Research Journal of Management Sciences _____ ISSN 2319–1171
Vol. **6(6),** 1-7, June **(2017)** *Res. J. Management Sci.*

Predictive model for movie's success and sentiment analysis

M. Prasanna Mohan Raj[1] and S. Aditya[2]
[1]School of Business, Alliance University, 19th Cross, 7th Main, BTM 2nd Stage, N.S. Palya, Bengaluru – 560 076, India
[2]Marketing & Business Intelligence, TVS next Pvt. Ltd., ASV Chandilya Towers, OMR, Chennai-600096, India
prasanna.mr@alliance.edu.in

Available online at: www.isca.in, www.isca.me
Received 6th April 2017, revised 2nd June 2017, accepted 5th June 2017

Abstract

The film industry is one of the biggest contributors to the entertainment industry and also it is characterized with its unpredictability in success and Failure. Film Industry has always amused everyone with its unpredictable success and Failure. The Indian scenario works a lot different than the western movies; a lot of importance is normally given to different parameters such as celebrity appeal, the movie album and others, which is an integral part of the movie itself; unlike, the western movies. This research looks into the inner details of watching a movie by splitting the research into three main components. First section is exploring the variables that influence the frequency of movie watch; second, developing a model to predict the success or failure. Finally, social network sentiment analysis is carried out through data mining to capture the audience sentiment and its impact on movie's success and failure. The research tries to look at the success or failure of a movie on a more holistic manner than trying to grade the performance of a movie over a few variables based on the previous research works on movie success prediction.

Keywords: Movie success prediction Model, Influencing variables for movie watching, Social sentiment analysis, word cloud, Discriminant Analysis.

Introduction

The film industry is one of the biggest contributors to the entertainment industry and also it is characterized with its unpredictability in success and Failure. Film Industry has always amused everyone with its unpredictable success and Failure. 'Hollywood is the land of hunch and the wild guess'. Research hints the unpredictably of the product demand and proves that 25% of total revenue of a movie earned from the first 2 weeks of receipts[1]. Litman was the first to develop a multiple regression model to predict the commercial success of movies[2].

The U.S movie industry generated 564 billion U.S. dollars in revenue by the end of 2014. It is predicted that the entertainment industry will grow to over 679 billion US dollars in value over the next four years. e. It is projected that 80% of the movie industry's profits over the last decade is generated from just 6% of the films released; 78% of movies have lost money of the same time period[3]. So lot of research have been carried out in predicting the movie success in United States. Indian Movie Industry is different from its counterparts in many ways especially in terms of different movies in different languages and low ticket prices. In 2013, the Indian film industry generated 1.68 billion U.S. dollars out of its total revenue of 2.07 billion from overseas and domestic box office collections[4]. According to the Deloitte and ASSOCHEM report, researchers predict that theatrical revenue in India will be increased from $1.78 billion to $2.13 billion by FY2017. That makes theatrical circuits account for some 74% of total income. TV and satellite revenues will grow at the rate of 15% per year[5]. India was, however, only the sixth largest film market by box office revenue, behind North America, China, Japan, the United Kingdom and France[6].

The Indian scenario works a lot different than the western movies; a lot of importance is normally given to different parameters such as celebrity appeal, the movie album and others, which is an integral part of the movie itself; unlike, the western movies. This research, looks into the inner details of watching a movie by splitting the research into three main components. First section is exploring the variables that influences the frequency of movie watch. Second, developing a model to predict the success or failure. Finally, social network sentiment analysis is carried out through data mining to capture the audience sentiment and its impact on movie's success and failure. The research tries to look at the success or failure of a movie on a more holistic manner than trying to grade the performance of a movie over a few variables[7]. Based on the previous research works on movie success prediction variables are identified and used in this research to develop the model[8].

This research paper tries to understand the scenario with more variables that are best suited for the Indian scenario. The need of this study is predominantly to study the different parameters that will influence a movie watch in the Indian Scenario.

Research objectives: This research mainly focuses on three components. i. Variables that influence the frequency of movie watch ii. Predicting the success of a movie by developing a model iii. sentiment analysis of selected two movies.

Methodology

This study is descriptive in nature. The research design comprises of both qualitative and quantitative techniques. The factors that influence the frequency of watching movie are identified by literature review and quantitative approach is used to predict the success of a movie by developing a model. Multiple Regression analysis and Discriminant analysis are used for developing prediction model to predict the success of a movie.

