ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Survey on Movies Popularity Prediction System Using Social Media Feature

Babita M. Jangid[1], Chaitali K. Jadhav[2], Swati M. Dhokate[3], Grish M.Jadhav[4,] Prof. G.M.Bhandari[5]

B.E. Students, Department of Computer Engineering, JSPM'S BSIOTR, Wagholi, Pune, Maharashtra, India[1,2,3,4]

Asst. Professor, Department of Computer Engineering, JSPM'S BSIOTR, Wagholi, Pune, Maharashtra, India[5]

ABSTRACT: Using Crowd-source based features from social media and Conventional features to predict the movies popularity. Now a day, use of social media has been increased widely. By using social media like Twitter, YouTube etc. users can posts their reviews about movies. The economists, investors and predictive analysts are very interested to predict the success of their movies. For predicting success as well as popularity of movies number of factors affected like actor, actress, invested budget, production house, genre, PG rating. In this project, machine learning algorithms are used for predictive analysis. Machine learning algorithms applied on conventional, collected from movies databases, and social media features (text comments on Tweets, YouTube). Mining the attributes and contents of social media gives us an opportunity to discover social structure characteristics, analyze action patterns qualitatively and quantitatively, and sometimes the ability to predict future human related events. Result of this project is predicts the success with control and use of sentiments form social media and other social media features. 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).

KEYWORDS: Energy Data Mining, Predictive Analysis, Classification, Regression.

I. INTRODUCTION

There are many users sharing their opinions and experiences via social media, there is aggregation of personal wisdom and different viewpoints. Such aggregation has limitations as viewpoints are subject to change with time. In a sense the social media prediction problem is paralleled by prediction of financial time series based on past history, which has its uses in trading. In general, if extracted and analysed properly, the data on social media can lead to useful predictions of certain human related events. Such prediction has great benefits in many realms, such as finance, product marketing and politics, which has attracted increasing number of researchers to this subject. Study of social media also provides insights on social dynamics and public health. A survey provides us perspective and is helpful for carrying out further research.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 IMDB1. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

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.

There are many forms of social media that include blogs, social networking sites, virtual social worlds, collaborative projects, content communities and virtual game worlds. Some forms of social media lack a social network. Thus in blogspot.com, which is a famous blog platform, there is no social links among bloggers.Social media has some or all of these seven function blocks: identity, conversations, sharing, presence, relationships, reputation, and groups. Different forms of social media have different points of focus. For example, collaborative projects such as Wikipedia mostly care about sharing and reputation, And in virtual game worlds, identity, presence, reputation, and groups are of the greatest concern.

Recently, social media played important role in unfolding newsworthy events. For example, in the aftermath of the Tohoku Earthquake in Japan people used social media to contact friends, exchange crisis information, and find necessary resources.To predict movie box-office with social media is one of the most studied area. In addition to the traditional predictionfactors, such as MPAA rating and number of screens, social media contents could also be effective to predict box-office. There are many reasons that predicting movie box-office a good subject for research. Firstly, there are volumes of data about movies and related social media. According to IMDB.com, more than 200 feature films, which originate in the U.S.A and have U.S.A box-office record, were released every year. Besides, movies are widely talked on social media. For example, there are more than 100,000 tweets for each monitored movie. Consequently, there is enough data to be analyzed.

Secondly, the box-office is easy to be accessed and estimated. On one hand, the gross income and opening weekend income is easily obtained from Internet Movie Database (IMDB). On the other hand, the income on opening weekend typically accounts for about 25% of total sales. So we could get the approximate box-office just after the opening weekend. In some cases, the prediction about the high-grossing movies is much accurate than that about low-grossing movies. Even though most researchers treat the box-office as continuous variable, sometimes discretization is applied to divide the box-office into classes according to their amount.

II. RELATED WORK

In [1] author 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 7independent 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 released1990 onwards to study and build a predictive model to get theexpected revenue. They achieved accuracy of 36.9%.

