Movie Success Prediction Using Data Mining

[1]Antara Upadhyay, [2]Nivedita Kamath, [3]Shalin Shanghavi, [4]Tanisha Mandvikar, [5]Pranali Wagh

[1,2,3,4]B.E Student, [5*]Asst. Professor

Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India

***Abstract*— The success rate of a movie is largely dependent on people's perception and opinions, through word-of-mouth. In today's digital world, word-of-mouth is prevalent in the form of reviews found online. The opening day and the first few weeks following a movie's release is crucial, and hence production houses place a lot of importance on moviegoers' opinions, and develop trailers and publicity strategies to sway public opinion. Taking this into account, the paper proposes the use of a review system for predicting the success rate of a movie. Moviegoers' opinions of a movie before and after the release of the movie will be determined using sentiment analysis. A custom dictionary will be developed comprising words commonly used in movie reviews, which will be mapped to their corresponding weight-age in order to score reviews on a scale of one to five, and accordingly classify the success rate of movies.**

***Keywords*—Movie Reviews; Sentiment Analysis; Classification.**

I. INTRODUCTION

A movie revenue depends on various components such as cast acting in a movie, budget for the making of the movie, film critics review, rating for the movie, release year of the movie, etc. Because of these multiple components there is no formula that helps us to provide analysis for predicting how much revenue a particular movie will be generating. However by analysing the revenues generated by previous movies, a model can be built which can help us predict the expected revenue for a particular movie. Such a prediction could be very useful for the movie studios which will be producing the movie so they can decide on different expenses like artist compensations, advertising of the movie, promotions in various cities, etc. accordingly. Plus it allows investors to predict an expected return-on-investment (ROI). Also, it will be useful for many movie theatres to estimate the revenues they would generate from screening a particular movie.

Now a day's, online review system has become one of the most important part of any business approach. Posting reviews online for products bought or services received has become a trendy approach for people to express opinions and sentiments, which is essential for business intelligence, vendors and other interested parties. Social media contains rich information about people's preferences. Our study proposes a decision support system for movie investment sector using data mining techniques.

In this research, we will be using our own customised dictionary where different words that users commonly use in reviews will be grouped together and will be assigned a specific rate based on the admin's choice. According to the calculated rate we will classify the movie into hit, average or flop. Through this project we aim to provide a data mining algorithm which gives the most accurate result for movie success prediction.

II. RELATED WORK

Javaria Ahmad et al. came up with a mathematical model to find the success rating of upcoming movies based on certain factors. [1]. The goal of this mathematical model was to provide a precise prediction of success, hence providing confidence to take holders in their investments. 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. Various variables were studied to provide the movie success prediction. Some of these variables 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. Their proposed model consisted of an algorithm that involved finding correlation between these various attributes using X2 analysis. The expected frequency for the correlation between genre and ratings was calculated as 64.39 with degree of freedom as 8.The correlation between actors and ratings was 11.57 with its degree of freedom being 1 whereas the correlation and degree of freedom between actors and genre was found to be 20.6 and 1 respectively.

Xiaohui Yu et al. used two models to predict sales performance of movies using online reviews [2]. They have used the Sentiment Probabilistic Latent Semantic Analysis (S-PLSA) model to summarize the sentiment information from online reviews and tweets and the ARSA model for predicting sales performance of movies using sentiment information and past box office performance. They predicted the success rate of movies by analyzing data from IMDB reviews and tweets. Their sales prediction for the movie "The Da Vinci Code" was approximately $66662.6 on the 1st day,$61102.31999999999 on the 2nd

day an $59990.26 on the 3rd.They further classified the reviews into positive, negative and neutral after that they set a simple metric P-N ratio and threshold value to predict the success of movies, i.e. Hit, Average and Flop.

