Integration of Classical Features, Social Media Response and Power of Stars to Predict the Box Office Sales

Mohammed Allama Hossain
Computer Science & Engineering
Institute of Engineering & Management
Kolkata, India
a2hossain2000@gmail.com

Rashika Daga
Computer Science & Engineering
Institute of Engineering & Management
Kolkata, India
rashikadaga@gmail.com

Saptarsi Goswami
Computer Science & Engineering
Institute of Engineering & Management
Kolkata, India
saptarsi.goswami@iemcal.com

***Abstract*— Predicting the box office success of a movie is a contemporary and relevant problem to solve given the size of Bollywood industry. Bollywood, with a high number of releases per year, spends a lot of money on promotion activities. Thus, it is critical to increase the average returns of movie and balance the cost thereafter. Most of the existing state of the art models and techniques for predicting box office sales focus either on classical feature or social media factors or influence of the stars. In this paper, we have integrated all these features and have built a model using random forest method. The empirical study consists of 95 movies from 2015 – 2016 timeframe. Our model produces 81.39% accuracy with 10-fold cross validation. The model is first of its kind as the success of the movie is predicted three days before the release considering the mentioned combinations of features.**

Keywords—Social Media; Box Office Sales; Statistical Model; Facebook; Influence of Stars; Random Forest; Predicting;

I. Introduction

According to the Dansk Industri International Business Development, a consulting unit of the Confederation of Danish Industry, Bollywood Industry based in Mumbai (former Bombay) generated revenues worth $3 billion(bn) in 2011 and this figure has been growing by 10 percent, Year on Year (YoY). It was expected to hit the mark of $4.5bn by 2016(International Business Times, 2013). This emerging industry employs over six million people in different roles [1]. The head count in the theatres in India was around 2.7 million(mn) in 2013 [2]. India produces about 1000 movies in a year, in Hindi language and other regional languages. Bollywood Industry is recognized as one of the growth-centric industries in India by the investors [3].

This research work aims to predict the range of Box Office Sales of Bollywood movies in Hindi Language three days before its release. The main aim of the paper is building a statistical model based on the classical features, social media responses and the influence of stars in the movie [6, 7]. Classical features include the title, genre, type, duration, cast of the movie, date of release of the movie [3], [5]. Social media response includes the reaction of the audience.

Data from Social Media networking sites like Facebook and Twitter has been used in the analysis of movies' success [8-10], [39], healthcare sector [4], stock market [34] and other diverse fields [12]. Actors, investors, producers, financers, directors, etc can use the predictive model to know the range of the financial success of a movie three days before its release. This will help them in making productive decisions on branding and promotion activities accordingly.

Bollywood movies rely heavily on promotion (online and offline) [13]. Due to these promotion activities, it creates a lot of buzz on social networking sites like Facebook, Twitter and Instagram. Peoples liking and reaction is an important role in determining the worth of the movie. It is commonly believed that online word-of-mouth or the online buzz creates a huge impact on the box office revenue [18]. Thereby, Facebook is considered for developing the statistical model because of the count of users (statistia.com, 2016) [19, 20]. To the best of our knowledge, research work has not been carried out yet considering the combination of parameters we have taken into account. Therefore, this is the first attempt in this domain.

II. Related Works

There are various models for predicting the box office sales of Bollywood movies. Barry R Litman is the pioneer in developing regression model to depict the financial success of theatrical movies [21]. In research work[22, 23] regression model was built on the hype (buzz) created in Twitter, along with sequel, star and category of the film . The model works on the polarity of the sentiment data generated by Twitter. Asur and Huberman were the pioneers in correlating the quantitative and qualitative aspects of tweets [23].

Wenbin Zhang and Steven Skiena have incorporated the news data with data obtained from Internet Movie Database (IMDb) to generate the box office revenue. They have designed Regression and k-Nearest Neighbor method to predict the result [25]. Subsequently, Zhang and Varadarajan have concluded that online reviews are important via models built using Support Vector Regression and Simple Linear Regression. They have developed models for traditional data set (IMDb), combined data set (IMDb and News data), and checked the accuracy of both the models. The model on combined data set gave a better accuracy [26]. Yoo, Kanter and Cunnings carried out another research work [27] on IMDb data.

