Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction

Nahid Quader[1], Md. Osman Gani[2] and Dipankar Chaki[3]
Department of Computer Science and Engineering
School of Engineering and Computer Science
BRAC University, Dhaka, Bangladesh
Email: [1]whereisnahidquader@gmail.com, [2]usmansujoy33@gmail.com, [3]joy.dcj@gmail.com

Abstract— **Movie industry is a multi-billion-dollar industry and now there is a huge amount of data available on the internet related to movie industry. Researchers have developed different machine learning methods which can make good classification models. In this paper, various machine learning classification methods are implemented on our own movie dataset for multi class classification. The main goal of this paper is to conduct performance comparison among various machine learning methods. We choose seven machine learning techniques for this comparison such as Support Vector Machine (SVM), Logistic Regression, Multilayer Perceptron Neural Network, Gaussian Naive Bayes, Random Forest, AdaBoost and Stochastic Gradient Descent (SGD). All of these methods predict an approximate net profit value of a movie by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Meta Critic. For all these seven methods, the system predicts a movie box office profit based on some pre-released features and post-released features. This paper analyzes the performance assessment of all these seven machine learning techniques based on our own dataset which contains 755 movies. Among these seven algorithms, Multilayer perceptron Neural Network gives better result.**

Keywords— Movie industry, box office success, machine learning, classification techniques, performance evaluation.

I. INTRODUCTION

Movie industry is a very big industry. For big business sectors, it is hard to take a decision where to invest. Predicting a movie's box office success is a very complex task and for that, at first we need to clarify the definition of success. The definition of success for a movie is relative, some movies are called successful based on its worldwide gross income, where some may not shine in business part but can be called successful for good critics' review, good rating and popularity. In this paper, we consider a movie's box office success based on its profit only. Researches show that almost 25% of movie revenue comes within the first or second week of its release [1]. Predicting a movie's profit is hard before its release. In this research, two types of features called pre-released features and post-released features are considered. Pre-released features are responsible for upcoming movies' box office success prediction while both pre-released and post-released features will take place to predict further more after its release. Instead of binary classification like only flop or blockbuster movies [2], we rather choose to classify a movie based on its box office profit in one of five categories ranging from flop to blockbuster. In this research, we have calculated two types of prediction, one is exact match which refers correct classification and the other is, one away which means taking consideration of one class up or one class down from a particular class along with the exact match [3]. The reason behind considering one away is, sometimes percentage of accuracy becomes low because of the marginal value of target classes. For prediction, seven machine learning methods Support Vector Machine (SVM), Logistic Regression, Multilayer Perceptron Neural Network, Gaussian Naive Bayes, Random Forest, AdaBoost and SGD are implemented. These classifiers are good enough for binary classification and some of them can be used for multi class classification. In our dataset, most of the data point are overlapping for different classes. That is why, our dataset has a very complex pattern. However, when data pattern is very complex, Neural Network consistently produces better result. We have applied all these methods on our dataset for prediction. With all features in consideration, from 755 movies, neural network correctly classifies 442 movies. If we consider one away prediction the number of correctly classified movies becomes 677. One away prediction means difference between predicted class and target class is 1. Among all these methods MLP Neural Net (58.5%) and SVM (55.3%) work better than others.

In the next section, we mention about different research works related to movie box office success prediction. In section 3, we briefly explain our dataset. Seven machine learning classification techniques have been discussed in section 4. We discuss about experimental results in section 5. At the end of our paper, in section 6, conclusion along with future works are written. Finally, we conclude our paper by mentioning some references.

978-1-5386-2307-7/17/$31.00 ©2017 IEEE

II. LITERATURE REVIEW

Previously many research works made prediction of a movie's box office gross based on data available at IMDb [4] - [6]. Few of them prioritized gross box office revenue [7] - [9]. Most of the cases, they used binary classification for prediction of a movie's success and classified as either flop or success. Success of a movie depends on many relevant issues like casting members, story of the movie, number of screens the movie will be shown etc. In some previous researches, prediction of a movie success was made based on available pre released data [3]. On the other side, few researches adopted different application of Nature Language Processing for sentiment analysis on reviews of audiences and movie's critics for forecasting the success of a movie [10], [11].

