An Exploratory Data Analysis of Movie Review Dataset

V. Vanitha, V. P. Sumathi, V. Soundariya

Abstract - The film industry plays a major role in the planetary or world-wide economy. It is the symbolic contributor to the global economy. Every year more than hundreds to thousands of movies are released to the public audience with the hope that the movies getting released will be the next block buster. According to the movie industry statistics, six to seven movies out of ten movies gets unprofitable, only one third of the movie gets success. The producers, studios, investors, sponsors in the movie industry are alike interested in predicting the box office success of the movie. This paper work is on analysing the film genre, the release date around holidays, the release month of movies, the languages and country with more movies from the movie review dataset. There are attributes (country, languages, genre, movie release date, budget and revenue) taken from the dataset and the derived attributes (release month of the movie derived from release date of movie and profit from budget and revenue) is analysed to determine the movie performance. The analysed data is plotted in graphs for statistical observation of the movie success.

Keywords: predicting box office success, block buster, film genre, genre count, release month, movie profit, and movie review dataset.

I. INTRODUCTION

Movie industry is the most evergreen industry in the field of entertainment Many artists from all over the world exhibit their talent to the viewers and win the hearts of the many people Only one third of the movie in a full ratio gets box office hit in a year and hits the revenue, the rest of movie remains unprofitable. Movies have several genres like comedy drama, thriller, documentary, biography, crime, family, fantasy, mystery, musical, horror, animation, action, romance, adventure and so on. A large amount of data has been accumulated related to movies. So, it takes huge time to organise this data and to predict further information related to it. The main motive of this paper is to analyse the factors that add credits to a movie's success rate. Viewers can rate the movies based on their likes and dislikes for the scale of 10. Generally, when a film receives a high positive response, its success rate gradually increases. So, the profit of the film is only based on the response of the viewers. Even there are times where the viewer's expectation for a movie is high before the release, but the success rate of the movie is low.

Several analyses and predictions have already been carried using the movies dataset by applying the concepts of machine learning, natural language processing, neural Networks and even exploratory data analysis. In the paper [1], performance comparison among seven machine learning techniques such as Support Vector Machine, logistic regression, Multilayer perceptron neural network, Gaussian Naïve Bayes, Random Forest, AdaBoost and Stochastic Gradient Descent is conducted and upon analysing historical data from different sources all the above methods has predicted an approximate net profit value of the film. Also the Multilayer perceptron neural network algorithm gives better result among the seven algorithms. This analysis predicts movie box office success based on pre-released and post released features and by finding the total number of audience in a year, the accurate result can be constructed. This paper [2] predicts the box office hit and popularity of a movie by analysing the sentiment of the twitter data by machine learning and natural language processing. Conclusively, it predicts the box office results by classifying the movie as Hit, Flop, and Average movies. In the paper [3], it is found that sentiment classifiers are severely dependent on domains or topics. In future, more work is needed on further improving the performance measures. The techniques used are Negation, calculating various metrics like precision, recall and F measure. In the paper [4], uses the messages of twitter to review a movie by using opinion mining. Best accuracy level is 77% achieved by SVM-PSO after data cleansing. Accuracy level of SVM-PSO still can be improved using enhancement of SVM. It is found that sentiment classifiers are severely dependent on domains or topics. Negation, calculating various metrics like precision, recall and F-measure. In future, more work is needed on further improving the performance measures. In the paper [5], the exploratory data analysis technique was applied over the call data records, the computed the participant is equal to ground truth value.

Methods of Exploratory Data Analysis are applied on the data to acquire appropriate results. The raw data for movies was obtained from Kaggle dataset. This analysis is based on the movies count in a month, type of genre which is most liked by the viewer's, genre count and profit rate of a movie, country with more number of movies, language with more number of movies and month with more number of movie releases.

A. Motivation

By this paper work, the producers, studios, investors, sponsors will find it easier for investing to the movies by knowing beforehand about the movies future performance (whether the movie will become success or not).

Manuscript published on 30 November 2018.
***Correspondence Author(s)**

V. Vanitha, Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.
V.P. Sumathi, Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.
V. Soundariya, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.