N. Quader et al. tried to predict a movie's box office success using SVM and Neural Networks [3]. They calculated two types of predictions, an exact match which refers to correct classifications and one away predictions which means taking consideration of one class up or down from a particular type along with the exact match. They predicted the success rate of movies by analysing data from IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. 15 features, including 6 pre-release features were used, such as critics' ratings, audience's ratings, star power, etc. Sentiment analysis of reviews was done using Microsoft Azure's Text Analytics API. Four kernels were used in SVM, from which linear kernel and Gaussian Radial Basis (GRB) kernel gave one away accuracies of 88.87% and 87.54% respectively, and exact accuracies of 56.16% and 55.36% respectively. Multi-Layer Perceptron (MLP) Neural network analysis gave exact prediction accuracy of 58.41% and one away prediction accuracy of 89.27%.It was hence concluded that MLP Neural network gave better predictions, and the most important pre-release features, in the order of importance, were the movie's budget, number of screens and release month.

Mehreen Ahmed et al. presented the comparison of Conventional Features with Social Media features in determining the popularity of movies[4].They considered a number of conventional features collected from IMDB such as Genre, Budget, Number of Screens and Sequel and other features taken from social media such as YouTube and Twitter. A number of experiments were performed using Linear Regression, Decision Tree (J48), Artificial Neural Network and Support Vector Machine. The best performance was calculated using J48, where sentiment score came up as the best discriminating attribute.They achieved best value of 77% and 61% using selected social media features for Rating and Income prediction respectively. While selected conventional features gave results of 76.2% and 52% for rating and income prediction. Moreover, while predicting the Rating using Linear Regression 95% accuracy was calculated. Their method showed that social media features such as Sentiment Score of tweets related to movies, Number of Views and Comments of movies' trailers on YouTube and twitter fan following can usefully be utilized to predict the popularity of movie. They had assumed that popularity is depicted by movie rating and gross income, and performed two set of experiments to predict these features individually. It was found that Social media features are better on conducting both set of experiments.

S. Shim et al. predicted a movie's opening weekend revenue using Twitter data [5]. Tweets were collected using Twitter's REST API. Features such as weather data from Weather Underground, number of theatres a movie was released and movie budget from Box Office Mojo were used. Twitter based features that were used were the number of tweets, re-tweets and favourites for a movie, positive and negative words used in the tweets, number of capital letters and punctuation marks, and emojis used. Linear Regression was used to predict the revenue and Cross Validation Leave One Subject Out (CV-LOSO) was used to test the accuracy of this model. On performing a correlation-based feature selection step to find and use the best set of features, 65% accuracy was achieved. In another attempt to improve the accuracy of prediction, an unsupervised layer was added ahead of the supervised predictive model and multiple predictive models were developed. For this, an unsupervised clustering stage was used to cluster the data samples into several groups, after which the best set of features for each cluster was extracted and an individual supervised regressor was trained for each cluster. The number of capital letters in each tweet and theatre count were found to be the clustering features providing the best accuracy, which was 75%.

Wutao Wei et al. used a hybrid method based on Bayesian forecasting theory and classic statistical learning techniques to predict box office revenue[6].The proposed method considered both movie attributes and actual office box data. Based on box office revenue data of a selected set of movies, a dynamic linear model (DLM) was trained, to forecast box office revenue of new movies. The DLM based method used in this work was augmented by a Bayesian framework and a dynamic filtering method such that model coefficients could be updated timely with actual box office data. The DLM trained with historical data accurately depicted dynamic dependencies on office box time series data, given classic attributes such as director, cast, etc. The modelling strategy was to primarily eliminate the reliance on non-salient features derived from other sources which were noisy to fit in an accurate model, and to further optimize model performance regarding actual incoming revenue data. The model fitting approach was in a Bayesian way so it did not depend on auxiliary data sources such as social media information which was usually difficult to cleanse. Compared to other time series modelling techniques, the proposed method took prior knowledge into consideration, which allowed an early prediction of real-time box office.

