MOVIE SUCCESS FORECAST AND REVENUE PREDICTION USING FUZZY SYSTEM

Rajat Singh[1], Pushpa Singh[2], Sriram Yadav[3]

[1]Rajat Singh, CSE, RGPV, Bhopal, M.P, (India)
[2]Prof. Pushpasingh, CSE, RGPV, Bhopal, M.P, (India)
[3]Prof. Sriramyadav, CSE, RGPV, Bhopal ,M.P, (India)

ABSTRACT

In recent years, social media has become omnipresent and essential for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped by official authorities. The present tutorial investigates techniques for social media modeling, analytics and optimization. First we present methods for collecting large scale social media data and then discuss techniques for coping with and correcting for the effects arising from missing and incomplete data. In this paper, we show how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based calculations. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media. We summarize an effort to predict box office revenues using Twitter data. It is hypothesized that an increased number of tweets about a movie before its release will result in increased box office revenue.

I. INTRODUCTION

This online social media websites have become a new trend in providing vital information regarding the success of an particular media. As far as our work is concerned it is related to movie performance in the box office. There are many aspects and approaches to this; one can be the use of fuzzy systems. Every day people on the internet create content and expose it to various other online users. This content is basically the views and comments of individual users through social media on their behalf. Now a day’s the entertainment industry intensely depends on the contents and comments created by users on social networking sites. We can predict the real world outcomes by just collecting the online contents from social networking sites. It has been found that if such systems are properly designed, they are more effective in predicting the outcome as compared to surveys and other online polls. Here we need not involve any market mechanisms for particular predictions. Also the future trends of a collective population can be predicted. In this paper we have considered the task of predicting the movie status on the box office using the tweets from the twitter. Twitter being the first choice of online users to express their views on any currently trending topics.it is one of the micro-blogging sites which has gained enormous popularity in recent times with huge user base, nearly about 25 million. These users regularly contribute to the creation and circulation of online content in form of their own views. Twitter was launched on July 13, 2006. It is an extremely popular online micro-blogging site. The maximum size limit of a message is about 140 characters. Each user has his/her own followers as well. A retweet can be considered as a forwarded post by another user. Retweet increases the impact of the post by further propagation.

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.

Our goals in this paper are as follows. First, we assess how buzz and attention is created for different movies and how that changes over time. Movie producers spend a lot of effort and money in publicizing their movies, and have also embraced the Twitter medium for this purpose. We then focus on the mechanism of viral marketing and pre-release hype on Twitter, and the role that attention plays in forecasting real-world box-office performance. Our hypothesis is that movies that are well talked about will be well-watched. After all sentiment analysis is done to know positive, negative or neutral feedback. This is done by storing keywords in the database.

II. RELATED WORK

There has been some prior work on analysing the correlation between blog and review mentions and performance. Gruhl and others [4] showed how to generate automated queries for mining blogs in order to predict spikes in book sales. And while there has been research on predicting movie sales, almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, MPAA rating, running time, release date, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. Joshi and others [5] use linear regression from text and metadata features to predict earnings for movies. Mishne and Glance [10] correlate sentiments in blog posts with movie box-office scores. The correlations they observed for positive sentiments are fairly low and not sufficient to use for predictive purposes. Sharda and Delen [3] have treated the prediction problem as a classification problem and used neural networks to classify movies into categories ranging from ’flop’ to ’blockbuster’. Apart from the fact that they are predicting ranges over actual numbers, the best accuracy that their model can achieve is fairly low. Zhang and Skiena [1] have used a news aggregation model along with IMDB data to predict movie box-office numbers. We have shown how our model can generate better results when compared to their method.
Movie domain is considered as the input, because information related to the movie and revenue information, are easily available. The movie information which is used for conducting experiments includes two factors.

1. The first factor is a set of blog reviews on movies which are of interest, collected from the web.
2. Second factor contains the corresponding revenue data for these movies.

III. FLOW FOR THE PROPOSED MODEL

Flow of the proposed model is as shown in the figure 1, i.e

1. We are collecting tweets of newly released movies from twitter in java using twitter.4j API. From the movies hashtag we are extracting tweets of different movies from twitter.
2. Selecting hashtag of particular movie from dropdown menu and then searching for the respective movie tweets, it will display tweets of that movie on display page as well as store it into sql database.