Also the audience will find it easier whether to buy tickets to watch the movie or not, by knowing whether the movie would be considerably worthwhile to spend money to watch in the theatre.

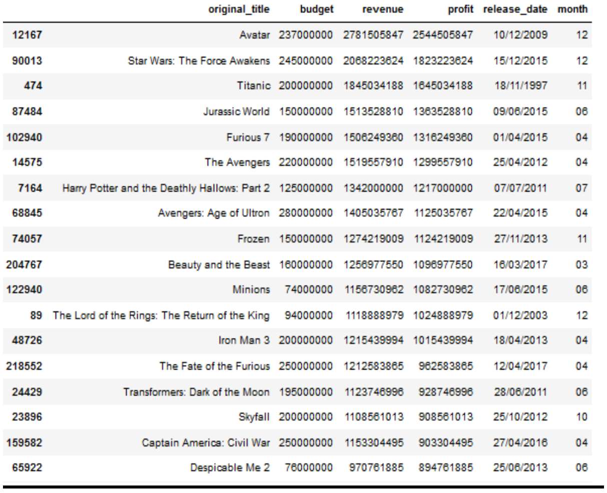
B. Objective

The main objective is to:

1) Find the most profited month of the year
2) Analyse the month with more number of movie releases
3) Analyse the relation between number of movies released in a month and the most profited month in a year.

II. MOVIE REVIEW DATASET DESCRIPTION

The dataset is captured from kaggle dataset and it contains 3, 50,000 movies. It contains nearly 15 years -17 years (between the years 1997 to 2017) of movies data. The dataset accommodates nearly 23 attributes. The dataset holds attributes such as movie Id, imdb_id, original title, overview, popularity, production companies, production companies number, production countries, production countries number, the type of movie genre, revenue of the movie, movie budget, languages of the movie, country name of the movie, release date of the movie, movie runtime, status, tagline, movie languages count, vote average and movie vote count. Upon utilizing the existing budget and the revenue attributes captured from the dataset, movie profit attribute is derived and the release month of the movie is derived from the release date of the movie attribute as shown in figure 2. The analysis is carried out with the type of movie genre, genre count of a movie, revenue of movies, country, language of movie, budget for movie release, profit of movie, vote count of movie, release month of the movie attributes. Also the analysis is constructed by capturing only the movie data of INDIA from the dataset and plotting the graphs by analysing the movie releases in INDIA and profit months of movies in INDIA.



	original_title	budget	revenue	profit	release_date	month
12167	Avatar	237000000	2781505847	2544505847	10/12/2009	12
90013	Star Wars: The Force Awakens	245000000	2068223624	1823223624	15/12/2015	12
474	Titanic	200000000	1845034188	1645034188	18/11/1997	11
87484	Jurassic World	150000000	1513528810	1363528810	09/06/2015	06
102940	Furious 7	190000000	1506249360	1316249360	01/04/2015	04
14575	The Avengers	220000000	1519557910	1299557910	25/04/2012	04
7164	Harry Potter and the Deathly Hallows: Part 2	125000000	1342000000	1217000000	07/07/2011	07
68845	Avengers: Age of Ultron	280000000	1405035767	1125035767	22/04/2015	04
74057	Frozen	150000000	1274219009	1124219009	27/11/2013	11
204767	Beauty and the Beast	160000000	1256977550	1096977550	16/03/2017	03
122940	Minions	74000000	1156730962	1082730962	17/06/2015	06
89	The Lord of the Rings: The Return of the King	94000000	1118888979	1024888979	01/12/2003	12
48726	Iron Man 3	200000000	1215439994	1015439994	18/04/2013	04
218552	The Fate of the Furious	250000000	1212583865	962583865	12/04/2017	04
24429	Transformers: Dark of the Moon	195000000	1123746996	928746996	28/06/2011	06
23896	Skyfall	200000000	1108561013	908561013	25/10/2012	10
159582	Captain America: Civil War	250000000	1153304495	903304495	27/04/2016	04
65922	Despicable Me 2	76000000	970761885	894761885	25/06/2013	06

