Predicting Movies' Box office result - A large scale study across Hollywood and Bollywood

Risko Ruus and Rajesh Sharma

Abstract Predicting movie sales figures has been a topic of interest for research for decades since every year there are dozens of movies which surprise investors either in a good or bad way depending on how well the film performs at the box office compared to the initial expectations. There have been past studies reporting mixed results on using movie critics reviews as one of the sources of information for predicting the movie box office outcomes. Similarly using social media as a predictor of movie success has been a popular research topic. We analyze the Hollywood and Bollywood movies from three years, which belong to two different geo as well as cultural locations. We used Twitter for collecting the wisdom of the crowd features (4.3 billion tweets, 1.41TB in compressed size) and used movie critics review scores from movie review aggregator sites *Metacritic* and *SahiNahi* for Hollywood and Bollywood movies respectively. In addition, we also used metadata about movies such as budget, runtime, etc. for the prediction task. Using three different machine learning algorithms, we investigated this problem as a regression problem to predict the movie opening weekend revenues. Compared to past studies which have performed their analysis on much smaller datasets, we performed our study on a total of 533 movies. In addition to r^2, we measured the quality of our models using MAPE and we find out that a model (Random Forest) based on all the three features (Metadata, Critics, Twitter) gives the best results in our analysis.
Keywords: Box office forecasting, machine learning, Twitter.

1 Introduction

Hundreds of movies are released every year in the world. However, not every movie turns out to be a commercial success. For example, only three or four major movies out of every ten major Hollywood movies produced are profitable [1]. Forecasting the box office results has been a big concern for the movie industry as early box office predictions help to make vital decisions regarding marketing budget allocation and distribution. Equally important is determining the best screen allocation for a movie in each country since empty seats mean bad business for movie studios and cinemas alike. However, past studies have shown it is difficult to predict the tastes

Risko Ruus, Rajesh Sharma
University of Tartu, Estonia, e-mail: risko.ruus@ut.ee, rajesh.sharma@ut.ee,

of moviegoers [2, 3, 4] and subsequently forecasting the box office results has been a big concern for the movie industry.

Litman was the first to study multivariate regression models [5, 6] for predicting the box office outcome of movies. Predictor variables considered in such research include the number of theaters the movie is scheduled to be released in, parental rating and the budget of the film. Many researchers consider predicting commercial movie success as a classification problem. For example in [7, 8, 9] movies are classified into different categories usually ranging from a flop to a blockbuster. These segments are created by using the movie production budget as an estimated figure for calculating how much a movie should make to earn its production costs back. The problem with this approach is that while recently many studios have started to reveal their film production budgets, the money spent on marketing is not disclosed and can influence the actual profitability of the movie significantly. Also as mentioned in [10], star actors are often paid a percentage of the movie profits and their salaries might not be included in the movie production budget figures making the movie budget deceptively low. For these reasons we have followed the example of studies such as [11, 12, 13] and consider predicting commercial movie success as a regression problem and predict the amount of money a movie is expected to earn after its opening weekend.

Most of the previous studies involving predicting movie success ahead of its release have worked by exploring either social media platforms such as Twitter [14], Wikipedia [15], Facebook [16], Google search queries [17] or have only analyzed movie expert's reviews [3, 18, 19, 20].

Social media content can be thought of as a very large collection of collective wisdom. When asking the right questions from such data, it is possible to make predictions about future outcomes and the question we will be asking is about predicting the box office outcome of upcoming movie releases [14]. In comparison, movie critics reviews refer to the views expressed by a smaller group of domain experts. In this work, we followed a holistic approach and used both social media platform (Twitter in our case) and movie critics for predicting the box office outcome of the movies. The models can be used by stakeholders, including distributors and movie theatre operators to make improved financial decisions when promoting the film at the *critical period*[1] of its release.

This work is an empirical study, which involves collecting all the necessary data for building prediction models for the Hollywood and Bollywood movies released between April 2015 and April 2018. For model building, we used three different types of features. The first we call as *Metadata* which includes general movie information e.g. budget and opening theatre count. The second set of features are called *Twitter* features which uses the hourly tweet rate from two weeks before the film's release and the sentiment score of the movie tweets. The third set of features, *Critics*, takes input from movie expert reviews. We evaluate prediction results using Linear Regression, Random Forest and XGBoost machine learning algorithms.