Sentiment analysis is carried out through "word cloud". "R" programming is adopted to carry out these data analysis. Judgmental sampling is adopted in this study. Primary data is collected from 187 respondents using questionnaire. Secondary data are collected from Face book using Self-written mining algorithms. Respondents are selected from four metro cities and Bangalore.

Theoretical foundations: Jonas Krauss and Stefan Nann[9] identify three variables namely intensity, positivity and time that predicts the movie being nominated for the Academy Awards and thereby winning it. Though, winning an Oscar has always been speculative; the model was almost accurate in predicting if the movie would be nominated for the academy awards. Authors calculate the levels of web buzz by mining discussions in movie related online forums, combining information about the structure of the social network with an analysis of the contents of the discussion. The paper further demonstrates the approach by predicting the success of movies based on the communication in the online community IMDb.com. Authors also discuss the finding over three relevant components namely- Discussion Intensity, Positivity & Time.

Discussion Intensity is mainly the number of mentions in the online forum, Positivity explores the quality of communication and time explores the time difference between different movies being mentioned in the Academy Awards. The main predictive variables; intensity and time are quite easy to be captured. Positivity variables are computed based on sentiment analysis with some of the highest ranked positivity phrases include – "Win", "Nominate", "Great", "Good", "Award", "Super".

Simonoff and Sparrow[10] identify the variables influencing the success of the movie; Genre of the film, such as Action, Children's, Comedy, Documentary, Drama, Horror, Science Fiction, or Thriller etc. rating of Motion Picture Association of America (MPAA), The origin country of the movie, classified as U.S or Non-U.S., the production budget of the film, Whether or not the movie was a sequel to an earlier movie. Timing of release of the movie such as before a long week end or during festivals and Academy Award (Oscar®) nominations for the film. This research identifies the star power i.e the presence of popular actor/actress playing lead role in the movie.

Zhang and Skiena[11] considered the news data is the best variable to predict movie success. According to the authors news data contain information about actors, directors and media news. The authors also used text map to analyze the news through which they have predicted the success or failure of a movie. The key findings of the research is movie news references are highly correlated with movie grosses, and sentiment measures including derived sentiment indexes are also correlated with movie grosses.

Mestyán et.al[12] proved that popularity of a movie can be predicted before its release by measuring and analyzing the activity level of editors and viewers of the respective movie in Wikipedia. Authors considered the activity level of editors and the number of page views by readers to assess the popularity of a movie. Linear regression model is developed to forecast the first weekend box office revenue.

Sharda and Delen[13] developed the model to predict movie's success by using neural networking. They converted the forecasting problem into a classification problem. The authors classified in to nine categories ranging from a 'flop' to a 'blockbuster.' MPAA Rating, Competition, Genre Special effects Sequel, Number of screens and start value are key variables identified by the authors and used in their forecasting model. Deniz and Robert[14] analyzed the data from one hundred and fifty top grossing movies of2010 and developed the movie prediction model based on its genre, MPAA rating, budget, star power, adaptation from another medium , sequels and remakes on total U.S. box-office revenue. Authors adopted multiple linear regression to develop the prediction model.

Prag and Casavant[15] proved the positive relationship between star power, and the box office performance and also points out that star power is the most significant variable which can predict the success of the movie. Critics rating, academy awards and production costs are also identified by the authors as positive determinants of movie prediction. King T.[16] reveals that the movie reviews has a high correlation with critic's reviews. The author identifies the different variables that will influence the gross box office revenue through a multiple regression model with variables such as "Opening Screens", "Metacritic Reviews", and "Documentary".

Ravid[17] carried out the research study on film revenue and return-on- investment (ROI) as functions of production cost and star actors. According to his model, large production costs of a movie may significantly increase film revenue, but do not increase the ROI. The model also suggests that movie stars increase revenue, regressions find star presence to be insignificant.

Joshi et.al.[18] carried out movie prediction in a different way. Authors have taken the forecasting problem as an application of NLP apart from the economic value of such predictions. Authors have utilized the text of critics' reviews to predict opening weekend revenue. They also considered metadata for each movie. The objective of this work is to identify specific

group of phrases that predict the movie-going tendencies. Authors have adopted linear regression from text and non-text (meta) features to predict gross revenue of a movie.