Paper by Liu suggested that word of mouth before the release and after release is significant in determining the aggregate box office revenue and weekly box office sales [28]. In paper [29], the critics rating, budget and stars influence is considered in predicting the box office effects of a film. It is demonstrated that negative reviews affect the performance more adversely than positive reviews enhancing it. However, Wyatt and Badger have conducted research on positive, negative reviews and non-reviews [31,32,33]. In three subsequent work they have established that positive and negative reviews hardly portray the interest of an individual

TABLE I. RELATED WORKS- METHODOLOGY AND DETAILS

Paper	Year of Publication	Data	Attributes	Prediction	Algorithm and Models
[38]	2013	Korea Box Office Information System (KOBIS)	Director Star Value Sequel Release Date Genre KMRB Rating	Gross Box Office Revenue for Domestic Films (Korean Films)	Regression Methods: Linear Model, Random Forests Model and Gradient Boosting Model
[8]	2015	YouTube Twitter and Wikipedia	Views of a movie trailer Tweets View and edit counts of a particular movie	Movie Success	Multi-variate Linear Regression
[9]	2015	Twitter	Tweets No. of tweets	Success of a movie and Box Office Sales of Bollywood Movies	Probabilistic Latent Semantic Analysis classification model Fuzzy Inference System Algorithm
[16]	2013	Wikipedia	No. of users No. of edits Collaborative rigor No of views	First Weekend Box Office Revenue	Linear Regression Model
[10]	2013	Twitter	Tweets sent in the critical Period (Before 2 weeks and after 4 weeks) Number of tweets	Movie Success	Lingpipe Sentiment Analyzer 8 gram language model
[14]	2009	Blogs	Movie references in Blogs Critic Ratings User Ratings	Movie Sales and User/ Critics Rating	Correlation, KL Divergence Method, and Clustering
[23]	2010	Twitter	No. of tweets per author Rate of tweets in the critical period (1^{st}, 2^{nd} and 3^{rd} week before release)	Movie Revenue Prediction	Regression
[40]	2016	IMDb and Trailer Videos	Budget Genre MPAA Rating Release Period Sequel Actor's experience First Week Screens	Success Prediction of Movies	Regression Analysis

and reviews with more content and information is more impactful than positive reviews. In paper by Chintagunta, Gopinath and Venkataraman [35], they have combined the online reviews with the local market feedback and proved that this aggregation gives a better accuracy. Huan Jianxiong and the other authors in their paper [36] have considered expert reviews, network based peer reviews, and non-network based peer reviews, along with control variables as the classical features and star power. However, they have not build a model until now neither given any kind of proposition. In research work [37], the authors applied neural network analysis and statistical modeling to predict the Box Office Sales. The overall accuracy was 36.9% and accuracy of 75.2% within a category.

Thus, our study is first of its kind in determining the range of Box Office Sales 3 days before the release using statistical analysis, while most have used sentiment analysis. And, none of the works have been carried out with the parameters considered in our study. Table I lists the works that have been carried out in this domain, along with the methodology and the type of data used.

III. DATA MINING AND DATA PREPROCESSING

A. Data – Type and Source

TABLE II. DATA ATTRIBUTES AND ITS SOURCE

Parameter	Source
Like Count of the Official Page of the movie in Facebook	Facebook API
Times Celebex Rating	www.timescelebex.com
Whether the movie is released during public holidays or festivals	Calendar-Wikipedia
Duration of the movie	IMDb
Genre of the movie	IMDb
Whether the movie is a Sequel	IMDb
Type of the Movie	BookMyShow
Total Nett Gross	Box Office India

The extracted data is structured in nature. The dataset was prepared for Bollywood movies released in the period of 2015 to 2016. Two hundred and four Hindi movies (including dubbed versions) were released in 2015 (Wikipedia, 2015.) However, not all the movies of 2015 were considered in the dataset due to the absence of an official movie page on Facebook [41]. Similar is the trend followed by movies release in 2016. Many movies do not maintain an official page on Facebook and hence, relevant data for the particular movies could not be obtained. In Table II, we have listed all the attributes along with the respective website/API that were used to collect the data from various online sources. Web Scrapping was done using R Language (Version Rx64 3.3.2) with data being stored in a csv file. Some packages used for scraping data from different websites include [45-47].

