ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.2

Movie Success Prediction based on Classical and Social Factors

Sachin Darekar[1], Pratik Kadam[2], Prajakta Patil[3], Chinmay Tawde[4]
Assistant Professor[1], BE Student[2, 3, 4]
Department of Information Technology
Bharati Vidyapeeth College of Engineering, Navi Mumbai, India

Abstract:
Like many innovations, the movie industry has been driven by advances in technology and is mainly dependent on customer approval and response. Social media such as Twitter, YouTube, IMDb , Wikipedia, etc are major platforms where people can share their views about the movies. Along with these social factors the integration of classical factors such as director, producer, cast, runtime and genre play a major role and affect the popularity of the movie. Thus, the overall success of an unreleased film can be accurately predicted by considering the classical factors as well as the user anticipation or feedback through social media channels. The classical as well as the social factors play a vital role in the success of movie. Prediction Models for the success of a movie will help to improve the business significantly. So, the blend of classical as well as social factors for movie success prediction is studied in this paper. This proposed system will help to achieve higher accuracy rate for movie success prediction which in-turn will be useful for users for better decision making.

Keywords: Classical factors, user anticipation, Preiction Models, social factors.

I. INTRODUCTION

Film industry is growing rapidly, whether it be a Hollywood film or a Bollywood movie. The movie industry worldwide produces a large number of movies every year. There is great evolution in movies due to digitization. A movie has to do well on Box Office in order to be profitable. However very few movie has been of great interest to all the economists as well as financial experts. Most of the studies performed for the prediction of movies use conventional attributes. This data is collected from online movie databases. The availability of this data from various social platforms like YouTube, IMDb and Wikipedia helps us to gauge society's reaction towards a particular movie. There are also various classical factors like director, producer, cast, runtime and genre that pay a vital role in determining a movie's success. The use of only social factors or only classical factors will fail to attain the required accuracy level for prediction of movie success. The classical factors are only used for quality analysis of a movie. While the social media responses are required to observe public anticipation and feedback towards a particular movie. The current predictive models available are based on various factors for assessment of the movie such as the classical factors such as cast, producer, director etc. or the social factors in form of response of the society on various online platforms. This methodology lacks to harvest the required accuracy level. Hence a better method is required. Our paper suggests that the integration of both the classical and the social factors to generate the result and the study of interrelation among the classical factors will lead to more accuracy. To achieve this, collecting the data scattered across internet is necessary and thus data on various platforms such as YouTube, Twitter, and Wikipedia etc. is taken into account along with the classical factors resulting in effective integration. In this paper, integration of both classical and social factors is done which will help us to identify strengths, risks and opportunities to make prediction about a particular movie. In this proposed system, blending these classical and social factors for higher accuracy will show the prediction for a movie whether it will be a success or flop on Box Office.

II. LITERATURESURVEY

Even though there are many factors that are responsible for a movie's success, and it is not always clear how they interact, this paper attempts to determine these factors through the different attributes, social media etc.and predictive analytics.

A. ROLE OF CLASSICAL ATTRIBUTES OFMOVIE

The classical movie attributes such as cast, director, producer, and genre play a crucial role in the movie's success. Dan Cocuzzo et al have used Naive Bayes and Support vector machine to predict the movie success. In Naive Bayes algorithm, they represented movie as independent combination of associated personas and attributes, which was given by, P(rating | movie) proportional to P(movie | rating) * P(rating) ,where P(movie | rating) is product of individual conditional probabilities for each persona. Jason van der Merweetal have built Linear Regression and Logistic Regression models. In linear regression least mean square method, specifically stochastic gradient descent, was used to learn the weight vectors. In order to include the movie title in the feature vector, the movie title was given a score. The movie title was included since the title of a movie does have an effect on the movie's success. To accomplish this, K- Means clustering was used. The accuracy was increased to 52% by implementing K-Means clustering on the titles. Nikhil Apte et al have implemented Linear Regression, K- means clustering, Weighted linear regression and Polynomial Regression algorithms. The authors have also considered effect of inflation rate on movie gross. This was done by dividing the global box office collection and the movie budget, by the values of the normalized price of a movie ticket for the year of its release and then multiplying it by the current normalized movie ticket price. Besides traditional movie attributes Jeffrey Simon off et al used additional variables for measuring star power. Linear regression for predicting movie grosses was used. Steven Yoo et al categorized the features into numeric, text and sentiment. The numeric features consist of budget, average rating, duration, user vote count and critics review count. The text based features consist of MPAA rating, director and genre. The sentiment feature consists of the sentiment score. Sharang et al have considered the features that can be used by producers prior to the beginning work on a movie. The variables considered are director, actors and genre. Also the audience rating was used as an extra criterion variable. Four different regression techniques, support vector regression, Ada boosted decision tree regression, gradient boosting regression and random forest regression were used in this project. The best performing regression method was boosted decision trees. Alec Kennedy has studied the interrelationship between the success of the movie and their critical reviews. The author concludes that along with effective marketing strategies and favorably good critical reviews, it is profitable to release a film. Jeffrey Ericson et al use only the attributes that are influential in the pre-release phase. They also tried to analyze the impact of the movie title on its success. Despite much effort with various approaches, predicting the financial success of a movie remains a challenging problem.