M. T. Lash and K. Zhao's [2] categorized their analysis into audience based, released based and movie based while developing their model using machine learning techniques. Their data source was Box Office Mojo and IMDb. They made the model by focusing on the movies released in USA and excluded all foreign movies. In [12], a neural network was implemented for movie box office gross prediction. They categorized the net profit in 9 classes and converted it into a classification problem. But a small number of features were considered in this research. On the other side Sivasantoshreddy et al. [13] focused on hype of twitter for the prediction [13]. Their hypothesis was the success of a movie mainly depended on its opening weekend income and hype it got from the audiences. Hype factor, number of screens where the movie was to be released and the price of tickets were taken in consideration in this research. In this work, they did not apply any application of NLP for analyzing the positivity and negativity of tweets. Furthermore, their model was based on very simple calculations.

M.H Latif and H. Afzal also used IMDb data in their model and they mentioned their data was not clean and inconsistent [14]. As a result, they applied central tendency to fill the missing data of different features. Jonas et al. [15] calculated intensity and positivity by analyzing sentiment of social network IMDb's sub forum Oscar Buzz. Their model failed to get a sentiment result when some words were used for negative meaning. In their work, they considered different awards for directors and casting of a movie but no category was taking in consideration. In the research [16], a model was made by analyzing news from different sources where data were generated by Lydia (high-speed text processing system for collecting and analyzing news data). They took high budget movies in consideration and applied regression and *k*-nearest neighbor. The significant problem was it could not identify the movie name from the news when common words were used as a movie name and it was unable to predict the success of a movie if there were no news about that movie.

In some early researches, neural networks were used for the prediction of a movie box office gross [3], [17]. While some other research models were based on social media, social networks and hype analysis [18] - [21]. In that cases, they figured out the sentiment of the audience reviews, the number of reviews and most of the cases their sources were IMDb sub forum Oscar Buzz, Twitter and YouTube. Audience reviews can be biased for a particular actor/actress and most of the previous works excluded movie critics' reviews. Furthermore, a large number of people did not consider the number of the screens in their analysis. For those reasons, the prediction accuracy will be doubtful and it is not possible to generate an appropriate result. For a movie, two types of data are available on the internet, one pre-released data including director, budget, cast members, the number of screens etc. The other one is post-released data which are available on IMDb, Box Office Mojo etc. In some cases, few researchers considered both types of data but very few of them were used in the prediction model.

III. DATA DESCRIPTION

Our dataset contains 755 movies released in between 2012 to 2015. We exclude recent movies as movies' information are changing every day. Our data sources are IMDb, Rotten Tomatoes, Metacritic and Box Office Mojo. Initially our dataset contains 3183 movies. Most of the features are missing for most of the movies. In many cases, movies budgets are unavailable. After removing those movies, we have data for 800 movies. Among these, budget of some movies were available but other features were not found. After excluding those movies, finally we have a dataset containing all the information of 755 movies. Table I shows the description of all features in our dataset. We use both pre-released and post-released features in our model. Total 15 features are used in our proposed model. Among these 15 features MPAA, cast star power, director star power, no of screen, release month and budget are pre-released features, rest of the features are post released.

TABLE I. DATASET DESCRIPTION OF ALL FEATURES

Features	Type	Description/Range of possible values
IMDb Rating	Float	0 to 10
Tomato Meter	Integer	0 to 100
Tomato Rating	Float	0 to 10
Audience Meter	Integer	0 to 100
Audience Rating	Float	0 to 5
Meta Score	Integer	0 to 100
MPAA	Integer	Value between 1 to 6 indicating G, PG, PG-13, R, NC-17, NR respectively.
Cast Star Power	Integer	Addition of all casts' lifetime gross income
User Review	Float	Sentiment value multiplied by no of reviews
Critics Review	Float	Sentiment value multiplied by no of reviews
IMDb Votes	Integer	Number of IMDb votes
Release Month	Integer	Between 1 to 12
Budget	Integer	Budget of a movie
No of Screen	Integer	Number of screens a movie released
Director Star Power	Integer	Addition of directors' lifetime gross income

