2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

***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. In this paper, we demonstrate 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 predictors. We further demonstrate how sentiments extracted from Twitter can be utilized to improve the forecasting power of social media.**

I. 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.

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].

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. One can also build models to aggregate the opinions of the collective population and gain useful insights into their behavior, while predicting future trends. Moreover, gathering information on how people converse regarding particular products can be helpful when designing marketing and advertising campaigns [1], [3].

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

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.

Next, we study how sentiments are created, how positive and negative opinions propagate and how they influence people. For a bad movie, the initial reviews might be enough to discourage others from watching it, while on the other hand, it is possible for interest to be generated by positive reviews and opinions over time. For this purpose, we perform sentiment analysis on the data, using text classifiers to distinguish positively oriented tweets from negative.

Our chief conclusions are as follows:

- We show that social media feeds can be effective indicators of real-world performance.
- We discovered that the rate at which movie tweets are generated can be used to build a powerful model for predicting movie box-office revenue. Moreover our predictions are consistently better than those produced by an information market such as the Hollywood Stock Exchange, the gold standard in the industry [4].

[1] http://www.twitter.com

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.63

- Our analysis of the sentiment content in the tweets shows that they can improve box-office revenue predictions based on tweet rates only after the movies are released.

This paper is organized as follows. Next, we survey recent related work. We then provide a short introduction to Twitter and the dataset that we collected. In Section 5, we study how attention and popularity are created and how they evolve. We then discuss our study on using tweets from Twitter for predicting movie performance. In Section 6, we present our analysis on sentiments and their effects. We conclude in Section 7. We describe our prediction model in a general context in the Appendix.

II. Related Work

Although Twitter has been very popular as a web service, there has not been considerable published research on it. Huberman and others [2] 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 [7] investigated community structure and isolated different types of user intentions on Twitter. Jansen and others [3] 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 do not perform 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 [9] 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 [15] 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 [6] 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.

III. Twitter

Launched on July 13, 2006, Twitter [2] is an extremely popular online microblogging service. It has a very large user base, consisting of several millions of users (23M unique users in Jan [3]). 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.

IV. Dataset Characteristics

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 movie, we used keywords present in the movie title as search arguments. We extracted tweets over frequent intervals using the Twitter Search Api [4], thereby ensuring we had the timestamp, author and tweet text for our analysis. We extracted 2.89 million tweets referring to 24 different movies released over a period of three months.

Movies are typically released on Fridays, with the exception of a few which are released on Wednesday. Since an average 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 behavior. 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 popularity fades and opinions from people have been disseminated.

Some details on the movies chosen and their release dates are provided in Table I. 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. For instance, for the movie *2012*, it was impractical to segregate tweets talking about the movie, from those referring to the year. We have taken care to ensure that the data we have used was

[2]http://www.twitter.com

[3]http://blog.compete.com/2010/02/24/compete-ranks-top-sites-for-january-2010/

[4]http://search.twitter.com/api/

Movie	Release Date
Armored	2009-12-04
Avatar	2009-12-18
The Blind Side	2009-11-20
The Book of Eli	2010-01-15
Daybreakers	2010-01-08
Dear John	2010-02-05
Did You Hear About The Morgans	2009-12-18
Edge Of Darkness	2010-01-29
Extraordinary Measures	2010-01-22
From Paris With Love	2010-02-05
The Imaginarium of Dr Parnassus	2010-01-08
Invictus	2009-12-11
Leap Year	2010-01-08
Legion	2010-01-22
Twilight : New Moon	2009-11-20
Pirate Radio	2009-11-13
Princess And The Frog	2009-12-11
Sherlock Holmes	2009-12-25
Spy Next Door	2010-01-15
The Crazies	2010-02-26
Tooth Fairy	2010-01-22
Transylmania	2009-12-04
When In Rome	2010-01-29
Youth In Revolt	2010-01-08

TABLE I
NAMES AND RELEASE DATES FOR THE MOVIES WE CONSIDERED IN OUR ANALYSIS.

disambiguated and clean by choosing appropriate keywords and performing sanity checks.

