International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206461

Design of a Movie Review Rating Prediction (MR2P) Algorithm

Oluwatofunmi Adetunji[1], Mamudu Hadiza[2], Nzechukwu Otuneme[3]
[1]Software Engineering Department, Babcock University, Nigeria
[2]Computer Science Department, Babcock University, Nigeria
[3]Computer Science Department, Wesley University, Nigeria

Article Info
Volume 6, Issue 4
Page Number: 423-432
Publication Issue :
July-August-2020

Article History
Accepted : 20 July 2020
Published : 10 Aug 2020

ABSTRACT

Entertainment is no longer just anything that we enjoy occasionally, with over two million spectators a day, the amount generated by the movie industry is huge. The movie sector is one of the biggest contributors to the entertainment industry's unpredictability in success and failure. The aim of this research work to design an efficient movie recommendation algorithm that will increase prediction accuracy, the Movie Review Rating Prediction (MR2P) was achieved through a systematic review of the existing movie success algorithm. This research work will enable movie stakeholders (producers, directors, crew, cast already in the movie industry or aspirants) to know the kind of movie to invest in which will, in turn, be beneficial in terms of higher profit.

Keywords : Movie Rating, Prediction Algorithm, Movie Success Prediction, Entertainment

I. INTRODUCTION

In the early 10,000 BC activities such as public executions, archery, and sword skills were used as entertainment forms, it soon moved to radios link for communication purposes. The pioneers of broadcast radio saw the benefit of the technology in providing one-way information, including entertainment content, to an oversized audience, and developed further with prerecorded music, vinyl records, magnetic tapes, and later on compact discs (CD) and other digital recording media (Ng, 2012).

Entertainment is no longer just anything that we enjoy occasionally, on an evening out or night in. In our connected world, entertainment is now at the tip of our fingers, and all around us, all the time (Quail, Razzano, & Skalli, 2007). Pre-produced entertainment can offer consumers pleasure by providing them access to filmed content such as fiction movies and series, documentaries, video clips, written content such as novels and poems, recorded content such as pop songs, classical compositions, movie soundtracks, and programmed content such as console games, massively multiplayer online games (MMOGs), and smartphone games (Hennig-Thurau, Houston, Hennig-Thurau, & Houston, 2019). The filmed content with special emphasis on movies will be considered in this project. It is believed movies will still be popular within the next thousand years

(Bhave, Kulkarni, Biramane, & Kosamkar, 2015), this is because it provides a way to relax and be in a world of your own with a genre that you love to read or see.

With over two million spectators a day, the impact of the film industry is formidable (Meenakshi, Maragatham, Agarwal, & Ghosh, 2018). The success of a movie is discovering what makes a movie successful in terms of being a major hit or a flop, the popularity factor of movie components can be used to predict the success of upcoming movies. Over the last decade, there has been a burgeoning of data due to social media, e-commerce, and the overall digitization of enterprises. These data are being exploited to make informed choices, predict marketplace trends, and patterns in consumer preferences (Subramaniyaswamy, Logesh, Chandrashekhar, Challa, & Vijayakumar, 2017). An attempt is made to predict the past as well as the future of a movie for business certainty or simply a theoretical condition in which decision making is without risk because the decision-makers (stakeholders) have all the information about the exact outcome of the decision before deciding to release the movie.

Loss of revenue by stakeholders is the main concern in the film industry. The number of available viewing logs and friendship networks is too limited to design effective recommendation algorithms for movies, thereby leading to a largely inefficient algorithm. Due to the inefficiency of some existing algorithms, there could be a loss in revenue by stakeholders, and inaccurate movie prediction mechanism. Hence, this research work is aimed at building a web application that will not only make new movies more intriguing to the general public but present an accurate recommendation algorithm for movie prediction.

This research work is aimed at designing an efficient movie recommendation algorithm that will increase prediction accuracy as well as implementing the designed algorithm. A systematic review of the existing movie success algorithm was carried out to accurately design a Movie Review Rating Prediction (MR2P) Algorithm. The designed algorithm was further implemented using web development tools such as JavaScript (JS), Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), MySQL, and python for the programming language.