We take in consideration of Motion Picture Association of America (MPAA) Rating, tomato critics' meter, tomato critics' rating, tomato audience score and tomato audience rating from Rotten Tomatoes, Meta score of Metacritic, IMDb rating from IMDb for a particular movie. We also count the number of viewers who rated the movies which is IMDB votes. Sentiment value of reviews and no of reviews are multiplied together and used as a single feature for both audience reviews from IMDb and Rotten Tomatoes along with critic reviews from Rotten Tomatoes. We calculate star power of actors, actresses and directors. Star power of a single artist is calculated by summing up the gross income of all movies done by that specific artist during his/her career. And Star power of a movie is the addition of those total gross value for all the casts involved in that movie. We also calculate directors' star power in the same manner.

In addition, we also consider the month of release, budget and the number of screens. This paper includes budget with inflation rate adjustment. Inflation rate is very important because value of money is changing over time, value of $100M five years ago is not same as now.

Instead of binary classification like "Flop" or "Blockbuster" movie, we rather choose to classify our target class into five classes. Table II describes our target class classification where lower class means lower profit, class 1 means flop movies and class 5 means blockbuster movies.

TABLE II. TARGET CLASS CLASSIFICATION

Target Class	Range (USD)
1	Profit <= 0.5M (Flop)
2	0.5M < Profit <= 1M
3	1M < Profit <= 40M
4	40M < Profit <= 150M
5	Profit > 150M (Blockbuster)

IV. METHODS

We have used seven machine learning methods for performance assessment of our dataset. All these methods are briefly described in this section. We implement most of the algorithms using python library Scikit Learn [22].

A. Logistic Regression

Logistic Regression is a kind of regression model where attributes are categorical and that is the reason to go for it. In our dataset we categorized most of the attributes [3]. Logistic regression is good for classification but for a small amount of data. In this case we have a very small dataset so we choose logistic regression as one of our methods for comparison. Furthermore, logistic regression works well with noisy data. Again it has low variance and so is less prone to over-fitting.

B. Support Vector Machine

Support Vector Machine is a very popular machine learning algorithm for pattern recognition and classification. SVM is best for binary classification but it also can produce a good result for multiclass classification [2]. SVM works very efficiently when there is a higher number of dimensions. We decide to use SVM on our dataset because of its power of detecting different classes by making vectors using hyperplanes. We use different kernel functions like Gaussian radial basis function (RBF), linear kernel and polynomial kernel of SVM to have better accuracy.

C. Random Forest

Random Forest is another famous machine learning algorithm; it is one kind of decision tree. Random forest is a very powerful predictive modelling method [2]. It can discover very complex dependences using more time in fitting but some time it is not very good at handling noisy data as well. Again it is able to make better accuracy when other may not make such result. The reason to add this algorithm is to see how it behaves on our dataset.

D. Gaussian Naive Bayes

Gaussian Naive Bayes is a very simple algorithm for classification and one of the biggest advantages is training and prediction speed [2]. It is very fast and works better for small dataset. Moreover, it is very easy to train and understand the results. This algorithm assumes every feature is independent, however, each feature of our dataset is not independent. Hence, this algorithm doesn't perform well in our dataset.

E. AdaBoost

Adaptive Boosting or AdaBoost is another popular machine learning method. AdaBoost is adaptive in a sense that it finds out some points on training data which are not well predicted. Those points have a significant error, by finding those points it can find out the weakest classifiers to make better classification. It is some time sensitive to label noise as it fits classification model to an exponential loss function. We choose this algorithm to see how it performs on our dataset.

F. Stochastic Gradient Descent

Stochastic Gradient Descent is also a well-known machine learning algorithm. It is basically a first order iterative optimization algorithm but it can be used in classification problem and for minimizing error. The problem of Gradient Descent is when we have a big and complex dataset it takes too much time to predict but with SGD it takes smaller steps for training set and its faster and better. We include it to know is it possible to have a better performance.

