, as well as when a movie will be released [12]. Subramaniyaswamy V., Vignesh Vaibhav M., Vishnu Prasad R., and Logesh R in their work classified data into three classes for big-budget film, medium budget, and low budget. SVM and multiple regression were used to predict box office success[13]. A study on Bollywood movie success had done by Garima Verma and Hemraj Verma they were classifying data into two classes hit and flop. In their research, they used different machine learning models and included music rating with IMDb rating and [14]. Marton Mestyan, Taha Yasseri, and Janos Kertesz in their work considered the activity level of an editor, and Wikipedia and the number of pages viewed by readers. They define various variables and apply the logistic regression model for data analysis[15]. Authors in their paper applies an ensemble classification method to review the film as well as film-related meta-data to generate more precise outcomes. The max voting method will predict film as a success, flop and neither flop nor success. When two or more classifiers will give the same outcome. The max voting provide an accuracy of 90% after the outcomes are combined [16]. In August 2016 a paper was published that also focus on movie success prediction using machine learning by predicting rating of a movie. Muhammad Hassan Latif and Hammad Afza in the work also uses machine learning as like previous paper to predict the success of a movie through movie rating[17].

III. PROPOSED METHODOLOGY

In this research, we follow a different approach to achieve a good accuracy. Fig.1 represents the overall scenario of methodology of this research work. Steps are given as follows:
Step 1: Collect data from different internet movie related sources.
Step 2: Clean dataset by removing duplicate values.
Step 3: Preprocess the data as like: missing value handling encoding.
Step 4: In this step, some features have been extracted and correlation feature selection method have been applied.
Step 5: After data preprocessing apply five ML algorithm to calculate accuracy. The steps 1 to 5 have been repeated several times until the models gives a good accuracy the highest accuracy achieved by model is near about 85 percent.

Step 6: All ML models have been validated after achieving satisfactory level of accuracy.

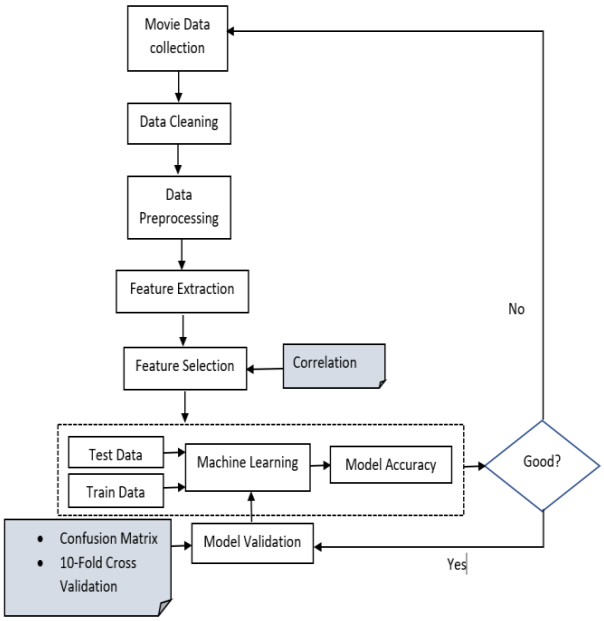

Fig. 1. Methodology of our proposed work

A. Data Preprocessing

Data preprocessing is very important before feed dataset into a machine learning model.

a) Data Cleaning: In this step, duplicate and unnecessary data have been removed. During the time of data scarping some garbage and inaccurate data inserted into the dataset are also removed or corrected. Data cleanind is important for ensure more accuracy of data because unnecessary and noisy data reduce model accuracy.

b) Handling Missing Value: The dataset sometimes contains missing/null values and those missing values also reduce model accuracy. Those null/missing values are not possible to handle or not possible to convert into other values. In this step we just remove those instances. There are several ways to deal with missing value. If the number of missing values are significant then it must be handled[18][19]. In this research the missing values are around 2 percent that why we have removed those values.

c) Splitting Multiple Genres and Construct Individual Column: In this research collected data genre column for some movies contains multiple genres for the same movie. In that case, individual columns have been generated for each genre.

d) Convert Some Column into Binary Value: For analysis purpose some columns converted into binary values, where columns are screening in different festivals, joint venture, and foreign actor. At primary stage these attributes values were yes and no.

e) Encoding Categorical Data: Scikit-learn library of python is very useful for data encoding, by using Scikit-learn library categorical features and instances were encoded, and converted into numeric values. Label encoding is a property of Scikit-learn library. Label encoder is an encoding method that converts any object(text) type data into an integer (numeric) by labeling the value between range 0 to n-1 classes. For applying the label encoding process "sklearn: a preprocessing import Label Encoder" this library need to import. Another popular encoding process, 'One-Hot' encoder also used in this research. This method encodes object value into a binary value. This research includes a binary classification problem, 'One-Hot' encoder is applicable and perform well for solving binary class classification problem.

f) Feature Scaling: Feature scaling is very much effective before feature selection but this is not necessary for all ML models. Feature scaling helps to normalize the data within a particular range. It is a standardization process of machine learing it also helps to improve the calculations of an algorithm because sometimes functions are not work well withour feature scaling. Not for all models, it is important but in this study used KNN model. For KNN model, feature scaling is required. Sklearn and StandardScaler have been used for feature scaling in this research.