Results and discussion

This study adopts "R" as a statistical tool to find out the solution for following key objectives. A Movie success is determined by the footfall, so the first component tries to understand the factors that influence the frequency of watching a movie. The second component, tries to understand the probability of a movie success vs. failure by using Discriminant analysis. The third component, tries to capture the audience sentiment on Movies using text mining sentiment analysis on Face book Movie pages of two movies.

The first component as discussed before tries to understand the different determinants of frequent movie watch. The data is obtained from a primary research involving variables such as Frequency of movie watch, online booking, Movie Purchase and ease of access. The model is developed by using linear regression analysis to identify the influencing factors for 'Frequency' of movie watch. "Frequency of movie going" is taken as dependent variable and the independent variables are "Online Booking", "Movie Purchase", "Ease of Access" and "peer preferences". These independent variables are derived from the literature review and pilot test.

Table-1: Multiple Regression Analysis.

Code:

```
Model2 =lm(indata$How.often.do.you.go.to.Movies. ~Online.Booking
+indata$Do.you.prefer.purchasing.your.favorite.movies. +
indata$Ease.of.Access.of.Theatres +indata$Peer.Preferences, data=indata)
summary(Model2)
```

Output:

```
##
Residuals:
     Min       1Q   Median       3Q      Max
 -2.1422  -0.5560   0.1187   0.4327   2.3568
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                      -0.88486    0.68822
## Online.Booking                                    0.59171    0.13259
## indata$Do.you.prefer.purchasing.your.favorite.movies.  0.10379    0.08110
## indata$Ease.of.Access.of.Theatres                 0.20444    0.10253
## indata$Peer.Preferences                           0.31557    0.08238
##                                                 t value Pr(>|t|)    ## (Intercept)                                      -1.286 0.202163

## Online.Booking                                    4.463 2.55e-05 ***
## indata$Do.you.prefer.purchasing.your.favorite.movies.   1.280 0.204249
## indata$Ease.of.Access.of.Theatres                 1.994 0.049484 *
## indata$Peer.Preferences                           3.831 0.000249 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

## Residual standard error: 0.8927 on 82 degrees of freedom
## Multiple R-squared:  0.4443, Adjusted R-squared:  0.4171
## F-statistic: 16.39 on 4 and 82 DF,  p-value: 6.661e-10
cor(indata$How.often.do.you.go.to.Movies., indata$Online.Booking)
## [1] 0.5548935
cor(indata$How.often.do.you.go.to.Movies.,indata$Ease.of.Access.of.Theatres)
## [1] 0.2741955

hist(indata$Online.Booking)
hist(indata$Ease.of.Access.of.Theatres)
```

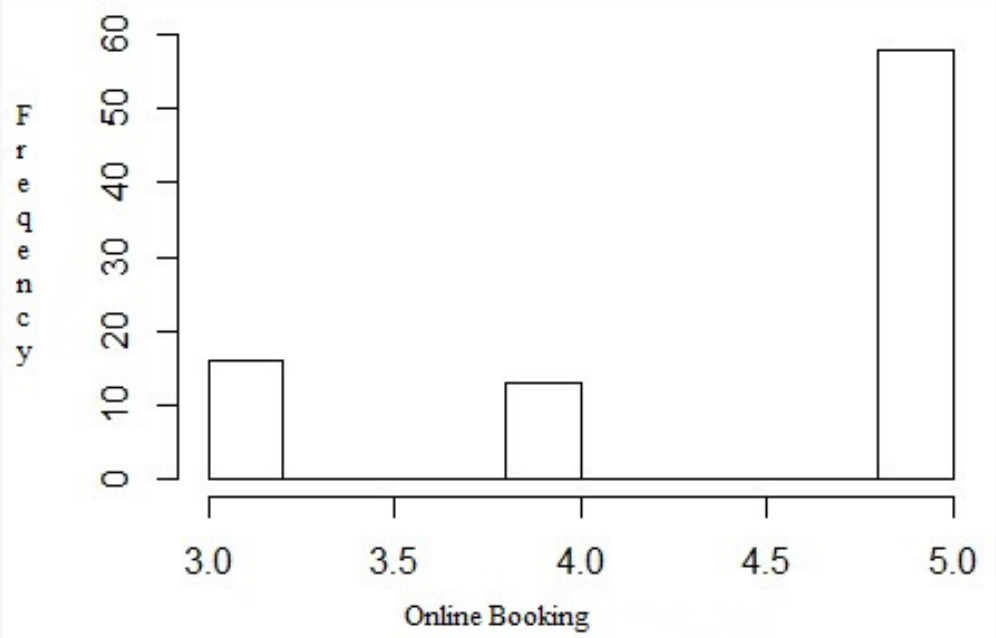

Figure-1: Histogram on Online booking Data.