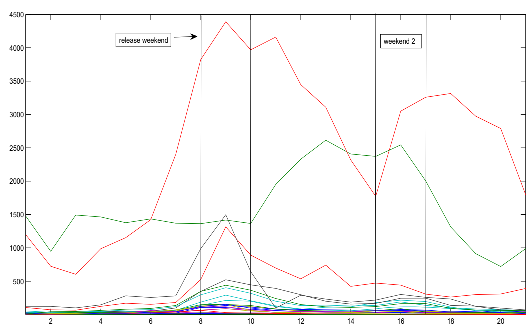

Fig. 1. Time-series of the rate of tweets over the critical period for different movies. Y-axis represents the rate of tweets (tweets per hour). The X-axis represents the days in the critical period.

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

Fig 1 shows the timeseries trend in the rate of tweets for movies over the critical period. The y-axis represents the traffic for tweet mentions of different movies. 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 2 shows how the number of tweets per unique author changes over time. We find that this ratio remains fairly consistent with a value between 1 and 1.5 across the critical period. 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 behavior from other networks [12]. 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.

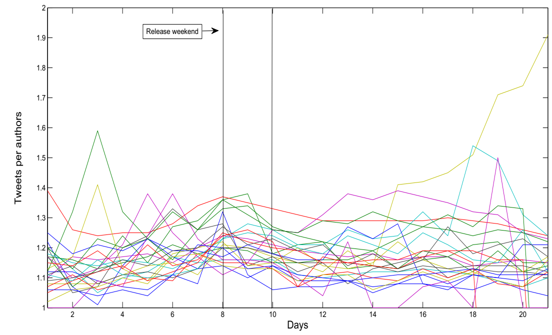

Fig. 2. Number of tweets per unique authors for different movies

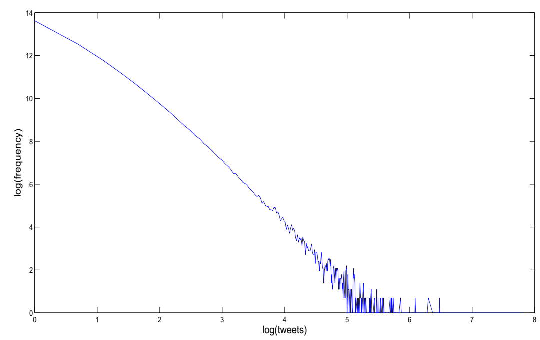

Fig. 3. Log distribution of authors and tweets.

V. ATTENTION AND POPULARITY

Our goal is to observe if the knowledge that can be extracted from the tweets can lead to reasonably accurate prediction of future outcomes in the real world.

The problem that we wish to tackle can be framed as follows. *Using the tweets referring to movies prior to their release, can we accurately predict the box-office revenue generated by the movie in its opening weekend?*

To build a predictive model for real-world effects on twitter, we first need quantifiable measures for capturing the attention and popularity that different movies receive. In this section,

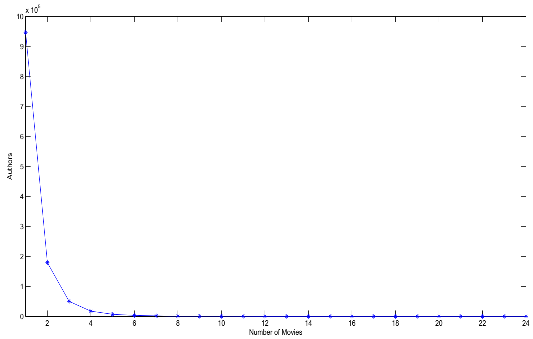

Fig. 4. Distribution of total authors and the movies they comment on.

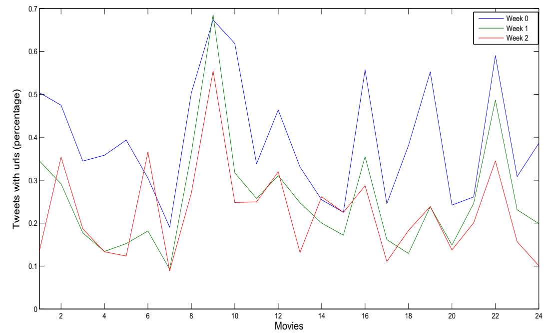

Fig. 5. Percentages of urls in tweets for different movies.