The study focused on factors such as actor, actress, director, movie, marketing budget, and release date, all these factors were considered in the database and existing statistics on each of them were derived. The proposed work includes certain terms and conditions which states but not limited to any loss in revenue which may occur due to ratings of some movies. Ticket pricing and availability of the movie at various theatres will not affect the site, various suggestions will be made on where tickets can be purchased. This research work will enable stakeholders (producers, directors, crew, cast already in the movie industry or aspirants) to know the kind of movie to invest in which will, in turn, be beneficial in terms of higher profit, furthermore knowing the futuristic success rate will help to know what content needs to be improved on.

II. LITERATURE REVIEW

2.1 Movie Success Rate

The film industry is one of the biggest contributors to the entertainment industry's unpredictability in success and failure (Raj & Aditya, 2017). Because of quick digitization and the rise of internet-based life the film business is developing significantly as the average number of movies produced per year is greater than 1000, therefore to make the movie profitable, it becomes a matter of concern that the movie succeeds (Bhave et al., 2015). The success rate is the fraction or percentage of success among several attempts, and also, the average task success rate can be calculated either per participant or per task that users complete correctly (Nielsen, 2006). Neural Networks have been extensively used in forecasting and prediction studies, it can, therefore, be employed for predicting the success and failure of the movies also (Sharma & Kaur, 2013). This study brings the understanding that the prediction of movie success is indeed possible with high percentages of accuracy, therefore, using a prediction engine, producers can evaluate beforehand if the movie is worth investing in and accordingly make their decisions. The accuracy of a predictive model depends a lot on the extraction and engineering of independent variables. When it comes to studying movie success, three types of features have been explored: audience-based, release-based, and movie-based features.

Factors Affecting Movie Success Rate

The following are the features of movie success rate according to (Lash & Zhao, 2016):

a) Audience Based Factors: This is about the audiences' potential reception of the movies. The more excited and hyped the audience is about a movie, the higher the revenue return is going to be.
b) Release Based Factors: This focuses on the availability of a movie and the time of its release. Many movies are targeted for release at peak times such as summer and other holiday breaks.
c) Movie Based Factors: This is directly related to the movie itself, the genre, the cast. The more popular the star actor, the higher the probability for movie success.

2.2 Review of related works

(Gaikar, Solanki, Shinde, Phapale, & Pandey, 2019), addressed the shortcomings of limited research in forecasting the power of social media in India by using Twitter data to predict the performance of Bollywood movies. Sentiment analysis and prediction algorithms were used to dissect the presentation of Indian films dependent on information acquired from web-based life locales. The creators utilized Twitter4j Java API for separating the tweets through verifying association with Twitter sites and put away the extricated information in MySQL database and utilized the data for sentiment analysis. The researchers were able to find out that the study suffers from the limitation of not having enough computing resources to crawl the data. The data mining technique for analyzing and predicting the success of the movie was addressed by (Meenakshi et al., 2018). The study aimed at developing a system based on data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain levels of uncertainty. An attempt is made to determine the past as well as the future of movies with the end goal of achieving movie success because investing is without undue risk. Through data collection, data cleaning, data transfer, data analysis, and prediction the researchers developed the best-suited algorithm to obtain data from IMDB. From the results, the researchers found that it was hard to apply data mining methods to the data in the IMDb dataset.

To help in predicting the success of movies in advance, (Kumar, Mehta, & Joy, 2019) aimed at developing a model-based system upon the data mining techniques. The methodology dealt with different stages of the project which consists of data collection, data pre-processing, generating training and testing dataset, model generation, prediction, and outcomes. The researchers predicted the success of their movie based on critics' scores. It was concluded that critics score is the best predictor of audience scores. However, the ratings gotten from the critics or audience were not so reliable.

(Shah, Kapadia, Samel, Saple, & Deshmane, 2019) developed a model, capable of predicting the box office financial success of a certain set of movies through specific variables and historical data. It was conceivable to presume that the level of accomplishment of the cinematographic income expectation was extraordinarily dependent on the typology of the needy variable utilized in the investigation. By collecting the number of likes, dislikes, and the view count of the trailer, release date, star ranking, Multiple Linear Regression algorithm was used on these features and the result was compared to an output that is already known to determine the accuracy. The observational model showed great measurable outcomes when the dependent variable was binary and interval. Nonetheless, for the multiclass forecast, the outcomes were a long way from the real world, adversely affecting the model.