B. Data Preprocessing

The data scrapped from online sources must be converted into usable binary or numeric data. This conversion is done for simplicity while applying it to various models. It is ensured that the effect of the data remains the same even after converting it.

1) Star Value

The entire cast is retrieved from IMDb. If at least one of the Actors in the cast list lie in the Top 10 list in the Times Celebex Rating 1 month prior to the release date, the Times Celebex Male is assigned 1 in this case, else 0. The same process is repeated for Actresses and the values are stored in a separate column named Times Celebex Female.

2) Facebook Likes

The Facebook likes are generated at 6 intervals starting from 3 months up to 3 days before the release date at various instants- 3 months, 2 months, 1 month, 15 days, 7 days and 3 days prior to the release. The relative increase or decrease is calculated, rounded off to 2 decimal places and stored in a separate column.

$$\text{Relative Percentage Increase} = \frac{\text{No. of likes in the period(m)} - \text{No. of likes in the period(n)}}{\text{No. of likes in the period(n)}} \times 100$$

where, (number of days in m< number of days in n)

The ratio of the like count of a movie at the beginning to the average like count of movies in that period is calculated. The data is converted into suitable logarithmic form using log transformation [15], [17], [30]. Similarly, the ratio of like count of a movie at the end (here, 3 days) to the average like count of movie in that period is computed and stored after applying the logarithmic transformation. The formula for calculation is stated below:

$$\log_{10} \frac{\text{Like count of a movie at a period n}}{\text{Average like count of all movies in period n}}$$

(Only the starting and end are considered)

The starting period of a movie is the first interval from 3 months to 3 days when the like count was non-zero. Similarly, the end period is 3 days before the release date for all the movies.

3) Category of the Film

The Category of the movie are retrieved as U/A or A or U from BookMyShow. It is again converted into binary variables. For Adult (A) movies, the isAdult flag is set to 1, else it is reset (set to zero) for U/A and U.

4) Genre

The genre has been divided further into 11 columns with names isAction, isComedy, isDrama, isHorror, isAdventure, isBiography, isThriller, isCrime, isMystery, isSports, isRomantic. The genre is compared with the list of genres. If a movie lies in a particular genre, the particular row for that column is set to one, and the others are filled with zeroes.

5) Holiday Effect

isHoliday is another parameter taken into consideration. The release date is increased by 1, 2 and 3 days. The dates are checked with the list of public holidays in Wikipedia for that year. If at least one of the dates fall on a holiday for more than 3 states, the isHoliday flag is set to 1, else 0.

6) Duration

The duration is broken down into 4 intervals- Less than 2 hours, 2-2.5 hours, 2.5-3 hours and greater than 3 hours.

7) Box Office Sales

After analyzing the box office sales for 95 movies, the range was divided into 3 classes. Class 1 comprises of movies with a net income less than 50 crores, Class 2 consists of movies with income greater than 50 crores and less than 100 crores and Class 3 with income greater than 100 crores. Since the data set is quite small due to non-availability of Official Facebook Pages, less number of classes was taken into consideration to maintain a better consistency and reduce the problem of over fitting.

IV. Proposed Methodology

The model was developed in R (Version Rx64 3.3.2) with the help of packages like [42 – 44].The data was split into training and test data. The analysis of features involves setting up the relationship among various features or parameters, which is important to the prediction of box office sales. We make use of the Random Forest method to predict membership of cases in the classes of a categorical dependent variable from their measurements on several predictor variables. Random Decision Forest is a modified version of Tree Bagging. It selects a random subset of features (feature bagging) while splitting a candidate, thereby reducing any kind of biasness in the model. Moreover, the method performs well as compared to Support Vector Machines, Neural Networks and helps in reducing the problem of over-fitting [11]. The Random Forest method also gives importance to all the variables in determining the required outcome.