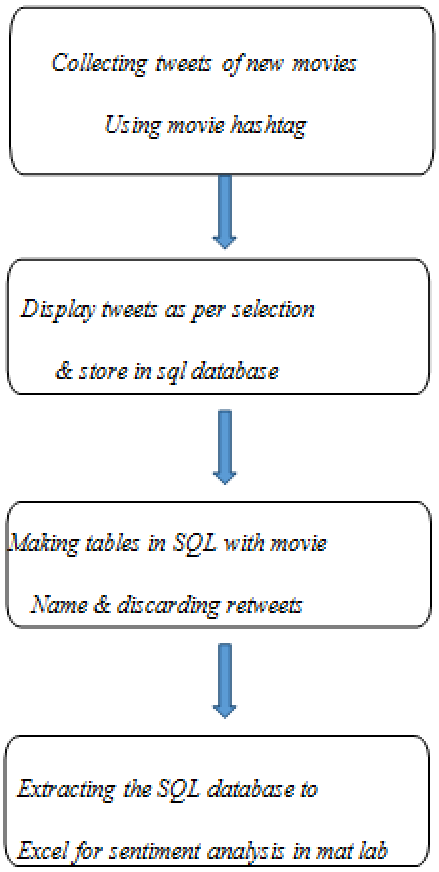
Collecting tweets of new movies
Using movie hashtag
Display tweets as per selection
& store in sql database
Making tables in SQL with movie
Name & discarding retweets
Extracting the SQL database to
Excel for sentiment analysis in mat lab

Fig.1 Flow of proposed system

3. Selecting hashtag of particular movie from dropdown menu and then searching for the respective movie tweets, it will display tweets of that movie on display page as well as store it into SQL database.
4. We are discarding retweets and storing only tweets into different tables named as respective movie_name in SQL database.
5. From SQL database we are extracting the SQL database file into excel sheet format for doing sentiment analysis in matlab.

Many existing models and algorithms for sentiment mining are developed for the binary classification problem, i.e., to classify the sentiment of a review as positive or negative. However, sentiments are often multi-faceted, and can differ from one another in a variety of ways, including polarity, orientation, graduation, etc. Therefore, for applications it is necessary to understand the opinions accurately. Here extraction of ratings starts with modelling sentiments in online reviews, which presents unique challenges that is not possible to be easily addressed by conventional text mining methods by classifying reviews as positive or negative, as most current sentiment mining approaches are designed for, does not provide a comprehensive understanding of the sentiments reflected in blog reviews [2].To organize the model of a variety of natures of complicated sentiments, sentiments are analysed which is embedded in reviews as a result of the combined role of a number of hidden factors. Since social media can also be construed as a form of collective wisdom, we decided to investigate its power at predicting real-world outcomes. Surprisingly, we discovered that the chatter of a community can indeed be used to make quantitative predictions that outperform those of artificial markets. These information markets generally involve the trading of state-contingent securities, and if large enough and properly designed, they are usually more accurate than other techniques for extracting diffuse information, such as surveys and opinions polls. Specifically, the prices in these markets have been shown to have strong correlations with observed outcome frequencies, and thus are good indicators of future outcomes [4], [5].

IV. DATASET ATTRIBUTES

The dataset that we used was obtained by crawling hourly feed data from Twitter.com. To ensure that we obtained all tweets referring to a movie, we used keywords present in the movie title as search arguments. We extracted tweets over frequent intervals using the Twitter Search Api , thereby ensuring we had the timestamp, author and tweet text for our analysis. We extracted around 1.2 million tweets referring to 9 different movies released over a period of two months.

Movie Name	Release Dates
The shaukeens	07-11-14
Kill Dill	14-11-14
Happy Ending	21-11-14
Ungli	28-11-14
Zid	28-11-14
Action Jackson	05-12-14
Lingaa	12-12-14
P.K	19-12-14
Tevar	09-01-15
Alone	16-01-15
Crazy cukkad Family	16-01-15
Sharafatgayitellene	16-01-15
Baby	23-01-15
Jai jawaan jai kisaan	23-01-15
Khamoshiyan	23-01-15
Dolly kidoli	05-02-15
Rocky handsome	05-02-15
Shamitabh	05-02-15
Detective byomkeshbakshi	13-02-15
Patel ki Punjabi shaadi	13-02-15
Roy	13-02-15
Badlapur	20-02-15
Meeruthiya gangsters	20-02-15
Guddurangeela	27-02-15

Table 1 Names & Release Dates of Movies

Movies are normally released on Fridays, with the exception of a few which are released on Wednesday. Since averages of 2 new movies are released each week, we collected data over a time period of 3 months from November to February to have sufficient data to measure predictive behaviour. For consistency, we only considered the movies released on a Friday and only those in wide release. For movies that were initially in limited release, we began collecting data from the time it became wide. For each movie, we define the critical period as the time from the week before it is released, when the promotional campaigns are in full swing, to two weeks after release, when its initial popularity fades and opinions from people have been disseminated. Some details on the movies chosen and their release dates are provided in Table 1. Note that, some movies that were released during the period considered were not used in this study, simply because it was difficult to correctly identify tweets that were relevant to those movies.

