International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Movie Success Prediction using Data Mining and Social Media

Ms. Kenvi Shah, Mr. Jigesh Kapadia, Mr. Yash Samel, Mr. Sarvesh Saple, Mrs. Pallavi Deshmane

[1,2,3,4]*Student, Dept. of Computer Engg. SAKEC, Mumbai*
[5]*Professor, Dept. of Computer Engg. SAKEC, Mumbai*

Abstract - *Feature films are a multi-billion industry. Here prediction of a movie's success is predicted based on its features like cast, genre of movie, month of release, run time, directors, producers etc. Based on multiple such features and with the database of previous movies statistics, machine learning algorithm like Linear Regression can predict the approximate ratings the movie can receive once it is actually released and hence classify a movie as a hit or a flop. A large amount of data representing feature films is maintained by the Internet Movie Database (IMDb). Due to the large number of films produced as well as the level of scrutiny to which they are exposed, it may be possible to predict the success of an unreleased film based on data.*

1. INTRODUCTION

Historical data of each component such as actor, actress, and director, composer that influences the success or failure of a movie is given due to its weightage. This proposed work aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. The system is used to predict the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making the success of the movie is without risk, because the decision maker, movie makers and stakeholders have all the information about the approximate outcome of the decision, before he or she makes the decision to release of the movie. In particular, we concentrate on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed.

1.1 Steps of System Flow (regression):

Input: Movie database, User input film with feature values

Output: Rating of user entered film

1. Feature selection

2. Feature normalization

3. Apply Multivariate Linear regression model on the dataset

4. Get prediction result

1.2 MULTIPLE LINEAR REGRESSIONS

The value that we want to predict is called the dependent variable. Let us consider number of explanatory or independent variables is p. So the variables can be denoted as x1,x2,....,xp. The regression line can be defined as $\mu y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp$. Thus, the model of multiple linear regression having n explanatory variables can be denoted as follows: $yi = \beta 0+\beta 1xi1 +\beta 2xi2 + \dots \beta pxip + \varepsilon i$ for i=1,2,...,n.

2. IMPLEMENTATION

The aim is collecting number of likes, dislikes and view count of trailer, release date, star ranking. Multiple Linear Regression Algorithm is used which was discussed above for the prediction of earnings of the movie. WEKA tool was used for choosing the best algorithm on a database with following Attributes/Features: star_power(Lead actors' rank), view_ count, like_count, dislike_count. The labels in this model are: day1_collection and lifetime_collection. When we plotted the each Feature against the label day1_collection, we got the following results in WEKA tool.

Table -1: Table 1(References 1)

Positive	Positive review of the movie
Negative	Negative review of the movie
Neutral	Neither positive nor negative reviews Mixed positive and negative reviews
	Unable to decide whether it contains positive or negative reviews Simple factual statements Questions with no strong emotions indicated

So we applied multiple linear regression to the data for which we knew the output/actual collection, to check the accuracy.e plotted the graph (actual vs. predicted values) using matplotlib library. Now once the movie is released we use its IMDb ratings and actual first day collection to predict the Lifetime collection of the movie. Here again we use Multiple Linear regression. We used the same dataset to check the accuracy and plotted the graph..

The goal is 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. In our case m=50000.

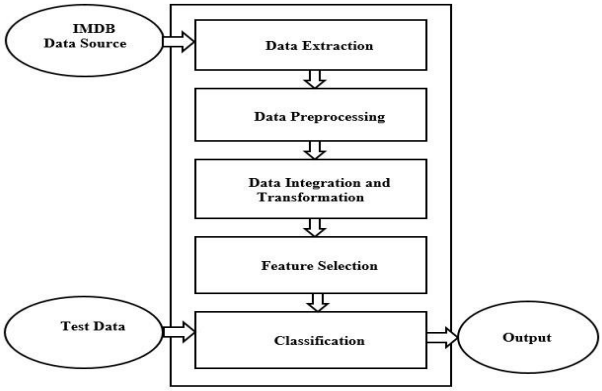

Fig -1: Methodology

This movie set will 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 we find features all the producers, directors, actors and actresses were rated based on their past performance at the Box Office.

