Using Crowd-source based features from social media and Conventional features to predict the movies popularity

Mehreen Ahmed[1], Maham Jahangir[2], Dr. Hammad Afzal[3]
Department of Computer Software Engineering
National University of Sciences and Technology (NUST),
Islamabad, Pakistan
[1]mahreenmcs@gmail.com, [2]mahamjahangir@yahoo.com,
[3]hammad.afzal@mcs.edu.pk

Dr. Awais Majeed[4], Dr. Imran Siddiqi[5]
Bahria University
Islamabad, Pakistan
[4]awais.majeed@bui.edu.pk, [5]imran.siddiqi@bui.edu.pk

Abstract— **Predicting the success of movies has been of interest to economists and investors (media and production houses) as well as predictive analysts. A number of attributes such as cast, genre, budget, production house, PG rating affect the popularity of a movie. Social media such as Twitter, YouTube etc. are major platforms where people can share their views about the movies. This paper describes experiments in predictive analysis using machine learning algorithms on both conventional features, collected from movies databases on Web as well as social media features (text comments on YouTube, Tweets). The results demonstrate that the sentiments harnessed from social media and other social media features can predict the success with more accuracy than that of using conventional features. We achieved best value of 77% and 61% using selected social media features for Rating and Income prediction respectively; whereas selected conventional features gave results of 76.2% and 52% respectively. More it was found that the blend of both types of attributes (conventional and those collected from social media) can outperform the existing approaches in this domain.**

Index Terms—Data Mining, Predictive Analysis, Classification, Regression

I. INTRODUCTION

Prediction of success in business has been of great interest to the economists and financial experts. With advent of data analytics, the prediction process has been made intelligent by considering the historical data and employing various data analytical techniques to infer the future events. Such studies have been performed in prediction of movies success as well where success and popularity is measured in terms of the Ratings (typically represented by a numeric number from 0-10) and Income. There have been a large number of studies reported in this domain due to reasons such as general interest of public in this popular medium of entertainment, non-requirement of domain experts as required in other domains such as medical and huge number of data freely available on Web resources such as IMDB[1].

Most of the studies performed for prediction of movies success use conventional attributes, collected from online movies databases. However, with advent of social media, public opinion has been harnessed about various events/entities from forums such as YouTube and Twitter. Similarly, for movies, social media websites have contributed a great amount to the popularity of movies. Now anyone can review, rate, comment or share their opinions about a movie online. Thus social media plays a vital role in predicting the success of a movie. Many researchers believe that one should consider the social factors along with the classical factors for this purpose. Among social media mediums, Twitter has gained remarkable popularity and usage lately. Thus making it a point of focus, for researchers to predict the movie success using sentiments or feedback collected via twitter. However, most of the studies performed in this domain have shown that sentiments about movies are not determining factor (or among the top factors) in predicting the success of movie while calculating it before release.

In our experiments, we considered a number of conventional features collected from IMDB such as *Genre, Budget, Number of Screens* and *Sequel* and other features taken from social media such as YouTube and Twitter. These latter type of attributes include *Aggregate Actor Followers* (equal to sum of followers of top 3 cast from Twitter), *Number of views*, *Number of likes* and *Number of dislikes* , *Number of comments* (all taken from official trailer of movies on YouTube) and *Sentiment Score* (from Twitter). We found that most discriminating feature that played vital role in classifying the Ratings and Income was *Sentiment score*. One of our proposed feature, *Aggregate Actor Followers*, performed better than traditionally used *Top Actor follower*s. Other features proposed including *Likes/Dislikes* on YouTube, *Sequel* (represented as a number from 1 to N) also played vital role in prediction. We used these attributes to measure (i) Rating of movie, (ii) Box Office success of movie. A number of experiments were performed using Linear Regression, Decision Tree (J48), Artificial Neural Network and Support Vector Machine. The best performance was calculated using J48, where sentiment score came up as the best discriminating attribute. We achieved best value of 77% and 61% using selected social media features for Rating and Income prediction respectively; whereas selected conventional features gave results of 76.2% and 52% respectively. Moreover, 95% accuracy was calculated while predicting the Rating using Linear Regression, while considering the predicted value in range of 1 as correct.