Features	Week 0	Week 1	Week 2
url	39.5	25.5	22.5
retweet	12.1	12.1	11.66

TABLE II
URL AND RETWEET PERCENTAGES FOR CRITICAL WEEK

Features	Adjusted R^2	p-value
Avg Tweet-rate	0.80	3.65e-09
Tweet-rate timeseries	0.93	5.279e-09
Tweet-rate timeseries + thcnt	**0.973**	9.14e-12
HSX timeseries + thcnt	0.965	1.030e-10

TABLE IV
COEFFICIENT OF DETERMINATION (R^2) VALUES USING DIFFERENT PREDICTORS FOR MOVIE BOX-OFFICE REVENUE FOR THE FIRST WEEKEND.

we will examine different modes by which attention and popularity can be measured on twitter, and evaluate their utility for predicting future trends.

A. Effect of Promotional Material : Urls and Retweets

Prior to the release of a movie, media companies and and producers generate promotional information in the form of trailer videos, news, blogs and photos. We expect the tweets for movies before the time of their release to consist primarily of such promotional campaigns, geared to promote word-of-mouth cascades. On Twitter, this can be characterized by tweets referring to particular urls (photos, trailers and other promotional material) as well as retweets, which involve users forwarding tweet posts to everyone in their friend-list. Both these forms of tweets are important to disseminate information regarding movies being released.

First, we examine the distribution of such tweets for different movies, following which we examine their correlation with the performance of the movies.

Table II shows the percentages of urls and retweets in the tweets over the critical period for movies. The actual distribution of urls in tweets for each movie over the critical period is shown in Fig 5. We can observe that there is a greater percentage of tweets containing urls in the week prior to release than later. This is consistent with our expectation.

Features	Correlation	R^2
url	0.64	0.39
retweet	0.5	0.20

TABLE III
CORRELATION AND R^2 VALUES FOR URLS AND RETWEETS BEFORE RELEASE.

In the case of retweets, we find the values to be similar across the 3 weeks considered. In all, we found the retweets to be a significant minority of the tweets on movies. One reason for this could be that people tend to describe their own expectations and experiences, which are not necessarily propaganda.

We want to determine whether movies that have greater publicity, in terms of linked urls on Twitter, perform better in the box office. When we examined the correlation between the urls and retweets with the box-office performance, we found the correlation to be moderately positive, as shown in Table III. However, the adjusted R^2 value is quite low in both cases, indicating that these features are not very predictive of the relative performance of movies. This result is quite surprising since we would expect promotional material to contribute significantly to a movie's box-office income.

B. Rate of Tweet Mentions

We have observed in Fig 1, that the rate of tweet mentions for movies differs quite significantly across the movies considered. We define the **tweet-rate**, as the *number of tweets referring to a particular movie per hour.*

$$Tweet\text{-}rate(mov) = \frac{|tweets(mov)|}{|hours|} \quad (1)$$

Our initial analysis of the correlation of the average tweet-rate with the box-office gross for the 24 movies considered showed a strong positive correlation, with a correlation coefficient value of 0.90. This suggests a strong linear relationship among the variables considered. Accordingly, we constructed a linear regression model using least squares of the average of

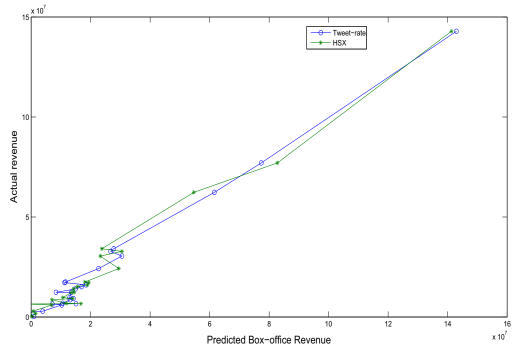

Fig. 6. Predicted vs Actual box office scores using tweet-rate and HSX predictors