Nahid Quader has used various machine learning classification methods which they implemented on their own movie dataset for multi class classification [7]. Comparison among various machine learning methods is the main goal of this paper. They implemented most of the algorithms using python library Scikit Learn. The seven machine learning tool that they used are: Logistics Regression, Support Vector machine, Random Forest, Gaussian Naïve Bayes, Adaboost, Stochastic Gradient Descent, Multilayer Perceptron Neural Network. Multilayer Perceptron Neural Network is the most powerful machine learning method among all of these giving an accuracy of 58.53%. MLP can handle very complex data pattern where other models are unable to detect any pattern and very good for a prediction model .All of these methods predict an approximate net profit value of a movie by analysing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Meta Critic. Based on some pre-released features and post-released features the system predicts a movie box office profit for all the seven

methods. This paper analyses the performance assessment of all these seven machine learning techniques based on our own dataset which contains 755 movies. Multilayer perceptron Neural Network gives better result among the seven algorithms used.

Quazi Ishtiaque Mahmud differentiated between positive and negative comments using Support Vector Machine and then used Statistical Reasoning to predict movie success [8]. By analysing public sentiment they could extract useful information for many things. They used a non linear RBF kernel for their sentiment classifier which achieved better accuracy than the classifiers that used linear kernels. Using this system they could predict whether a movie will be successful or not with an accuracy of 90.3%. Using their sentiment classifier they achieved stable results in both of the popular datasets for movie reviews thus thier classifier is fairly stable and it's not prone to a specific dataset.

Adarsh Tadimari et al. aimed at investigating the impact of the content of a movie's trailer on its initial financial success[9].They created a database consisting of movie trailers, and various metadata associated with them. The success metric that they choose was a movie's box office collection in the first weekend. They collected trailers for 474 American movies, released during the years of 2010 to 2014. For each movie, they used only a single trailer, even though multiple trailers were available. Their database contains trailers of 111 movies from 2010, 104 from 2011, 102 from 2012, 80 from 2013 and 77 from 2014. These were the movies for which the estimated budget and the opening weekend sales information was available on the Internet Movie Database (IMDb).They also obtained metadata for 474 movies from IMDb. They used Linear Support Vector Machine(SVM)to predict the category 22of genre with features from the audio-visual content using 5x2 Cross Validated Paired t-test. They have shown that taking into account the content information present in movie trailers, can improve the current state-of-the-art in success prediction of movies. They have designed a number of audio-visual features, and have shown that these features especially the visual features were as powerful as some of the metadata, such as genre and MPAA rating.

M. S. Rahim et al. attempted to predict the gross income of movies by mining data relating to movies' trailers on YouTube [10]. The data was collected from Box Office Mojo using the Html Agility Pack(HAP) for crawling the website, and from YouTube using the YouTube Data API, following which Open Refine, a powerful tool for data processing, was used to remove inconsistencies from the dataset. Various movie related features such as year of release, genre, film studio, opening weekend income, total income and release date were used. Additionally, movie trailer related features such as upload date, number of views, likes, dislikes, comments, like : dislike ratio and age of the trailer were used. To improve the fitness of the regression analysis, two models were created: one using Spearman's rho correlation technique and the other using Pearson's correlation method. Linear Regression, Polynomial Regression, Gradient Boosted Tree and Simple Regression Tree were applied. It was found that Linear Regression and Gradient Boosted Tree were the most suitable of the four methods, for both the models.