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

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

In 5], authors predicted the first weekendbox office revenue for movies released in 2010. They used adata set of 312 movies collected from BoxOffice Mojo and theattributes including views count, editors' count, number ofedits and collaborative rigor from Wikipedia articles. Theopening weekend revenue and number of theatres screenswere also included. They applied linear regression and got anaccuracy of 0.94 one month prior to release date of the movie.

In [6] author used an existing data set of 2009, 2012 moviesprovided by SNAP, a Stanford university research group.They collected the tweets text, id, username, time and methodfrom Twitter API and searched for the relevant movie tweets.Ling pipe sentiment analyzer was used for Sentiment analysison the tweets, to classify movies as hit, flop and average. Anaccuracy of 64.4% was computed as tweets can have noisydata and the analyzer used was not suitable for tweets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

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 24movies 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.

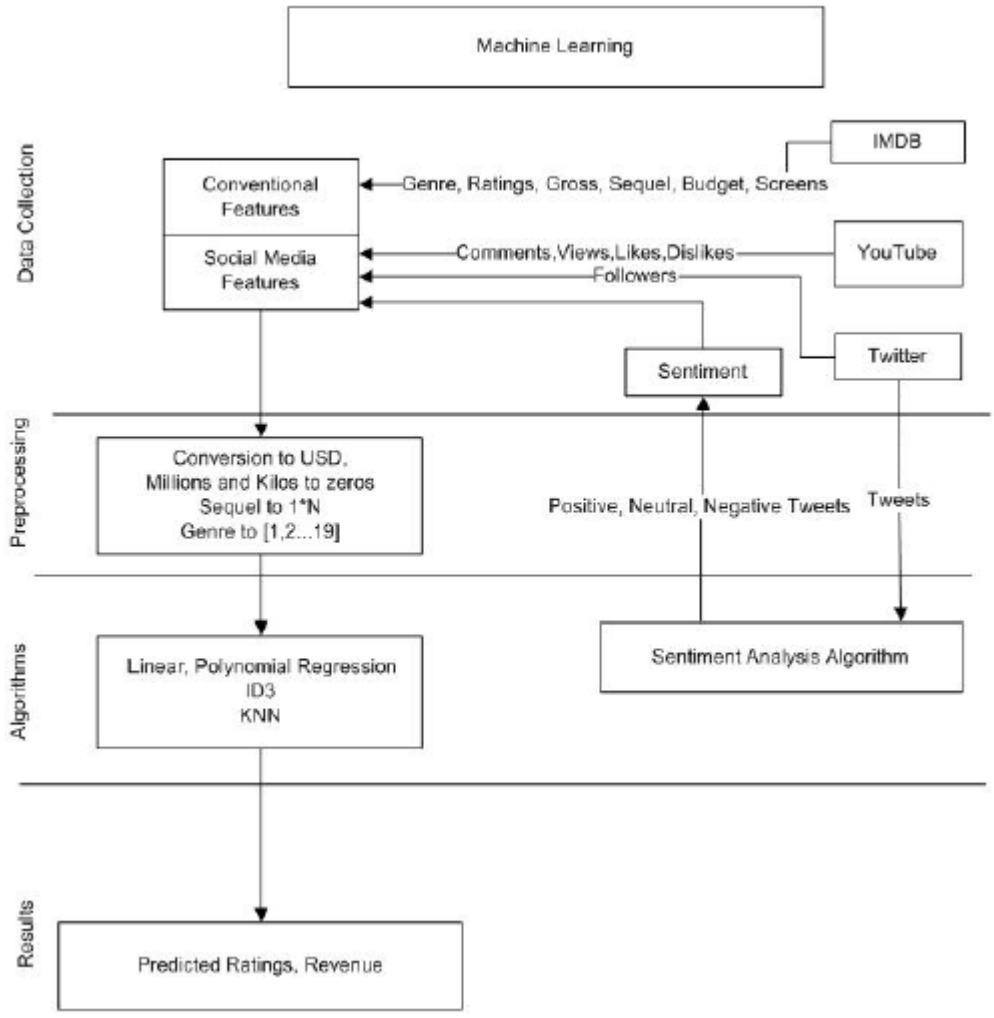

FIG 1: SYSTEM MODEL- PREDICTIVE MODEL OF MOVIES POPULARITY USING SENTIMENT ANALYSIS AND OTHER CONVENTIONAL FEATURES

In [8] the authors analyzed the effect of twitter on moviesales by partitioning the user tweets with more than 400followers as Type 1 and less than 400 followers as Type 2.They collected the total tweets, positive, negative tweets ratioand intention ratio and found that Type 2 tweets have a higherimpact on movie performance. They conclude that usingeffective sentiment analysis algorithms to classify tweets cannot give you proper result and it was challenging also.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

