Prediction of Box Office Success for Movies

Prof. Manisha Gahirwal
Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Email: manisha.gahirwal@ves.ac.in

Akshay Wagh
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Email: akshay.wagh@ves.ac.in

Akash Indani
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Email: akash.indani@ves.ac.in

Shital Chaudhari
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Email: shital.chaudhari@ves.ac.in

Shreya Sherugar
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Email: shreya.sherugar@ves.ac.in

***Abstract*—Social websites play an important role to get data about what a particular person thinks about something. In this project we try to predict the box office success of movies. We analyze different factors like the comments on social websites, budget of the movie, ratings of actors and directors and other external factors to generate the revenue in.**

Keywords—Social, predict, sentiment analysis, revenue, facepager.

I. INTRODUCTION

Facebook, YouTube, Twitter and other such social networking websites are usually used by all people to share their views around the world via Internet. People usually feel free to comment on this websites. The lag of communication is simplified using this.

In today's world, prediction is done on various subjects. The future prediction is really an amazing task. So, having a craze for movies, we tried to create a prediction model for predicting the success of the movies and hence predicting its revenue. As soon as the trailer or teaser for a movie is launched, the viewers start watching it and share their comments and views on different social networking sites.

The first task is to collect comments about the movie before it is released. Collecting database a tough work. As data sets are not simply available, hence by using different techniques of web crawling and API's like google API, twitter API we will be receiving bunch of data. Also collecting the data from facebook using software like facepager.The redundant data will be will be removed by preprocessing the data. And by applying sentiment analysis we can find the polarity of the comment using the positive and negative words dictionary. In this dictionary we can either add new words or remove unwanted words.

The system then segregates the positive and negative data using sentiment polarity and after getting the polarity scale, the Revenue for that movie is predicted.

II. RELATED WORK

Data mining domain was required for learning how the prediction model works. Various Algorithms and prediction model was initially analyzed which were used in data mining domain for predicting the Rain forecasting, Natural Disaster and so on. Various groups are trying to use these planning in different fields.

Nagamma P+ [1] applied machine learning technique to gather data from IMDB(http://www.imdb.com/). Collected set of words were examined as positive and negative and processed straightforward. He selected a set of keywords(love, amazing, great, fantastic, and so on). The clustering algorithm used was K-means and DB-SCAN. Then he performed Sentiment analysis in which text preprocessing and text transformation was performed. But the above process gave less accuracy, so to improve the accuracy, fuzzy clustering was used. Auto regression model was used for sales prediction of current day using the previous day collection.

Sameer[2] performed hype analysis, in this paper, regression method was used for predicting the box-office success. The data was collected using Twitter API on hourly basis. The paper focuses on multiple linear regression model in which more than one explanatory variable is used to predict the one variable. It is a time bound model.

They used models like Three-tier architecture, Multiple linear regression model, Critical period, Contributing factors. The three tier architecture used the presentation layer which contained user requirements. User authentication was required for using these models. To increase the accuracy, different contributing factors considered were attention seeking of audience, star- cast, and category and holiday effect.

Yonsei[3] predicted the Ticket sales for movies using reviews for Korean movies. This paper was an implemented paper and actual results were found about the ticket sale. They collected there data sets from Korean website (http://www.kobis.or.kr) which was under Korean movie council.

They purely used Hadoop language of big data analytics. A scale was made for direction, music, acting, and player value and through which the sale was predicted. They processed their way in a systematic way, The Online review and its star rating collected was first Hadoop based and on the other side they collected the Box office data and using Search volume scale was preprocessed . The preprocessed result was then considered with external factors and further was passed into the prediction model consisting of linear model, Support vector model, artificial neural network (which was built using R language).

The paper gave a clear approach between various prediction models and also helped in gathering data from various sites.

III. PROPOSED WORK

The paper focuses on predicting the revenue of movies.

A. Data Collection :

Firstly we analyzed the input data to the system and how to collect the data set. Collecting data from various domains was not an easy task. As discussed earlier, many website API and repositories were checked for real time data of movies. Many software were analyzed which can be used to gather data set or comments and posts from social networking websites. The bulk source of data to gather was Facebook, Twitter and YouTube, where people leave their comment for every post or trailer. The YouTube and Twitter data was collected using their respective API's and Facebook data was collected using facepager software. The reason for collecting data from such different sources is due to limitations. YouTube API provide maximum 100 comments for each video.