G. Multilayer Perceptron Neural Network

Multilayer Perceptron Neural Network is the most powerful machine learning method among all of these. MLP can handle very complex data pattern where other models are unable to detect any pattern and very good for a prediction model [12]. In this research, we use Scikit learn and Keras with tensorflow respectively to build and implement our MLP model [22], [23]. Our MLP model has three hidden layers with 15 features as input. For small datasets, the performance of multilayer perceptron neural network is not thriving.

V. EXPERIMENTAL RESULT AND DISCUSSION

All these machine learning methods are good as classifiers. Some methods are good for small dataset like ours for example Naive Bayes or Logistic Regression but they are not good for recognizing complex pattern, where Neural Network and other methods like SVM works better. Among these seven methods, MLP Neural Network gives the best result. From 755 movies, our MLP model is able to predict 442 movies correctly and 677 movies if we consider one away prediction.

One away prediction means the difference between predicted and target class is 1. For example, suppose a movie is classified as class 5, means it is a blockbuster hit. But the prediction result is class 4. That means our classifier predicted one class less than the true value. For Exact match prediction it will be considered as classification error but if we take one away prediction in consideration, it will be accepted as correct result. We have used two types of features, pre-released features for upcoming movies' prediction and all features which includes both pre-released and post-released features for prediction after opening weekend. In Fig. 1 and Fig. 2 we can see different performance from all these algorithms considering both exact and one away prediction. Fig 1 shows performance comparison for all features where Fig. 2 is for only pre-released features.

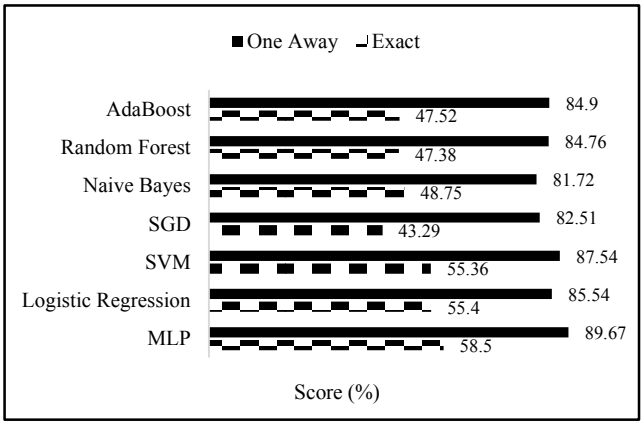

Figure 1. Performance comparison of different methods for all features.

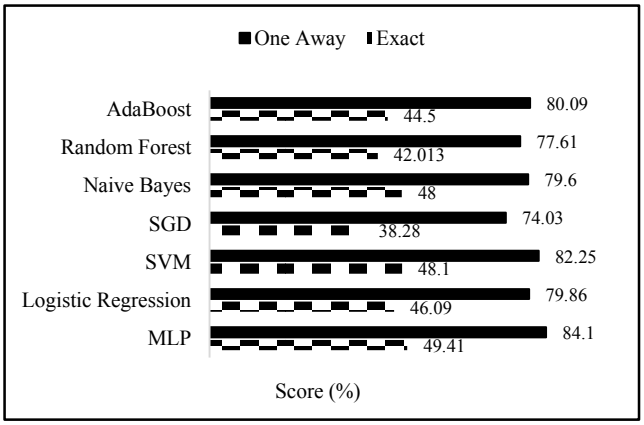

Figure 2. Performance comparison of different machine learning methods for pre-released features.

Among all these features, MLP gives better result. The reason is MLP is good for complex data pattern recognition. One of the major problem in our dataset is there are many data point which are overlapping, which is shown in Fig. 3. For this problem it is hard for algorithms to learn the pattern effectively. However, 58.53% exact accuracy and 89.67% one

