Box Office Revenue Prediction Using Dual Sentiment Analysis

Prashant Rajput, Priyanka Sapkal, and Shefali Sinha

***Abstract*—Twitter is amongst the most widely used social networking website and it is also a reliable source of mass opinion. Success of a movie can be predicted by analyzing tweets and examining the impact of movie on the mob. Pre-release buzz can also be captured through tweets. This knowledge helps in predicting the success of a movie and its approximate revenue. In this paper, Dual Sentiment Analysis (DSA) is used for sentiment analysis of tweets that avoids sentiment classification problems and improves performance. Along with sentiment analysis of tweets, contribution of other factors such as star cast, holiday effect, sequel and genre are also considered. Finally, multivariate linear regression is performed on all above-mentioned factors to predict the Box Office revenue of a movie. The results show that this proposed system performs better while providing better accuracy.**

***Index Terms*—Natural language processing, sentiment analysis, opinion mining, machine learning, social media.**

I. Introduction

From the past few years, there has been an increase in the demand for prediction applications in fields such as sports, entertainment and politics. Such predictions can be done using Social media knowledge. Social networking websites are used by many people, of almost all age groups across the globe. They make it easier to connect and share views with other people. Mass opinion from such websites can be used to predict success of movies. In this paper, Twitter data is used to predict how a movie performs post its release. But processing Twitter data is difficult because of its ungrammatical structure. Bag-of-words is typically used for text classification, but its performance is limited. Hence, Dual Sentiment Analysis (DSA) can be performed on tweets extracted from twitter and the results of which can be used to determine whether the people are in favor of or against a movie. This helps in predicting the Box Office Revenue of upcoming movies. DSA model considers both the original and reverse of a tweet [1]. DSA considers the positivity as well as negativity of both the original and reverse tweet. Dual prediction helps in predicting the final sentiment of the tweet. Using this, polarity ratio is calculated. Along with polarity ratio, other factors such as the holiday effect, star cast, sequel and genre of the movie are also considered while predicting Box Office Revenue. These factors play a key role in predicting the revenue since a movie releasing on a holiday, having good star cast and the genre is particularly favored by audience at the Box Office.

Manuscript received March 26, 2017; revised August 10, 2017.

Prashant Rajput is with Computer Science Department, University of California, Los Angeles (UCLA), United States (e-mail: prashanthrajput@ucla.edu).

Priyanka Sapkal is with Persistent Systems, India (e-mail: priyanka_sapkal@persistent.com).

Shefali Sinha is with State Bank of India (SBI), India (e-mail: shefali.sinha@sbi.co.in).

II. Related Work

A. Sentiment Analysis

Significant amount of work is done in the field of data mining to extract the polarity of texts. However, they follow a very straightforward technique of sentiment analysis which may sometimes lead to wrong interpretation of sentiments. One of the traditional methods for sentiment classification is Bag-Of-Words (BOW) model. The BOW model is typically used for text representation. Here, a text (sentence) is represented as a bag of its words, disregarding grammar and even word order but keeping multiplicity. The statistical machine learning algorithms (such as Naive Bayes, maximum entropy classifier, and support vector machines) are then employed to train a sentiment classifier [2]. The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier. It does not take into consideration the order of words or semantics. According to the levels of granularity, tasks in sentiment analysis can be divided into four categorizations: document-level, sentence-level, phrase-level, and aspect-level sentiment analysis. Focusing on sentence level, there two methods namely term-counting method and machine learning method. But, sometimes two sentences with opposite sentiment are also considered same in BOW. This results in a problem known as 'Polarity Shift'. Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. There are sentiment analysis methods that have tried to avoid polarity shift problem. One of them consists of directly reversing the sentiment of polarity-shifted words, and then adding up the sentiment score, word by word [3]-[6]. There are proposed models that perform negation by searching for specific part-of-speech tag patterns [7]. Another way is to add 'not' to words, in the scope of negative word. So, in the text, I don't like this hero, the word 'like' is considered as an unfamiliar word similar to 'not' [8]. But such methods have poor accuracy of sentiment analysis. For example, there are proposed models that perform negation by searching for specific part-of-speech tag patterns or use syntactic parsing to capture three types of valence shifters (negative, intensifiers, and diminishers). Their results showed that handling polarity shift improves the performance of term-counting systems significantly, but the

doi: 10.18178/ijmlc.2017.7.4.623

improvements upon the baselines of machine learning systems are very slight.