III. METHODOLOGY

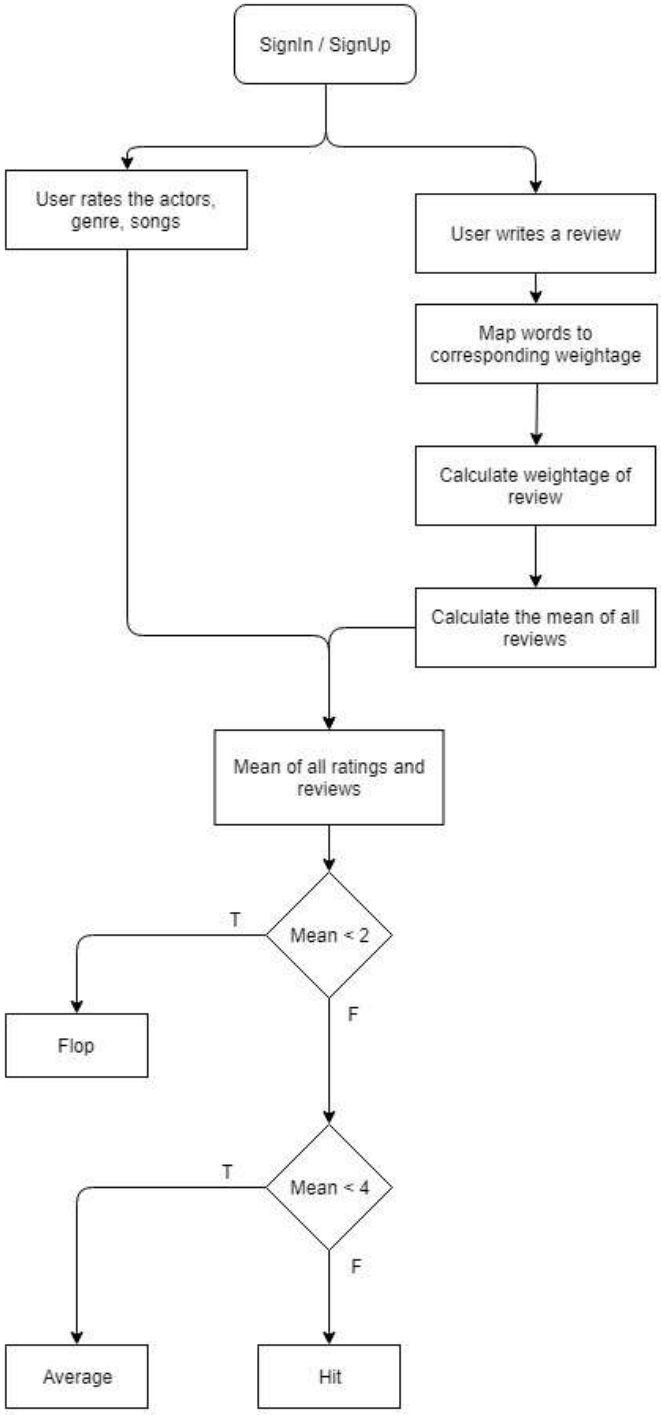
The proposed system is a website where registered users can write reviews for movies. The website will be developed using HTML, CSS and JavaScript on the front-end, and PHP and MySQL on the back-end. It will allow uses to register using their name, mobile number and password. The password will be securely hashed using the MD5 hashing algorithm.

The administrator of the website will be able to add movies which can be reviewed by users. The movie's poster as well as a link to the movie's trailer will be given. Users will be able to type in their reviews as well as rate the movie on a scale of one to five stars. The system will then use the rating and the review for predicting the success of the movie. Sentiment analysis will be used to gauge the opinions of the reviewers.

Sentiment analysis is used to quantify opinions about a particular topic or product. For this, the proposed system will assign weightages to words which are commonly found in reviews, such as 'good', 'boring', etc. Weightages will be assigned on a scale of one to five, where one indicates the negative extreme and five indicates the positive extreme. For example, words like 'fantastic', 'brilliant' and 'fascinating' will be assigned a weight of five, while words like 'worst', 'pathetic' and 'disgusting' will be assigned a weight of one. In this manner, an algorithm will be developed which will determine the score of each sentence, and then of each review, to determine the user's opinion of the movie. Thus, a custom dictionary will be used which will map adjectives commonly used in reviews to their corresponding weights. An average of scores assigned to each review of a movie will be taken. Additionally, users will also be asked to rate various movie parameters like genre, star cast, songs, etc. on a scale of one to five. The average of these ratings will then be calculated.

Finally, the average of the reviews and the average of the ratings will be used to calculate the overall rating by calculating the mean of the two. This will be mapped to possible outcomes, such as 'Hit', 'Flop' and 'Average' which will indicate the movie's likeability and hence, the rate of success.