Facepager (version 3.6) collected the data about viewers comment on particular post or trailer. The node was to be selected which consists of the name of the movie and object consisting of comments, likes or posts about that node were fetched. The database file is formed and also the data is stored in excel sheet (.csv). Many data sets related to the movie were collected and stored.

Twitter and YouTube data is collected using R-Studio and Python respectively, it uses Twitter API key and YouTube API key to search comments and the output is saved in .CSV format.

The internal factors director and actor ratings are collected by web crawling the rotten tomatoes website. The R-script is written in R-studio for collecting this data.

B. Internal and External Factors:

Internal factors such as actor ratings, director ratings, assosciation between different actors and directors is also very important for increasing the accuracy of the system.
Along with the multiple internal factors there are many external factors which have impact such as

- Holiday effect: If the movie is released on a series of holidays then obviously the viewers will be more and hence the revenue will increase.
- Adult: If the movie is only for adults or it is for kids also is also very important.

Other factors like budget and movie ratings are also considered for prediction.

C. Sentiment Analysis:

The data collected is not in proper manner for preprocessing, as it contains many redundant tokens, for example: "The movie has excellent story, must watch" in this case only the word Excellent has meaning which can be used for preprocessing rest all words are to be eliminated i.e. only particular words were required for preprocessing.

Sentiment analysis is required for removal of redundant data. It gathers meaningful words from sentences and arranges the data. R-studio is the software which we will be using for sentiment analysis, which gives the polarity of the words as positive and negative viewing the comments and comparing it with the data dictionary. Sometimes a situation arises about a word whether it should be consider as positive or negative, at that time neutral polarity is used which is compared finally along with the positive/negative polarity result.

It splits the sentence in words and then search those words in positive and negative words dictionary. And the calculates the polarity by subtracting positive words and negative words, if the answer is positive then polarity is positive, if negative then negative otherwise neutral.

D. Methodologies Proposed:

First we directly applied neural networks and predicted the results but we got the accuracy only about 55 percent which is very less. But then we realized preprocessing is required and then we performed clustering. The preprocessing of the data required proper algorithm, and hence various models were studied. The model selected were K-means and Artificial Neural Network.

K-Means Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.
K-means clustering is a method of classifying/grouping items into k groups (where k is the number of pre-chosen groups). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid.

A centroid is "the center of mass of a geometric object of uniform density", though here, we'll consider mean vectors as centroids.

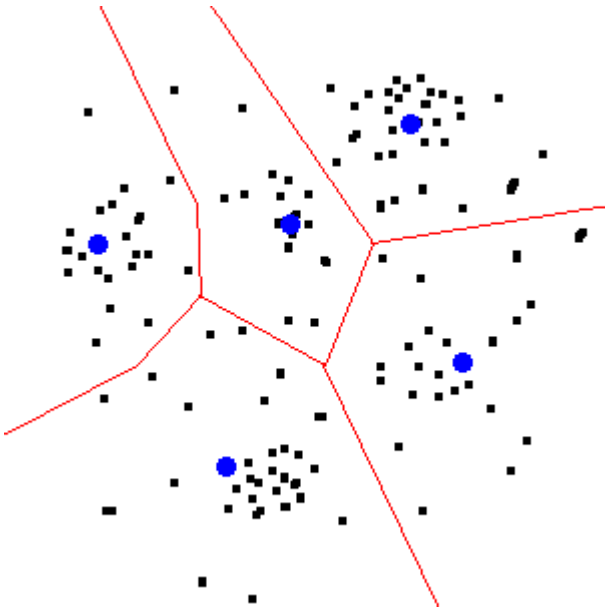


Figure 1. A clustered scatter plot. The black dots are data points. The red lines illustrate the partitions created by the k-means algorithm. The blue dots represent the centroids which define the partitions.