After the development of working model on the training set, we use a statistically sound method called k-fold Cross validation to estimate the accuracy of the model on the test data. In k-fold Cross Validation Method, the entire dataset is divided into k mutually exhaustive subsets. One subset is chosen in a random order and the other k-1 subsets act as the training data. For estimating the accuracy of the model on the test data, k Cross Validation was applied with the value of k as

10. The k-fold Cross Validation is pessimistically biased for lower folds, with low variance and less computational costs and at higher folds, it is almost unbiased with higher variance and high computational cost. The estimate obtained is considerably good at 10 stratified, even if computational power allows using more folds [24]. Fig. 1. shows the entire working of the model starting from extraction to generation of the result in steps.

Fig. 1. Flowchart depicting the sequence of operations to obtain the result arranged in bottom to top fashion

IV. Result and Analysis

The classification tree obtained after applying the random forest method on the training set depicts the error associated with the random tree generated by the model (Fig. 2.).

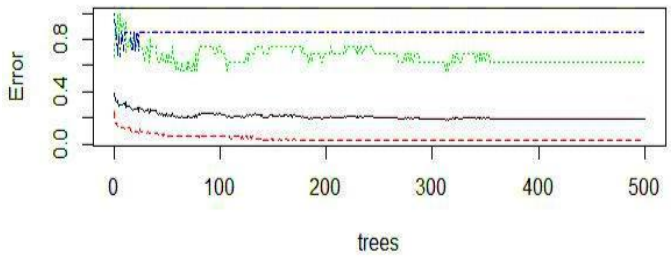

Fig. 2. Graph depicting the error rate across decision trees

In the above figure, the red solid line denotes the error associated with class 1 which is 0 – 50 crores, this line shows very less deviation from the bottom because of the availability of large cases in the data set. The green solid line is used to depict the error associated with the class 50 – 100 crores and the blue one delineates the error associated with class 3. The error tied with class 3 is maximum because in the dataset of 95 movies, only 7 movies lie in 100 crores club. The black solid line denotes the overall "out of bag error".

A. Importance of Features

Facebook has more than 1 billion registered accounts with 1.71 billion active users (statistia.com, 2016). The count of active users exceeds those of Twitter [19] [20], Google+ and Instagram [20]. Indians lead in the usage and consumption of Facebook with around 195 million users from India (statistia.com, 2016). Thereby, Facebook is considered for developing the statistical model because of the count of its users. In the current study, the seven important features depicted in Fig. 4., five of them are directly related to the Facebook Like Pattern. This supports the reason for choosing Facebook Likes as an important attribute in predicting the range using our statistical analysis. Facebook Likes before 15 days seems to play a major role in classification, followed by likes 7 days before the release. Increase from 2-to1 month is also noted as an important trend for analysis. The increase in Facebook likes from 7 days to 3 days before the release was an important parameter as this is the critical period that takes into account all major promotional campaigns for Bollywood movies.

The ratio of Facebook likes at the end (3 days prior to the release), namely log end/average is an important parameter in determining the range as depicted clearly by Fig. 4. The ratios corresponding to three different classes are well distributed as showed in Fig. 3. There is a slight overlapping in certain cases. But, the attribute when combined with other parameters gives an accuracy of 81.39% (k=10). The trend between the ratio of Facebook Likes and the range can be easily determined from Fig. 3. and this is a crucial factor for our statistical analysis.

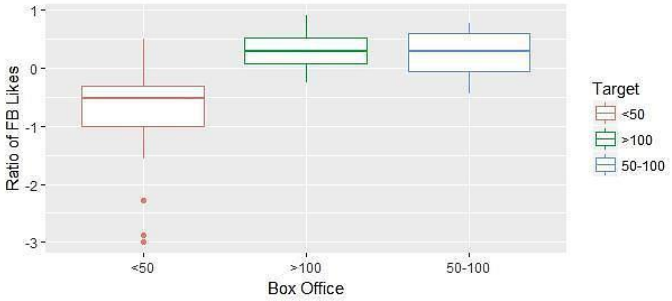

Fig. 3. Relationship of Ratio of Facebook Likes and Box Office Range (Social Media Response)

The buzz around every actor was collected in quantitative format from Times Celebex for a specific period before the release of the film to predict how much the audience is looking forward to the actor's film. The above stated factor (Rating of Male Actors) is the second important feature and helped in increasing the accuracy of the model as compared to other works, also depicted by Fig. 4. This can also be validated on the basis of the evaluation as well as visual analysis. For example, Akshay Kumar managed to stay in the Top 10 actor list given by Times Celebex for most of the time during the two year period during which he gave out hit films like Airlift, Rustom and Gabbar is Back, however he could not make the cut for a couple of months and that's when his movie Singh is Bliing did not do well. Also it was noted that the effect of a Bollywood actor on the financial success was more than a female actress' effect.