The research on analyzing social media community sentiment Score for prediction of success of Bollywood movies was carried out by (Ranjan & Sood, 2018). The research aimed to develop a model used in predicting the box office collection for Bollywood movies. The model presented in the study used betweenness centrality to detect communities in the datasets and the sentiment analysis of the largest community. When compared with actual Box-office collection, it gives 80% accuracy in categorizing movies as a hit, flop, or an average movie. The study dissected the issues of community detection and sentiment mining from film tweets to foreseeing the film execution as a hit, normal, or flop. Although the forecast proved to be accurate it was not possible to conduct larger research because there is not a solitary site that gives this information to all the Bollywood motion pictures. In conclusion, these various researches suffered from such gaps as not having enough data sets resources to crawl the data leading to inefficient algorithms and negative effects on prediction models. In a bid to bridge these gaps the research work aims at developing an efficient algorithm that will increase prediction accuracy.

III. REVIEW OF RELATED ALGORITHMS

An algorithm is a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation (Patel, 2018). Computers use algorithms to list the detailed instructions for carrying out an operation. In terms of efficiency, various algorithms can accomplish operations or problem solving easily and quickly. Below are the reviews of some existing movie algorithms.

A. **Linear Regression Algorithm** (Shah et al., 2019)

The algorithm aimed to collect the number of likes, dislikes, and the view count of a trailer, release date, star ranking, and so on. Multiple Linear Regression Algorithm was used for the prediction of earnings of the movie. WEKA tool was used for choosing the best algorithm. Once the movie is released, we use its IMDb ratings and actual first-day collection to predict the lifetime collection of the movie. The goal was to define a relationship between the prediction value and the features by solving for the linear coefficients, θ that best map the features to the prediction value. Where the ratings have been collected in a vector Y. Y is a (m x 1) vector (where m=50000). The movie set was to be pruned to select a set of features that have been found to make a major impact on the success or failure of a film. After the identification, all the producers, directors, actors, and actresses were rated based on their past performance at the Box Office. Sentiment analysis was carried out on the tweets after ‘noise’ was removed.

Algorithm (Steps of System Flow)

Input: Movie database, User input film with feature values

Output: Rating of user-entered film

Step 1: Feature selection: This is the process of automatically or manually selecting the movie features which contribute most to the prediction variable.

Step 2. Feature normalization: This is the process of making the value of each feature standardized.

Step 3. Apply Multivariate Linear regression model on the dataset: Multivariate Linear Regression (LR) is when more than one predictor variable is used.

Step 4. Get prediction result

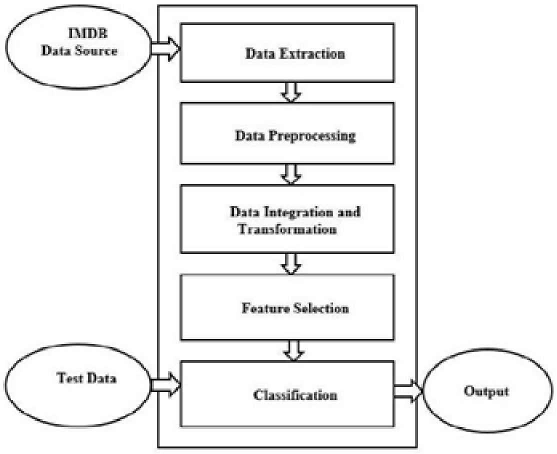

Figure 1 Linear Regression Algorithm (Shah et al., 2019)

This goal of this research work was to develop a model, able to predict the box office financial success of a certain set of movies through specific variables and historical data. The empirical model demonstrated good statistical results when the dependent variable was binary and interval. However, in regards to the multiclass prediction, the results were very far from reality, negatively influencing the model

B. **Weightage Algorithm** (Antara, Nivedita, Shalin, Tanisha, & Pranali, 2018)

In this research, a custom website and algorithms for predicting the success class of a movie such as a flop, hit, or average was developed. In doing this, a custom dictionary of words in which common and important words used in reviews were stored according to the weightage assigned to them by the administrator. With the help of sentiment analysis, weightages will be assigned on a scale of one (1) to five (5), where 1 indicates the negative extreme, and five (5) indicates the positive extreme. A rating segment to rate movies based on the parameters such as genre, star, cast, songs, and so on was developed to capture movie ratings. The average of the reviews and ratings can then be used to calculate the overall rating by calculating the mean of the two (reviews and ratings). This can be mapped to possible outcomes, such as 'Hit', 'Flop', and 'Average'.

