ISSN 2395-1621

Prediction of Movies Box Office Performance Using Social Media

[#1]PRATIK DHANAWADE, [#2]RITESH PATIL, [#3]REENA CHOUGULE, [#4]KIRTI ADWANT

[1]dhanawadepratik@gmail.com
[2]riteshpatil36@gmail.com
[3]reenachougule93@gmail.com

[#1234]Department of Computer Engineering

G H Raisoni College of Engineering And Management, Wagholi, Pune.

ABSTRACT

Social media such as Twitter and YouTube have been used for sharing contents and comments on all types of subjects by millions of people on a daily basis. It is clear that businesses have a strong interest in tapping into these huge data sources to extract information that might improve their decision making process. For example, predictive models derived from social media for successful movies may facilitate filmmakers making more profitable decisions. The topic of movies is of considerable interest in the social media user community. Posting online reviews is a new trend set for people to share with other users their opinions and sentiments toward products and services. E-commerce websites provides the venues and facilities for people to publish their reviews. Those online reviews present a wealth of information. In this project, reviews of viewers from movie fertility are collected and processed. When the movie trailer releases various reviews are posted by users on social media sites. We are mining those reviews and predicting performance of the movie. These predictions will be used by shareholders and box office for the movie business. We label the prediction in three classes, Hit, Neutral and Flop.

Keywords: Twitter, YouTube, Online reviews, ,Opinions, Sentiments , Mining ,Prediction.

ARTICLE INFO

Article History

Received: 28h November 2016

Received in revised form :

28th November 2016

Accepted: 1st December 2016

Published online :

2nd December 2016

I. INTRODUCTION

In this era social media is becoming more popular where netizens can express themselves, gives reviews etc. data generated through social media is nearly 10TB per day. With increase in such large amount of data it is necessary to develop a system which will make use of such large amount of data to perform analysis and predict future with social networking. So we are developing a system which makes use of twitter data for predicting box office collection of movie. This system include Natural Language Processing domain of computer science and scientific study of human language i.e. linguistics which is related with the interaction or interface between the human (natural) language and computer. Opinion mining or Sentiment analysis refers to a broad area of Natural Language Processing and text mining. It is concern not with the topic a document is about but with opinion it expresses that is the aim is to determine the attitude (feeling, emotion and subjectivities) of a speaker or writer with respect to some topic to determine opinion polarity. Initially it was applied for classifying a movie as good or bad based on positive or negative opinion. Later it expanded to star rating predictions, prediction of box office collection of movie. Application Of Sentiment Analysi Mouth publicity is the process of conveying information from person to person and plays a major role in customer buying decisions. In commercial situations, consumers share attitudes, opinions, orreactions about businesses, products, or services with other people. People trust on families, friends, and others in their social network. Research also indicates that people appear to trust opinions from people outside their immediate social network, such as online reviews. This is where Sentiment Analysis comes into play. Availability of opinion rich resources like online review

sites, blogs, social networking sites have made this "decision-making process" easier for us because we can get more reviews about product or services from consumers all across the world. With explosion of social networking platforms consumers have a power by which they can share opinions. Major companies have realized these consumer voices affect shaping voices of other consumers. Sentiment Analysis thus finds its use in Consumer Market for Product reviews, Marketing for knowing consumer attitudes and trends, Social Media for finding general opinion about recent hot topics in town, Movie to find whether a recently released movie is a hit.

II. LITERATURE SURVEY

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 the metadata collected for 2005-2009 movies. The dataset consisted of 1718 movies. The authors used seven meta data 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, meta data 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) rom 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 [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 an d got an accuracy of 0.94 one month prior to release date of the movie.

In [6] author 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. Ling pipe 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.

