International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

# MOVIE SUCCESS PREDICTION USING DATA MINING

**Payal Patel[1], Rasika Mate[2], Vrushali Thakare[3], Rohan Adhikari[4], Swapnil Kharat[5]**

[1,2,3,4,5]*Department of Electronics & Telecommunication Engineering, Shivajirao S. Jondhle College of Engineering & Technology Asangaon, Maharashtra, India.*

---

**Abstract**— *Now a day's social media plays very important role in varies domains. There are number of recourses available on the Internet to express the opinions, ideas emotion and interests. Blogs are most popular way for the peoples to express opinion. Web Blog Mining which is the efficient and effective way of analyzing the sentiments of consumer reviews pertaining to specific products becomes desirable and essential. Blogs provides information but it hard to reach information automatically because blogs are full of un-indexed and unprocessed text that reflects the opinions of people. To grab people's idea sentimental opinion mining is the best efficient way to mine their blogs. This study covers the sentimental web mining approach to understand people's opinions about reviews web blogs. This is the efficient and effective way of analyzing the sentiments of peoples review.*

***Keywords***—**Data Mining, Movie, Movie Review.**

## I. INTRODUCTION

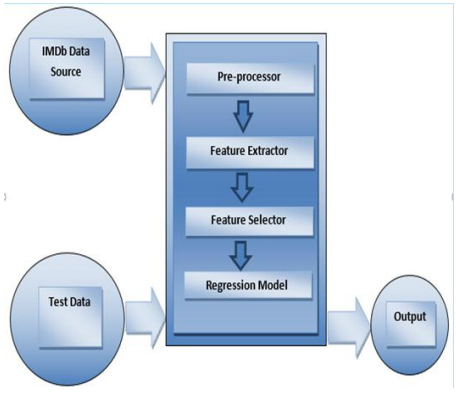
Movies is the most convenient way to entertain peoples. However only few movies get higher success and are ranked high. Many movies are produces by the movie industry in a year. A movie revenue depends on various components such as cast acting in a movie, budget for the making of the movie, film critics 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 analyzing the revenues generated by previous movies, a model can be built which can help us predict the expected revenue for a particular movie. As we know in today's world the movie is one of the biggest source of entertainment and also for business purposes. . To expend this business further we need the technology through which we can predict the success rate of the movie. If we were able to predict the movie success rate in the correct manner then it will be easy for the businessman to get higher profit from it and also if the prediction shows the success rate is low of certain movie then it helps those businessmen to improve the content of the movie so that they can get higher revenue from it. Success rate of movies, models and mechanisms can be used to predict the success of a movie.

## II. LITERATURE SURVEY

In 2004, M Saraee, White and Eccleston [1] performed analysis of online movie resource of over 390,000 movies and television shows. In 2006, Ramesh Sharda and Delen [2] worked with predicting financial success of movies even before the movie is released Classification approach is used where the movies were categorized from flop to blockbuster. Facts and relationships among alternatives can be made by making use of data mining. In August 2007, Yun-Qing Xia, [3], worked with "The Unified collocation Framework for Opinion Mining", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong. In 2009, W.Zhang and Skeina [4] worked on utilizing news analysis to make movie predictions. In 2010, Suhaas Prasad [5] worked with "Using Social Networks to improve Movie Ratings predictions, Dept. Electronics Eng. Stanford University, California.

## III. PROPOSED METHODOLOGY

The proposed system deals with different stages of the project which consists of data collection, data preprocessing, generating training and testing dataset, model generation, prediction and outcomes. These all methods prevents us from getting any irrelevant data which further keeps our outcomes more relevant and accurate for the prediction. Here we collected dataset from IMDB which consist of 32 attributes and 651 tipples. Further steps are explained below:-




### Preprocessing:

In this stage dataset is prepared for applying data mining technique. Before applying data mining technique, pre-processing methods like cleaning, variable transformation and data partitioning and other techniques for attribute selection must be applied. After pre-processing we have attributes or variables for each movie. As the data is taken in the raw format from IMDb it is first required to be pre-processed. To overcome

missing value scenario central tendency method is used both mean and median and later the duplicate items are removed. Pre-processing is the crucial phase for the project as it mainly focuses on the working of the algorithm. As the data is now pre-processed next comes data integration and transformation in which the alpha numerical data need to be converted to the numerical data as it is required for regression model.