Algorithm (Steps of System Flow)

Input: User reviews and ratings for Actors, Genres and Songs

Output: Rating values in form of Emojis

```
Step 1: Users rate movie parameters like Genre, Acting, and so on.
Step 2: User enters Review for the movie.
Step 3: For each Word in Review
    Step 3.1: If Word is present in the custom dictionary, then assign the corresponding weightage.
    Step 3.2: Else, if Word doesn't match the words in the custom dictionary then assigned a default value.
Step 4: Calculate weightage of the review as
    Weightage of Review = Weightage of each word/ number of words
Step 5: Calculate the mean weightage of all reviews
Step 6: Find the overall rating as mean of ratings of movie parameters and mean weightage of all reviews.
Step 7: If the overall rating results add up to be 2 or less than 2 then it is a Flop
    If the overall rating is more than 2 but less than 4 then it is Average
```

If the overall rating is 4 or more than 4 then it is a Hit

Step 8: The rating values will be shown with the predicted result in form of Emojis.

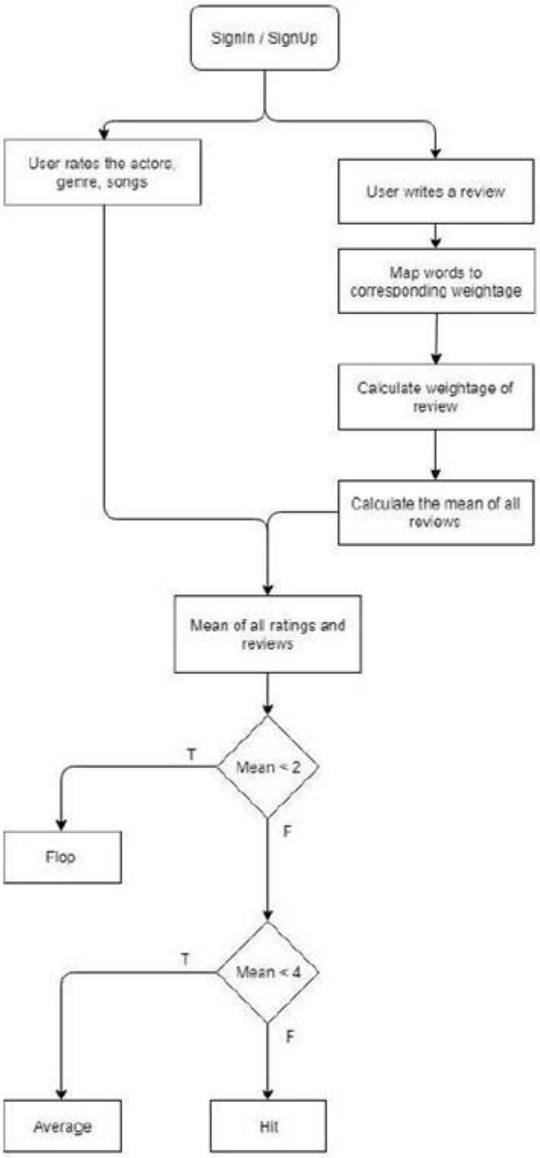
SignIn / SignUp
User rates the actors, genre, songs
User writes a review
Map words to corresponding weightage
Calculate weightage of review
Calculate the mean of all reviews
Mean of all ratings and reviews
T
Mean < 2
F
Flop
T
Mean < 4
F
Average
Hit

Figure 2 Weightage Algorithm Flow Chart (Antara et al., 2018)

The deigned algorithm which was further implemented in a web application enabled customers and movie owners to get reviews of movies within a month. This can further give them insights on what needs to be taken into consideration in future movies. However, this algorithm did not put into consideration festive periods, holidays, special weekends, and so on which can cause a surge in the number of reviews and thus affect the decision of stakeholders in future movies.

C. Random Forest by (Dhir & Raj, 2018)

The researchers wanted to dig deeper into the business side of movies and explore the economics behind what makes a successful movie. They wanted to examine the trends among films that lead them to become successful at the box office. The data set they utilized to train and test the model was gotten from *Kaggle.com* which included information about several movies on IMDb such as titles, directors, genres, countries of origin, and so on. Ensuring the dataset was representative and suitable for analysis, the dataset was prepared and structured. Lastly, the researchers made a comparison of several machine learning (ML) algorithms using the same dataset. The ML algorithms that were compared are Support Vector Machine (SVM), Random Forest (RF), Ada Boost (AB), Extreme Gradient Boost (XG Boosting) and K-Nearest Neighbors (KNN).