FLOWCHART



ALGORITHM

STEP 1 : User rates movie parameters like Genre, Songs, Acting ,etc.

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 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.

IV. Conclusion And Future Scope

In this project we will prepare a custom website and algorithms for predicting the success class such as flop, hit, average of the movies. For doing this we will prepare a custom dictionary of words in which the common and important words used in reviews are stored and according to the weightage of the words the ratings will be predicted. Based on the weight age of these words the movie will be labelled as average, hit or flop. This application will help us find out the review of the new movie within a week or month. Since we are preparing a custom dictionary the results will prove to be accurate.

To achieve this, individual moviegoers can be treated as potential customers and then careful data research will be undertaken to ascertain which of these potential customers are most likely to influence the opinion of others. In this regard, it will be beneficial to factor in both prospective audience members and also theatre owners, who might have their own strategies regarding the scheduling of particular movies to expand their occupancy and profits. Another element which will be factored into this analysis to achieve more accurate results is seasonality. A detailed study can be provided into the fortunes collected from various movies that are released on a day which coincides with specific events like major festivals, bank holidays or weekends. Usually, every week has several different movies being released, yet people are more likely to opt to watch just one movie over a particular week rather than watching all movies. Therefore, there will be a need to look into ratios rather than absolute values to derive accurate results for prediction which not only drive profit, but also play an important part in the various ways of attracting people to opt watching a specific movie instead of any other movie, which includes marketing and promotional strategy of the movie. Due to this methodology, user can easily decide whether to book the ticket in advance or not. The long-term gain from this approach is that any kind of movie like Hollywood, Bollywood, etc can be reviewed on the website. In future our website can be used for reviewing sports events and music concerts and also for reviewing product sales, etc.

V. References

[1] Javaria Ahmad,Prakash Duraisamy,Amr Yousef,Bill Buckles, "*Movie Success Prediction Using Data Mining*", in 8th International Conference On Computing, Communication And Networking Technologies (ICCCNT),3 -5 July, 2017.

[2] Xiaohui Yu, Yang Liu, Jimmy Xiangji, Huang Aijun An, "*Mining Online Reviews and Tweets for Predicting Sales Performance and Success of Movies*", in International Conference on Intelligent Computing and Control Systems (ICICCS) 2017.

[3] N. Quader, Md. O. Gani, D. Chaki, Md. H Ali, "*A Machine Learning Approach to Predict Movie Box-Office Success*", in 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017.

[4] Mehreen Ahmed,Maham Jahangir,Dr. Hammad Afzal,Dr. Awais Majeed,Dr. Imran Siddiqi,"*Using Crowd-source based features from social media and Conventional features to predict the movies popularity* ", in IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2,2015.

[5] Shim, Steve, and Mohammad Pourhomayoun. "*Predicting Movie Market Revenue Using Social Media Data.*" 2017 IEEE International Conference on Information Reuse and Integration (IRI) (2017).

[6] Wutao Wei,Le Zhang,Qi Ding,Bingrou Zhou,"*Dynamic Bayesian Predictive Model for Box Office Forecasting*", in IEEE International Conference on Big Data (BIGDATA),2017.

[7] Nahid Quader , Md. Osman Gani, and Dipankar Chaki, "*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), 7-9 December 2017, Khulna, Bangladesh

[8]Quazi Ishtiaque Mahmud , Asif Mohaimen, Md Saiful Islam, Marium-E-Jannat, "*A Support Vector Machine mixed with statistical reasoning approach to predict movie success by analyzing public sentiments*", in 2017 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017

[9] Adarsh Tadimari, Naveen Kumar,Tanaya Guha,Shrikanth S. Narayanan, "*Opening Big In Box Office? Trailer Content Can Help*", in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2016.

[10] Rahim, Md Shamsur, A. Z M Ehtesham Chowdhury, Md. Asiful Islam, and Mir Riyanul Islam. "*Mining Trailers Data from Youtube for Predicting Gross Income of Movies."* 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (2017).