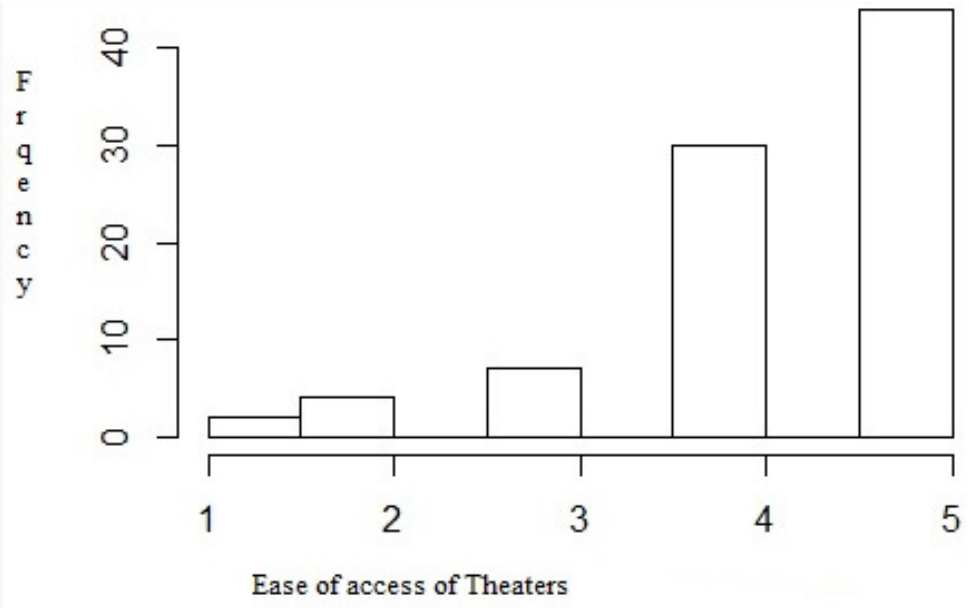

Figure-2: Histogram on Ease of Access of Theaters.

Multiple Regression Model: Frequency of Movie watch = -0.88484 + 0.59171 * (Online Booking) + 0.20 * (Ease of Access) + 0.315 * (Peer Preference).

"Online Booking", "Ease of Access" and "Peer preference" are identified as the significant variables influencing frequent movie going. The R-Squared value is 44.43% which is acceptable. The Correlation between the variable Frequency and Online Booking is 50.12%; Correlation between the variable Frequency and Ease of Access is 45.27%. Peer Preference and Online Booking are the most significant variables followed by Ease of Access.

Discriminant analysis model: The second component tries to understand the classification of Movie into "Success" or "Failure". Discriminant Analysis is used to develop the model. The independent variables are identified from the literature review. Those independent variables are "Celebrity Appeal", "Movie Review", "Special Effects", "Movie Promo", "controversy", "Movie Album". The discriminant model developed is to test a movie's success.

Table-2: Discriminant analysis to predict movie's success.

```
ldat =table(indata$Factor, predictions$class)
sum(diag(ldat))/sum(ldat)
## [1] 0.8275862
classAgreement(ldat)
## $diag
## [1] 0.8275862
##
## $kappa
## [1] 0.2737896
##
## $rand
## [1] 0.7113071
##
## $crand
## [1] 0.2098389
## Group means:
##
Celebrity.AppealMovie.ReviewSpecial.EffectsMovie.Promo
## Failure      3.411765    3.882353      3.882353    4.588235
## Success            3.771429      4.157143            4.214286
4.342857
##       Controversy Movie.Album
## Failure   2.941176   3.823529
## Success   3.242857   4.328571
## Coefficients of linear discriminants:
##                    LD1
## Celebrity.Appeal  0.27381165
## Movie.Review      0.03106478
## Special.Effects   0.28813698
## Movie.Promo      -0.92351552
## Controversy       0.05119296
## Movie.Album       0.77813038
```

The prediction $ posterior gives us the prediction for each observation for "Success" and "failure". The "ldat" and the Class Agreement tries to predict the model accuracy which falls to around 82.70%.