B. ROLE OF SOCIAL ATTRIBUTES OFMOVIE

To get the higher accuracy in the estimation of the box office collection, all possible criterion should be considered. For this, response of users on social media should be taken into account along with the classical factors. Part of the hypothesis of the project is that the anticipation and social media feedback helps to predict movie success as discussed by Gloor et al. They generated the feedback for the movies in three ways: using web searches, using blog searches and using posters on movie forums. In addition to determining the feedback the author performed sentimental analysis on IMDb forums to gather the general mood towards the movie. An important step is to measure the movie title's relative importance on web and other such forums. The user feedback or movie popularity can be estimated through sentiment analysis of twitter data. Twitter, a micro blogging website plays an important role by conveying information about user feedback and preferences. It can be done by measuring the extent of positive or negative words in tweets. Vasu Jain in his work tries to predict the movie popularity from sentiment analysis of tweets. The data fields for each tweet such as tweet id, user name, tweet text, time of tweet are stored. The author tries to classify the movie into three categories: hit, flop, average. Lyric Doshi, in his project explored the effectiveness of collective intelligence, social network analysis and sentiment analysis in predicting trends by mining publicly available online data sources. To determine general sentiment about movies, the author considered posts from IMDb forums, Oscar Buzz, Film General. The author used single variable and multi variable linear regression models which proved to be effective. Another effective measure is determining the Wikipedia metrics associated with a particular movie. It signifies the user interest or anticipation towards a movie. MartonMestyan et al in their paper have considered Wikipedia metrics consisting of the following parameters such as V: Number of views of the article page, U: Number of users, being the number of human editors who have contributed to the article, E: Number of edits made by human editors on the article, and R: Collaborative rigor of the editing train of the article. The authors used multivariate linear regression. Alexander Jagar et al developed visual analytics tool based on tweets and IMDb data. The authors in this project displayed the tweets' content as a graph structure to get feeling for actors, associations and sentiments. The MooVis tool was then used to get an overview about the movie itself. A popular approach to predicting box office success was developed by Google this approach utilizes Google's vast corpus of search data to predict box office performance using query volume. .Movie success prediction through YouTube metrics is another important way. The metrics such as view count or likes, a particular movie trailer gets can be influential for predicting the box office performance. Eldar'sSadikov et al implemented a model for analysis of comprehensive set of features extracted from blogs for prediction of movie sales. The authors used the blog data set from spin3r.com for comprehensive list of features that deal with movie references in blogs.

III. PROPOSED SYSTEM

The overall methodology is shown in Figure 1. Our system is comprised of two major modules, namely Data Collector and Predictive Engine. Data Collection is the more significant task of the two; it involves data collection from IMDB, Twitter and YouTube and then pre-processing that involves dealing with maximum values, data transformation (converting currencies etc.) and calculation of sentiment analysis score for each of the tweet etc.

Data Collection and Prediction Model are explained below:

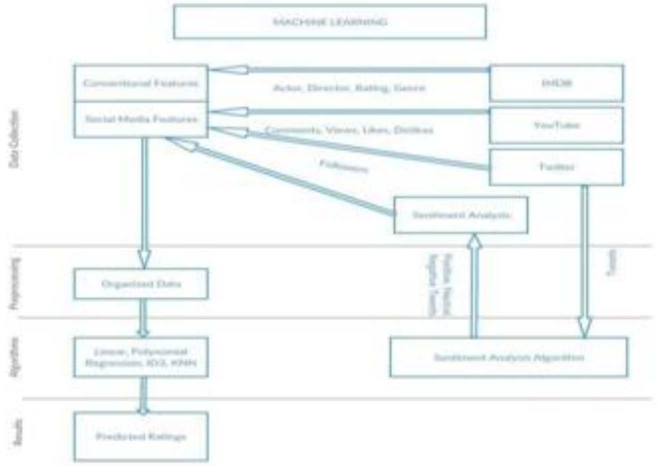

Figure.1. Architectural Diagram for Movie Success Prediction Based On Classical and Social Factors