all tweets for the 24 movies considered over the week *prior to their release*. We used the box-office revenue values as the dependent variable and the tweet-rate as the explanatory variable as:

$$Rev(mov) = \beta_0 + \beta_1 * Tweet\text{-}rate(mov) \tag{2}$$

The β values above represent the regression coefficients. We obtained an adjusted R^2 value of 0.80 with a p-value of $3.65e-09***$, where the '***' shows significance at 0.001, indicating a very strong predictive relationship. Notice that this performance was achieved using only one variable (the average tweet rate). The real box-office revenue information was extracted from the Box Office Mojo website [5].

The movie $Transylmania$ that opened on Dec 4th had easily the lowest tweet-rates of all movies considered. For the week prior to its release, it received on an average 2.75 tweets per hour. As a result of this lack of attention, the movie captured the record for the lowest-grossing opening for a movie playing at over 1,000 sites, making only $263,941 in its opening weekend, and was subsequently pulled from theaters at the end of the second week. On the other end of the spectrum, two movies that made big splashes in their opening weekends, *Twilight:New Moon* (making 142M) and *Avatar* (making 77M) had, for their pre-release week, averages of 1365.8 and 1212.8 tweets per hour respectively. This once again illustrates the importance of attention in social media.

Next, we performed a linear regression of the time series values of the tweet-rate for the 7 days before the release. We used 7 variables each corresponding to the tweet-rate for a particular day. An additional variable we used was the number of theaters the movies were released in, $thcnt$.

$$Rev(mov) = \beta_0 + \sum_{i=1}^{7} \beta_i * Tweet\text{-}rate_i(mov) + \beta_{th} * thcnt \tag{3}$$

The results of the regression experiments are shown in Table IV. Note that, in all cases, we are using only data available prior to the release to predict box-office for the opening weekend.

[5]http://boxofficemojo.com

Predictor	AMAPE	Score
$Reg_{nobudget}+nReg_{1w}$	3.82	96.81
Avg Tweet-rate + thcnt	1.22	98.77
Tweet-rate Timeseries + thcnt	**0.56**	**99.43**

TABLE V
AMAPE AND SCORE VALUE COMPARISON WITH EARLIER WORK.

Comparison with HSX:
To compare with our tweet-based model, we used the Hollywood Stock Exchange index. The fact that artificial online markets such as the Foresight Exchange and the Hollywood Stock Exchange are good indicators of future outcomes has been shown previously [4], [5]. The prices in these markets have been shown to have strong correlations with observed outcome frequencies. In the case of movies, the Hollywood Stock Exchange (http://www.hsx.com/), is a popular play-money market, where the prices for movie stocks can accurately predict real box office results. Hence, to compare with our tweet-rate predictor, we considered regression on the movie stock prices from the Hollywood Stock Exchange, which can be considered the gold standard [4].

From the results in Table IV, it can be seen that our regression model built from social media provides an accurate prediction of movie performances at the box office. Furthermore, the model built using the tweet rate timeseries *outperforms* the HSX-based model. The graph outlining the predicted and actual values of this model is also shown in Fig 6, outlining the utility of harvesting social media.

Comparison with News-based Prediction:
In earlier work, Zhang and others [6] have developed a news-based model for predicting movie revenue. The best-performing method in the aforementioned work is the combined model obtained by using predictors from IMDB and news. The corresponding R^2 value for this combined model is 0.788, which is far lower than the ones obtained by our predictors. We computed the AMAPE (Adjusted Mean Absolute Percentage/Relative Error) measure, that the authors use, for our data. The comparative values are shown in Table V. We can observe that our values are far better than the ones reported in the earlier work. Note however, that since historical information on tweets are not available, we were able to use data on only the movies we have collected, while the authors in the earlier paper have used a larger database of movies for their analysis.