Step 1: Data Extraction – this involves getting the dataset from *Kaggle.com.*

Step 2: Data Pre-Processing – this involves eliminating unnecessary data from raw extraction.

Step 3: Feature Extraction – here, relevant attributes were extracted which includes general pre-production information such as genre, language, information about actors, directors, and so on.

Step 4: Feature Selection – Desired variables that served as possible predictors were selected.

Step 5: Classification Model – five ML algorithms SVM, RF, AB, XG Boosting, and KNN served as the classification model.

Step 6: Test Data Comparison – for each ML algorithm, the test data were passed to determine the one with the best level of accuracy

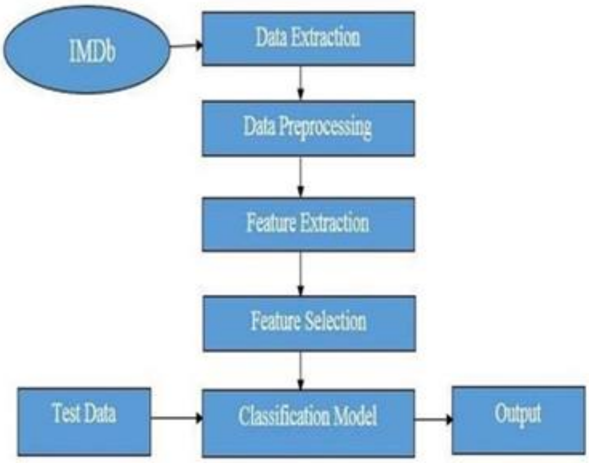

Figure 3 Algorithm Workflow (Dhir & Raj, 2018)

The result showed that Random Forest (RF) gave the best accuracy followed by Gradient Boost. RF takes a random sample of data as input and selects the subset of features randomly which makes it robust enough. Figure 4 shows the accuracy level of all models compared. However, the researchers failed to include social media sources such as comments on twitter and YouTube which could have a huge impact on the prediction of movie success.

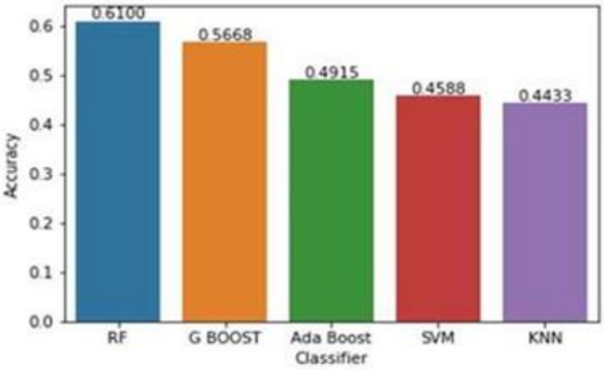

Figure 4: Bar Chart showing the accuracy of compared models (Dhir & Raj, 2018)

IV. Movie Review Rating Prediction (MR2P) Algorithm

The designed algorithm is embedded in a developed web application that enables users (movie viewers) to write reviews for movies based on certain parameters. The web application was developed using HyperText Markup Language (HTML), JavaScript (JS), and the Python programming language. The administrator's module on the other hand can upload movies, movie trailers, movie posters to the database for the general public's view. The user module has a landing page where both registered and unregistered users can preview the web application features, view reviews, watch trailers, and so on. Upon registration and successful login, users can then write a review for a particular movie as well as give a star rating to a movie. The reviews and ratings will serve as input to the algorithm which will be used in predicting the success of a movie.

Afterward, sentiment analysis will be used to computationally identify and categorize the opinions expressed by a user on a particular movie as either positive, negative, or neutral. For the MR2P algorithm to accurately handle this, a customized dictionary containing a mapping of common adjectives used in movie reviews to a corresponding value on a scale of one to five (1-5). Words such as 'fantastic', 'wow' will possess higher value compared to words such as 'disgusting', 'rubbish'. The MR2P algorithm will further assign an average value to each review which will categorically state the user's opinion on a particular movie, afterward, the average value of all reviews *(A_{review})* for a particular movie will be gotten as shown in Equation 1.

$$\mathrm{A}_{review} = \sum_{i=0}^{n} \frac{Value_i}{n} \text{ .... } \textit{Equation 1}$$