This can be supported by the fact that movies like Jazbaa, Jai Gangaajal, Shandaar and Dolly ki Doli with Aishwaryaa Rai Bachchan, Priyanka Chopra, Alia Bhatt and Sonam Kapoor respectively who made it to the Top 10 actresses in Times Celebex List 1 month prior to the release of the film could not deliver films with big box office sales. The

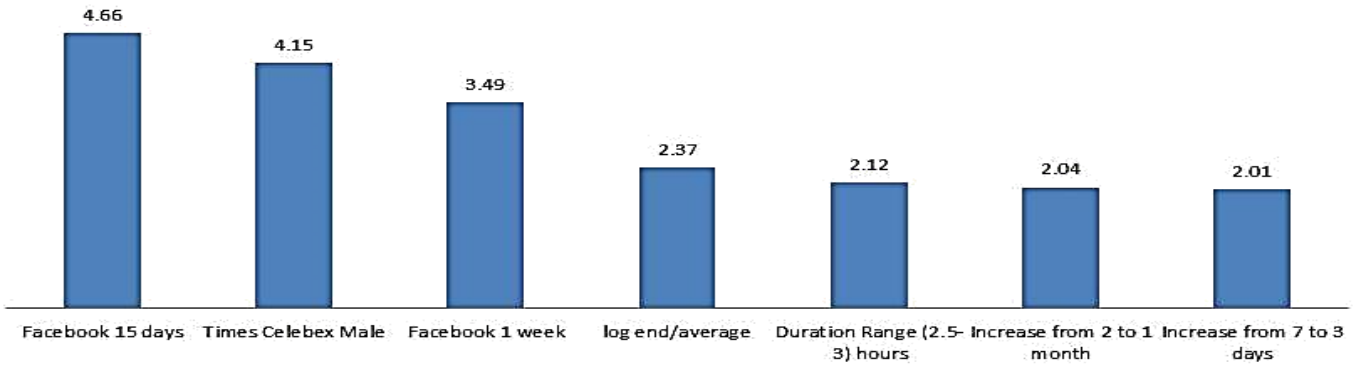

Fig 4. Bar Graph depicting the MeanDecrease in Gini coefficient of 7 important features in the analysis

same trend was observed from the importance of features deduced from the Mean Decrease in Gini coefficient generated by the Random Forest Model. Mean Decrease in Gini coefficient measures the contribution of each variable to homogeneity of leaves and nodes in Random Forest method. The value of MeanDecrease in Gini for actors was as high was 4.15, whereas it was 0.56 for actresses, thereby, establishing the importance of the two. Fig. 4. represents the MeanDecrease in Gini for 7 most important parameters deduced by the Random Forest Method.

B. Model Performance Evaluation

Accuracy of the Random Forest Method is estimated using k-fold cross validation technique as discussed earlier in Section III. For estimating the accuracy via k-fold cross validation, we assumed the value of k to be 10. Accuracy was selected using the optimal model for which the value of mtry was 44, giving an accuracy of 81.39%. Then, we measured the accuracy for different values of k (3,5,7,8,10). 5- fold cross validation showed the highest accuracy as 87.57%. The Random Forest Method thus gives a decent accuracy based on the parameters taken into consideration.

TABLE III. CONFUSION MATRIX FOR 10 FOLD CROSS VALIDATION

	Class 1	Class 2	Class 3	Class Error
Class 1	70	2	0	0.027
Class 2	7	6	3	0.625
Class 3	2	4	1	0.857

Table III represents the confusion matrix for the aggregated 10- fold cross validation Random Forest classifier results on the test data. A confusion matrix is suitable for representing the data in tabular form for the classification models. The rows represent the predicted value while the columns determine the reference or the comparison. The value in the intersection of rows and columns gives information about the correct number of samples in which the classification model predicted results matched the actual value. The error was least for Class 1 and maximum for Class 3. The reason behind this is the dataset and the distribution pattern. Out of 95 movies, Range for 72 movies lied in Class 1, whereas Class 3 witnessed 7 movies in the dataset only.

