Detailed Descriptive and Predictive Analytics with Twitter Based TV Ratings

[1] **Amrapali Mhaisgawali,** [2] **Dr Nupur Giri**

[1, 2] Vivekanand Institute of Technology, Chembur, Mumbai, India

Abstract- In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. This paper demonstrates about how social media content can be used for descriptive and predictive analytics. In particular, chatter from Twitter.com was used to find the contribution of mobile device usage in different cities of India giving predictive analytics.

***Keywords* – Twitter, TV Rating, Microblogging, Sentiment Analytics.**

1. Introduction

Social media has exploded as a category of online discourse where people create content, share, bookmark and network at a prodigious rate. Examples include Facebook, MySpace, Digg, Twitter and JISC listservs on the academic side. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry.

In the case of social media, the enormity and high variance of the information that propagates through large user communities presents an interesting opportunity for harnessing that data into a form that allows for specific predictions about particular outcomes, without having to institute market mechanisms. [9].

Twitter [10] receives approximately 190 million tweets. These large numbers of tweets contains opinion about products or services that users uses. The tweets considered in this paper are the tweets related to the TV which are broadcasted by different channels in India

.

2. Related Work

Although Twitter has been very popular as a web service, there has not been considerable published research on it. Huberman and others [3] studied the social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions. Java et al [1] investigated community structure and isolated different types of user intentions on Twitter. Jansen and others [4] have examined Twitter as a mechanism for word-of-mouth advertising, and considered particular brands and products while examining the structure of the postings and the change in sentiments. However the authors did not performed any analysis on the predictive aspect of Twitter.

There has been some prior work on analyzing the correlation between blog and review mentions and performance. Gruhl and others [5] 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 [10] use linear regression from text and metadata features to predict earnings for movies. Mishne and Glance [6] 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 [8] 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 [11] 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 other method.

3. Twitter

Launched on July 13, 2006, Twitter is an extremely popular online microblogging service. It has a very large user base, consisting of several millions of users. It can be considered a directed social network, where each user has a set of subscribers known as followers. Each user submits periodic status updates, known as tweets that consist of short messages of maximum size 140 characters. These updates typically consist of personal information about the users, news or links to content such as images, video and articles. The posts made by a user are displayed on the user's profile page, as well as shown to his/her followers. It is also possible to send a direct message to another user. Such messages are preceded by the recipient's screen-name indicating the intended destination. A retweet is a post originally made by one user that is forwarded by another user. Retweets are useful for propagating interesting posts and links through the Twitter community.

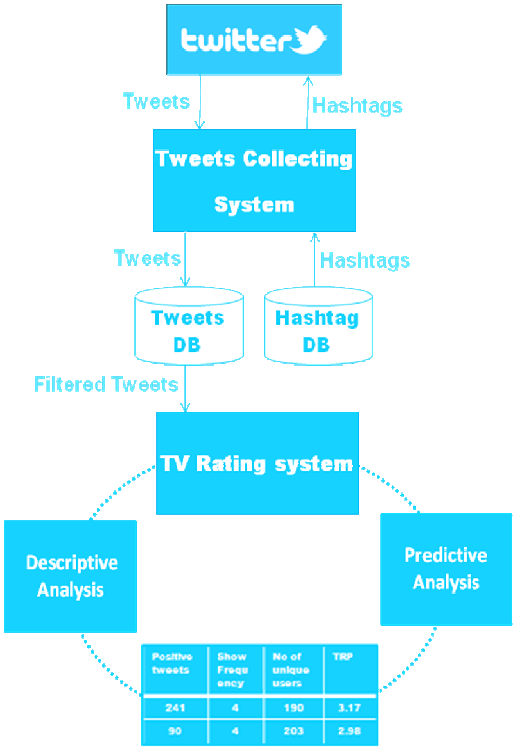
Twitter has attracted lots of attention from corporations for the immense potential it provides for viral marketing. Due to its huge reach, Twitter is increasingly used by news organizations to filter news updates through the community. A number of businesses and organizations are using Twitter or similar micro-blogging services to advertise products and disseminate information to stakeholders.