[1] http://www.imdb.com/

978-1-5090-1893-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SmartCity.2015.83

IEEE computer society

II. Literature Review

For many years, researchers have been investigating and generating predictive models for the movies performance. They have used conventional features as well as social media features to predict the popularity of movies. However, most of the studies have shown that conventional features are better in predicting the success of movies. A few studies performed in this domain are briefly described in this section.

A study [1] was conducted over a span of five years (1998-2002) in which the authors classified nine classes from flop to blockbuster. They applied neural network algorithm on 7 independent variables and found that number of screens, high technical effects and high star value contribute a great deal to a movie's success. K-Means clustering, Polynomial and Linear Regression [2] was applied on 2510 movies released 1990 onwards to study and build a predictive model to get the expected revenue. They achieved accuracy of 36.9%.

Another study [3] applied Text regression on critics' film reviews to predict the opening weekend revenue for the metadata collected for 2005-2009 movies. The dataset consisted of 1718 movies. The authors used seven metadata features including Movie Running Time (in minutes) , Budget, the number of opening weekend screens, genre, MPAA rating, opening time (whether summer or holiday), total number of actors, high grossing actors count and whether the movie had any Oscar winning actors and directors. Similarly three types of text features were extracted from the metadata features. For the first weekend release revenue metadata features gave an accuracy of 0.521 and the amalgamation of text, metadata features gave even better results.

In [4] researchers proposed the idea to integrate classical and social media factors to improve the prediction accuracy of the movie success. They collected classical attributes (genre, budget etc.) from IMDB and social attributes (Tweets, views) from social websites like YouTube, Twitter. The study suggests that by increasing the data set, a higher accuracy than the one obtained (70%) through linear regression, can be achieved.

In a similar study [5], authors predicted the first weekend box office revenue for movies released in 2010. They used a data set of 312 movies collected from BoxOffice Mojo and the attributes including views count, editors' count, number of edits and collaborative rigor from Wikipedia articles. The opening weekend revenue and number of theatres screens were also included. They applied linear regression and got an accuracy of 0.94 one month prior to release date of the movie.

Author [6] used an existing data set of 2009, 2012 movies provided by SNAP, a Stanford university research group. They collected the tweets text, id, username, time and method from Twitter API and searched for the relevant movie tweets. Lingpipe sentiment analyzer was used for Sentiment analysis on the tweets, to classify movies as hit, flop and average. An accuracy of 64.4% was computed as tweets can have noisy data and the analyzer used was not suitable for tweets.

Sitaram and Bernardo [7] investigated if tweets prior to the release of a movie can predict the opening weekend revenue. They used Twitter API to extract 2.89 million Tweets for 24 movies of 2009. They concluded that the effect of promotional tweets was negligible while the tweet mentions per hour for a movie predicted accurate box office results. After predicting first weekend revenue they calculated the subjectivity and polarity of the movies by applying sentiment analysis on the tweets. Although the sentiments did improve the results, they were not as important as the tweet rate.

In [8] the authors analyzed the effect of twitter on movie sales by partitioning the user tweets with more than 400 followers as Type 1 and less than 400 followers as Type 2. They collected the total tweets, positive, negative tweets ratio and intention ratio and found that Type 2 tweets have a higher impact on movie performance. They conclude that using effective sentiment analysis algorithms to classify tweets cannot get you perfect results and can be challenging.

Sentimental Analysis [9] was performed by authors on the YouTube comments to classify movies as hit, neutral or flop for 35 movies for the year 2013. The attributes included i) genre, ii) director, actor and actress popularity, iii) view and comment count, iv) sequel movie and v) sentiments.

