ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.4

Pre-Release Analysis and Success Prediction of Movies Box Office Performance

Jeetiksha Chandiramani[1], JayeshNagpal[2], RohitWadhwa[3], JayashreeHajgude[4]
BE. Student[1, 2, 3], Assistant Professor[4]
Department of Information Technology
Vivekanand Education Society's Institute of Technology, Mumbai, India

Abstract:
Use of socially generated "big data" to access information about collective states of the minds in human societies has become a new paradigm in the emerging field of computational social science. A natural application of this would be the prediction of the society's reaction to a new product in the sense of popularity and adoption rate. However, bridging the gap between "real time monitoring" and "early predicting" remains a big challenge. In the recent past, machine learning algorithms have been used effectively to identify interesting patterns from volumes of data, and aid the decision making process in business environments. In this project, we aim to use the power of such algorithms to predict the pre-release box-office success.

Keywords: Machine learning, Sentiment analysis, Social media data collection, Linear Regression, Box-Office Revenue Prediction, Data Analysis.

I. INTRODUCTION

There are many ways of predicting whether a particular movie will be succeed or not, one of the factors is comparing the revenue spent in making of the movie with the earnings. The important parameters are the cast and crew, the sentiments of the people, cult index of the movie. Sentiment Analysis searches sentiment bearing words like adjectives and then each sentiment is classified as positive, negative or neutral. Large volumes of data are generated on social media. Such data can be used to identify interesting patterns using machine learning techniques. Similarly, data related to movies like genre, trailer, actor, director etc can be used to predict a movie's box office performance before it's released. Web mining would be performed to obtain statistics like genre, cast popularity, etc for every movie. Every source site from which data has to be obtained would be checked for its authenticity in order to filter out the unreliable data and only the reliable sources would be used for obtaining statistics. A algorithm would be used to predict upcoming movies based on the data mined. Using the results provided by the software and that inputted by the user, the software would provide the user, success rate of that movie and sentiments about that movie. Part purpose of this project is to get an insight in the emotions of people. The emotions describe what will be the reaction of people be on occurrence of certain events, the data further can used to be predict the kind of earnings and other benefits that will be for a particular show or a movie.

II. LITERATURE/TECHNIQUES STUDIED

There were various methods proposed for sentiment analysis and prediction. Denecke et al. used SentiWordNet for sentiment analysis, in which a triple of polarity scores are assigned i.e., a positivity, negativity and objectivity score. The LingPipe Classifier treats language identification as a problem and for each language to be identified, a set of training texts are used. It assigns probability to a sequence of words by a means of probability distribution and aims to predict probability of natural sequences[6]. Linear Regression algorithm basically is built on the fact that takes into account the linear interdependencies of various factors (predictors) as per the responses been recorded. Sitaram Asur et al. constructed a linear regression model using least squares of the average of all tweets for the 24 movies considered over the week prior to their release. This performance was achieved using only one variable i.e. the average tweet rate[5]. Krushikanth R. Apala et al. used the K-Means tool from Weka. The movies were grouped into three classes: Hit, Neutral, and Flop. Before grouping, two possibilities were considered for assigning weights to the attributes. One way is to consider all attributes being equal and using a uniform weight assignment, the other is to give a higher weight to the sentiment attribute and treat the rest of attributes as equal[8].
Non-Linear regression algorithm basically takes the interdependency between the predicates and generate the responses taking into consideration the fact that predicators may sometimes not always be directly correlated with each other and may show signs of interdependence[4].

III. DATA COLLECTION

We gathered data from official trailers of movies with pre-release dates from year 2000 till year 2014 on global markets. We identified these movies from the following website. A. Wikipedia - The titles used in making data for the project were collected from wikipedia and then the titles were fetched to Boxmojo (Table I).

Table I.