4. Dataset Characteristic

The dataset that we used was obtained by crawling a regular feed of data from Twitter.com. To ensure that we obtained all tweets referring to a TV show, we used hash tags of TV show as search arguments. Tweets were extracted over frequent intervals using the Twitter Search API 1.1, thereby ensuring the timestamp, author and tweet text for analysis. We extracted 46,304 tweets referring to 31 different TV shows over a period starting from week 11-2014 to week 20 2014.,Twitter is a micro-blogging site targeting for various real life applications. Thus, in order to use the Twitter platform as a back-channel for TV rating, we should find out tweets related TV programs. In order to look for those TV-related tweets, the hashtags which are popularly used on the site were used as an index to enable retrieval by other services or users. Generally, Twitter users can simply create a hashtag by prefixing a word with a hash symbol "#hashtag." For instance, various hashtags, such as #balikavadhu, #BALH, #CID had been created for different TV shows. PHP platform was used to collect tweets from twitter database. System collects tweets for the chosen TV shows hashtags and tweets get stored into the tweets database. The system which is described in this paper is real time system which will pull the relevant tweets four times in a day from the twitter database. A job was set for collection of tweets for the said frequency.

5. System Architecture

Fig 1 shows the block diagram of the system architecture. Hashtags are stored in the hashtag table in database. Hashtags were passed to twitter.com and responses get stored in the database, these tweets are called as raw tweets. Raw tweets then get preprocessed and stored in the filtered table. Further analytics were done on filtered tweets as well as on raw tweets.



Positive tweets	Show Frequency	No of unique users	TRP
241	4	190	3.17
90	4	203	2.98

Fig 1 System Diagram

6. Data Processing

Data processing was designed as 4 stage process consisting of Extraction, Transformation, Cleaning, and Sentiment Analysis. In the extraction phase, tweets were extracted from twitter.com. In transformation phase, extracted tweets were filtered based on the relevancy of tweets for analysis and only filtered tweets were stored into filter table. In cleaning phase filtered tweets were processed further to remove the stop words, replacing short words by actual words and replacing emotional flavor words by actual words. Stop-word are nothing but the words which are not be used for analysis. Some stop-words are 'the', 'we', 'which', 'who' etc. All words within the tweets with emotional flavor were replaced with actual words. Example: if tweets contain "sweeeeeeeeeeeeet" then it was replaced by "sweet". Short words were replaced by actual words by using the short words dictionary.
Re-tweets were not considered for calculating TV ratings as these does not necessarily involve any viewing by the user but is just forwarding the tweets to their followers. Hence all re-tweets were also removed in the transformation phase.

A. Finding Relevant Tweets

Relevant tweets are the tweets which were twitted by people who actually watch the TV show. So only those tweets were considered which contain words related to "watch" and "view". In cases where words related to "watch" and "view" were not present in the tweet another condition was added based on the actor or actress name for the TV show. Features used in this classification were usage of proper nouns of cast and crew of TV shows. If such nouns were used, the tweet would mostly be "TV-viewing". List of cast and crew of all TV shows was extracted by parsing proper nouns from web pages of site Wikipedia.

B. Sentiment Analysis

Cleaned and transformed tweets using above conditions were used for sentiment analysis. Each of the tweets was classified as positive, negative or neutral. "Datum Box Machine learning API" [12] was used for this classification. This API classifies the tweet as positive, negative or neutral. Only positive tweets were considered for finding out the TV ratings.

C. Calculate TV Ratings

Following formula [2] was used to calculate TV ratings

$$Rating = \log\left(\left(\sqrt{Tweets / Frequency}\right) \times Users\right)$$

Where
Tweets – Number of positive tweets for a particular show.
Frequency – Weekly broadcasting Frequency of a particular show.
Users – Number of unique a user who has contributed the tweets which includes positive, negative and neutral tweets.