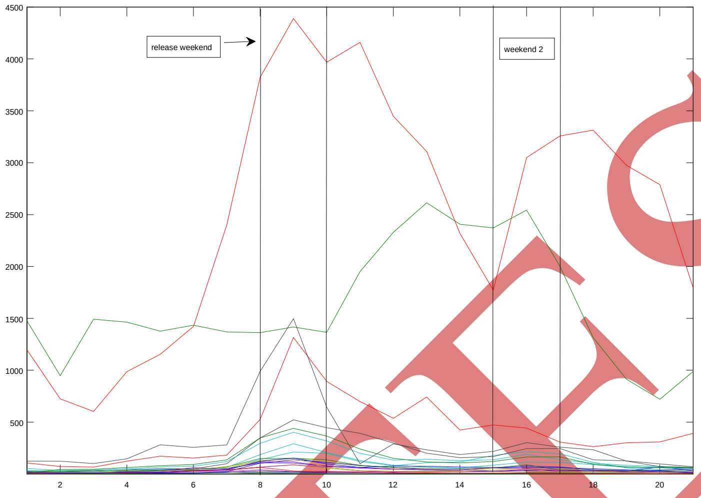

Fig.2 Time-Series of Tweets over the Critical Period for Different Movies

The total data over the critical period for the 24 movies we considered includes 2.89 million tweets from 1.2 million users.

Fig 2 shows the time series trend in the number of tweets for movies over the critical period. We can observe that the busiest time for a movie is around the time it is released, following which the chatter invariably fades. The box-office revenue follows a similar trend with the opening weekend generally providing the most revenue for a movie.

Fig 3 displays the distribution of tweets by different authors over the critical period. The X-axis shows the number of tweets in the log scale, while the Y-axis represents the corresponding frequency of authors in the log scale. We can observe that it is close to a Zipfian distribution, with a few authors generating a large number of tweets. This is consistent with observed behaviour from other networks [7]. Next, we examine the distribution of authors over different movies. Fig 4 shows the distribution of authors and the number of movies they comment on. Once again we find a power-law curve, with a majority of the authors talking about only a few movies.

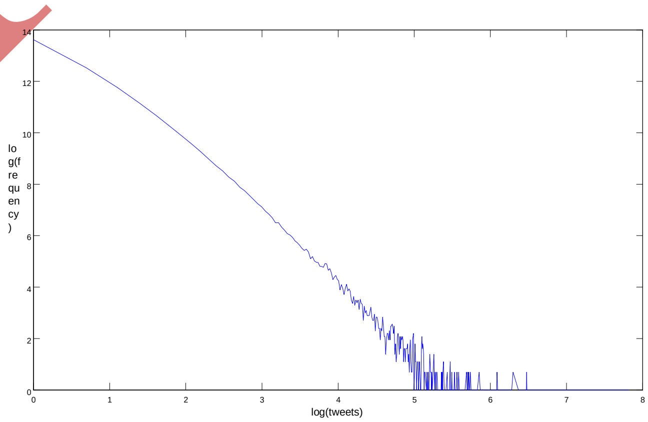

Fig.3 Log Distribution of Authors and Tweets

V. SENTIMENT ANALYSIS

Next, we would like to investigate the importance of sentiments in predicting future outcomes. We have seen how efficient the attention can be in predicting opening weekend box-office values for movies. Hence we consider the problem of utilizing the sentiments prevalent in the discussion for forecasting.

Sentiment analysis is a well-studied problem in linguistics and machine learning, with different classifiers and language models employed in earlier work [8], [9]. It is common to express this as a classification problem where a given text needs to be labeled as Positive, Negative or Neutral. Here, we constructed a sentiment analysis classifier using the Ling Pipe linguistic analysis package which provides a set of open-source java libraries for natural language processing tasks. We used the Dynamic LM Classifier which is a language model classifier that accepts training events of categorized character sequences. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators. To obtain labeled training data for the classifier, we utilized workers from the Amazon Mechanical Turk 3. It has been shown that manual labeling from Amazon Turk can correlate well with experts [6]. We used thousands of workers to assign sentiments for a large random sample of tweets, ensuring that each tweet was labeled by three different people. We used only samples for which the vote was unanimous as training data. The samples were initially preprocessed in the following ways:

- Elimination of stop-words
- Elimination of all special characters except exclamation marks which were replaced by $< EX >$and question marks ($< QM >$)
- Removal of URL's and user-ids
- Replacing the movie title with $< MOV >$

We used the pre-processed samples to train the classifier using an n-gram model. We chose n to be 8 in our experiments. The classifier was trained to predict three classes - Positive, Negative and Neutral. When we tested on the training-set with cross-validation, we obtained an accuracy of 98%. We then used the trained classifier to predict the sentiments for all the tweets in the critical period for all the movies considered.