SENTIMENT ANALYSIS

Text Processing**:**

The text of each tweet includes a lot of words that are irrelevant to its sentiment. For example, some tweets contain URLS, tags to other users, or symbols that have no meaning. In order to better determine a tweet's sentiment score, before anything else we had to exclude the "noise" that occurred because of these words. For this to happen, we relied on a variety of techniques using the Natural Language ToolKit (NLTK) for.

A. Tokenization

Firstly, we divided the text by spaces, thus forming a list of individual words per tweet. We then used each word in the tweet as features to train our classifier.

B. Removing Stopwords

Next, we removed stopwords from the list of words. 's Natural Language ToolKit library contains a stopword dictionary, which is a list of words that have neutral meaning and are inappropriate for sentiment analysis. To remove the stopwords from each text, we simply checked each word in the list of words against the dictionary. If a word was in the list, we excluded it from the tweet's text. The list of stopwords contains articles, some prepositions, and other words that add no sentiment value (able, also, or, etc.)

C. Twitter Symbols

It is not uncommon that tweets may contain extra symbols such as "@" or "#" as well as URLs. On Twitter, the word following an "@" (mentions) symbol is always a username, which we also exclude because it adds no value at all to the text. Words following "#" (hashtags) are not filtered out because they may contain crucial information about the tweet's sentiment. They are also particularly useful for categorization since Twitter creates new databases that are collections of similar tweets, by using hashtags. URLs are filtered out entirely, as they add no sentiment meaning to the text and could also be spams.

D. Training Set Collection

To train a sentiment analyzer and obtain data, we were in need of a system that could gather tweets. Therefore, we first collected a large amount of tweets that would serve as training data for our sentiment analyzer. In the beginning, we considered manually tagging tweets with a "positive" or "negative" label.

E. Training the classifiers

Once we had collected a large tweet corpus as training data, we were able to construct and train a classifier. Within this project we used two types of classifiers:

A. Feature Extraction

B. Feature Filtering

3. CONCLUSION

This project had as a main goal to develop a model, able of predicting the box office financial success of a certain set of movies through specific variables and historical data. It was possible to conclude that the percentage of success of the cinematographic revenue prediction is quite different based on the typology of the dependent variable used in the study. The empirical model demonstrated good statistical results when the dependent variable was binary and interval. However, in what regards the multiclass prediction, the results were very far from reality, negatively influencing the model

ACKNOWLEDGEMENT

A lot of effort and study have been put to make this project. This would not have been possible without the genuine support and assistance provided by the people whom we approached during the various stages of writing this Paper. We would like to express gratitude to our academic supervisor for her advice, counseling, direction and help. Without her guidance & methodology of works, this would not have been possible.

We would also must thank to all the faculty members of College for all of their direct and indirect encouragement and assistance in this work, my batch mates and friends who have provided me with their valuable suggestions throughout the year. Their cooperation, suggestion, guidance and sincere encouragement played significant role throughout our working period.

REFERENCES

[1] The International Journal of Soft Computing and Software Engineering [JSCSE],) "Prediction of Movie Success using Sentiment Analysis of Tweets" Vasu Jain, Department of Computer Science, University of Southern California, Los Angeles, CA, 90007, vasujain@usc.edu

[2] Basuroy, S., Chatterjee, S. and Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. Journal of Marketing, 67(4), 103–117.

[3] What Determines Box Office Success of a Movie in the United States? Proceedings for the Northeast Region Decision Sciences Institute, (757), 447. Deuchert, E., Adjamah, K. and Pauly, F. (2005).

[4] Data mining in course management systems: Moodle case study and tutorial. Computers & Education. https://doi.org/10.1016/j.compedu.2007.05.016 Gennari, J. H., Langley, P. and Fisher, D. (1989).

[5] What Determines Box Office Success of a Movie in the United States? Proceedings for the Northeast Region Decision Sciences Institute, (757), 447. Deuchert, E., Adjamah, K. and Pauly, F. (2005).