B. Feature Extraction and Feature Selection

Data was extracted from IMDb, Wikipedia, and so many internet sources. These sources contain many relevant and irrelevant information about a movie. Before extracting, relevant features have been point out. In this part, some features have been reconstruct from existing features to improve result. In movie data collection and dataset preparation section, we have described the steps of create a new feature through brainstorm also regenerate features from existing features and modify the dataset several times. Total 24 features have been found in some steps. First of all, four leading actors name for each movie was scrapped from IMDb but third and fourth leading actors was reduced model accuracy. Then four features have been rearranged and convert into three. Those are leading actor, leading actress, and leading negative role and this rearrangement performed manually. Actor, actress and villain these newly created features increase the accuracy of model. Some features have been removed before correlation analysis such as: title, year etc. Title and year are significant for a movie but not related to the movie status

C.Correlation

Correlation analysis is a statistical method, that has been used to analyze degree of association of two quantitative variables. Correlation indicates two types of relation between variables, those are positive and negative. When two variables have a strong relationship with each other and value changing in the same direction, this indicates a positive correlation between variables. On the other hands a negative correlation means the variables value is not changing in same direction that means one value is increasing while another value is decreasing. Correlation technique is not appropriate for all kinds of data. If the data is quantifiable and data in which numbers are represent something meaningful and informative, correlation is effective there. To determine the relation and affinity between all the variables with each other, correlation analysis is used.

After analyzing the correlation among features some positive and negative relation have been identified. Correlation heatmap shown in Fig.2. Heatmap lighter and darker colors indicate positive and negative relations respectively. TABLE I. shows our final features with description after correlation analysis

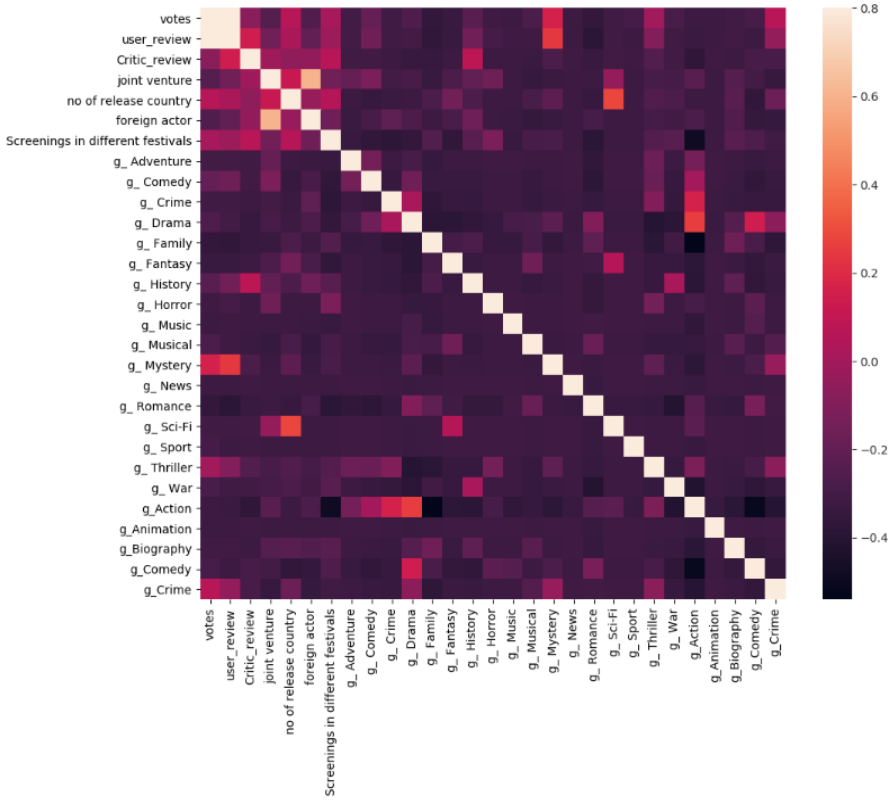

Fig. 2. Correlation heatmap

Positive correlations are given as follows

- g_Drama wit g_Action
- User review with g_Mystery
- Foreign actor with Joint venture
- User_review with votes
- g_Sci-Fi with Number of release country

Negative correlations are given as follows

- Screening in different festival with g_Action
- g_Action with g_Comedy
- g_Family with g_Action