7. Descriptive Analytics

Dataset used for descriptive analytics were considered from week 11 of 2014 to week 20 of 2014.
In this period, 46,304 tweets were collected for the relevant hashtags. After filtering the terms related to "watch" and "view", relatively small tweet dataset (20,558 tweets) was obtained. These potential tweets were written by 17,162 distinct users.

A. Advertisement Preference

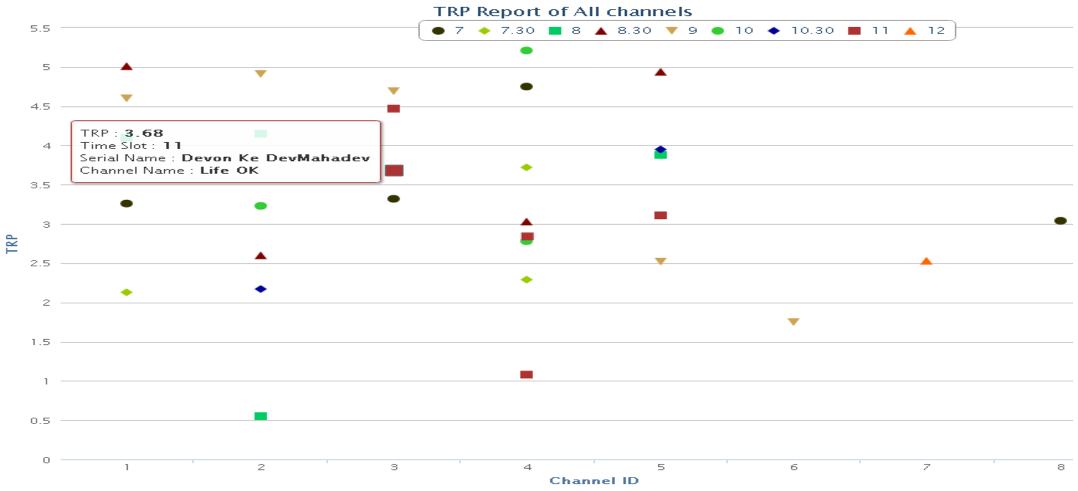


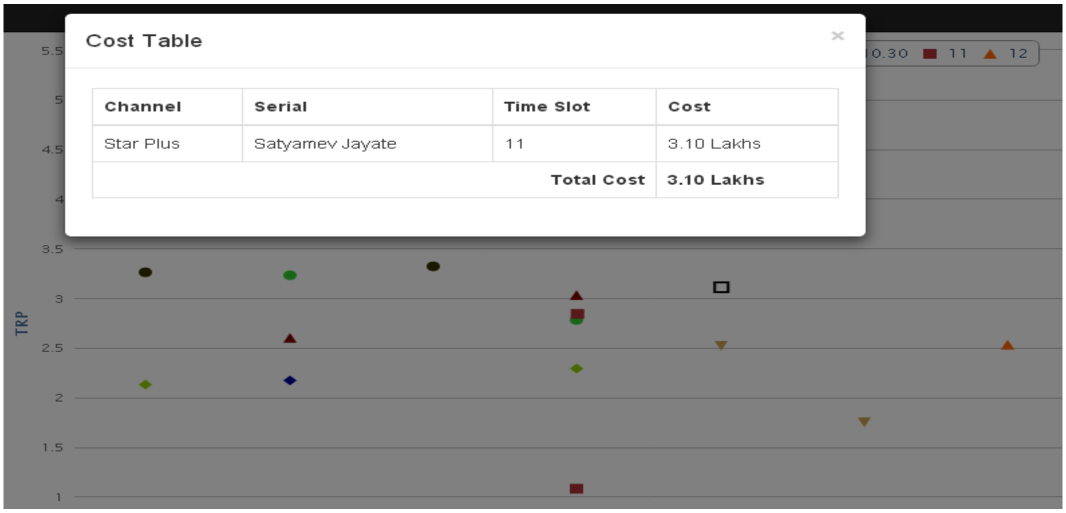
Fig 2 Advertisement Preference Graph

Advertisement preference graph is scatter plot and it shows the TRP (Television Rating Point) for different TV shows for different channels. Graphs were drawn using Highcharts [13]. Details of TRP, Time slot, Show name and channel name can be viewed by placing cursor to a point on scatter plot as shown in fig 2.