[1] We use the same definition for the critical period as [14]. It is defined to be between a week before the movie is released until two weeks from its release date. This is usually the time when most of the promotional budget is being spent on various forms of advertising.

This paper has following contributions:

1. **Wisdom of the crowd and experts:** Our empirical study shows that people's collective wisdom (gathered from Twitter) when combined with the critics' review and metadata about movies can help to predict movie opening weekend box office results better.
2. **Large scale study:** To the best of our knowledge this research is made on the largest amount of Hollywood and Bollywood movies.
3. **Hollywood & Bollywood:** The work offers a unique cross-cultural comparison of box office predictions for Hollywood and Bollywood - the two of the world's biggest movie markets.

Rest of the paper is organized as follows. In Section 2 we give a brief overview of previous related research regarding predicting movie box office results. Section 3 focus on describing the data collection process for predicting the final results. An overview of our empirical results is in chapter 4. Finally, Section 5 describes our overall contribution and proposes some directions for future research.

2 Related works

In this section, we provide an overview of the past research done on predicting the success of movies. We look at works which have used either social media or critics movie reviews as a source for predicting box office revenue.

2.1 From Social Web Platforms

Before the rise of the internet most of the dependent variables used for predicting movie box office outcome, have been based on movie metadata e.g., its genre, parental rating and actors which as reported by [21] can explain approximately 60% of the variances.

With the rise of dedicated communities for movie lovers, blogs and various web services, researchers have been looking for additional sources of information, which could help predict the movie economical success even better. For example, [12] were able to predict box office revenue from 600,000 blog entries with a relative error of 26.21%. Authors of [22] have compared the predictive power of tweet sentiment analysis and online movie review sites such as *imdb* and Rotten Tomatoes[2] and find that Twitter users are more positive in their reviews compared to the dedicated review site's ratings.

Some studies like [8] have compared the prediction sources of different web resources and social networks, namely IMDb, Twitter, and Youtube. They find that the popularity of the leading actress estimated by the followers count the actress has on Twitter is a strong predictor, but the sentiment score from movie trailer comments does not help to determine the financial success of a movie.

[2] https://rottentomatoes.com

In a novel study, [14] have shown that data from Twitter, in particular, the average hourly tweet rate and sentiment analysis of the tweets can be used to predict movie box office outcomes using a simple linear regression model $r^2(t) = 0.98$ at the release night of the movie). However [15] does point out in Fig. 5 of their work that the paper of [14] achieves such a high score because most of the 24 movies considered are commercial successes, which the model is capable of predicting better than movies with low or moderate success.

In their work on 312 movies [15] show that movie box office performance can be estimated from the activity levels of Wikipedia articles about the movie before its release. Similarly to Wikipedia activity levels, Facebook official movie fan page activity is used as a prediction feature in [16]. Predictions from social media can be made not only about movie's financial success as [23] were able to rate movies very close to their IMDb star rating using tweets from Twitter and comments from YouTube. For predicting Academy Award nominations and movie box office results, [24] show successful results using movie comments from IMDb users as a possible source of information. A whitepaper from Google [17] on 99 movies released in 2012 shows that Google search volume explains 70% of the variance in the opening weekend box office performance of the film.

Research involving predicting movie profitability is not only limited to Hollywood releases. For example, Korean researchers in [25] have studied their local market on a dataset of 212 domestic movies using metadata and features from multiple social media networks. Similarly predicting movie box office success on the Chinese domestic market has been researched by [26] using 57 movies with 5 million tweets collected from the Sina Weibo microblog[3]. The only previous study on predicting the box office results of Bollywood movies that uses features from social media is done only on 14 movies by [27].