Table 1: Experimental results of movies popularity prediction reported in literature

Ref	Number of Movies	Data Set Source	Features Extracted	Algorithm Used	Predicted Class	Accuracy	Bias
[1]	834	ShowBiz Data	Sequel, Screens, MPAA, Star value, Genre, Technical Effects, Competition	Neural Network	Revenue	36.9%	No Bias
[8]	63	BoxOfficeMojo, Twitter	Followers, Total tweets, Tweet ratio, type I & II tweets ratio, Positive & negative tweets ratio, intention tweets ratio	Linear Regression	Revenue	93% precision, 99% recall	Tweets showing intention of watching were used
[5]	312	BoxOfficeMojo, Wikipedia, Wikimedia Toolserver	View count, editors count, edits count, collaborative rigor, revenue, screens count	Linear regression	Opening Weekend Revenue	0.94 coefficient of determination R2 (t) for few days before release	Only 1 month before release
[6]	30	Twitter	Tweet sent Time, Tweets number, positive ratio, negative ratio and profit ratio	Sentiment Analysis	Movie Success (Hit, Flop, Average)	**64.4% accuracy**	No Bias

Popularity was assessed on the basis of Twitter followers. Views and Comments were collected from the YouTube movie trailers for the movies. K means algorithm was used for clustering the movies as hit, flop or neutral. In the experiment they applied weight to the sentiment attribute and thus known as non-uniform weighted. And the attributes given equal weights as uniform weighted. An accuracy of 89% was calculated for uniform weighted with Naïve Bayesian and 100% with J48. For non-uniform weighted attributes they got an accuracy of 86% with Naïve Bayesian and 94% with J48. According to them, the actress popularity was the most important feature and sentiment was disregarded by J48. The results they got were for a small data set of 35 movies, so their findings might be inconclusive.

Table I shows the summary of the work done on predicting movies performance.

III. METHODOLOGY

The overall methodology is shown in Figure 1. Our system is comprised of two major modules, namely Data Collector and Predictive Engine. Data Collection is the more significant task of the two; it involves data collection from IMDB, Twitter and YouTube and then pre-processing that involves dealing with maximum values, data transformation (converting currencies etc.) and calculation of sentiment analysis score for each of the tweet etc. Data Collection and Prediction Model are explained below:

A. Data Collection

Data Collector is the major module as it retrieves information about movies from diverse sources including movies web sites i.e. IMDB, generic web resource i.e. Wikipedia, and social media including YouTube and Twitter. Furthermore, we used sentiment analysis libraries to get the sentiment score for different movies. As the data, that we are interested in such as followers count on twitter, ratings on IMDB etc. continuously changes, therefore, we collected the latest data from these web resources by using APIs and scrappers instead of using already available movies datasets.

A.1 Feature Extraction

In total, we collected data of 12 features for each movie. The features can further be categorized as conventional features and social media features. Conventional features are those that are typically available on movies resource websites (such as IMDB). These include

1. Genre
2. Budget
3. Number of Screens
4. Sequel
5. Ratings
6. Gross Income

Genre: 19 values of Genre such as Action, Adventure and Drama etc. are used. Each nominal value of genre was mapped onto numeric value from 1-19 in order to improve the performance of learning algorithms.