B. Prediction Applications

At present, there are applications that already apply various techniques to predict box office revenues of upcoming movies. One such method is to analyze the activity on Wikipedia pages. They consider the number of page views, number of page edits and number of editors contributing to the article. Also, there are methods that consider the search volume i.e. the number of searches for a movie using which the success is predicted. This information can be insufficient to predict Box Office Revenue. Another method is to mine through the screenplays. This method uses sentiment analysis and it needs knowledge about good and bad dialogues/comments as well, which form the main characters. Moreover, there is one more such method that predicts movie revenue based on the frequency of tweets [9]. The above systems do not consider additional information such as genre, holiday effect, sequel and star cast.

III. Method

The proposed method consists of three important components namely tweet processing, dual sentiment analysis and multivariate linear regression. Regression model is used to generate output based on six factors. Data of sample movies (training data) is used to calculate the regression coefficients. These coefficients are then used to calculate revenue for upcoming movies based on the following equation.

$$R = \alpha_1 P + \alpha_2 H + \alpha_3 A + \alpha_4 B + \alpha_5 C + \alpha_6 G + \alpha_6 S + \varepsilon$$

where, P indicates Polarity Ratio, H indicates Hype, A indicates Actor, B indicates Actress, C indicates Holiday effect, G indicates Genre, S indicates Sequel, ε indicates Error factor. Each of the mentioned phases is discussed in following sections. Fig. 1 illustrates the overall process.

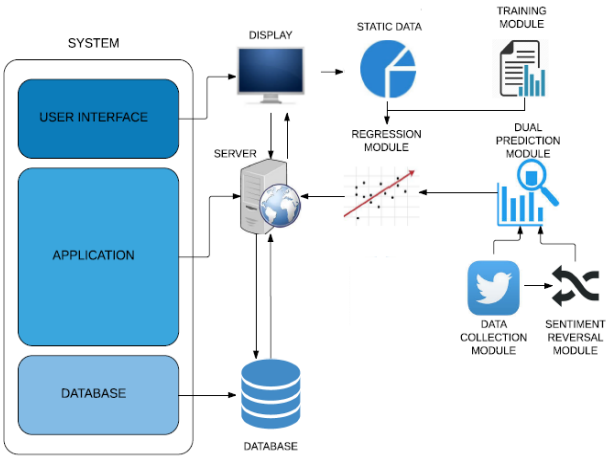

Fig. 1. Process of predicting Box Office revenue.

A. Processing Tweets

To examine the impact of a movie on public, tweets only concerning that movie are filtered and captured using Twitter streaming API. Tweets obtained are not always in the correct form and sometimes do not follow grammatical structure. Such an input to the sentiment analyzer will produce incorrect results. Hence, processing of tweets is of utmost importance. This includes the following steps.

1) Eliminating URLs, short forms and special symbols. For example,

Original Tweet: #FawadKhan's another killing look for #KapoorAndSons promotions. https://t.co/s5sMSqNiBu

Cleaned Tweet: FawadKhan's another killing look for KapoorAndSons promotions.

2) Correcting spellings and removing unnecessary reiterating letters in a word. For example,

Original Tweet: OH MY GODDDDDD..!! #aalia looks greaaaat in #KapoorAndSons

Cleaned Tweet: OH MY GOD..!! aalia looks great in KapoorAndSons

Stemming, to reduce the inflected or derived words to their root word. For example,

Original Tweet: The rocking #KapoorAndSons trio! They are cutest and craziest people https://t.co/3Q49mdPwE2

Cleaned tweet: The rocking KapoorAndSons trio! They are cute and crazy people

The steps mentioned above produce tweets with correct grammatical structure. This enables the sentiment analyzer to correctly interpret tweets and produces better output. It is also beneficial for correctly identifying parts of speech.

B. Dual Sentiment Analysis

In this stage, the processed tweets are analyzed for being positive or negative to calculate the overall polarity of tweets. Dual sentiment analysis (DSA) is used for this purpose. It consists of three important stages such as Sentiment analysis, Reverse sentiment analysis and Dual prediction.

1) Sentiment analysis

Sentiment analysis of tweet is performed using sentiment analyzer (API). The output of the analysis categorizes the tweet for being positive or negative and moreover giving it a measure for sentiment confidence. Sentiment confidence simply indicates how positive or negative the tweet is. The sentiment score is further used to calculate the final sentiment of the tweet.

2) Reverse sentiment analysis

DSA considers not only how positive/negative the original tweet is, but also considers how negative/positive the reversed tweet is. This helps in addressing the polarity shift problem. Hence, the original tweet is reversed. In order to reverse the original tweet, parts of speech of the tweet are identified. Antonyms of the adjectives are obtained from lexical database such as WordNet. Sentiment reversal consists of the following steps.

- Each tweet is checked for containing a negation word like ‘don’t’ or ‘not’.
- The negation words are removed and the words immediately following them are kept as unaltered.
- Remaining words are also kept unaltered.