TABLE I. DATASET FEATURES AFTER FEATURE SELECTION

Feature Name	Feature Description
Genre	Category or specific type of the movie (drama, mystery, sci_fi, action and so on)
Director	Name of the director of the movie
Actor	Name of leading male actor
Actress	Name of leading female actor
Villain	Name of leading negative role of the movie
Writer 1	Name of leading story writer of the movie
Votes	Votes for movie given by viewers
User_review	Number of comments given by user
Critic_review	Number of comments given by critic
Foreign_actor	Any part played by foreign actor or not
Joint venture	Movie jointly directed by more than one countries or not
Number of release country	Number of movie released countries
Screeningsin different festivals	A particular movie have been exhibited in different festival or not

C. ML Models

Five algorithms have been applied after feature selection:

- Support vector machine (SVM)
- K-nearest neighbor (KNN)
- Decision tree
- Random forest
- Logistic regression

For each ML algorithm, we calculate target variables three times for each of the different classes.

- Binary class classification – Yes or No
- Triple class classification – Excellent, Good, Bad.
- Four class classification - Excellent, Very Good, Good, Bad.

D. Validation

In our dataset data of all classes are not balanced, so only accuracy is not the basis of an appropriate performance measurement. In this study validation and evaluation have been done by using the 10-fold cross validation and confusion matrix. Confusion matrix had been used for measuring the performance of our machine learning classification model. It reveals the correctly classified and misclassified values made by our classification model. The matrix is NxN; we use three different values of N, which are 2, 3 and 4.

IV. Result Analysis And Discussion

A. Discussion

Each ML model have been applied three times and result indicate that triple class classification accuracy is higher than binary class and four class. Random Forest shows the best classification result for triple class classification. First of all, we set two categories of the movie for finding the status a movie. For better accuracy, data have been analyzed again and set four categories of movie status.

But the accuracy was lower than the previous one. There was a big problem with the middle two categories: very good and good. For example, if one director makes two movies with the same actor, actress, villain, and director, category of one could be good and another could be very good. In this situation, this model become confused and accuracy will be low. Then the movies have been categorized into three categories and achieved more accuracy than the previous classification. The class classifier increases the accuracy by reducing the number of false predictions. Fig. 3 represents a comparison chart that shows all model's accuracy.

B. Results

The classification results are summarized in TABLE II. The result table shows accuracy, precision, recall, and f-score for all models with three different class classifications. In this research we have executed each model with three different number of classes. The amount of data for all classes is not same. In this situation accuracy is not the perfect measure that is the reason to calculate precision, recall, and f-score. Results of four class classification are poor among the three models because of more imbalanced data. In our research work, tree-based algorithms are performing better than SVM, KNN, and Logistic regression.

TABLE II. Classifications results with precision recall and f-Score

ML models	Classification	Accuracy (%)	Precision (weighted average)	Recall (weighted average)	F-Score (weighted average)
SVM	Binary Class	68.20%	64.42	64.55	64.48
	Triple Class	70.93%	66.97	67.14	67.04
	Four Class	47.39%	40.01	39.21	39.39
KNN	Binary Class	67.63%	63.33	53.86	49.04
	Triple Class	70.97%	70.15	68.01	68.06
	Four Class	43.93%	32.30	34.02	32.92
Logistic Regression	Binary Class	64.73%	59.83	62.76	59.37
	Triple Class	73.25%	69.74	67.91	68.52
	Four Class	46.82%	36.28	37.02	36.51
Decision Tree	Binary Class	75.14%	72.07	73.01	71.42
	Triple Class	80.23%	76.86	78.15	77.23
	Four Class	49.13%	41.005	40.97	40.85
Random Forest	Binary Class	70.52%	69.60	58.59	57.40
	Triple Class	84.88%	83.39	80.84	81.81
	Four Class	46.82%	35.39	37.08	36.09

C. Comparison graph

The comparison graph indicates accuracy of three different classes for five ML models. One thing is common for all models, that is the triple class achieves highest accuracy, binary class achieves second highest but accuracy status of four class is always being lower among all. After analysis it has been found triple class can predict the movie status more accurately than other classes. Because accuracy and performance of triple class classifier is better than the other class classifiers. The performance and accuracy of four class classifiers is lowest than other two classifier because of imbalanced data and more false prediction.