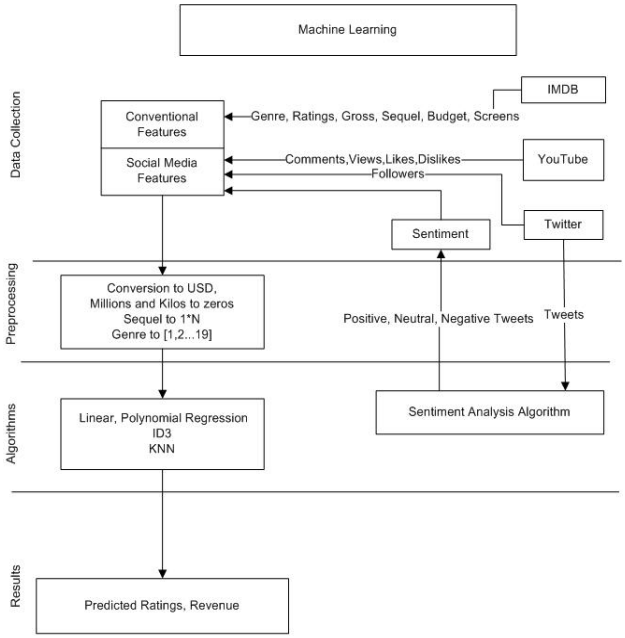

FIG 1: SYSTEM MODEL- PREDICTIVE MODEL OF MOVIES POPULARITY USING SENTIMENT ANALYSIS AND OTHER CONVENTIONAL FEATURES

Sequel: The value represents whether the movie is sequel or an individual. It is a numeric number that ranges from 1 to N. 1 shows that movie is first release (no sequel), whereas 2 represents that the movie is 2nd in a sequel; E.g. Pirates of Caribbean: Dead Man's Chest is 2nd in sequel, therefore it is assigned the value of 2.

Ratings: The value of Ratings ranges between 1 to 10 with 1 being lowest and 10 the highest. These values are collected from IMDB.

Gross Income, Budget and Number of Screens: Gross world-wide income and Budget for each movie is collected from IMDB. The values are converted into USD (if available in other currencies). Number of screens on which movie was initially launched in US is also considered.

From social media, we collected following features for each movie.

1. Aggregate Actor Followers
2. Number of views
3. Number of likes
4. Number of dislikes
5. Number of comments
6. Sentiment Score

Aggregate Actor Followers: Number of followers on twitter is used. Initially, we considered only the top actor; however, the data got too sparse as there are many actors who do not have twitter account, therefore, we used the followers count of top 3 cast.

Number of Views and Comments: The number of views and comments of trailer of movies on YouTube are calculated.

Number of likes and dislikes: Similar to the number of views and counts, number of Likes and Dislikes of trailers on YouTube are considered.

Sentiment Score: This feature is represented by signed integer value. 0 represents neutral sentiment, "+"sign shows the positive sentiment whereas number shows the magnitude. Similarly "–"sign shows negative sentiment. Sentiment score is calculated by retrieving all tweets related to each movie, assigning the sentiment score to each of them and then aggregating the score. It is calculated as below:

$$\textbf{\textit{Sentiment}}\ (M_i) = S_1 + S_2 \ldots S_n$$

Where,
Sentiment (M_i) represents the sentiment score of a movie i.
S_n represents the sentiment score of a tweet n,
and n is the total number of tweets linked with movie M_i

B. Creation of Prediction Model

As described in previous section, a total of 12 attributes were extracted for each movie. Among them, 10 features (Genre, Budget, Screens, Sequel and all 6 attributes from social media) are used to predict the value of Ratings and Gross Income. As discussed before, we have assumed that popularity is depicted by Rating or Income; therefore we have performed 2 sets of experiments, first for Rating and second for Income. Performance of Conventional features and Social media oriented features are measured separately and by combining them. The experimental setup is described below:

A. Experiment 1a

In this experiment, values of ***Ratings*** are predicted using all other attributes except Gross Income (as gross income is not available before release). As ***Rating*** is a continuous numeric attributes, we have performed Linear Regression in order to predict the values.

B. Experiment 1b

In this experiment, we have converted the values of Rating into 4 bands as shown in Table 2. The movies with Rating value ranging between 0-4.9 is assigned label Poor and so on. As we have converted numeric attribute into ordinal, we were able to apply a number of algorithms including Decision Tree (J48), SVM and ANN. Best performance was measured using J48.

TABLE 2: MAPPING THE LABELS TO RATING VALUES OF MOVIES

Ratings	Assigned Label
0-4.9	Poor
5-6.4	Average
6.5-8.9	Good
9-10	Excellent

C. Experiment 2

In this experiment, we developed a model to predict the value of Income. As Income is a continuous numeric attribute with values having a large range, we converted the values of income into ordinal and assigned the labels as shown in Table 3.