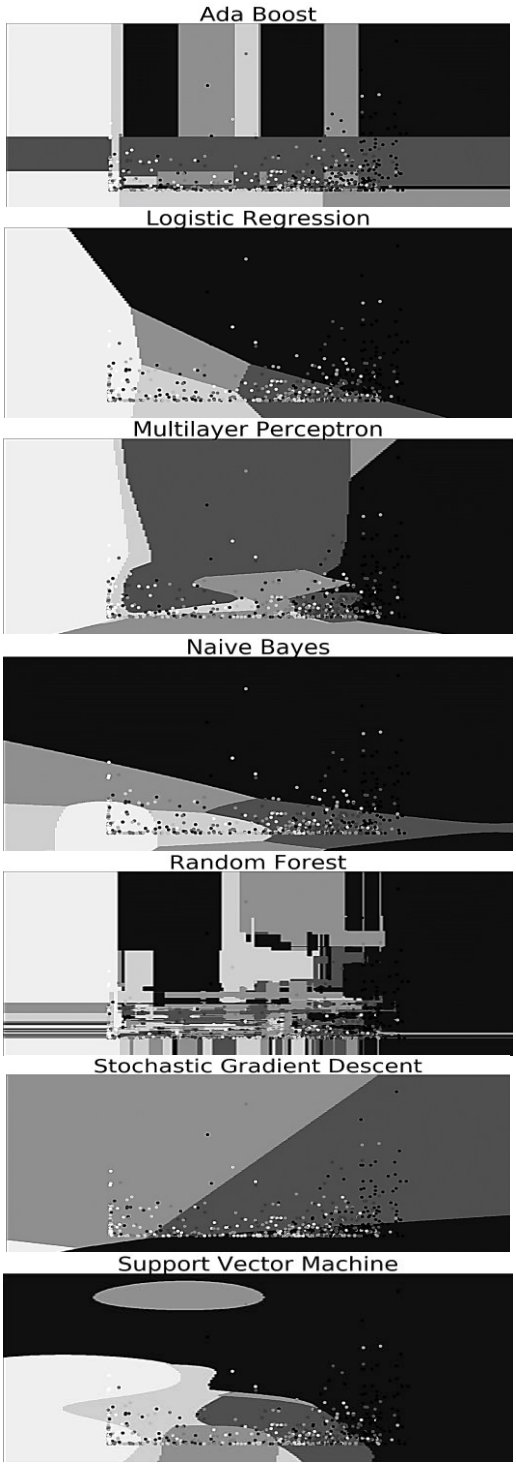

Figure 3. Data overlapping in different methods.

TABLE III. ACCURACY (IN PERCENTAGE) OF EACH FOLD FOR SEVEN ALGORITHMS (ALL FEATURES)

Algorithms	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Final
MLP	58.45	63.93	62.33	55.94	67.17	54.67	51.86	57.81	63.52	49.62	**58.53**
SVM	53.24	46.75	62.34	59.74	50	57.34	53.34	54.67	54.05	61.12	**55.26**
SGD	48.05	44.15	40.25	46.75	42.10	42.67	36	44	50	38.89	**43.29**
Logistic Regression	54.54	50.64	53.24	62.34	55.26	54.67	50.67	52	59.45	61.12	**55.4**
Random Forest	46.75	42.85	58.45	55.84	38.15	45.34	44.0	50.67	47.29	44.45	**47.38**
Ada Boost	50.64	38.96	55.84	50.64	50	42.67	41.34	53.34	45.94	45.34	**47.47**
Naive Bayes	37.67	51.94	51.04	52.84	55.26	49.34	40	48	50	51.38	**48.75**

away accuracy is a very good score for our MLP model comparing to other researches [3], [12]. Other methods do well, like SVM and Logistic Regression. SVM is a very powerful classifier but the main problem is separating data regions. While our data points are overlapping with each other, often it becomes hard for SVM to separate the region perfectly where logistic regression also does a very good job. Logistic Regression is good for categorical data and small dataset, both fits on our dataset. We can see performance of SGD, Ada Boost, Random Forest and Naive Bayes are poor related to MLP, SVM and Logistic Regression.