Success/Failure = 0.27 * (Celebrity Appeal) + .03* (Movie Review) + 0.288 (Special Effects) – 0.92 * (Movie Promo) + 0.05 * (Controversy) + 0.77 * (Movie Album).

The group means suggest that both failure and success movies both had almost equal celebrity appeal; Movie Album is identified as one of the key determinants of movie success.

Sentiment analysis: The third Component tries to study the sentiment of the users on Facebook; using R as the statistical machine. Mishne and Glance[19] applied sentiment analysis techniques to analyze pre-release and post-release blog posts about movies and proved the strong correlation between actual revenue and sentiment based metrics. Zhang and Skiena[20] used a news aggregation system to identify entities and obtain domain-specific sentiment for each entity in several domains. They used the aggregate sentiment scores and mention counts of each movie in news articles as Predictors.

The algorithm is written using the pre-existing R library files called RCurl, Rfacebook and SnowballC. The algorithm identifies the sentiment of the user groups across two movies 'Baahubali' and 'Udta Punjab' ''Word Cloud' is generated for those movies. The data comprises of 40 posts from the Facebook page of Baahubali. From each post, the data on 100 likes and comments have been retrieved. The data comprises of 15 posts from the Facebook page of Udta Punjab. From each post, the data on 100 likes and comments have been retrieved. The libraries are called and the appropriate user token is taken is used for accessing the data (The user token is not displayed because of the user confidentiality).The data comprises of 40 posts from the Facebook page of Baahubali. From each post, the data on 100 likes and comments have been retrieved.

Sentiment analysis of Movie 1: Baahubali

Code:

```
fb_page_1<-
getPage(page='BaahubaliMovie',token=accessToken, n=40)
## 40 posts
post_1<-getPost(post=fb_page_1$id[5], n=100,
token=accessToken)
post_1$post
```

Corpus: A corpus function is now created to start the data clean-ups. The following are the data clean-ups that has been worked on: Numbers have been removed, Stop words have been removed, Stemming has been removed, Low frequency terms have been removed and DTM Matrix is created to mark the occurence of the unique words.

```
corpus_1 =Corpus (VectorSource (post_1$comments))
inspect (corpus_1[1])
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
##
## $from_id
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 432
corpus_1 =tm_map(corpus_1, removeNumbers)
corpus_1 =tm_map(corpus_1, removeWords,
c("cancer","cancers",stopwords("english")))
corpus_1 =tm_map(corpus_1, stemDocument)
corpus_1 =tm_map(corpus_1, PlainTextDocument)
DTM =DocumentTermMatrix(corpus_1)
findFreqTerms(DTM, lowfreq=60)
sparse_1 =removeSparseTerms(DTM, 0.80)
sparse_1 =removeSparseTerms(DTM, 0.90)
allComments =as.data.frame(as.matrix(sparse_1))
```

Word Cloud: A word Cloud is now created, to reflect the most unique words in the comments from the Facebook posts.

Figure-3: Word cloud for the movie 'Bahubali'.

The positive sentiment can be easily sensed through some of the words such as "great", "Sir", "best" from the word cloud of the movie 'Bahubali'. The word "wait" also gives us about the curiosity of respondents about its second part.

Sentiment analysis of Movie-2: Udta-Punjab: The libraries are called and the appropriate user token is taken is used for accessing the data (The user token is not displayed because of the user confidentiality).The data comprises of 15 posts from the Facebook page of Udta Punjab. From each post, the data on 100 likes and comments have been retrieved.