The steps mentioned above produce reverse tweet. Sentiment of a reverse tweet is also obtained using the same sentiment analyzer which provides its sentiment and confidence.

3) Final sentiment analysis

Considering both, original as well as reverse sentiment of a tweet, its final sentiment is calculated. Positivity of the original tweet is considered with the negativity of the reverse tweet for calculating the final positivity of a tweet. Similarly, the negativity of original tweet is considered with the positivity of the reverse tweet for determining the negativity of the tweet. Let $p(+|x)$ be the probability of the tweet x being positive and $p(-|\bar{x})$ be the probability of the reverse of tweet x being negative. Similarly, let $p(-|x)$ be the probability of the tweet x being negative and $p(+|\bar{x})$ be the probability of the reverse of tweet x being positive.

$$p(+|x, \bar{x}) = (1-\propto).\ p(+|x) + \alpha\ .\ p(-|\bar{x})$$

$$p(-|x, \bar{x}) = (1-\propto).\ p(-|x) + \alpha\ .\ p(+|\bar{x})$$

C. Multivariate Linear Regression

To carry out multivariate linear regression, seven independent variables are considered. The objective of the regression model is to generate such coefficients which fit the known data set into an equation with minimum error. Training phase uses data set of past movies including their gross revenue. The general format of a multivariate linear regression equation is,

$$Y = b_1X_1, +\ldots\ldots + b_kX_k, + e$$

where, b_k is the coefficient for the kth factor X_k and e is the error factor. When the gross revenue for a new movie is to be calculated, the coefficient of a factor is multiplied with the value of the factor of that movie. Finally, all the products are summed up to calculate approximate revenue.

There are 7 factors considered while predicting the revenue as mentioned below:

1) Polarity ratio

Polarity Ratio is calculated by the formula

$$\text{Polarity} = \text{Pos}/\text{Neg}$$

where,

Pos = Total number of positive tweets,

Neg = Total number of negative tweets. It conveys the overall sentiment for that movie.

2) Hype

Hype measures the reach of a movie among the masses. Success of a movie also depends on the promotional activities. A movie that is well talked about has more chances of being successful. Hype is used to measure the same. The following formula has been used to calculate hype:

$$\text{Hype} = (\text{No. of distinct users}/\ \text{total no. of users who tweeted}) + \text{Rate of Tweets}$$

Each tweet has user-name associated with it. Number of distinct users can be calculated by counting distinct user-names. To calculate rate of tweets, the number of tweets collected per hour is considered.

3) Actor

There are cases when a movie does well only because of its strong star-cast. Therefore, it is important to consider the lead characters in predicting the success of a movie. For this, the follower count of the actors on twitter is used. Based on the number of followers, there are five categories and these categories are assigned weights accordingly. Above 5 million: 1, 1 - 5 million: 0.8, 1 lakh - 1 million: 0.6, 10K - 1 lakh: 0.4 and Below 10K: 0.2.

a) Holiday effect

If a movie releases on a holiday, it is likely to collect more revenue. A dummy variable is assigned value 1 if movie release day is a holiday, else 0.

b) Genre

Per the IMDB ratings, movies are categorized based on their genre. Here also 6 genres are considered. 6 dummy variables are assigned weights for each of the genres. Drama/Sci-Fi: 1. Romantic: 0.8, Comedy: 0.64, Thriller: 0.48, Action: 0.32 and Biography/Sport: 0.16.

c) Sequel

For sequel, a dummy variable same as holiday effect is used. It is assigned value 1 if the movie is a sequel, else 0.

IV. Implementation

A. Data Collection Module

In this module, Administrator has the control to start Data collection module through user interface. This runs for a fixed period and before terminating, it stores the data in database.

1) While time elapsed is greater than given certain amount of time.
2) Initialize the Twitter connection.
3) Receive Tweets.
4) Initialize Database connection.
5) Store tweet text, user name, location, created at in database.
6) Close database connection.

B. Data Processing Module

This module is used to Clean the data(tweets), perform Sentiment Analysis and Reverse Sentiment Analysis. Also, it is triggered by user action through User Interface.

1) Initialize Database Connection.
2) Fetch Tweets from database.
3) Clean the data.
4) Calculate Sentiment and Confidence of Original text.
5) Reverse the Original Text.
6) Calculate Sentiment and Confidence of Reverse Text.
7) Calculate the Dual Prediction Score.
8) Calculate Polarity Ratio.
9) Calculate Hype.
10) Store the above results in the database.

C. Regression Module

This module is used to calculate final revenue of the movie. Moreover, this module is called after processing all tweets stored in the database and gets updated every time new tweets are added in the database.