We implement all these methods using 10-fold cross validation which is the best way to test. Here, in Table III, we include all ten folds' accuracy along with the final accuracy for all seven algorithms. Data overlapping is the main reason for such performance of those algorithms shown in Fig. 3, it shows budget in X axes and number of screens in Y axes relations with the classification areas of different algorithms. In Fig. 3, different level of darkness represents specific class and their belonging data points. Here we can see lots of specific colored points in other colored regions which indicates the classification errors and it is occurring because of data overlapping. Points are overlapping on each other that is why we cannot see all the data points. Different data points of different classes are overlapping and classifiers become confused in determining their actual classes. We show this problem in 2D graph as it is a good way to understand the problem. We select budget and number of screens as those are the most important features and available before the release of movies, although other features can be used. In Fig. 3 different shades of dark regions are different from each other for different algorithms as classification calculations are different. Each color represents a class in Fig. 3 and the colors of data points are same as their belonging class. For instances, if white color area represents class 1 then all white data points belong to class 1. Again if you find some white points in other colored areas, that means they are incorrectly classified by the classifier.

Here in Fig. 4 and Fig. 5 visualize the precision, recall and f1 scores of all seven methods. Fig. 4 is for all features and Fig. 5 shows these scores for only pre-released features. These scores give a good comparison of performance assessment between different methods. Precision, recall and f1 scores are calculated differently. Before the explanation we need to know some important terms, they are true positive,

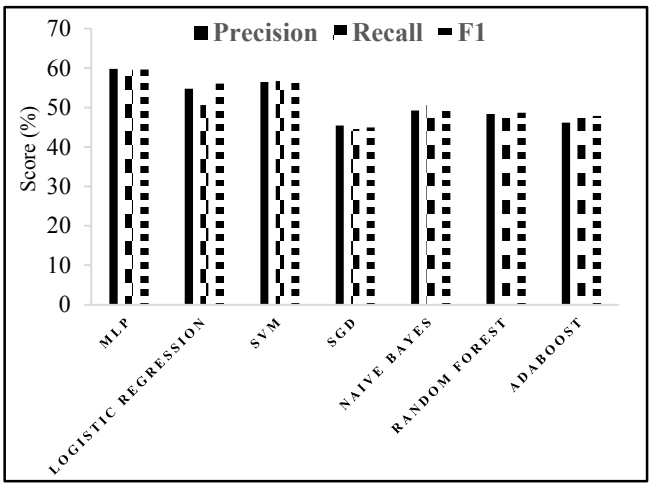

Figure 4. Precision, recall and F1 score for all features.

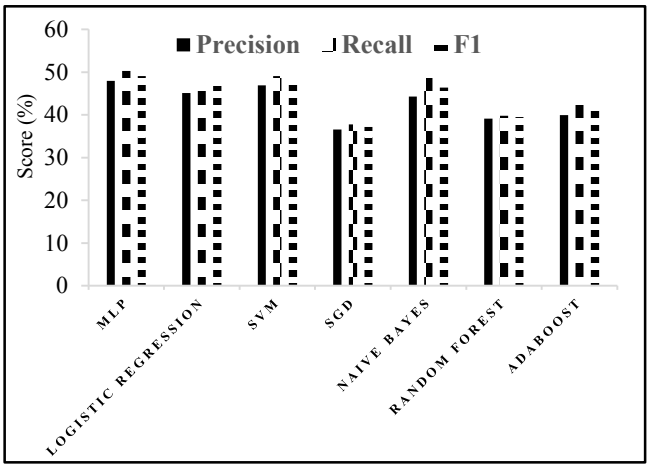

Figure 5. Precision, recall and F1 score for pre-released features.

true negative, false positive and false negative. True positive means where test class and predicted class both are true, true negative means both test class and predicted class are false, false positive means test class is false but predicted class is true and false negative means test class is true where predicted class is false. So precision formula is, TP/TP+FP. Which gives us how many selected classes are relevant where recall tells us how many relevant classes are selected. Formula of recall is TP/TP+FN. F1 score is the harmonic mean of precision and recall, it can be write as 2*(precision x recall)/precision + recall. Among all these algorithms MLP and SVM gives the best result. These scores are calculated based on exact prediction accuracy rather than one away prediction accuracy.