Alongside, the average rating of casts, genre, song, and the overall star rating will be gotten *(A_{rating})*. Because the number of movie viewers increases sporadically during festive periods, public holidays, and so on, the MR2P algorithm takes into consideration the period of release before the mean value is calculated. The mean of A_{review} *and* A_{rating} for a particular movie will then be used to categorize a movie as a 'Hit', 'Average', or 'Flop' as shown in Equation 2.

$$Mean\ Value = \frac{A_{reviews} + A_{ratings}}{2} \text{ .... Equation 2}$$

The outcome of this will indicate the likeability and attractiveness of viewers to the movie, based on this, the movie marketers, movie promoters, and other decision-making stakeholders can then decide whether to further invest in the movie or do otherwise. Figure 5 shows the flowchart of the MR2P algorithm.

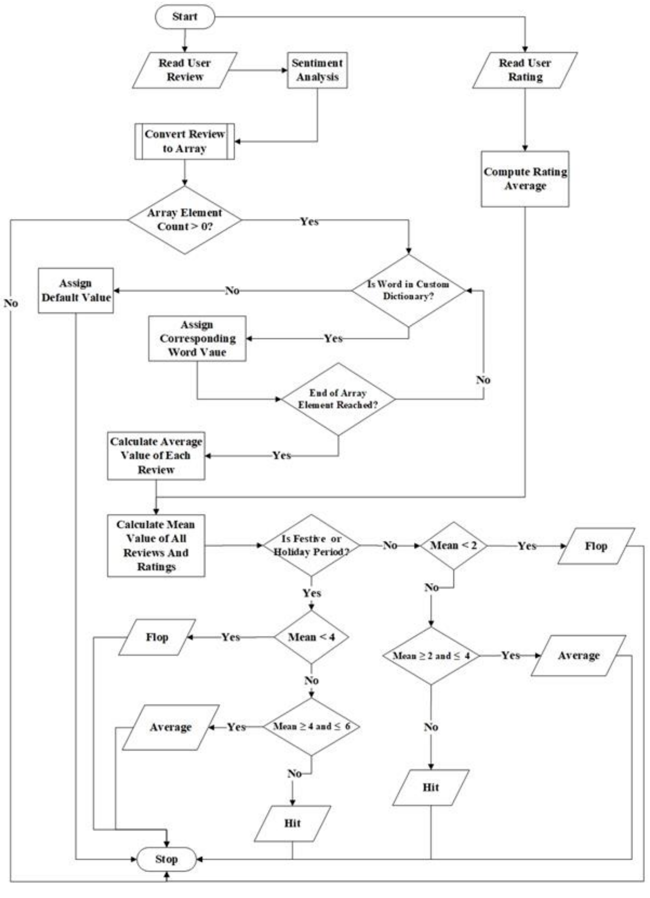

Figure 5 : MR2P Algorithm Flowchart (Researcher's Algorithm)

4.1 Steps of System Flow

Below is the system flow for the MR2P Algorithm.

```
Input: User Reviews and Star Ratings
Output: Movie Success Prediction
Step 1: Begin
Step 2: Read: Users' reviews
Step 3: Sentiment Analysis is applied to each review.
        Step 3.1: Each review is transformed into an array denoted by ARR
Step 4: For each ARR
        If word is in custom dictionary, Then:
                Set word = corresponding value
        Else
                Set word = default value
        [End of If-Else]
Step 5: Repeat Step 5 for i=0 to n
Step 6: Read ARR[i]
        Set $A_{review} = \sum_{i=0}^{n} \frac{Value_i}{n}$
        [End of for loop]
Step 7: Read: Average Star Ratings ($A_{ratings}$) of movie
Step 8: Set $Mean\ Value = \frac{A_{reviews} + A_{ratings}}{2}$
Step 9: If it is a festive period
        If mean < 4, Then:
                Write: "Flop"
        Else If mean ≥ 4 and ≤ 6, Then:
                Write: "Average"
        Else
                Write: "Hit"
        [End of If-Else-If Else]
Step 10: If it is not a festive period
        If mean < 2, Then:
                Write: "Flop"
        Else If mean ≥ 2 and ≤ 4, Then:
                Write: "Average"
        Else
                Write: "Hit"
        [End of If-Else-If Else]
Step 11: End
```

V. SUMMARY AND CONCLUSION

This project accomplished the improvement of a movie achievement expectation framework utilizing sentiment analysis. Posting reviews for purchased items and other online products has now become a popular trend for individuals to communicate assessments and assumptions to decision-makers. One of the benefits derived from the developed system is that any sort of movies such as Nollywood, Hollywood, or Bollywood can be inspected by movie stakeholders and decision-makers to confidently finance the right movie project. In future the sites can be utilized for investigating games and music shows and for exploring item deals, and so on. Furthermore, including a wider range of algorithms and algorithm configurations could also help to improve current accuracy.