Figure 2 shows profit and month extracted from the existing attributes

III. MOVIE REVIEW DATASET ANALYSIS AND VISUALIZATION

Movies dataset holds information about the directors, award categories, year in which the movie is released, details about crew members, Production Company, language of the movie, country in which the movie is released and so on. For this analysis, only limited fields are taken from the dataset. The fields are movie title, release date, budget, country, language, revenue, average vote and vote count. From the above-mentioned fields, using release date, month and year from it is derived to analyse the total number of movies month-wise. Using budget and revenue fields, profit for each movie is derived. Pandas library and matplot library is used for extracting the fields from the dataset and plotting the graph using the attributes for each of the objective to provide better visualisation of the analysis.

C. Most Profited Month Of The Year

There are approximately 190 countries all over the world. Every year in each month, more number of movies are released in many different languages in many different countries the world. Though more number of movies are released every year, not all the movies acquires success and brings profit. There are specific months in year which gives higher profit rate than the other months. The studios always keeps an eye on the historical box office performance for every month and weekend of the year. This analysis is to identify the most profited month in the year from the complete dataset.

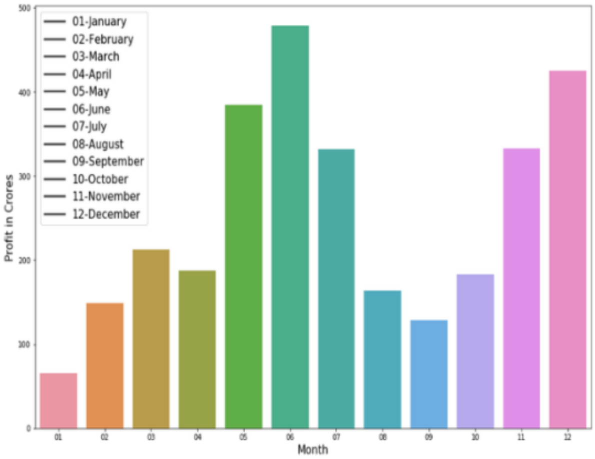


Figure 3.1 shows most profited month in a year

The above graph infers that among all the months in the year, the sixth month, June has the higher number of profit and June month is the most profited month all over the world in the movie production. Also, from this analysis, it is concluded that on an average, mid of the year during the summer seasons i.e., from May to July is the period where the profit gets increased more. It's been true that a strong turnout is a reflection of demand or overall consumer behaviour. This is why you will see an action blockbuster on the first weekend of summer. Holidays are vital for obvious reasons. The December month is the second profited month. During these month a more number of holidays and family movies are released more when family time is more. The May month is third most profited month which is the month for summer holiday seasons for schools and employees in India.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Maybe during holidays, people turn out to movies more often for the entertainment. More people buying ticket and watching movies during these seasons increases the revenue of the movie, which in turn increases the profit of the movie. This may be a reason for movie released in mid of the year getting more profited.

In the above bar graph x-axis represents the month of movie release and y-axis represents the profit of movie in crores. June month represents the most profited month according to the analysis.

D. Month With More Number Of Movie Releases

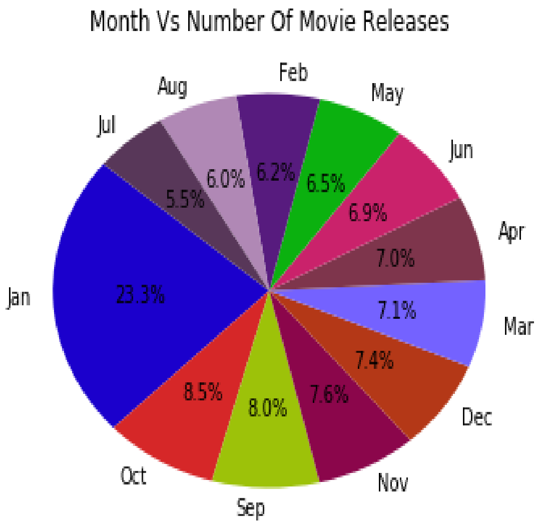


Figure 3.2.a. shows month with more number of movie releases

As shown in the above graph, it represents which month has the highest release of the movies. From the dataset, by retrieving the movie month release from the release date of the movie from the dataset, then analyse which month has the highest number of movie releases According to the analysis, January has the highest number of 71142 movie releases (23.3% movie releases). October has the second highest number of movie releases, September stands third with movie releases.