2.2 From Expert Movie Reviews

Predicting movie box office outcome using critic reviews as a source has attracted less attention from researchers compared to using social media platforms. The authors of [3] look at expert reviews and find confirmation to the common belief that positive reviews help box office performance and bad reviews have a negative impact on the sales. Some research has also done on the textual data of critic's movie reviews like [11] who use movie earnings text analysis on pre-release reviews and metadata features available before movie's release for predicting the opening weekend box office results.

In comparison to above, in his study on movies released in 2003 in the U.S. [19] finds that Metacritic.com scores do not have a strong relationship with the gross earnings of the films. Rotten Tomatoes ratings are used by [28] to find the critic scores to have a positive and significant effect on the movie box office revenue although it is much smaller when compared to independent variables like the number of opening screens and the budget of the movie.

[3] https://www.weibo.com

The aggregate movie critic score impact on movie box office revenue is studied by [18], and they find it to have a small positive effect. However, they do report that the impact is more influential on the total gross revenue of the movie and weaker for predicting the opening weekend earnings. However, authors of [29] find in similar to [3] and in opposite to [18] in their study focusing on individual movie critics, that critics act as more influencers rather than predictors. For a Bollywood movies study, authors of [20] look at both the online user-generated and the expert reviews from daily newspapers and find that volume and valence from both sources have had a positive effect on the financial success of movies.

3 Dataset Description

In this section, we describe various sources of the datasets being used for analysis.

3.1 Movie Selection

We considered Hollywood and Bollywood movies released between April 10th, 2015 and April 6th, 2018. For the sake of consistency, we focused only on the films that are released on Fridays. For Hollywood movies, we only included movies, which had a wide release from its first release day that is a film which runs in 600 or more cinemas [30]. If a movie had a limited release initially, but later went into a wide release then we did not include that in our work. For Bollywood movies, since we did not find any definition for a wide release, thus, we did not apply any such selection criteria for them.

3.2 Tweets collection

Our tweets had two main sources. The first one being the Twitter itself. Following the approach of recent papers like [31, 27] we used the unique hashtags to match a tweet to a movie. This approach has the benefit of being able to find tweets about a movie with a non-unique title like *Sisters* when people have marked them with a hashtag such as #SistersMovie[4]. When inspecting the official Twitter pages of such films, we found that the movie studios often pick the main hashtag for the movie and use it consistently in their marketing campaigns. When such tweets reach their audience, then they tend to use the same hashtag in their own tweets. In our work, we also decided to identify tweets by the hashtags that were used most often to refer to the movie the tweet was about.

For historical tweets, monthly dumps of the *Spritzer* version of the Twitter Streaming API by Archive Team[5] were used as a second source of tweets. Authors in [32] studied the Spritzer version of the Twitter stream on a number of datasets to see if there is any sampling bias in the stream. They found these dumps to be suitable for conducting research experiments and the sampling ratio measured on their

[4] https://twitter.com/sistersmovie

[5] https://archive.org/details/twitterstream

datasets was on an average of 0.95%. The total size of our tweet set downloaded from archive.org was 1.41TB in compressed format containing 4.3 billion tweets.

After gathering and validating the tweets, we had to find the Hollywood and Bollywood movies released during these years and look up the right hashtags for each film from the web. For finding the relevant Hollywood and Bollywood movie release dates we used the Box Office Mojo and Box Office India websites and collected the movies which release date fitted into our historical tweet set timeline. Finding hashtags for the films was again a manual process of looking at the official Twitter pages of the movie and searching for the most popular hashtags people had been using when tweeting about the film. If a tweet did not contain any hashtags or did not contain hashtags about films, then we skipped processing it. Further, if the tweet included any movie hashtags we were interested in, then the number of movies the tweet was about was calculated. If the tweet had hashtags for multiple distinct films, then we discarded the tweet since we could not determine, which movie the tweet was mostly about. Finally, the tweet referring to a single film was stored and assigned to the movie. A total of 281,322 tweets mentioning hashtags for Hollywood and Bollywood movies were extracted from the dataset containing all the tweets.