If any producer, who wants to give his advertisement in TV show, can chose particular time slot, for particular channel. Producer can chose the time slot whose TRP is high so that there will be more chances to sell his product or service.

i) Cost of advertisement for single show:

While choosing the particular time slot for particular channel, cost of advertisement per 10 second for show is most important. This application will also give the cost for particular show on a particular channel, by selecting the point on scatter plot as shown in Fig 3.
Fig 3 shows cost of TV show "Satyamev Jayate" on channel 'Star plus' and time slot is 11 p.m. Cost for this show is 3.10 lakhs. Costs shown in the cost table are considered by us for designing the application and are not the actual cost.



Cost Table

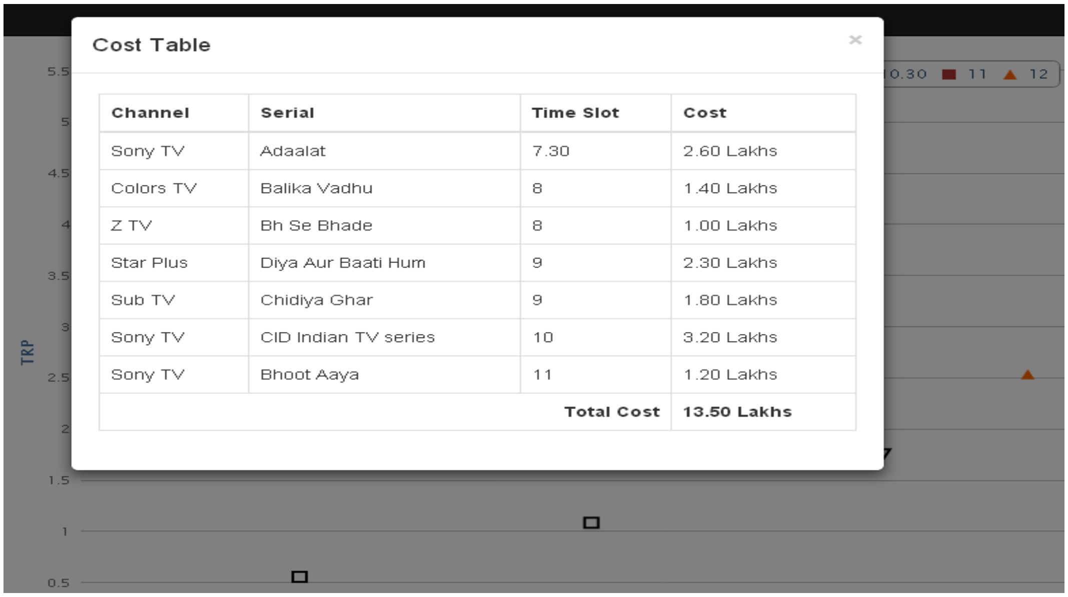
Channel	Serial	Time Slot	Cost
Star Plus	Satyamev Jayate	11	3.10 Lakhs
		Total Cost	3.10 Lakhs

Fig. 3 Cost of Advertisement for Single Show

If cost is bearable for producer then he can chose the above time slot, otherwise he can chose another time slot.

ii) Cost of advertisement for multiple shows:

If producer wants to give advertisement in multiple shows then he can chose multiple shows and application will give the individual cost as well as total cost for all selected TV shows as shown in fig 4.