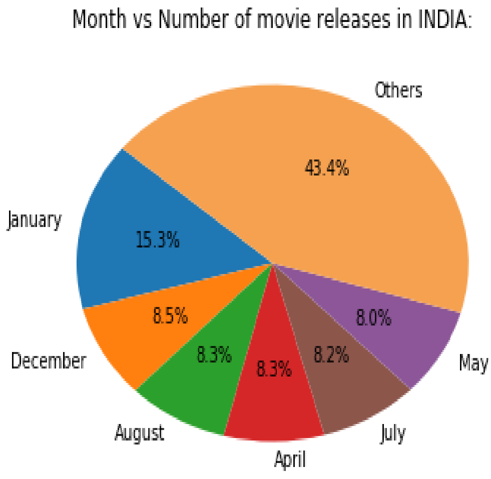


Figure 3.2.b. shows month with more number of movie releases in INDIA

The above pie-chart shows the percentage of movie releases in every month in India. As per the analysis, the January month has the highest percentage of movie releases in India with 15.3% movie releases. The December has the second highest percentage of 8.5 % movie releases. During the month of December, more number of family and holiday based movies are mostly released. August and April month has 8.3% of movie releases. The July month has 8.2% movie releases and other months cover 43.4% movie releases. By the above chart, the starting of the year January has the highest number of movie releases.

E. Number Of Movies Versus The Profit Of The Month

Releasing a movie is not a big fact, but releasing the movie on the correct season is little bit difficult. In this, the analysis is carried out for predicting the favourable month for releasing a movie, if it wants to be a block buster movie. The historical box office performance of movie for every month serves as a seasonality index for the studios to give them a look at the potential revenue for a weekend to face the potential of their own films and the other films slated for that date. The studios want to make sure that they are not putting a film with tons of effort on a month when people just don't turn out to the movie.

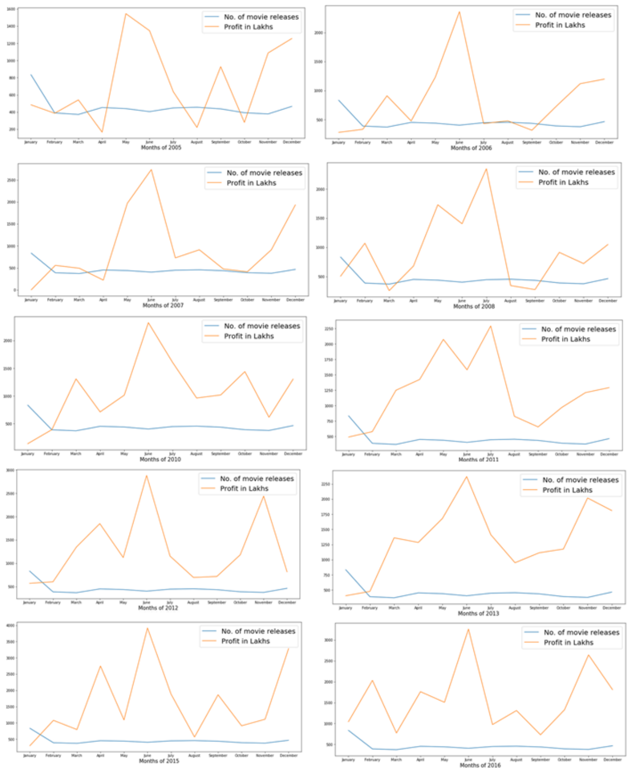


Figure 3.3 shows relation between Number of Movies and Profited Month in a year (Years between 2005-2016)

This analysis is carried out for analysing the historical box office performance of movie for every month. This is to find the relationship between the profited month in a year and the number of movies releases. Analysis is made for 12 years between the years 2005 – 2016. For each and every year, the number of movies released and profited month in particular year is analysed by line graph to provide better visualisation of the analysis.