C. Predicting HSX prices

Given that social media can accurately predict box office results, we also tested their efficacy at forecasting the stock prices of the HSX index. At the end of the first weekend, the Hollywood stock exchange adjusts the price for a movie stock to reflect the actual box office gross. If the movie does not perform well, the price goes down and vice versa. We

Predictor	Adjusted R^2	$p-value$
HSX timeseries + thcnt	0.95	4.495e-10
Tweet-rate timeseries + thnt	**0.97**	2.379e-11

TABLE VI
PREDICTION OF HSX END OF OPENING WEEKEND PRICE.

Weekend	Adjusted R^2
Jan 15-17	0.92
Jan 22-24	0.97
Jan 29-31	0.92
Feb 05-07	0.95

TABLE VII
COEFFICIENT OF DETERMINATION (R^2) VALUES USING TWEET-RATE TIMESERIES FOR DIFFERENT WEEKENDS

conducted an experiment to see if we could predict the price of the HSX movie stock at the end of the opening weekend for the movies we have considered. We used the historical HSX prices as well as the tweet-rates, individually, for the week prior to the release as predictive variables. The response variable was the adjusted price of the stock. We also used the theater count as a predictor in both cases, as before. The results are summarized in Table VI. As is apparent, the tweet-rate proves to be *significantly better* at predicting the actual values than the historical HSX prices. This again illustrates the power of the buzz from social media.

D. Predicting revenues for all movies for a given weekend

Until now, we have considered the problem of predicting opening weekend revenue for movies. Given the success of the regression model, we now attempt to predict revenue for all movies over a particular weekend. The Hollywood Stock Exchange de-lists movie stocks after 4 weeks of release, which means that there is no timeseries available for movies after 4 weeks. In the case of tweets, people continue to discuss movies long after they are released. Hence, we attempt to use the timeseries of tweet-rate, over 7 days before the weekend, to predict the box-office revenue for that particular weekend. Table VII shows the results for 3 weekends in January and 1 in February. Note, that there were movies that were two months old in consideration for this experiment. Apart from the time series, we used two additional variables - the theater count and the number of weeks the movie has been released. We used the coefficient of determination (adjusted R^2) to evaluate the regression models. From Table VII, we find that the tweets continue to be good predictors even in this case, with an adjusted R^2 consistently greater than 0.90. The results have shown that the buzz from social media can be accurate indicators of future outcomes. The fact that a simple linear regression model considering only the rate of tweets on movies can perform better than artificial money markets, illustrates the power of social media.

VI. 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 [13], [14]. 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 LingPipe linguistic analysis package [6] which provides a set of open-source java libraries for natural language processing tasks. We used the DynamicLMClassifier 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 [7]. It has been shown that manual labeling from Amazon Turk can correlate well with experts [11]. 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

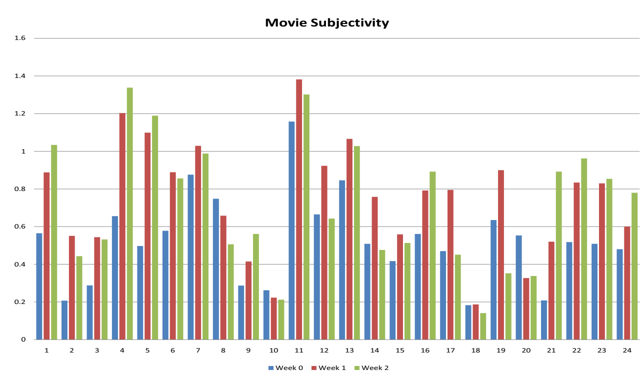

Fig. 7. Movie Subjectivity values

ways:

[6]http://www.alias-i.com/lingpipe
[7]https://www.mturk.com/

Predictor	Adjusted R^2	$p - value$
Avg Tweet-rate	0.79	8.39e-09
Avg Tweet-rate + thcnt	0.83	7.93e-09
Avg Tweet-rate + PNratio	0.92	4.31e-12
Tweet-rate timeseries	0.84	4.18e-06
Tweet-rate timeseries + thcnt	0.863	3.64e-06
Tweet-rate timeseries + PNratio	**0.94**	1.84e-08