3.3 From Expert Review Aggregator Sites

Critics' movie reviews are usually published a few days before or on the public release date of the movie, which leaves enough time to influence the movie-goers decision whether to go and see the film or not. Similar to previous work done in studies [19, 33, 34], we decided to use movie review aggregator scores and review counts as an input variable for predicting the box office outcome. For Hollywood movies, we collected movie review scores from the critic score aggregator website *Metacritic*[6] and for Bollywood, we gathered the review scores from the movie info portal *SahiNahi*[7]. The main reason we picked these review sites was that compared to many competitor review sites we investigated, these two had scores available for the most movies in our dataset. Also as mentioned before, *Metacritic* had been used in a number of past studies. Although we did not find any articles, which had used *SahiNahi* scores as an input variable for box office score predictions, but at the same time we did not find any other Bollywood movie critic aggregate site scores having been used either.

3.4 From Movie Revenue Information Sites

General movie information e.g. runtime, genre and the box office results for Hollywood movies was collected from *Box Office Mojo*[8] website which is often used as a source of financial movie information in similar studies to ours [14, 15]. In the case of Bollywood, we collected the data from movie information portal *Box Office In-*

[6] https://www.metacritic.com

[7] https://www.sahinahi.com/

[8] https://www.boxofficemojo.com/

dia[9]. Since for Bollywood movies the parental rating information was not available from *Box Office India*, we gathered the information from Times of India daily news website[10] which includes movie reviews for most of the Bollywood movies. For us, the most interesting data points were the number of theatres the movie was released in, the opening weekend gross domestic income and the budget of the movie.

3.5 Data Cleaning

Unfortunately we did not end up having all the features for every movie we collected available. For example, for some Hollywood and Bollywood movies, the budget info had not been disclosed. Because we use the budget as one of the predictor variables then movies with no budget information were discarded from further study. Also for a few movies like *The Bounce Back*, the Metascore was not available because there were not enough critic reviews about the movie available for Metacritic to generate an aggregated score. After the cleaning was applied, there were 347 Hollywood and 186 Bollywood movies in our study for a combined total of 533 movies.

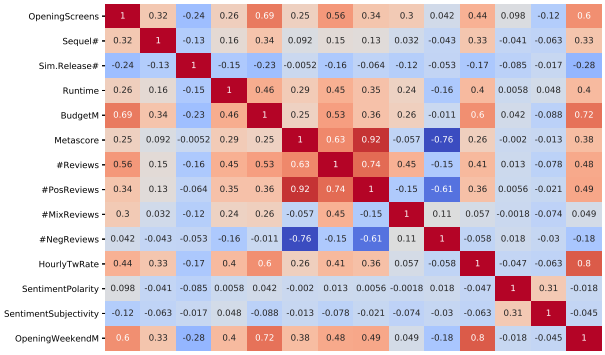


Fig. 1: Feature correlations for the Hollywood dataset

3.6 Exploratory data analysis

The heatmaps on Fig. 1 and Fig. 2 show numeric feature correlation information, which can give us strong hints for understanding which variables could be important for predicting the opening weekend box office. In case of Hollywood on 1 the top three positively correlated features are the number of tweets (0.80), budget (0.72) and the number of theaters (0.60), which all indicate quite strong correlations. We expect these features to be also useful for regression models for predicting the movie revenue. The top three negatively correlated features are the number of releases on the same weekend (-0.28), the number of negative reviews (-0.18) and tweet senti-

[9] https://boxofficeindia.com/

[10] https://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews

ment subjectivity (-0.045). The negative correlation here does not necessarily mean that a feature will not be useful for making box office predictions. On the opposite, the moderate negative correlation of releases on the same weekend variable hints at the expected outcome that more movies opening at the same weekend compete for the same general population to go and see their film and the more movies there are to choose from the less they make on average compared to films that have none or few competitors. It can also hint that sometimes smaller movies try not to compete with big blockbuster movie releases and will release on a different weekend to avoid the strong competition from the hit movies. The weak correlation with the negative review count also shows that the more negative reviews the film has, the less money it is likely to make.