### Test Dataset:

Training dataset is a set of attributes used to fit the parameters of the model. The model like naive Bayes classifier is trained on the training dataset using supervised learning method like gradient descent or stochastic gradient descent. In practice, the training dataset often consist of pairs of an input vector and the corresponding output vector (or scalar), which is commonly denoted as the target. The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation. Finally, the test dataset is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. When the data in the test dataset has never been used in training like cross-validation, the test dataset is also called a holdout dataset.

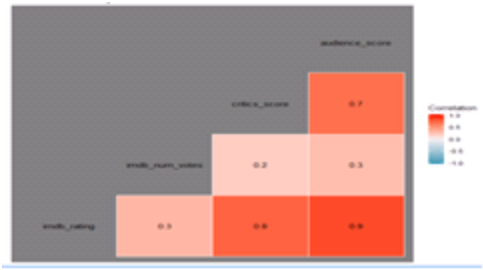
### Data Analysis:

In data analysis, all selected attributes are analyzed on the basis of different Factor that help us to gather most accurate outcome for further stages. Selected features for Analysis are as follows: imdb_rating, imdb_num_votes, critics_rating, critics_score, Audience rating and audience score. On the basis of these attributes, we are generating various Visualized graphs for analyzing the best possible attribute among these for further predictions.

An actor, actress, or director who has won an Oscar award is a great motivation to analyze movie success. Movie who has won an Oscar has the same weight as an actor or a director in making a movie popular.

### Model Generation:

In simple terms, modeling is a simplified, mathematically-



formalized way to approximate reality and optionally to make predictions from this approximation. The Statistical model is the mathematical equation that is used. Representing a quantity by an average and a standard deviation is a very simple form of statistical modeling.

## IV. EXPERIMENTAL RESULTS

The output is shown below –

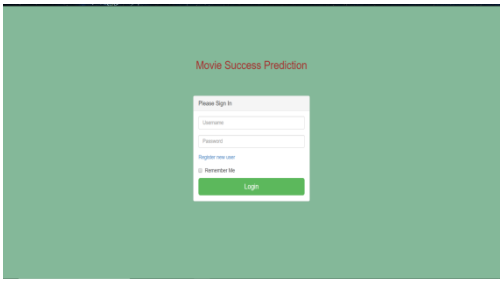



Figure No 4.1: Login Page

Above figure 4.1 shows the login page of our project from which we can enter in our project window.

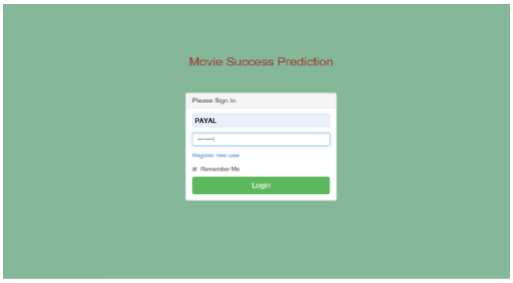



Figure No 4.2: Login Page (Entering username and password)

Above figure 4.2 also shows the login page of our project from which we can enter in our project window by entering username and password.


Figure No: 4.3 Homepage of project

Above figure shows the homepage of our project on which we can see 3 options: Dashboard, Search, about us.

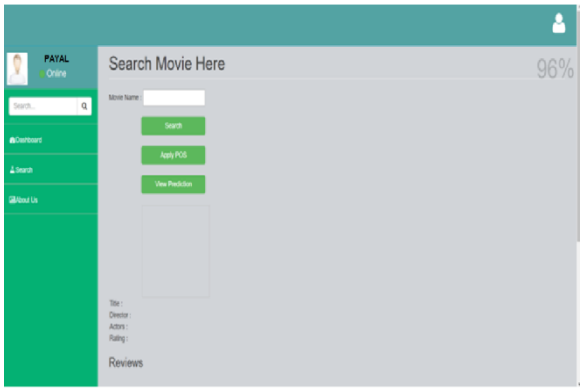



Figure No 4.4: Homepage for searching movie

Above figure 4.4 shows the option of search on which we can enter any upcoming movie name and when we hit search button related result will be shown.