V. CONCLUSION

Most research works that have been carried out in this particular field have been based mostly on classical features and sentiment analysis of textual data. The novelty of this work is that it firstly, integrates various factors like rating of actors, Facebook responses and classical features and secondly the prediction is available before the release of the movie. Rating of actors turned out to be a very important factor because in several countries, certain actors enjoy a strong fan base which becomes a major deciding factor of the success of a movie. The random forest model produces a classification accuracy of 87.57% with 5-fold cross validation. One problem that was faced throughout the research work was the limited availability of data-set and this was because relatively fewer number of movies maintained an official Facebook page but this problem will seize to exist in the coming years because the promotion teams are realizing that the digital world (such as Facebook) is becoming an omnipresent tool for marketing. Also the API does not allow to access data prior to 2 years, however as the testbed contains recent movies only it does not have any bearing on the results and conclusion of the said study.

REFERENCES

[1] Gupta,(2 9), “Brand Bollywood”, available at: http://theviewspaper.net/brand-bollywood/

[2] Niall McCarthy (2 4), “Bollywood: India's Film Industry By The Numbers [Infographic]” available at http://www.forbes.com/sites/niallmccarthy/2014/09/03/bollywood-indias-film-industry-by-the-numbersinfographic/#6d6639df7bf0

[3] Fetscherin, Marc. "The main determinants of Bollywood movie box office sales." Journal of global marketing 23, no. 5 (2010): 461-476.

[4] Bamwal AK, Choudhary GK, Swamim R, Kedia A, Goswami S, Das AK. Application of twitter in health care sector for India. In Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on 2016 Mar 3 (pp. 172-176). IEEE.

[5] Overdorf, Jason. "Bigger Than Bollywood." Newsweek International 10 (2007).

[6] Elberse, Anita. "The power of stars: Do star actors drive the success of movies?." Journal of Marketing 71, no. 4 (2007): 102-120.

[7] Karniouchina, Ekaterina V. "Impact of star and movie buzz on motion picture distribution and box office revenue." International Journal of Research in Marketing 28, no. 1 (2011): 62-74.

[8] Bhave, Anand, Himanshu Kulkarni, Vinay Biramane, and Pranali Kosamkar. "Role of different factors in predicting movie success." In Pervasive Computing (ICPC), 2015 International Conference on, pp. 1-4. IEEE, 2015.

[9] Gaikar, Dipak Damodar, Bijith Marakarkandy, and Chandan Dasgupta. "Using Twitter data to predict the performance of Bollywood movies." Industrial Management & Data Systems 115, no. 9 (2015): 1604-1621.

[10] Jain, Vasu. "Prediction of movie success using sentiment analysis of tweets." The International Journal of Soft Computing and Software Engineering 3, no. 3 (2013): 308-313.

[11] Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.

[12] Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." ICWSM 10 (2010): 178-185.

[13] Liu, Yong. "Word of mouth for movies: Its dynamics and impact on box office revenue." Journal of marketing 70, no. 3 (2006): 74-89.

[14] Sadikov, Eldar, Aditya G. Parameswaran, and Petros Venetis. "Blogs as Predictors of Movie Success." In *ICWSM*. 2009.

[15] Bartlett, J. W., and C. Frost. "Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables." Ultrasound in Obstetrics & Gynecology 31, no. 4 (2008): 466-475.

[16] Mestyán, Márton, Taha Yasseri, and János Kertész. "Early prediction of movie box office success based on Wikipedia activity big data." PloS one 8, no. 8 (2013): e71226.

[17] Osborne, J. "Notes on the use of data transformations." Practical Assessment, Research and Evaluation 9, no. 1 (2005): 42-50.

[18] Walsh, Ekaterina. "Entertaining young net surfers." Forrester Research, Cambridge, MA (2000).