Cost Table

Channel	Serial	Time Slot	Cost
Sony TV	Adaalat	7.30	2.60 Lakhs
Colors TV	Balika Vadhu	8	1.40 Lakhs
Z TV	Bh Se Bhade	8	1.00 Lakhs
Star Plus	Diya Aur Baati Hum	9	2.30 Lakhs
Sub TV	Chidiya Ghar	9	1.80 Lakhs
Sony TV	CID Indian TV series	10	3.20 Lakhs
Sony TV	Bhoot Aaya	11	1.20 Lakhs
		Total Cost	13.50 Lakhs

Fig. 4 Cost of Advertisement for multiple shows

Depends on the total cost for all selected TV shows he can select or deselect the TV shows whichever he wants.

B. Serial Type Wise Comparison

Comparison can also be done depends on serial type. TV shows can be categorized as ‘Historic’, ’Reality shows’, ‘Drama’, ‘Horror’, ‘Crime’, ‘Comedy’ etc.
Depends on these categories TV shows are distinguished as shown in fig 5.

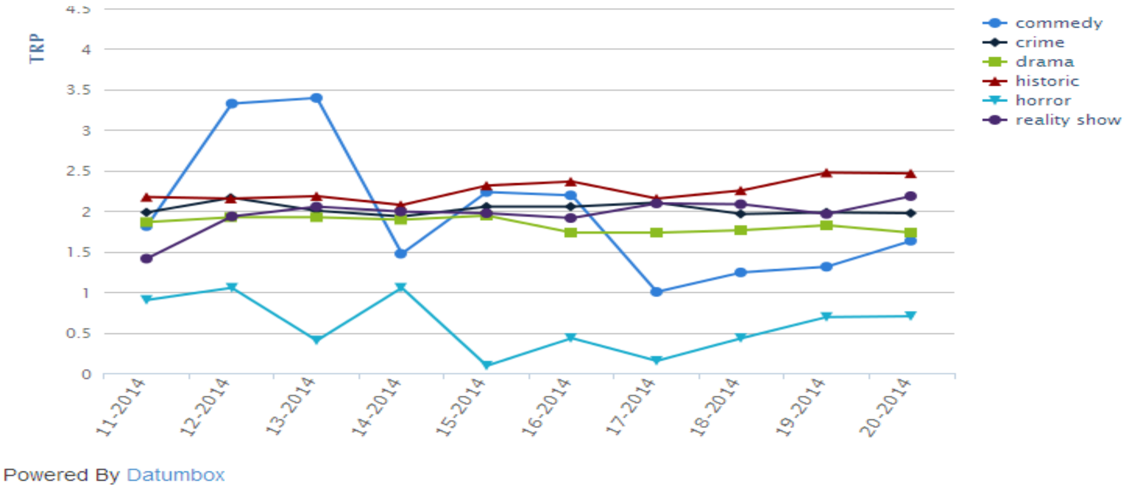


Fig. 5 Serial Type wise Comparison

All the TV shows which are coming under category drama or comedy or crime etc are grouped together and then by taking the average of all TV shows TRP were generated and plotted in particular week.

Graph shows that TRP of TV shows which were coming under category comedy fluctuate every week. It was not constant in all week. In week 12-2014 and 13-2014 TRP was highest among all category. And in week 17-2014 to 20-2014 TRP was in range from 1 to 1.5.

TRP of TV shows which were coming under category ‘Historic’, ‘Drama’, ‘Reality Show’ and ‘Crime’ was in range from 1.5 to 2.5. But in all the weeks ‘Historic’ TV shows were more popular than any other TV shows. TRP of ‘Historic’ TV shows were in range from 2 to 2.5 in all weeks. TRP of TV shows which were coming under category ‘Horror’, was in range from 0 to 1.05 in all weeks. So ‘Horror’ TV shows were not much popular.

8. Predictive Analytics

Predictive analytics can predict what will happen in future. Twitter data can be used by mobile companies for launching the new mobile device.