Algorithmic steps for k-means clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

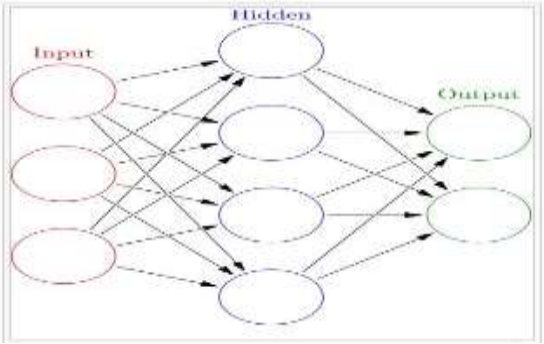
where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

ANN:

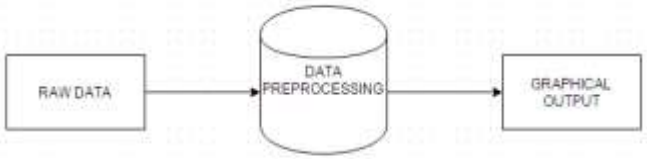
ANN is generally described as system of interconnected neurons which exchange some data between them. Those neurons have a weight as a numeric value which can be tuned according to the previous experience. This makes the neural system adaptive to the inputs and make it capable of learning.



Neural which is the class of statistical model possesses the characteristics such as adaptive weights and capability of approximating the non-linear functions.

E. Proposed Design and Working

The working model is discussed in brief, consider the figure



Basic proposed model:
The reviews from social networking websites are collected. The data is initially redundant, to select only meaningful words we use sentiment analysis.

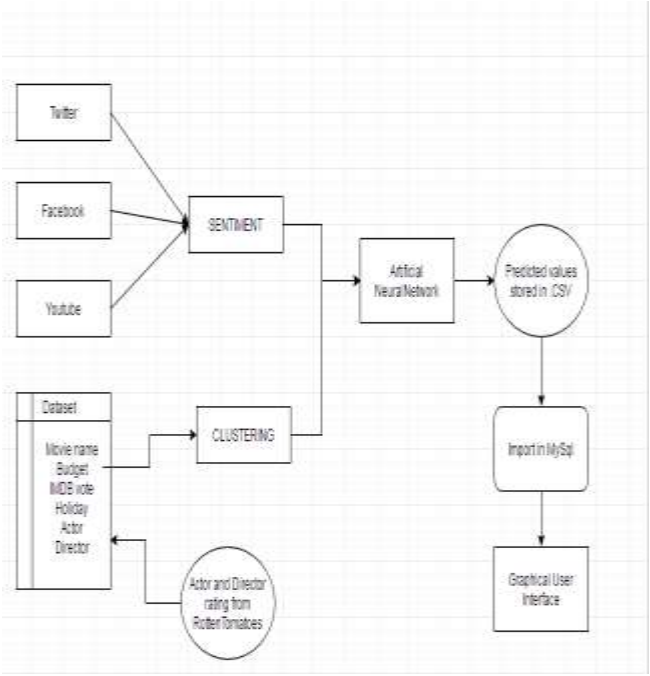
The data preprocessing step apply proper algorithm on the data and calculates the revenue for first three week.

The revenue calculated is shown in graphical format (bar graph or linear graph).

Flow of proposed diagram:
The paper overcomes the revenue prediction drawback from previous papers.

Comments, posts for a movie is collected from social networking websites like Facebook, Twitter, YouTube with the help of Twitter API, Facebook API, YouTube API and uses software named Facepager. The data is collected and stored as database file and .csv file. The Software or API retrieves a token for each user and with the help of that token data is fetched.

Many contributing factors are also consider like star-cast, director, sequel, and so on. Through which people might get more attracted towards the movie and contribute in the revenue.



The data set that will be retrieved would contain redundant data. The task of prediction only deals with proper data, hence for removal of redundant data from the data set we use Sentiment Analysis. The sentiment analysis is used with a data dictionary which is used to match the meaning of the comment's word. R-studio tool is used for doing this process.

The data is then preprocessed using suitable algorithm.

R-studio uses packages and code for calculating the polarity of positive and negative words and also for calculating the revenue of the movie. It also considers the contributing factor discussed earlier. The output resulted from preprocessing algorithm is then passed into the neural network and considering all factors the revenue will be calculated.

The output is calculated on the basis of trailer, the result of revenue is then compared with the actual box-office collection. The result is calculated and the revenue data is stored further in database which can be further viewed in by UI provided.

IV. Analysis

The result for a movie is given on the webpage as the name is submitted. The output would compare the predicted revenue and the actual revenue of the box office.

Error in the prediction can be calculated using Root Mean Square Error (RMS).