III. PROPOSED SYSTEM

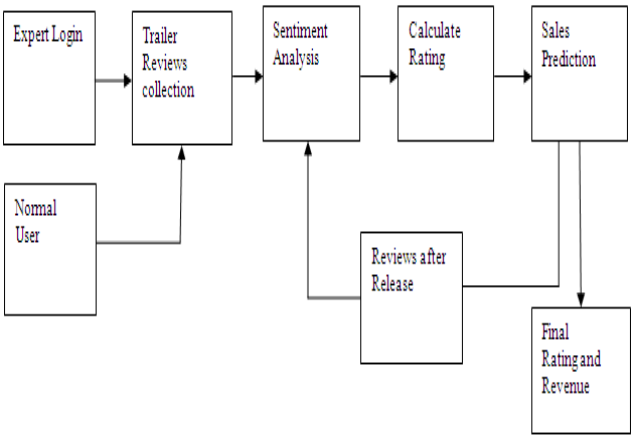
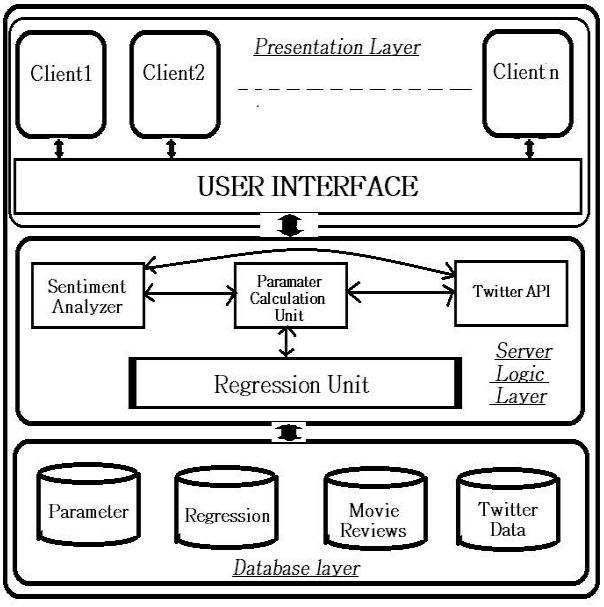


Fig 1. Data flow of proposed system



SYSTEM ARCHITECTURE

Fig 2. System architecture

Linear Regression:

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.

The data can then be used to fit a linear regression model using least squares. The parameters of the model include:

• A : rate of attention seeking
• P : polarity of sentiments and reviews
• D : distribution parameter

Let y denote the revenue to be predicted and ϱ the error. The linear regression model can be expressed as :

$$y = \beta a * A + \beta p * P + \beta d * D + \varrho \quad (4)$$

where the β values correspond to the regression coefficients. The attention parameter captures the buzz around the product in social media.

Sentiment analyzer:

Natural Language Processing (NLP) is a technique that facilitates easy pre-processing of input text. Pre-processing refers to the cleaning and normalization of text to make sentiment analysis.

1. Words: Removal of stop words such as a, an, the, this which are very common and do not determine the sentiment of the text is carried out.
2. Punctuation: Punctuation marks such as commas and periods must be removed from the input text. 3. Duplicate Words: Duplicate words deviate the overall sentiment of the text. Duplicate words must be removed to make the input text for sentiment analysis.
4. Repeated Characters: Repeated characters in words such as “looooonnng” deviate the meaning of the original word. Thus words with repeated characters must be brought to their normal form.
5. Internet Acronyms and Emoticons: The use of emoticons and acronyms on the Internet such as ASAP, AFAIK prove a major problem while analyzing sentiment of the input text. A dictionary of common acronyms is maintained and cross-checked with the input text. The acronyms are expanded to their intended format.
6. URLs: Twitter API provides all the URLs present in the Tweet. The URLs do not change the sentiment of the input Tweet and thus must be removed from the Tweet. After treating the input Tweet with the above methods the input Tweet becomes ready for analysis. Only the required words which will make a difference to the sentiment are considered.

Movie review:

We have focused on movies in this study for two main reasons.

• The topic of movies is of considerable interest among the social media user community, characterized both by large number of users discussing movies, as well as a substantial variance in their opinions.

• The real-world outcomes can be easily observed from box-office revenue for movies.

IV. CONCLUSION

In this article, we have shown how social media can be utilized to forecast future outcomes. Specifically, using the rate of chatter from almost 3 million tweets from the popular site Twitter, we constructed a linear regression model for predicting box-office revenues of movies in advance of their release. We also analysed the sentiments present in tweets and demonstrated their efficacy at improving predictions after a movie has released.

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 (WIIAT), 2010 IEEE/WIC/ACM InternationalConference on. Vol. 1. IEEE, 2010.