TABLE 3: MAPPING THE LABELS TO INCOME VALUES OF MOVIES

Income (in Millions)	Assigned Label
Below 0.1	Flop
0.1 to 9	Average
10 to 90	Success
More than 100	Block-Buster

IV. RESULTS AND DISCUSSION

The results of experiments are described below.

A. Experiment 1a

In this experiment, values of Ratings are predicted. We performed experiments with different evaluation strategies such as percentage split, 10 Fold Cross Validation Method; however, best results were obtained with 80% Split of Training and Testing data. In order to measure accuracy, we have performed two calculations: $Accuracy_1$ which is measured as:

$$Accuracy_1 = \frac{\text{Number of Movies with exact prediction of Rating}}{\text{Total Number of Predictions}}$$

and $Accuracy_2$ which is measured as:

$$Accuracy_2 = \frac{\text{Number of Movies with approx prediction of Rating}}{\text{Total Number of Predictions}}$$

In $Accuracy_2$, True Positives are calculated as Number of instances where predicted value is in range of 1 to that of Actual Value. For example, predicted rating value of 6 is considered correct if actual value is [5-7].

TABLE 4: PERFORMANCE EVALUATION OF LINEAR REGRESSION TO PREDICT THE ACCURACY OF CONVENTIONAL FEATURES AND SOCIAL MEDIA FEATURES

	Accuracy1	Accuracy2
Social Media Features (SMF)	43%	**95%**
Conventional Features (CF)	36.4%	**89.2%**

The results are summarized in Table 4 and illustrated in Figure 2.

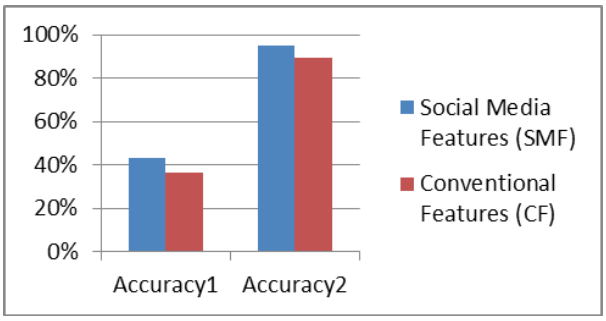

FIG 2: PERFORMANCE EVALUATION OF CONVENTIONAL FEATURES AND SOCIAL MEDIA FEATURES IN PREDICTING RATINGS

B. *Experiment 1b*

In this experiment, values of **Bands of Ratings** are predicted. 80% Split of training and testing data gave best result for evaluation. Accuracy is calculated as:

$$Accuracy = \frac{\text{Number of Movies with correct prediction of BandofRating}}{\text{Total Number of Predictions}}$$

TABLE 5: PERFORMANCE EVALUATION OF DECISION TREE (J48) TO PREDICT THE BAND OF RATINGS USING CONVENTIONAL FEATURES AND SOCIAL MEDIA FEATURES

	Accuracy
Social Media Features (SMF)	**77%**
Selected SMFs	**77%**
Conventional Features (CF)	**47%**
Selected CFs	**76.7%**

Accuracy is measured by considering all Social Media Features (SMFs), selected SMFs, all Conventional Features (CFs) and selected CFs. Selected features contain the best 3 features which were chosen by Feature Selection method of Principal Component Analysis. In SMFs, *Sentiment score* was best while *Likes/Dislikes* were the worst. Similarly, three best features were considered for CFs. *No of Screens*, *Genre* and *Budget* came as best three features.

Experiments showed that results of *SMFs* were slightly better than *CFs.*

C. *Experiment 2*

In this experiment, values of **Bands of Popularity** as shown in table 3 are predicted. Evaluation is measured using 80% split of training/testing data. Accuracy is calculated as:

$$Accuracy = \frac{\text{Number of Movies with correct prediction of Band of Income}}{\text{Total Number of Predictions}}$$

Results are summarized in Table 6. As shown, social media features outperformed conventional features, giving the best value of 61% with selected social media features.