Minimum the error rate, maximum is the accuracy would be between the predicted and actual revenue.

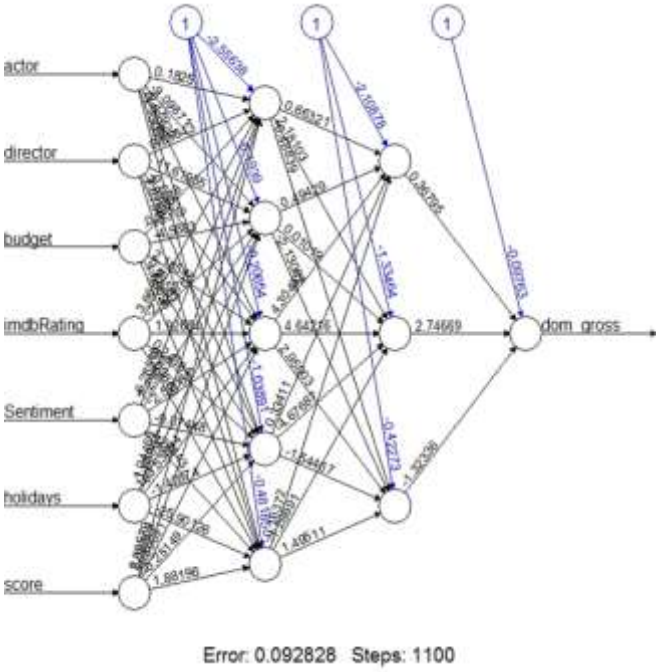
Users or customer could only view the output of the movie and not the inner process of the model.

V. Algorithm

Step 1: Fetch data from Facebook, twitter and YouTube. Also fetch data by web crawling Rotten tomatoes website.
Step 2: Storing information in organized manner using json (java script object notation).
Step 3: Applying text preprocessing on the data to remove the redundant data.
Step 4: Performing Sentiment analysis to calculate the polarity of each user review.
Step 5: Calculating other variable internal factors such as star-cast, music, director, etc and external factors such as holiday effect, festival effect, etc
Step 6: Based on above analysis predicting the revenue for the movie.
Step 7: Storing Results in database such that it can readily available whenever requested by the user.

VI. Result and Analysis

The UI elaborates the working of the whole system.
The Box-Office page contains top 5 upcoming movie predicted Revenue .It has a Home, Movies, Coming soon and about us menus.
Movies menu contains a drop down list of movies whose actual revenue is present in our dataset and the predicted revenue for each movie is done using R-Studio and stored in the database.
Selecting any of the movie gives the Predicted Revenue of that movie.
Coming Soon menu has images of the upcoming movies whose Prediction would be shortly done in coming days and will be updates on the HOME menu.



The Ann plots the neural graph considering the input factors and gives the predicted domestic gross.

VII. CONCLUSION

The sales performance of the near future is predicted using box-office data, internal factors such as star-cast and their association and external factors such as online reviews, movie ratings, holiday effect.

We learned prediction model like ANN. R-Studio software was used for prediction of revenue.

The users of the system will be the viewers and the managing team of box-office. Our prediction model forecasts the revenue and can determine the ideal screen number based on our projected sales and also explains that how proper use of social networking websites be an accurate indicator for future outcomes.

VIII. FUTURE SCOPE

We predicted the revenue of the movie. But using this revenue one can suggest the box-office mangers that how many screens should be allotted to the particular movie. Also best show timings can be predicted. This analysis will maximize their profit.

IX. REFERENCES

[1] Seonghoon Moon, Suman Bae, Songkuk Kim, "Predicting the Near- Weekend Ticket Sales Using Web-Based External Factors and Box-Office Data", Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences, Aug 2014

[2] Sameer Thigale , Tushar Prasad , Ustat Kaur Makhija , Vibha Ravichandran, "Prediction of Box Office Success of Movies Using Hype Analysis of Twitter Data"

[3] Nagamma P∗, Pruthvi H.R†, Nisha K.K‡ and Shwetha N H, "An Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction", International Conference on Computing, Communication and Automation (ICCCA2015)

[4] Bhave, A.; Kulkarni, H.; Biramane, V.; Kosamkar, P, "Role of different factors in predicting movie success", International Conference Pervasive Computing (ICPC), 2015