In x-axis, the month name is taken and in y-axis, the movie count and movie profit is scaled. In the year 2005, the month of May month has the higher profit rate and December month has a profit rate little lower than May month. In the year 2006, the June month has the highest profit rate. In the year 2007,

2010 the month of May to July has the highest profit. In the year 2008 and 2011, July month has the highest profit rate. In the year from 2012 to 2016, June month peaks the highest profited month. The number of movie releases is considerably high during the starting month of the year. Though, the January month has the higher number of movie releases than any other month, the profit of movies on January month is relatively very low. The March and September months during 2007, 2008, 2012, 2015, 2016 years has a very low profit. On an average, the March month wasn't a lucrative period for movie success. The month of November to December seems to be little profitable month next to the mid months of the year.

During the year (2005, 2007, 2012, 2013, 2016) the November to December month seems to be average profited month compared to other months.

From the analysis of 12 years data, it is clear that always the mid months of the year like May, June and July has relatively highest profit compared to other months. Even when the movies are released less in number, the profit seems to be considerably high during these months. This infers that movie been released during these seasons in a year can make up the movie profit.

F. Language With More Number Of Movie Releases

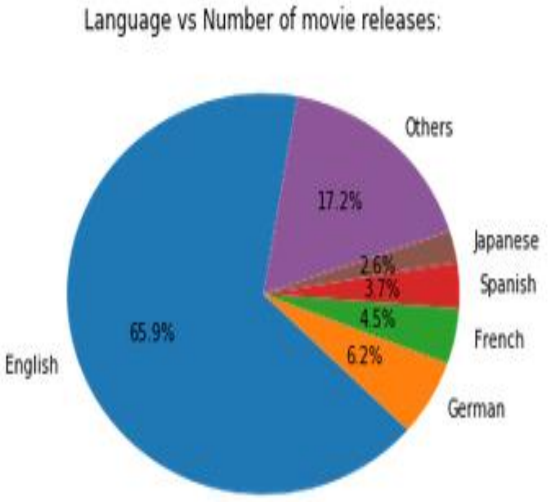


Figure 3.4 shows language with more number of movie releases

In each and every language, a large number of movies are releasing annually. For instance, a Hollywood film typically takes 36 months to plan and 12 months to execute; where as an Indian film takes 6 months in planning and 18 months in execution. Indian film makers say that what they lose out in planning, they make up on account of time and cost-efficient execution. By releasing a large number of movies, the experience and the perfection in the movies will be increased. From the dataset, the total languages is considered and top five languages with highest percentage of movies and other remaining languages are analysed and represented in pie-chart. This representation depicts which language has highest number of movies.

G. Country With More Number of Movie Releases

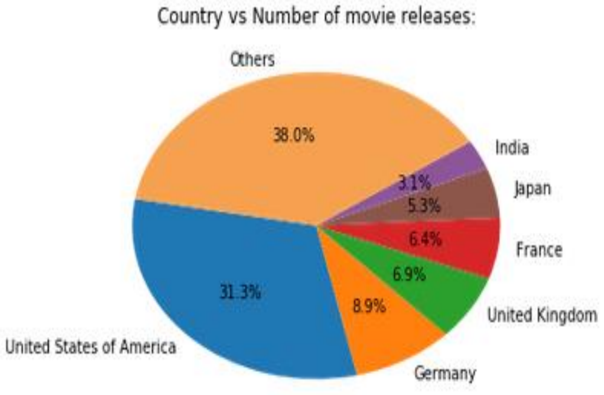


Figure 3.5 shows the country with more number of movie releases

The movie success contributes a major change in world economy. Each country releases movies every year. There are countries which release movies on a large scale compared to other countries. An analysis is done to depict which country releases more number of movies. From above pie chart, United Nations has the highest percentage of movie releases with 31.3% among all the countries in the world. It depicts that more number of Hollywood films are released every year than any other movies. Germany stands second in more number of movie releases and top six countries with the highest number of movie releases is listed and other countries are grouped into 'others' and they hold 38% of movie releases.