TABLE 6: PERFORMANCE EVALUATION OF DECISION TREE TO PREDICT THE BAND OF INCOME USING CONVENTIONAL FEATURES AND SOCIAL MEDIA FEATURES

	Accuracy
Social Media Features (SMF)	35%
Selected SMFs	**61%**
Conventional Features (CF)	42%
Selected CFs	52%

Figure 3 summarizes the performance of Social Media features and Conventional features in determining the Ratings and Incomes. As shown, Social Media Features (SMF) and their selected subset outperformed conventional features in both set of experiments.

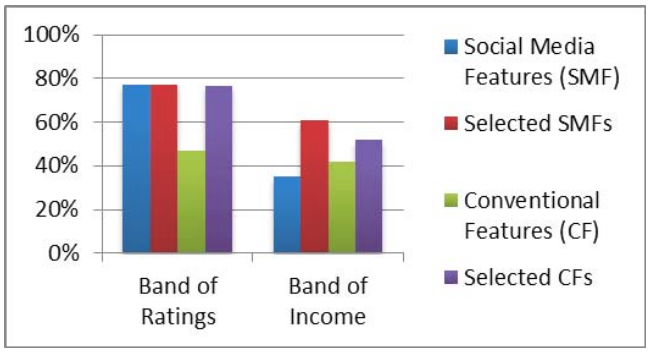

FIG 3: OVERALL COMPARISON OF CONVENTIONAL FEATURES AND SOCIAL MEDIA FEATURES IN PREDICTING RATINGS AND INCOME.

V. CONCLUSIONS

This paper presents the comparison of Conventional Features with Social Media features in determining the popularity of movies. Our experiments showed that social media features such as Sentiment Score of tweets related to movies, Number of Views and Comments of movies' trailers on YouTube and fan following on twitter can usefully be utilized to predict the popularity of movie. We assumed that popularity is depicted by movie rating and gross income, and performed two set of experiments to predict these features individually. In both set of experiments, it was found that social media features are better. This is in contrast to findings of some other research works reported in this domain, where social media features were not deemed useful as compared to other features such as number of screens and genre of movies. However, most of these studies were performed either on small set of data (less than 40 movies), or they considered tweets of only 1 week/1 month before release. Our experiments showed that, on larger datasets, with no consideration of time while collecting statistics from twitter and YouTube, social media features give better performance.

REFERENCES

[1] Sharda, R., & Delen, D. (2006): "Predicting box-office success of motion pictures with neural networks". Expert Systems with Applications, 30(2), 243-254.

[2] Nikhil Apte, Mats Forssell, and A. Sidhwa, "Predicting Movie Revenue". 2011.

[3] Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010). "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 (pp. 293-296). Association for Computational Linguistics.

[4] Bhave, A., Kulkarni, H., Biramane, V., & Kosamkar, P. (2015). "Role of different factors in predicting movie success". In Pervasive Computing (ICPC), 2015 International Conference on (pp. 1-4). IEEE.

[5] Mestyán, Márton, Taha Yasseri, and János Kertész. "Early prediction of movie box office success based on Wikipedia activity big data." PloS one 8.8 (2013): e71226.

[6] Jain, Vasu. "Prediction of Movie Success using Sentiment Analysis of Tweets." The International Journal of Soft Computing and Software Engineering3.3 (2013): 308-313.

[7] Asur, Sitaram, and Bernardo Huberman. "Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

[8] Rui, Huaxia, Yizao Liu, and Andrew Whinston. "Whose and what chatter matters? The effect of tweets on movie sales." Decision Support Systems 55.4 (2013): 863-870.

[9] Apala, Krushikanth R., et al. "Prediction of movies box office performance using social media." Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.