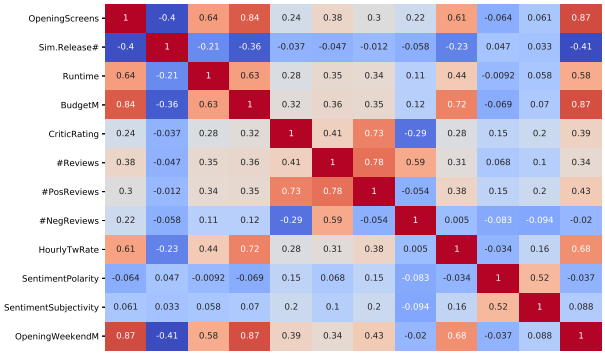


Fig. 2: Feature correlations for the Bollywood dataset

4 Prediction Analysis

The aim of this work is not to identify the best model based on r^2 but rather based on MAPE. We first report model performance using only the Metadata features and then we create also models where we combine Metadata with Critics and Twitter features. We run the experiment using Linear Regression, Random Forest and XGBoost algorithms and use k-fold cross-validation for calculating the average model performance using k=10 folds. We use the default hyperparameters and do not perform any parameter tuning in this experiment to optimize for better scores.

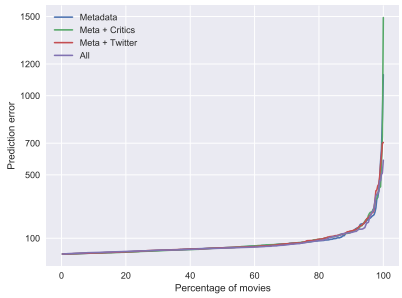
4.1 Hollywood

Table 1 (Columns 2 and 3) shows the performance metrics for different models on the Hollywood movies dataset. We can see that the models with the most features report also the best performance. On average the model using Random Forest algorithm and all the available features predicts with roughly 64% of error. To understand this score and explore it further we gathered the predicted results from each fold and then plotted all the movie predictions by their absolute prediction error.

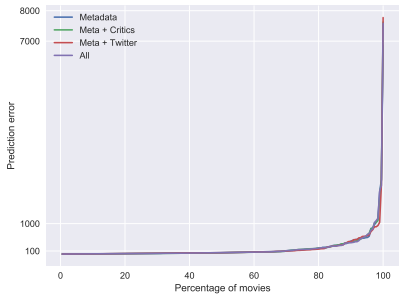
Figure 3a illustrates this experiment and shows that more than 80% of movies have an error of less than 100%, but there are a few outliers, which the model predicted with a very large error. Interestingly, the models with critics features (0.767 r^2, 69% MAPE) perform worse compared to the Twitter counterpart (0.739 r^2, 64% MAPE) in case of the Random Forest algorithm. After inspecting the few movies, which had large errors between the predicted and actual values, we noticed that most of such movies were independent films such as *The Bronze (2015)*, which made $400k during its opening weekend. Our models predicted it would make much more since it was the least earning movie in the dataset, but the dataset did not include many similar movies to learn how to predict revenues so low.

Table 1: Model performance for Hollywood and Bollywood movies. Within a column, boldface shows the best result for a metric.

Model		**Hollywood**		**Bollywood**	
		r^2	MAPE	r^2	MAPE
Metadata	Random Forest	0.700	71.524	0.759	108.268
	XGBoost	0.702	73.625	0.831	115.203
	Linear Regression	0.464	167.757	0.795	183.336
Metadata ∪ Critics	Random Forest	0.767	69.183	0.793	91.012
	XGBoost	0.709	75.782	0.862	90.506
	Linear Regression	0.527	164.487	**0.863**	184.479
Metadata ∪ Twitter	Random Forest	0.739	64.081	0.803	89.338
	XGBoost	0.724	65.499	0.849	86.507
	Linear Regression	0.686	111.766	0.783	177.653
All	Random Forest	**0.777**	**64.007**	0.777	**80.011**
	XGBoost	0.748	65.435	0.862	86.34
	Linear Regression	0.715	111.745	0.855	184.292



(a) Hollywood opening weekend box office absolute prediction errors using Random Forest algorithm



(b) Bollywood opening weekend box office absolute prediction errors using Linear Regression