H. Genre With More Number Of Movie Releases

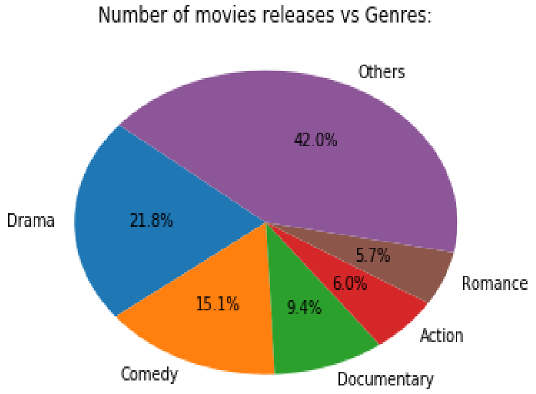


Figure 3.6 shows the genre with more number of movie releases

In this graph, there is different kind of genre taken from the dataset and top 5 genres that has the highest count of particular genre movies is taken and the other genres are grouped into attribute named 'other' and then graph is plotted between the top five genre and the other genre.

In this, the other type of genres apart from the top five genres such as horror, romance, action, thriller, western, family, crime based movies etc…shows the highest percentage of 27.5%, the drama depicts the 23.6% from the top five genres. The documentary movies shows 20.0%, the comedy movie shows 17.7%, music based movie shows 6.3%, and animation movie gives 5.0%.

IV. CONCLUSION AND FUTURE WORK

From the analysis, it is concluded that considering seasons for movie release is a concern. If a specific type of movie released on a specific month and if it has performed well, similar aspiring movies in genre, production size etc. may be slotted onto that weekend of that month in the future. And also each and every year, release dates for movies are affected by some events in the news and also by bigger events like the Olympics, elections, the World cup and so on and so forth. Big media events like cricket and other sports, presidential elections takes in people's money/time in watching it and people’s attention absolutely and affects the movie participation. So, studios releasing the movie has to consider the seasonality of the movie release for the movie profit. It has been analysed from the historical data that the months like May, June and July have higher profit when compared with other months of the year. It is also found that in these mid months of the year,

the number of movies released is comparatively less. The March month, September is not considered to be profitable month for movie releases as the profit of these months is comparatively very low. It has been analysed that the January month, the start month of the year has the highest

number of movie releases worldwide. But though, it has more number of movie releases,

the profit of the movie is not higher. So, based on this prediction, the studios and directors can make sure that upon releasing a movie with tons of potential on these specific months when more people turn out to the movie.

The type of genre people liked to watch is to be considered. Though more movies are released not all people watches all kinds of genres. Upon analysing which type of genre is liked by people, producing a movie with such genre makes more people turn out to the movie. The movie success is independent to the factors like the budget of the movie, it doesn’t make profited movie even when lot of budget is invested in a movie.

From this analysis, it’s clear to the point that to make a movie to be successful, the studios, releasing the movie has to release the favourable genre movie of people’s interest on favourable months between May to June or also between November and December months. This concept can also be extended to the production companies upon producing a favourable genre movie of people’s interest which turns to be profitable.

REFERENCE

1. Quader, N., Gani, M. O., & Chaki, D. (2017, December). Performance evaluation of seven machine learning classification techniques for movie box office success prediction. In Electrical Information and Communication Technology (EICT), 2017 3rd International Conference on (pp. 1-6). IEEE.
2. Jain, V. (2013). Prediction of movie success using sentiment analysis of tweets. The International Journal of Soft Computing and Software Engineering, 3(3), 308-313.
3. Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal, 2(6), 282-292.
4. Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 453-462.
5. Sumathi, VP, Kousalya, K, Vanitha, V, Cynthia, J, (2018), ‘ Crowd estimation at a social event using call data records’, Int. J. Business Information Systems, Vol 28, No. 2, pp 446-461.