Id	title	distributor	Genre	leadActor
3412	Pollock	Sony Classics	Drama	Ed Harris
3231	Jesus' Son	Lions Gate	Drama	Billy Crudup
3406	Panic	Roxie	Crime Drama	William H. Macy
3423	Requiem for a Dream	Artisan	Drama	Ellen Burstyn
2988	Brown's Requiem	Unknown	Unknown	Michael Rooker
3375	The House of Mirth	Sony Classics	Period Drama	Gillian Anderson
3182	But I'm a Cheerleader	Lions Gate	Comedy	Natasha Lyonne
3354	An Everlasting Piece	DreamWorks	Unknown	Barry McEvoy
3472	You Can Count on Me	Paramount Classics	Drama	Laura Linney

B. BOXmojo - BOXmojo tracks box office revenue in a systematic, algorithmic way, and publishes the data on its website. Using Boxmojo we generated rev_totalGross, Rev_opening, Num_theaters, Distributor, Genre_bomojo, Runtime, Prod_budget, rating and the titles for the same were taken from wikipedia as mentioned above(Table II)

Table II.

rev_opening	num_theaters	runtime	rev_postOpening	runtime_mins	releaseYear	rev_opening_ADJ	rev_totalGross_ADJ
44244	2	1 hrs. 57 min.	22462994.31	117	2000	71625.19245	22534619.5
37089	1	1 hrs. 45 min.	3352328.228	105	2000	60042.19245	3412370.42
18006	4	1 hrs. 28 min.	2012760.99	88	2000	29149.33585	2041910.326
64770	2	1 hrs. 42 min.	9422774.871	102	2000	104854.0755	9527628.946
3077	2	1 hrs. 44 min.	6253.827298	104	2000	4981.256604	11235.0839
48770	7	2 hrs. 20 min.	7896683.705	140	2000	78952.18868	7975635.894
60410	4	1 hrs. 24 min.	5682564.441	84	2000	97795.81132	5780360.252
9128	8	1 hrs. 43 min.	182375.5026	103	2000	14777.02642	197152.529
118170	8	1 hrs. 51 min.	23875841.5	111	2000	191301.6226	24067143.12

Figure.1. shows a graph of gross revenue versus distributor.

Figure 1.

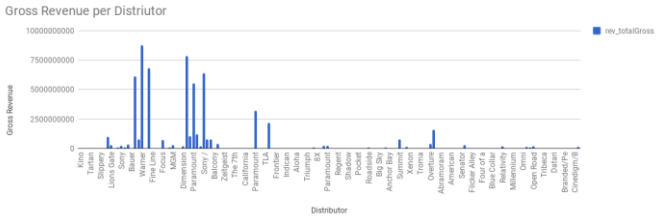
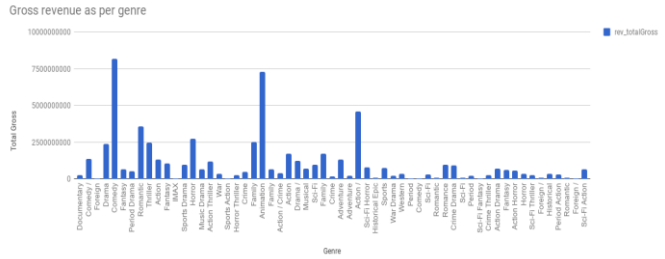


Figure.2. shows a graph of gross revenue versus genre.

Figure 2.



C. TMDB - Using TMDB we collected actors/actress popularity, directors popularity(Table III).

Table III.

leadActor	Actor popularity	director	Director popularity
Ed Harris	9.482665	Ed Harris	9.057137
Billy Crudup	5.55511	Alison Maclean	1.000003
William H. Macy	4.371691	Henry Bromell	2.325409
Ellen Burstyn	4.039423	Darren Aronofsky	4.87431
Michael Rooker		Jason Freeland	1.000799
Gillian Anderson	6.719518	Terence Davies	1.000554
Natasha Lyonne	3.087959	Jamie Babbit	1.030467
Barry McEvoy	1.000197	Barry Levinson	2.244897
Laura Linney	4.728341	Kenneth Lonergan	2.812408

IV. ARCHITECTURE

I.Data:

Twitter will be used to get the views of people on the particular movie. The movie box will give the trailer id for every movie and this id will be help us get the comments on the trailers. The data received through Twitter will be in different languages, the data will go through machine translation before further processing. The comments and tweets received will go through machine translation to be converted to English, whereas the English will remain as it is. The TMDb movie database is one the most comprehensive resources containing detailed information about almost any film ever made. It has a vast amount of data containing valuable information about general trends in films.