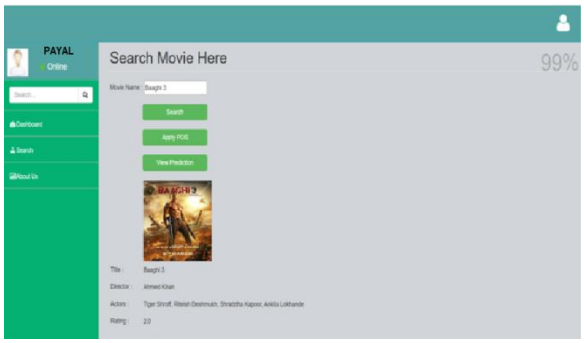



Figure No 4.5: Homepage of searching movie

Above figure 4.5 shows an example of an upcoming movie (Baaghi 3) below that we have 3 buttons, Search, Apply POS, View Prediction from which when we hit Apply POS button we can see how much content is grammatically positive as well as negative. On third button View Prediction we can see prediction of movie which will predict the movie in form of percentage.

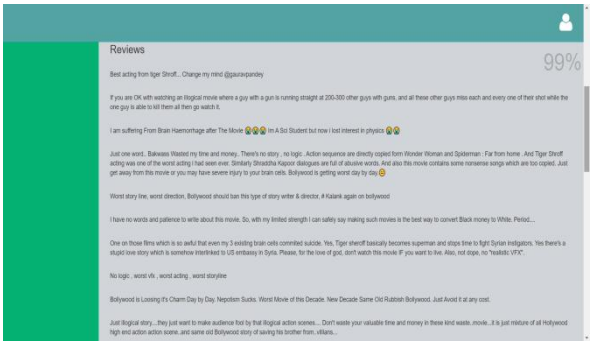



Figure No 4.6: Showing review of particular movie

In above figure 4.6 we can see when we hit search button we can also see reviews of customers which has been posted on official website or on social media.

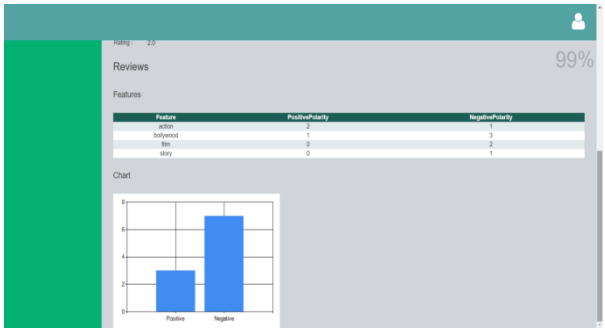



Figure No 4.7: Showing review rating of particular movie

In above figure 4.7 we can see when we hit search button we can also see reviews of customers which has been posted on official website or on social media and also get the reviews ratings in features – For example : action – positive 2 and negative 1 etc. and shows the chart that the movie is all over positive (hit) or negative (flop).

## V. CONCLUSION

In this project we are just trying to determine if there is any association between Different attributes present in our dataset. Here, our main aim is to find association between Numeric type attributes that is used s a scoring systems and how we can use this association for Prediction. As a result we found that critics score is strongly positive relationship between critics score and audience score. And we can also conclude that critic's score are best predictor of audience scores. Thus, we can predict our movies success on the basis of critics score. In future, we can add many attributes as our predictors and build model for that attributes to perform prediction. Here, we can assume that if we have movie gross score and movie net profit along with movie manufacturing cost, then we can build a stronger model for movie success prediction. In future, we can apply other machine learning algorithms for movie success prediction.

## VI. FUTURE SCOPE

Overall, we have found that it is difficult to apply data mining techniques to the data in the IMDb. The data needs extensive cleaning and integration, and this consumed a large proportion of the time available for this analysis. In addition, much of the data is in textual rather than numerical format, making mining more difficult. Much of the source data could not be integrated at all, without using natural language processing techniques. Despite these problems, we performed some useful data mining on the IMDb data, and uncovered information that cannot be seen by browsing the regular web front-end to the database.

## REFERENCES

[1] M Saraee, White and Eccleston performed analysis of online movie resource of over 390,000 movies and television shows in 2004.
[2] I Ramesh Sharda and Delen worked with predicting financial success of movies even before the movie is released Classification approach is used where the movies were categorized from flop to blockbuster, 2006. Facts and relationships among alternatives can be made by making use of data mining.
[3] Yun-Qing Xia, worked with "The Unified collocation Framework for Opinion Mining", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, in August 2007,
[4] W.Zhang and Skeina worked on utilizing news analysis to make movie predictions, 2009.
[5] Suhaas Prasad worked with "Using Social Networks to improve Movie Ratings predictions, Dept. Electronics Eng. Stanford University, 2020 at California.