VI. REFERENCES

[1]. Antara, U., Nivedita, K., Shalin, S., Tanisha, M., & Pranali, W. (2018). Movie Success Prediction Using Data Mining. International Journal of Engineering Development and Research, 6(4), 198–203.

[2]. Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015). Role of different factors in predicting movie success. 2015 International Conference on Pervasive Computing: Advance Communication Technology and Application for Society, ICPC 2015. https://doi.org/10.1109/PERVASIVE.2015.7087152

[3]. Dhir, R., & Raj, A. (2018). Movie Success Prediction using Machine Learning Algorithms and their Comparison. ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications, 385–390. https://doi.org/10.1109/ICSCCC.2018.8703320

[4]. Gaikar, D., Solanki, R., Shinde, H., Phapale, P., & Pandey, I. (2019). Movie Success Prediction Using Popularity Factor from Social Media. International Research Journal of Engineering and Technology, 6(4), 5185–5190. Retrieved from www.irjet.net

[5]. Hennig-Thurau, T., Houston, M. B., Hennig-Thurau, T., & Houston, M. B. (2019). The Fundamentals of Entertainment. In Entertainment Science (pp. 41–57). https://doi.org/10.1007/978-3-319-89292-4_2

[6]. Kumar, S., Mehta, A., & Joy, P. (2019). Movie Success Prediction using Data Mining. Vellore Institute of Technology.

[7]. Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. Journal of Management Information Systems, 33(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

[8]. Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. Journal of Physics: Conference Series, 1000(1). https://doi.org/10.1088/1742-6596/1000/1/012100

[9]. Ng, S. (2012). A Brief History of Entertainment Technologies. Proceedings of the IEEE, 100(SPL CONTENT), 1386–1390. https://doi.org/10.1109/JPROC.2012.2189805

[10]. Nielsen, B. (2006). Order Determination in General Vector Autoregressions. In Time Series and Related Topics (pp. 93–112). https://doi.org/10.1214/074921706000000978

[11]. Patel, M. (2018). Data Structure and Algorithm With C (1st ed.). Educreation Publishing, New Delhi.

[12]. Quail, C., Razzano, K., & Skalli, L. (2007). Vulture Culture: The Politics and Pedagogy of Daytime Television Talk Shows. Peter Lang Publishing Inc., New York.

[13]. Raj, M. P. M., & Aditya, S. (2017). Predictive Model for Movie's Success and Sentiment Analysis. Research Journal of Management Sciences, 6(6), 1–7.

[14]. Ranjan, S., & Sood, S. (2018). Analyzing Social Media Community Sentiment Score for Prediction of Success of Bollywood Movies. International Journal of Latest Engineering and Management Research, 3(2), 80–88.

[15]. Shah, K., Kapadia, J., Samel, Y., Saple, S., & Deshmane, P. (2019). Movie Success Prediction using Data Mining and Social Media. International Research Journal of Engineering and Technology, 6(3), 188–190.

[16]. Sharma, P., & Kaur, M. (2013). Classification in Pattern Recognition: A Review. International

Journal of Advanced Research in Computer Science and Software Engineering, 3(4), 2277–128. Retrieved from www.ijarcsse.com

[17]. Subramaniyaswamy, V., Logesh, R., Chandrashekhar, M., Challa, A., & Vijayakumar, V. (2017). A Personalised Movie Recommendation System Based on Collaborative Filtering. International Journal of High Performance Computing and Networking, 10(1–2), 54–63. https://doi.org/10.1504/IJHPCN.2017.083199

Cite this article as :

Oluwatofunmi Adetunji, Mamudu Hadiza, Nzechukwu Otuneme, "Design of a Movie Review Rating Prediction (MR2P) Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 423-432, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206461
Journal URL : http://ijsrcseit.com/CSEIT206461