5.1 Subjectivity

Our expectation is that there would be more value for sentiments after the movie has released, than before. We expect tweets prior to the release to be mostly anticipatory and stronger positive/negative tweets to be disseminated later following the release. Positive sentiments following the release can be considered as recommendations by people who have seen the movie, and are likely to influence others from watching the same movie. To capture the subjectivity, we defined a measure as follows.

$$Subjectivity = \frac{|Positive\ and\ Negative\ Tweets|}{|Neutral\ Tweets|} \qquad (2)$$

When we computed the subjectivity values for all the movies, we observed that our hypothesis was true. There were more sentiments discovered in tweets for the weeks after release, than in the pre-release week. Fig 7 shows the ratio of subjective to objective tweets for all the movies over the three weeks. We can observe that for most of the movies, the subjectivity increases after release.

5.2 Polarity

To quantify the sentiments for a movie, we measured the ratio of positive to negative tweets. A movie that has far more positive than negative tweets is likely to be successful.

$$PNratio = \frac{|Tweets\ with\ Positive\ Sentiment|}{|Tweets\ with\ Negative\ Sentiment|} \quad (3)$$

Fig 8 shows the polarity values for the movies considered in the critical period. We find that there are more positive sentiments than negative in the tweets for almost all the movies. The movie with the enormous increase in positive sentiment after release is *Ungli*(5.02 to 9.65). The movie had a lukewarm opening weekend sales (34cr) but then boomed in the next week (39cr), owing largely to positive sentiment. The movie *Zid*had the opposite effect. It released in the same weekend as *Ungli*and had a great first weekend but its polarity reduced (6.29 to 5), as did its box-office revenue (25cr to 9cr) in the following week.

Considering that the polarity measure captured some variance in the revenues, we examine the utility of the sentiments in predicting box-office sales. In this case, we considered the second weekend revenue, since we have seen subjectivity increasing after release. We use linear regression on the revenue as before, using the tweet-rate and the PN ratio as an additional variable. The results of our regression experiments are shown in it. We find that the sentiments do provide improvements, although they are not as important as the rate of tweets themselves. The tweet-rate has close to the same predictive power in the second week as the first. Adding the sentiments, as an additional variable, to the regression equation improved the prediction to 0.92 while used with the average tweet-rate, and 0.94 with the tweet-rate time series. It shows the regression p-values using the average tweet rate and the sentiments. We can observe that the coefficients are highly significant in both cases.

VI. 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 then showed that the results outperformed in accuracy those of the Bollywood Stock & media Exchange and that there is a strong correlation between the amount of attention a given topic has (in this case a forthcoming movie) and its ranking in the future. We also analyzed the sentiments present in tweets and demonstrated their efficacy at improving predictions after a movie has released.

While in this study we focused on the problem of predicting box office revenues of movies for the sake of having a clear metric of comparison with other methods, this method can be extended to a large panoply of topics, ranging from the future rating of products to agenda setting and election outcomes. At a deeper level, this work shows how social media expresses a collective wisdom which, when properly tapped, can yield an extremely powerful and accurate indicator of future outcomes.

REFERENCES

[1] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. In Web Intelligence, pages 301304, 2009.

[2] T. Hofmann, “Probabilistic Latent Semantic Analysis,”Proc. Uncertainty in Artificial Intelligence (UAI), 1999.

[3] Ramesh Sharda and DursunDelen. Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, vol 30, pp 243–254, 2006.

[4] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins. The predictive power of online chatter. SIGKDD Conference on Knowledge Discovery and Data Mining, 2005

[5] Mahesh Joshi, Dipanjan Das, Kevin Gimpel and Noah A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression NAACL-HLT, 2010.

[6] Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of EMNLP, 2008.

[7] Fang Wu, Dennis Wilkinson and Bernardo A. Huberman. Feeback Loops of Attention in Peer Production. Proceedings of SocialCom-09: The 2009 International Conference on Social Computing, 2009.

[8] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis Foundations and Trends in Information Retrieval, 2(1-2), pp. 1135, 2008.

[9] NamrataGodbole, ManjunathSrinivasaiah and Steven Skiena. LargeScale Sentiment Analysis for News and Blogs. Proc. Int. Conf. Weblogs and Social Media (ICWSM), 2007.

[10] Mishne and N. Glance. Predicting movie sales from blogger sentiment. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.