4.2 Bollywood

Similar to Hollywood we can see for Bollywood from Table 1 that Random Forest algorithm achieved the best (80% MAPE) performance with all features included,

but interestingly the best r^2 score (0.863) was reported by a Linear Regression model where Metadata features were used together with Critics features. However, this model with best r^2 achieved the worst performance (184% MAPE). This illustrates that when comparing model performance, picking the best model based only by r^2 score might not lead to the best performing model in practice. Figure 3b shows the prediction errors using this model resulted in. There are outliers with prediction errors nearly 8000% such as the movie *Uvaa* which affects the MAPE value a lot for all four models with different feature combinations. Since we included also some less popular movies, but their overall distribution in the dataset was not very high, our models do not predict low box office income accurately and tend to overestimate the predictions.

4.3 Hollywood vs. Bollywood

Since Hollywood and Bollywood movie markets are quite different we cannot compare the prediction errors using metrics like MAE and RMSE, but the r^2 and MAPE values are still comparable. In our experiment on the Bollywood dataset, the r^2 values are higher, which indicates that more variance in the predicted opening weekend revenue is explained by the dependent variables we used for predicting. However, the reported MAPE values are larger for Bollywood than for Hollywood models. This can be explained by a few hard-to predict outlier Bollywood movies, which have significantly larger errors than the outliers in case of Hollywood. In the case of Bollywood dataset, outliers have also a bigger total impact since there were twice as many movies for Hollywood in our study. The difference between MAPE errors for Critics and Twitter models is larger for Hollywood movies than it is for Bollywood. For Bollywood movies, we are looking only at tweets in English and did not consider tweets in Hindi. This means we are capturing a larger sample of Tweets for Hollywood movies which benefits the performance of Hollywood Twitter-based models compared to Bollywood.

5 Conclusion and Future Work

Movie sales prediction has been an interest to many researchers as they often carry huge investments. In this work, we investigated the movie sales prediction problem from two different perspectives. Firstly by analyzing the reviews given by movie critics. Secondly, we focus on the wisdom of the crowd, collected using social media platform, Twitter.

Our reported r^2 scores are not as high as some of the earlier works have reported, but it is worth noting that most of such papers use statistical Ordinary Least Squares method without cross-validation on a small set of movies with similar budgets and coming from major studios. In addition to reporting r^2 scores, we report MAPE values, which give a more practical overview of model prediction performance. The results of our prediction analysis show that adding more features will generally improve model performance. Similarly, in both Hollywood and Bollywood dataset ex-

periments, roughly 80% of movies our models were able to predict with 100% of error or less.

The number of movies with their related tweets was the highest we have seen studied so far, but future work should include more movies to build even more effective machine learning models. In our current work, we used the movie critic aggregator scores as a general sentiment polarity score for the movies. For future work we propose to extract different aspect-level sentiment information from movie reviews similar to [35]. Separate aspect-level sentiment scores e.g. for acting, directing, music could all be used as features for the prediction model.

6 Acknowledgments

This work has been supported in part by the SoBigData project (under grant agreement no. 654024)

References

1. Harold L Vogel. *Entertainment industry economics: A guide for financial analysis*. Cambridge University Press, 2014.
2. Mohanbir S Sawhney and Jehoshua Eliashberg. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, 15(2):113–131, 1996.
3. Suman Basuroy, 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(4):103–117, 2003.
4. Yong Liu. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89, 2006.
5. Barry R Litman. Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, 16(4):159–175, 1983.
6. Barry R Litman and Linda S Kohl. Predicting financial success of motion pictures: The'80s experience. *Journal of Media Economics*, 2(2):35–50, 1989.
7. Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.*, 30(2):243–254, February 2006.
8. Krushikanth R Apala, Merin Jose, Supreme Motnam, C-C Chan, Kathy J Liszka, and Federico de Gregorio. Prediction of movies box office performance using social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 1209–1214. IEEE, 2013.
9. Nahid Quader, Md Osman Gani, Dipankar Chaki, and Md Haider Ali. A machine learning approach to predict movie box-office success. In *Computer and Information Technology (ICCIT), 2017 20th International Conference of*, pages 1–7. IEEE, 2017.
10. Jeffrey S Simonoff and Ilana R Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000.
11. Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.
12. Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, Daniel Krause, and Patrick Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pages 276–280. IEEE, 2010.
13. Lee Yoong Hon. Expert versus audiences opinions at the movies: Evidence from the north-american box office. *Marketing Bulletin*, 25:1–22, 2014.