VI. CONCLUSION AND FUTURE WORKS

A movie's box-office success depends on many parameters not only on some features related to movies. It also depends on other factors like number of audience. Also their appearance in theaters depends on the political and economic stability of a country. If economic and political condition of a country is not stable, then it does not matter how well the movie is made, there will be no one to watch that movie. So including a country GDP as an attribute is a good option for further analysis. We also suggest to analyze and include the number of audience for analysis. We can get number of annual audience by using total ticket sold in a particular year. Including these attributes will make the prediction more accurate.

In our research we exclude genre and sequel of movies. Forecasting the success of a sequel movie is hard, as some movies become successful just because of the fame from previous movie sequel. Previously other research works also excluded sequel movies [3]. In early days, most of the researches either considered only pre-released features [3], [16] or post-released features [5], [6] for their prediction, but we take both types of features in consideration for both upcoming movies' prediction and prediction after opening weekend. Our main goal for this research is to show how most popular machine learning algorithms act on movie related data for box office success prediction.

REFERENCES

[1] J. Valenti (1978). Motion Pictures and Their Impact on Society in the Year 2000, speech given at the Midwest Research Institute, Kansas City, April 25, p. 7.

[2] M. T. Lash and K. Zhao, "Early Predictions of Movie Success: The Who, What, and When of Profitability," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 874–903, Feb. 2016.

[3] R. Sharda and E. Meany, "Forecasting gate receipts using neural network and rough sets," in *Proceedings of the International DSI Conference*, 2000, pp. 1–5.

[4] J. S. Simonoff and I. R. Sparrow, "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers," *Chance*, vol. 13, no. 3, pp. 15–24, 2000.

[5] A. Chen, "Forecasting gross revenues at the movie box office," *Working paper, University of Washington, Seattle, WA*, June 2002.

[6] M. S. Sawhney and J. Eliashberg, "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," *Marketing Science*, vol. 15, no. 2, pp. 113–131, 1996.

[7] D. Gregorio, "Prediction of movies box office performance using social media," *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, 2013.

[8] S. Gopinath, P. K. Chintagunta, and S. Venkataraman, "Blogs, Advertising, and Local-Market Movie Box Office Performance," *Management Science*, vol. 59, no. 12, pp. 2635–2654, 2013.

[9] M. C. A. Mestyán, T. Yasseri, and J. Kertész, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," *PLoS ONE*, vol. 8, no. 8, 2013.

[10] B. Pang and L. Lee, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, pp. 79–86.

[11] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[12] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.

[13] A. Sivasantoshreddy, P. Kasat, and A. Jain, "Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining," *International Journal of Computer Applications*, vol. 56, no. 1, pp. 1–5, 2012.

[14] M.H Latif, H. Afzal "Prediction of Movies Popularity Using Machine Learning Techniques", National University of Sceinces and technology, H-12, ISB, Pakistan.

[15] K. Jonas, N. Stefan, S. Daniel, F. Kai "Predicting Movie Success and Academi Awards through Sentiment and Social Network Analysis" University of Cologne, Pohligstrasse 1, Cologne, Germany.

[16] W. Zhang and S. Skiena, "Improving Movie Gross Prediction through News Analysis," *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2009.

[17] T. G. Rhee and F. Zulkernine, "Predicting Movie Box Office Profitability: A Neural Network Approach," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016.

[18] J. Duan, X. Ding, and T. Liu, "A Gaussian Copula Regression Model for Movie Box-office Revenue Prediction with Social Media," *Communications in Computer and Information Science Social Media Processing*, pp. 28–37, 2015.

[19] L. Doshi, J. Krauss, S. Nann, and P. Gloor, "Predicting Movie Prices Through Dynamic Social Network Analysis," *Procedia - Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6423–6433, 2010.

[20] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, "Predicting movie Box-office revenues by exploiting large-scale social media content," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1509–1528, Feb. 2014.

[21] Z. Zhang, B. Li, Z. Deng, J. Chai, Y. Wang, and M. An, "Research on Movie Box Office Forecasting Based on Internet Data," *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, 2015.

[22] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[23] F. Chollet, Keras, 2015, GitHub repository, https://github.com/fchollet/keras. [Accessed: March-2016]