II.Data Attributes:

The data collected from TMDB will tell us the most successful genres in the recent past. Thus helping us determine rate of success. The popularity of director, leading actor and leading actress of a movie is represented by their followers count on Twitter. The popularity will be obtained from TMDB. A movie is a sequel if it is a successor of an already released movie. The sequel list is obtained from TMDb. The success of previous movie can help reach the goal. The dataset will be processed using a set of positive and negative words and then classified as positive, negative or neutral. The comments will then be divided into different categories.

III. DATA PREPROCESSING:

Data Preprocessing is an essential step, since the real world data is often incomplete, inconsistent or the data may be noisy. Data pre-processing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set. We will use Weka to perform preprocessing on the collected data.

IV. SENTIMENT ANALYSIS:

Sentiment Analysis will be performed using Text Blob which will give the sentiment polarity score. The polarity value ranges between [-1, 1]. So a tweet has Positive sentiment when it's polarity is greater than 0 and negative sentiment when it's polarity is lesser than 0. When the sentiment polarity is exactly 0 the tweet is said to have neutral polarity.

V. Prediction:

Linear Regression will predict a movie's success before its release. It takes into account the linear interdependencies of various factors (predictors) as per the responses been recorded. It will generate a scatterplot which will show how much the movie's performance will be affected based on the parameters taken into consideration. The regression model will be constructed using title, actor, director, distributor, total gross revenue, opening revenue, popularity of directors and actors, running time, production budget,etc. We need to create dummies for parameters like title, actor, director, distributor. The dummies can be created using Pandas library method given below:

t=df.title.str.get_dummies()

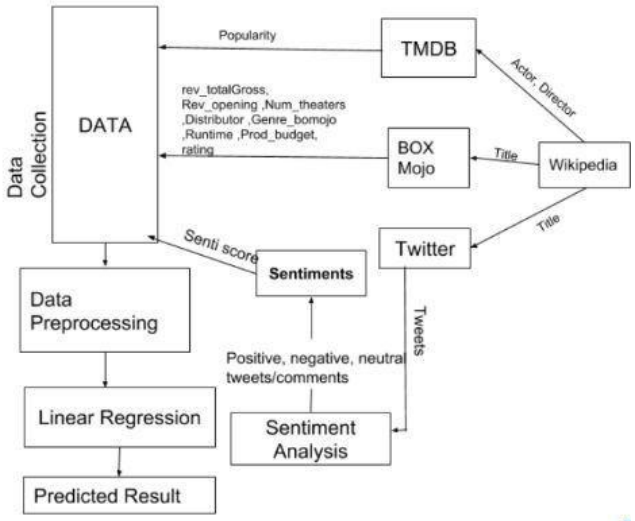


Figure 3

V. ALGORITHMS USED

I. Linear Regression:
The linear regression consists of a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient. In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients.

VI. CONCLUSION

The system implemented, will make use of Sentiment Analysis and Regression algorithm to determine a movie's success before its release. Classification will help to get accurate results about success or failure of a movie. Our approach can be extended to predict the success of videos, songs, books, etc.

VII. REFERENCES

[1] Pre-release Box-Office Success Prediction for Motion Pictures - Rohit parimi et al.https://link.springer.com/chapter /10.1007/978-3-642-39712-7_44

[2] Simonoff and Sparrow, 2000 Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. CHANCE, 13(3):15–24.https:// www. tandfonline.com/doi/abs/10.1080/09332480.2000. 105 42216

[3] Multilingual Sentiment Analysis via Text Summarization - Rupal Bhargava et al. (2017).https://ieeexplore. Ieee .org/ document/7943126/

[4] Pre-Release Success Quotient Prediction of Movies - Harsh Taneja et al. (2013).
https://www.ijsr.net/archive/v5i10/ART20161745.pdf

[5] Predicting the Future With Social Media- Sitaram Asur et al. (29 Mar 2010).https://www.researchgate.net/ publication/ 45909086_Predicting_the_Future_with_Social_Media

[6] Using SentiWordNet for Multilingual Sentiment Analysis - Kerstin Denecke et al. (6 August 2008).https://ieeexplore. ieee.org/document/4498370/

[7] How to Predict Social Trends by Mining User Sentiments- Iuliana Chepurna et al. (2015).https://link. springer.com/ chapter/10.1007/978-3-319-16268-3_29

[8] Prediction of Movies Box Office Performance Using Social Media- Krushikanth R. Apala et al. (2013)https:// ieeexplore.ieee.org/document/6785857/