14. Sitaram Asur 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*, volume 1, pages 492–499. IEEE, 2010.
15. Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, 8(8):e71226, 2013.
16. Wan-Hsin Tang, Mi-Yen Yeh, and Anthony JT Lee. Information diffusion among users on facebook fan pages over time: Its impact on movie box office. In *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*, pages 340–346. IEEE, 2014.
17. Reggie Panaligan and Andrea Chen. Quantifying movie magic with google search. *Google WhitepaperIndustry Perspectives+ User Insights*, 2013.
18. Jehoshua Eliashberg and Steven M Shugan. Film critics: Influencers or predictors? *The Journal of Marketing*, pages 68–78, 1997.
19. Timothy King. Does film criticism affect box office earnings? evidence from movies released in the us in 2003. *Journal of Cultural Economics*, 31(3):171–186, 2007.
20. Rakesh Niraj and Jagdip Singh. Impact of user-generated and professional critics reviews on bollywood movie success. *Australasian Marketing Journal (AMJ)*, 23(3):179–187, 2015.
21. Byeng-Hee Chang and Eyun-Jung Ki. Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics*, 18(4):247–269, 2005.
22. Felix Ming Fai Wong, Soumya Sen, and Mung Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 61–66. ACM, 2012.
23. Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke. Predicting imdb movie ratings using social media. In *European Conference on Information Retrieval*, pages 503–507. Springer, 2012.
24. Jonas Krauss, Stefan Nann, Daniel Simon, Peter A Gloor, and Kai Fischbach. Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS*, pages 2026–2037, 2008.
25. Taegu Kim, Jungsik Hong, and Pilsung Kang. Box office forecasting using machine learning algorithms based on sns data. *International Journal of Forecasting*, 31(2):364–390, 2015.
26. Ting Liu, Xiao Ding, Yiheng Chen, Haochen Chen, and Maosheng Guo. Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications*, 75(3):1509–1528, 2016.
27. Dipak Damodar Gaikar, Bijith Marakarkandy, and Chandan Dasgupta. Using twitter data to predict the performance of bollywood movies. *Industrial Management & Data Systems*, 115(9):1604–1621, 2015.
28. Stephanie M Brewer, Jason M Kelley, and James J Jozefowicz. A blueprint for success in the us film industry. *Applied Economics*, 41(5):589–606, 2009.
29. Peter Boatwright, Suman Basuroy, and Wagner Kamakura. Reviewing the reviewers: The impact of individual film critics on box office performance. *Quantitative Marketing and Economics*, 5(4):401–425, 2007.
30. BoxOfficeMojo. Bob Office Tracking By Time. `http://www.boxofficemojo.com/about/boxoffice.htm`, accessed 2018-04-15.
31. Steve Shim and Mohammad Pourhomayoun. Predicting movie market revenue using social media data. In *Information Reuse and Integration (IRI), 2017 IEEE International Conference on*, pages 478–484. IEEE, 2017.
32. Yazhe Wang, Jamie Callan, and Baihua Zheng. Should we use the sample? analyzing datasets sampled from twitters stream api. *ACM Transactions on the Web (TWEB)*, 9(3):13, 2015.
33. Shyam Gopinath, Pradeep K Chintagunta, and Sriram Venkataraman. Blogs, advertising, and local-market movie box office performance. *Management Science*, 59(12):2635–2654, 2013.
34. Thorsten Hennig-Thurau, Mark B Houston, and Gianfranco Walsh. Determinants of motion picture box office and profitability: an interrelationship approach. *Review of Managerial Science*, 1(1):65–92, 2007.
35. Rajesh Piryani, Vedika Gupta, and Vivek Kumar Singh. Movie prism: A novel system for aspect level sentiment profiling of movies. *Journal of Intelligent & Fuzzy Systems*, 32(5):3297–3311, 2017.