People used different devices for tweeting. They can use Android phone, iPhone, Samsung phone, Computer browser i.e. web etc for purpose of tweeting. Tweets extracted from twitter.com, also contains the city from where the tweets came and the source of tweets, i.e which device was used to tweet.

Based on this prediction can be done for the launching of new mobile device for a specific city

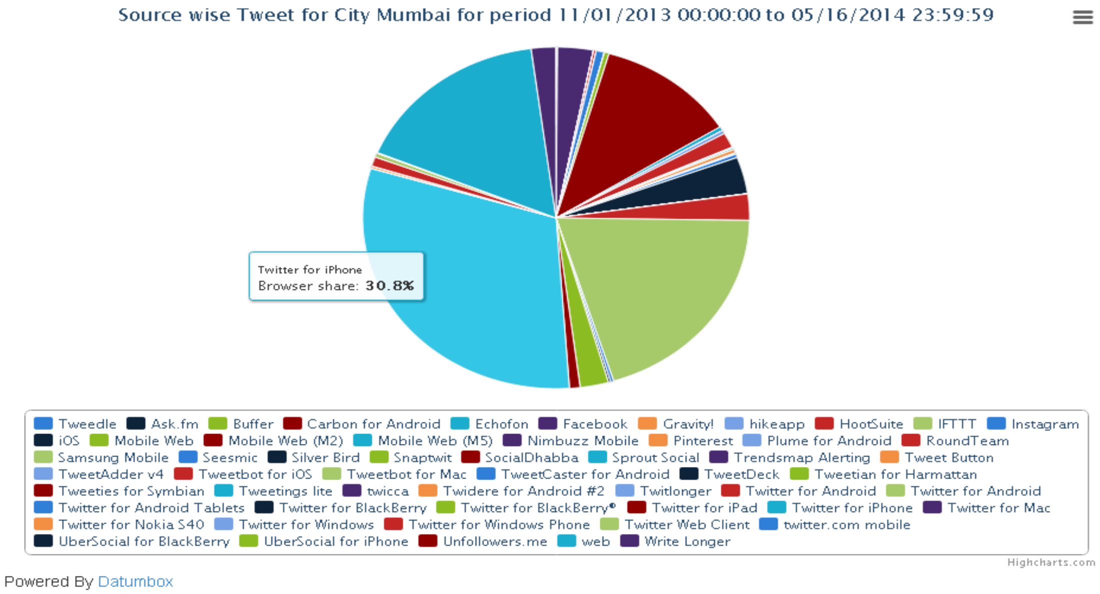


Fig. 6 Source Wise Tweets

Prediction was done on total raw tweets collected for Mumbai city for period from 01/11/2013 to 16/05/2014. Raw tweets were the tweets including re-tweets also. Fig 6 shows that 30.8 % tweets were done from iPhone. So any new mobile from iPhone can be launched in Mumbai because in this city more users are using iPhone for tweeting on twitter.

9. Conclusion

In this paper we had described the detail descriptive and predictive analytics on large amount of twitter data. We had described how advertisement slot can be selected depends upon time slot as well as cost of advertisement per 10 seconds. We also predicted that any new mobile device can be launched in which city.

Acknowledgement

We thank Mr. Lavin for his help in getting started, and for guidance along the way. We would also like to acknowledge to Mr. Vilas for completing the project.

References

[1] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65, 2007

[2] Amrapali Mhaisgawali, Dr. Nupur Giri Twitter based TV Rating Analysis with TV -viewing and Sentiment Analysis", Current Trends in Information Technology ISSN: 2249-4707 Volume 4, Issue 1 www.stmjournals.com

[3] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. First Monday, 14(1), Jan 2009.

[4] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Societyfor Information Science and Technology, 2009

[5] 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.

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

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

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

[9] Sitaram Asur, Bernardo A. Huberman "Predicting the Future With Social Media" 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

[10] Twitter: http:/twitter.com.

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

[12] http://www.datumbox.com/machine-learning-api

[13] http://www.highchart.com