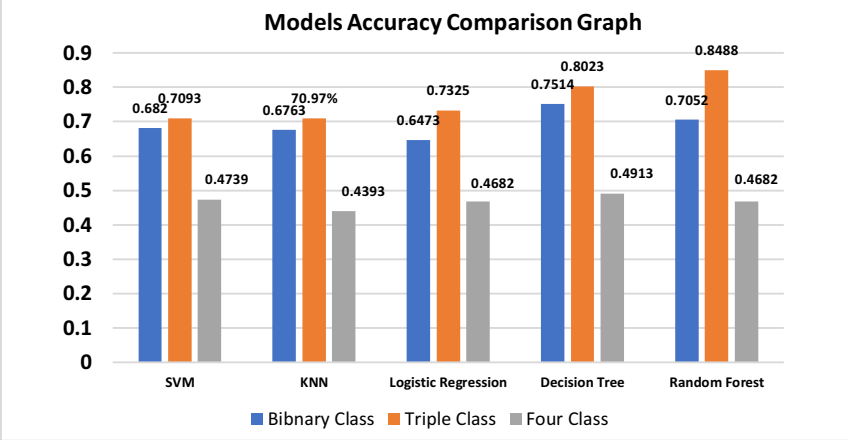

Fig. 3. Five models accuracy comparison graph

V. Conclusion And Future Scope

The purpose of this research is to evaluate the status of Dhallywood movies. In this research, one major problem has been addressed and that is the unavailability of data. To solve this problem, we apply the different techniques to manage the available data for a better outcome. We compare our target class with rating and find out a very little affinity between rating and movie status. In previous many of the studies, predicted success was based on the box office revenue, user review, etc. This study focuses on informative data analysis to accurately predict movie status. In this research, we have found a huge impact of actor, actress and villain in the Dhallywood movies. After modifying dataset with these features, the accuracy has been increased. This study prepares the platform for researchers who wants to go for further investigate and discover other facts and features to measure the movie's popularity. Besides that, the popularity of movies is not only depending on the features of a particular movie but also depends on the culture, thought, taste and choice of people of a certain country. That's the reason, a movie is excellent, very good, good or bad, it's totally depends on viewers acceptance. So, it is necessary to find out influenced features those are make a movie popular. Movie audiences' number rely on many parameters, like the current situation and economic stability of a country. If proper data, proper rating, and revenue can be found then in future the accuracy will be improved. In future, this research work can be extended by increasing both the number of movies and effective features in the dataset and some other ML models can be applied to the movie data and also can establish a constructive compression between existing work results with the future one.

References

[1] Jeffrey S. Simonoff and Ilana R. Sparrow "Predicting movie grosses: Winners and losers, blockbusters and sleepers" vol. 13, no. 3, pp. 15–24, 2000.

[2] S. Gopinath, "Blogs, Advertising, and Local-Market Movie Box Office Performance"vol.59, no.12, pp. 2635–2654, 2013.

[3] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, "Movie Success Prediction Using Data Mining," pp. 2015–2018, 2017.

[4] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," *2018 First Int. Conf. Secur. Cyber Comput. Commun.*, pp. 385–390, 2018.

[5] N. Quader, "A Machine Learning Approach to Predict Movie Box-Office Success," pp. 22–24, 2017.

[6] A. Bhave, "Role of Different Factors in Predicting Movie Success," vol. 00, no. c, 2015.

[7] L. Chen, C. Chi, and L. Huang, "Exploring contextual factors from consumer reviews affecting movie sales : an opinion mining approach," *Electron. Commer. Res.*, no. 0123456789, 2019.

[8] L. Zhou, "Movie Review Mining : a Comparison between Supervised and Unsupervised," vol. 00, no. C, pp. 1–9, 2005.

[9] 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 Prediction," pp. 933–937, 2015.

[10] A. Samad, H. Basari, B. Hussin, I. G. Pramudya, and J. Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization," *Procedia Eng.*, vol. 53, pp. 453–462, 2013.

[11] R. Niraj and J. Singh, "Impact of user-generated and professional critics reviews on Bollywood movie success," *Australas. Mark. J.*, pp. 1–9, 2015.

[12] M. Lash, S. Fu, S. Wang, and K. Zhao, "Early Prediction of Movie Success — What , Who , and When," vol. 1, pp. 345–349.

[13] V. Subramaniyaswamy, V. V. M, V. P. R, and R. Logesh, "Predicting Movie Box Office Success using Multiple Regression and SVM," *2017 Int. Conf. Intell. Sustain. Syst.*, no. Iciss, pp. 182–186, 2017.

[14] G. Verma and H. Verma, "Predicting Bollywood Movies Success Using Machine Learning Technique," *2019 Amity Int. Conf. Artif. Intell.*, pp. 102–105, 2016.

[15] T. Yasseri, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," vol. 8, no. 8, 2013.

[16] Athira M D and Lakshmi K S "Movie success prediction using ensemble classifier," pp. 20–24, 2020.

[17] M. H. Latif and H. Afzal, "Prediction of Movies popularity Using Machine Learning Techniques," vol. 16, no. 8, pp. 127–131, 2016.

[18] M. M. Rahman and D. N. Davis, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets."

[19] M. H. Cartwright, M. J. Shepperd, and Q. Song, "Dealing with Missing Software Project Data," pp. 1–12, 2003.