[19] Rani Molla, (2 6), "Social Studies: Twitter vs. Facebook" available at https://www.bloomberg.com/gadfly/articles/2016-02-12/social-studies-comparing-twitter-with-facebook-in-charts

[20] Justin Kerby,(2 6), "Here"s How Many People re On Facebook, Instagram, Twitter and Other Big Social Networks" available at http://www.adweek.com/socialtimes/heres-how-many-people-are-on-facebook-instagram-twitter-other-big-social-networks/637205

[21] Litman, Barry R. "Predicting success of theatrical movies: An empirical study." The Journal of Popular Culture 16, no. 4 (1983): 159-175.

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

[23] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, pp. 492-499. IEEE, 2010.

[24] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In Ijcai, vol. 14, no. 2, pp. 1137-1145. 1995.

[25] Zhang, Wenbin, and Steven Skiena. "Improving movie gross prediction through news analysis." In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 301-304. IEEE Computer Society, 2009.

[26] Zhang, Zhu, and Balaji Varadarajan. "Utility scoring of product reviews." InProceedings of the 15th ACM international conference on Information and knowledge management, pp. 51-57. ACM, 2006.

[27] Yoo, Steven, Robert Kanter, and David Cummings. "Predicting Movie Revenue from IMDb Data." (2011).

[28] Kimmons R (2011) Understanding collaboration in Wikipedia. First Monday 16:12.

[29] Basuroy, Suman, Subimal Chatterjee, and S. Abraham Ravid. "How critical are critical reviews? The box office effects of film critics, star power, and budgets." Journal of marketing 67, no. 4 (2003): 103-117.

[30] Parsons, Helen M., Christian Ludwig, Ulrich L. Günther, and Mark R. Viant. "Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation." *BMC bioinformatics* 8, no. 1 (2007): 234..

[31] Wyatt, Robert O., and David P. Badger. "How Reviews Affect Film Interest and Evaluation." (1984).

[32] Wyatt, Robert O., and David P. Badger. "To toast, pan or waffle: How film reviews affect reader interest and credibility perception." Newspaper Research Journal 8, no. 4 (1987): 19-30.

[33] Wyatt, Robert O., and David P. Badger. "Effects of information and evaluation in film criticism." Journalism & Mass Communication Quarterly67, no. 2 (1990): 359-368.

[34] Kumar S, Maskara S, Chandak N, Goswami S. Empirical Study of Relationship between Twitter Mood and Stock Market from an Indian Context. International Journal of Applied Information Systems (IJJAISS). 2015;8:1-5.

[35] Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman. "The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets." Marketing Science 29, no. 5 (2010): 944-957.

[36] Huang, Jianxiong, Wai Fong Boh, and Kim Huat Goh. "From A Social Influence Perspective: The Impact Of Social Media On Movie Sales." InPACIS, p. 79. 2011.

[37] Sharda, Ramesh, and Dursun Delen. "Predicting box-office success of motion pictures with neural networks." Expert Systems with Applications 30, no. 2 (2006): 243-254.

[38] Song, Jongwoo, and Suji Han. "Predicting Gross Box Office Revenue for Domestic Films." Communications for Statistical Applications and Methods20, no. 4 (2013): 301-309.

[39] Mishne, Gilad, and Natalie S. Glance. "Predicting Movie Sales from Blogger Sentiment." In AAAI spring symposium: computational approaches to analyzing weblogs, pp. 155-158. 2006.

[40] Tadimari, A., Kumar, N., Guha, T. and Narayanan, S.S., 2016, March. Opening big in box office? Trailer content can help. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 2777-2781). IEEE.

[41] Awl, D., 2010. Facebook me! A guide to socializing, sharing, and promoting on Facebook. Pearson Education.

[42] Piotr Romanski and Lars Kotthoff (2016). FSelector: Selecting Attributes R package version 0.21.

[43] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7.

[44] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[45] Pablo Barbera, Michael Piccirilli and Andrew Geisler (2016). Rfacebook: Access to Facebook API via R. R package version 0.6.6.

[46] Duncan Temple Lang and the CRAN team (2016). RCurl: General Network(HTTP/FTP/...) Client Interface for R. R package version 1.95-4.8.

[47] Hadley Wickham and Winston Chang (2016). devtools: Tools to MakeDeveloping R Packages Easier. R package version 1.12.0.