TABLE VIII
PREDICTION OF SECOND WEEKEND BOX-OFFICE GROSS

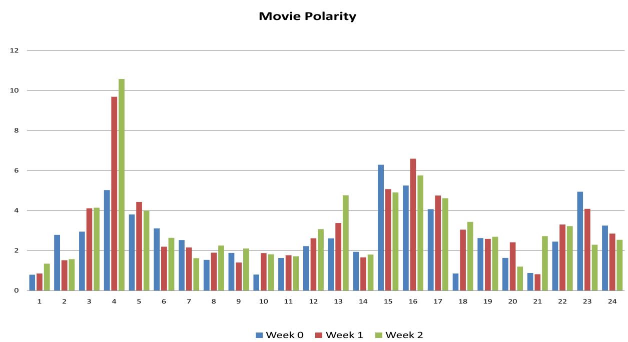

Fig. 8. Movie Polarity values

- Elimination of all special characters except exclamation marks which were replaced by $< EX >$ and question marks ($< QM >$)
- Removal of urls 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.

A. 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.

$$Subjectivity = \frac{|Positive\ and\ Negative\ Tweets|}{|Neutral\ Tweets|} \tag{4}$$

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.

Variable	$p - value$
(Intercept)	0.542
Avg Tweet-rate	2.05e-11 (***)
PNRatio	9.43e-06 (***)

TABLE IX
REGRESSION USING THE AVERAGE TWEET-RATE AND THE POLARITY (PNRATIO). THE SIGNIFICANCE LEVEL (*:0.05, **: 0.01, ***: 0.001) IS ALSO SHOWN.

B. 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.

$$PNratio = \frac{|Tweets\ with\ Positive\ Sentiment|}{|Tweets\ with\ Negative\ Sentiment|} \tag{5}$$

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 *The Blind Side* (5.02 to 9.65). The movie had a lukewarm opening weekend sales (34M) but then boomed in the next week (40.1M), owing largely to positive sentiment. The movie *New Moon* had the opposite effect. It released in the same weekend as *Blind Side* and had a great first weekend but its polarity reduced (6.29 to 5), as did its box-office revenue (142M to 42M) 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 PNratio as an additional variable. The results of our regression experiments are shown in Table VIII. 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 timeseries. Table IX 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.

VII. 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 Hollywood Stock 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.

VIII. Appendix: General Prediction Model for Social Media

Although we focused on movie revenue prediction in this paper, the method that we advocate can be extended to other products of consumer interest.

We can generalize our model for predicting the revenue of a product using social media as follows. We begin with data collected regarding the product over time, in the form of reviews, user comments and blogs. Collecting the data over time is important as it can measure the rate of chatter effectively. 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 ϵ the error. The linear regression model can be expressed as :

$$y = \beta_0 + \beta_a * A + \beta_p * P + \beta_d * D + \epsilon \qquad (6)$$

where the β values correspond to the regression coefficients. The attention parameter captures the buzz around the product in social media. In this article, we showed how the rate of tweets on Twitter can capture attention on movies accurately. We found this coefficient to be the most significant in our experiments. The polarity parameter relates to the opinions and views that are disseminated in social media. We observed that this gains importance after the movie has been released and adds to the accuracy of the predictions. In the case of movies, the distribution parameter is the number of theaters a particular movie is released in. In the case of other products, it can reflect their availability in the market.

IX. Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant # 0937060 to the Computing Research Association for the CIFellows Project.

References

[1] Jure Leskovec, Lada A. Adamic and Bernardo A. Huberman. The dynamics of viral marketing. *In Proceedings of the 7th ACM Conference on Electronic Commerce*, 2006.

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

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

[4] D. M. Pennock, S. Lawrence, C. L. Giles, and F. Å. Nielsen. The real power of artificial markets. *Science*, 291(5506):987–988, Jan 2001.

[5] Kay-Yut Chen, Leslie R. Fine and Bernardo A. Huberman. Predicting the Future. *Information Systems Frontiers*, 5(1):47–61, 2003.

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

[7] 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.

[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] 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.

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

[11] 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.

[12] 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.

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

[14] Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena. Large-Scale Sentiment Analysis for News and Blogs. *Proc. Int. Conf. Weblogs and Social Media (ICWSM)*, 2007.

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