Code:

```
fb_page_2<-getPage(page='Udta-Punjab',token=accessToken,
n=35)
## 15 posts
post_2<-getPost(post=fb_page_2$id[5], n=100,
token=accessToken)
post_2$post
```

Corpus:

```
corpus_2 =tm_map(corpus_2, removeNumbers)
corpus_2 =tm_map(corpus_2, removeWords,
c("cancer","cancers",stopwords("english")))
corpus_2 =tm_map(corpus_2, stemDocument)
corpus_2 =tm_map(corpus_2, PlainTextDocument)
DTM =DocumentTermMatrix(corpus_2)
sparse_2 =removeSparseTerms(DTM, 0.80)
sparse_2 =removeSparseTerms(DTM, 0.90)
allComments_2 =as.data.frame(as.matrix(sparse_2))
```

A word Cloud is now created, to reflect the most unique words in the comments from the Facebook posts.

```
wordcloud(colnames(allComments_2),
colSums(allComments_2),
scale=c(3,1),
random.color=TRUE,
colors=brewer.pal(8,"Dark2"),
random.order=FALSE,rot.per=.25)
```

Figure-4: Word cloud for the movie 'Udta-punjab'.

The word group like "overrated" and "harsh" show negative sentiments for Udat Punjab. The success of the movie can be attributed to its "controversies".

Conclusion

The research has identified the weightage of few variables for its success and failure. The key revolves around how well the marketing team is able to design a strategy that works well amidst its storyline and the analysis of the research. Location based promotions can increase the accessibility factor and thereby encourage the individual to watch the movie. The Marketing team of a production house may try to get the best and the most trusted critic to right a review about their movie before its launch. Research also finds a high correlation between user reviews and critics review. This review by a critic can help increase the opening week's revenue. The right mix of marketing spend across different parameters can help gain a good visibility for the movie.

References

1. Litman B.R. and Kohl Linda S. (1998). Predicting Financial Success of Motion Pictures. *Journal of Media Economics,* 2(2), 35-50.
2. Litman B.R. (1983). Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, 16(4), 159-175.
3. South-Africa, Nigeria and Kenya (2015). Entertainment and media outlook: 2015-2019., https://www.pwc.co. za/en/ assets/pdf/entertainment-and-media-outlook-2015-2019.pdf, Accessed on April 2017.
4. Indian films' box office collection to be USD 3.7 Bn in 2020 (2016). Business Standard. http://www.business-standard.com/article/pti-stories/indian-films-box-office-collection-to-be-usd-3-7-bn-in-2020-116092500337_1.html, Accessed on April 2017

5. Deloitte (2014). Digitization & Mobility: Next Frontier of Growth for M&E 2016. https://www2.deloitte.com /content/dam/Deloitte/in/Documents/technology-media-telecommunications/in-tmt-digitization-n-mobility-noexp.pdf , Accessed on March 2017.

6. Top 10 Film Countries by Box Office (2013). http://www.filmcontact.com/americas/united-states/top-10-film-countries-box-office, Accessed on April 2017

7. Elberse A. and Eliashberg J. (2002). The drivers of motion picture performance: the need to consider dynamics, endogeneity and simultaneity. *proceedings of the Business and Economic Scholars Workshop in Motion picture Industry Studies. Florida Atlantic University,* 1-15.

8. Neelamegham R. and Chintagunta P. (1999). A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Science*, *18*(2), 115-136.

9. Krauss J., Nann S., Simon D., Gloor P.A. and Fischbach K. (2008). Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis. ECIS, 2026-2037.

10. Simonoff J.S. and Sparrow I.R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3), 15-24.

11. Zhang W. and Skiena S. (2009). Improving movie gross prediction through news analysis. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01,* 301-304. IEEE Computer Society.

12. Mestyán M., Yasseri T. and Kertész J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, 8(8), 12-26.

13. Sharda R. and Delen D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243-254.

14. Deniz B. and Hasbrouck Robert B. (2012). What Determines Box Office Success of A Movie in the United States. *Chistoper Newport University*, 1-11.

15. Prag J. and Casavant J. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. *Journal of Cultural Economics*, 18(3), 217-235.

16. King Timothy (2007). Does film criticism affect box office earnings? Evidence from moviesreleased in the U.S. in 2003. *Journal of Cultural Economics*, 31(3), 171-186.

17. Ravid S.A. (1999). Information, blockbusters, and stars: a study of the film industry. *The Journal of Business*, 72(4), 463-492.

18. Joshi M., Das D., Gimpel K. and Smith N.A. (2010). Movie reviews and revenues: An experiment in text regression. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 293-296.

19. Mishne G. and Glance N.S. (2006). Predicting Movie Sales from Blogger Sentiment. AAAI spring symposium: computational approaches to analyzing weblogs, 155-158.

20. Zhang W. and Skiena S. (2009). Improving movie gross prediction through news analysis. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 301-304.