A. Integration of classical factors and social media interactions for improving overall accuracy rate

Along with the classical factors or the main movie attributes, considering the user anticipation or feedback improves the prediction success rate.

a. YouTube

Prior to movie release, movie teaser or trailer is available on YouTube. So YouTube hits or views of a movie trailer provide a opportunity to predict the popularity of movie and hence the success. YouTube provides API for accessing data related to particular video.

b. Twitter

Sentiment analysis of tweets on twitter can contribute to improve accuracy of model. Through requisite API the sentiment analysis of tweets gives insight about user feedback or response for a particular movie.

c. Wikipedia

From Wikipedia, the view and edit counts of a particular movie can be obtained which provides information about popularity of a movie.

Phases throughout the development process:-

1. Data Collection

Data Collector is the major module as it retrieves information about movies from diverse sources including movies web sites i.e. IMDB, generic web resource i.e.Wikipedia, and social media including YouTube and Twitter. Furthermore, we used sentiment analysis libraries to get the sentiment score for different movies. As the data, that we are interested in such as followers count on twitter, ratings on IMDB etc. continuously changes, therefore, we collected the latest data from these web resources by using APIs and scrappers instead of using already available movies datasets.

2. Data Pre-processing

The data acquired needs to be stored systematically in the database so that it can be used as the training or the testing dataset. This data base acquired initially can be considered as the raw data which is not directly applicable as it may have many redundancies, incomplete data and other inconsistencies. Some of the data might be in the form of lists while other can be present as the API network calls. Hence converting all this data to one single usable format is necessary. Data preprocessing involves the conversion of the raw data acquired previously to the usable data. Initially this involves the conversion of all the data in one single uniform format such as SQL database. In the later step among all the data acquired only the relevant data is to be stored in the database. This involves removal of all the redundant data such as removal of the entry of the movie tuple which has some of its classical factors missing or removal of the data that is out of the scope such as movies released before 1990 or those movies which are not released under Hollywood or Bollywood. Once only the relevant data is stored in the database we convert it to the directly applicable dataset by normalizing the database. Here we introduce various linking factors among entries and improve the accessibility of the database.

3. Algorithms

Linear Regression:-In statistics, linear regression is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X.

Polynomial Regression:- In statistics, it is just a form of regression analysis in which an nth degree polynomial in x is a model of the relationship between the independent the dependent variable x andy.

ID3:- ID3 (Iterative Dichotomiser3) is an algorithm invented by Ross Quinlan used to generate a decision tree from adataset.

KNN:- An simple algorithm that classifies new cases based on a similarity measure based on the stored test cases.

4. Results

As a result of the data mining algorithms performed, results would be produced consisting of the predicted ratings of a movie. These ratings would consist of prediction of social as well as critics’ ratings. On the basis of these two factors, a success rate of the movie can be predicted

IV. CONCLUSION

In business, predictive analytics models generate interesting patterns from historical and current data to identify various strengths, risks and opportunities to make prediction about future events. This paper documents the interrelationships established between various classical factors and social signals used while implementing the predictive model for predicting the total box office collections and critical rating for a particular movie. The results show that the prediction model built using integration of classical as well as social factors can achieve higher accuracy rate. Because the model built can predict the success of movie before its release, it can be used by movie stakeholders for better decision making.

V. REFERENCES

[1]. Himanshu Kulkarni; “Relationships between Classical Factors, Social Factors and Box Office Collections”, IEEE, 2016.

[2]. Mehreen Ahmed; “Using Crowd-source based features from social media and Conventional features to predict the movies”, IEEE, 2015.

[3]. Vasu Jain; “Prediction of movie success using sentiment analysis of tweets”; SCSE2013.

[4]. Dan Cocuzzo, Stephen Wu; “Hit or Flop: Box Office Prediction for Feature Films”; Stanford University, 2013.

[5]. Yafenglu, Robert Kruger, Student Member, IEEE; “Integrating Predictive Analytics and Social Media” IEEE,2014.

[6]. Adarsh Tadimari, Navin Kumar, Tanaya Guha; “ Opening Big in Box Office? Trailor content can Help”, IEEE, 2016.

[7]. P. Gloor, J. Krauss, S. Nann, K. Fischbach and D. Schoder; “Web science 2.0: Identifying trends through semantic social network analysis.”;In IEEE Conference on Social Computing, Vancouver, August2009.

[8]. Vasu Jain; “Prediction of movie success using sentiment analysis of tweets”; SCSE2013.