III. PROPOSED SYSTEM

A. *USING SOCIOLOGICAL THEORY TO INTERPRET PREDICTORS:*

Currently, researchers choose predictors using the trial and error method. We know neither why these predictors are better than others, nor how these predictors could predict the result. Not knowing the background logic between these metrics and the final prediction result, we just use a collection of metrics to be trained on test data, find out which ones have the highest coefficients, and use them to compose the prediction model. As well as lacking a solid supporting theory, we cannot be sure that one model, which works well in one case, could be applied to other situations with the same accuracy. That's why some models show good performance in one election prediction, but completely fail in another one. To guarantee our model has good performance in all cases, we need to know the logic and theory behind the model.

B. *TRYING MORE PREDICTION METHODS:*

Most researchers use simple methods such as linear regression analysis. These methods are known to work well under some conditions. Social media is produced on a complex system and thus more likely than not the predictors and prediction outcomes have non-linear correlation. Furthermore, combination of methods might lead to breakthrough. In such combination, a surface learning agent, such as instantaneously trained neural networks, quickly adapts to new modes and emerging trends on social media. And a deep learning agent focuses on long-term patterns. In a nutshell, we should try some non-linear methods and find out the suitable methods and/or combinations for each prediction realms.

C. *MODELING ON PREDICTIONS WITH SOCIAL MEDIA:*

We are far from knowing everything about social media. For instance, there are different kinds of prediction objects which show different features. Taking recommendation adoption as an example, the recommendation on DVDs is more likely to be accepted than that on books. But there is still no universal accepted conclusion about why these differences exist. This lack of understanding adds to the difficulties of modelling. Formal modelling could be necessary and helpful to understand and investigate the features and behaviours of prediction techniques.

D. *SEMANTIC ANALYSIS SYSTEM FOR SOCIAL MEDIA*:

Although semantic analysis is not a necessary part of the prediction methods, it is frequently used. Thus the accuracy of semantic analysis is critical to the prediction performance. Semantic analysis could be based upon lexicon or previous statistics. In terms of lexicon, compared with natural and formal English language, social media content has similar structure, but many different words, such as "lol", which short for "laughing out loud". This Internet slang affect the semantic analysis system, because the lexicon in most existing systems is designed for well-written English. Besides, Internet slang evolves quickly and chaotically. The SOM, which sometimes is used to construct thesaurus as an unsupervised or semi-supervised clustering method, could be helpful in this issue. These methods firstly label some posts manually and then use statistical model, such as naïve Bayes classifier, to mark other posts according to statistical features of labelled ones. In some forms of social media, such as microblogging, the length of post is so short that it shows no significant statistical characteristics.

IV. CONCLUSION

In this paper, we presented a survey of prediction using social media. We also gave an overview of16prediction factors and methods and listed challenging problems and areas for further research. Although prediction using social media is only an emerging research topic and its results have relatively low accuracy, it has created a new way for us to collect, extract and utilize the wisdom of crowds in an objective manner with low cost and high efficiency. 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' trailerson 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

REFERENCES

1. Sharda, R., &Delen, D. (2006): "Predicting box-office success of motion pictures with neural networks". Expert Systems with Applications, 30(2), 243-254.277
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 forComputational 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 Conferenceon (pp. 1-4). IEEE.
5. Mestyán, Márton, TahaYasseri, and JánosKertész. "Early prediction of movie box office success based on Wikipediaactivity big data." PloS one 8.8 (2013): e71226.
6. Jain, Vasu. "Prediction of Movie Success using Sentiment Analysis of Tweets." The International Journal of SoftComputing 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 InternationalConference on. Vol. 1. IEEE, 2010.