1) Initialize Database Connection.
2) Fetch training regression coefficients from database.
3) Fetch movie data from database.
4) Calculate movie's regression coefficients.
5) Calculate revenue.

6) Store the above result in the database.

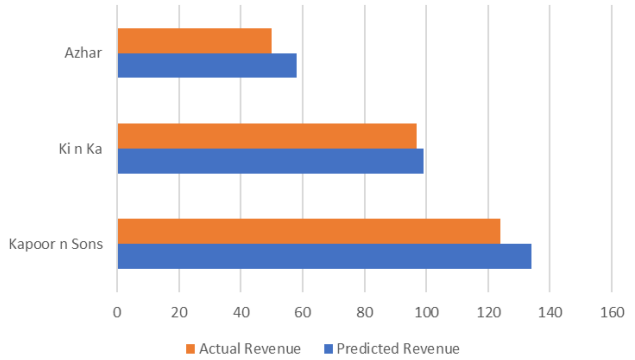

Fig. 2. Graph comparing predicted and actual revenue. X-axis represents a movie, Y-axis represents revenue in crore(INR).

TABLE I: MOVIES AND THEIR ASSOCIATED FACTORS

		Kapoor n Sons	Ki n Ka	Azhar
1	Actor	0.6	0.6	0.6
2	Actress	1	1	0.8
3	Genre	1	0.8	0.16
4	Holiday	0	0	0
5	Hype	0.85	0.5	0.6
6	Polarity Ratio	10	7	5
7	Predicted Revenue	134	99	58
8	Sequel	0	0	0

V. RESULTS

Table I shows the numerous factors contributing in prediction of revenue. The movie 'Kapoor n sons' had polarity ratio 10. It shows that the number of positive tweets about this movie is ten folds that of negative tweets, which indicates positive response from the audience. Hence, it had greater chances of success. Also, it had a strong star cast and a high rated genre. Considering all these factors, regression is carried out and total revenue of 134 Cr was predicted by the system. The movie's actual Box Office collection is 124 Cr as shown in the graph.

'Ki and Ka' had a considerably good polarity ratio, star-cast, average hype and no holiday effect. It's predicted revenue is 99 Cr and actual collection is approximately 97 Cr. 'Azhar' was a sports autobiography. Its predicted revenue is 58 Cr. The gross revenue of the movie till date is 50 Cr. Fig. 2 illustrates the same.

From the above results, the proposed method has 85% - 90% accuracy. The results prove that the proposed method has the potential to outperform existing methods. Accuracy of this method improves as more training datasets are provided.

VI. CONCLUSION

This paper puts forward a system that is efficient in predicting revenues of upcoming movies. DSA overcomes the drawbacks of traditional systems by addressing the polarity shift problem. The accuracy of system is increased by considering other factors such as sequel, genre, star-cast and holiday effect. Moreover, this system can also be used to predict election results, success of consumer product before its launch and can be incorporated in various other domains.

REFERENCES

[1] R. Xia, F. Xu, C. Q. Zong, Q. M. Li, Y. Qi, and T. Li, "Dual sentiment analysis: considering two sides of one review".
[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Language Process*, 2002, pp. 79-86.
[3] Y. Choi and C. Cardie, "Learning with compositional semantics as structural inference for subsentential sentiment analysis," *in Proc. Conf. Empirical Methods Natural Language Process*, 2008, pp. 793-801.
[4] S. Kim and E. Hovy, "Determining the sentiment of opinions," *in Proc. Int. Conf. Comput. Linguistic*, 2004, pp. 1367-1373.
[5] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters".
[6] L. Polanyi and A. Zaenen, "Contextual lexical valence shifters," in *Proc. AAAI Spring Symp. Exploring Attitude Affect Text*, 2004, pp. 1-10.
[7] J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou, "Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews," in *Proc. Conf. Int. Soc. Knowl. Org*., 2004, pp. 49-54.
[8] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards".
[9] P. T. Barthelemy, Devin, and C. Mandal, *Using Twitter Data to Predict Box Office Revenues*.

Prashant Rajput received his bachelor's degree in computer engineering from University of Pune, India. He is pursuing his Master's degree in computer science from University of California, Los Angeles, USA.

His research interest includes security and machine learning

Priyanka Sapkal received her bachelor's degree in computer engineering from University of Pune, India. She is working as a software engineer in persistent systems, India.

Her research interest lies in the field of machine learning, data mining and big data analysis.

Shefali Sinha received her bachelor's degree in computer engineering from University of Pune, India. She is working as an assistant system manager at State Bank of India, India.

Her research interest lies in the field of machine learning and data mining.