Predicting the Future With Social Media

Kenny Richard
Social Computing Lab HP Labs Palo Alto, California
Email: kenny@gmail.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 further utilized to improve the forecasting power of social media

Keyword: *Social media*

A. INTRODUCTION

Social media has exploded as a category of online discourse where people create content, share it, bookmark it 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 market s have been shown to have strong correlations with observed outcome frequencies, and thus are good indicators of future outcomes In the case of social media, the enormity and high vari-ance 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 par-ticular products can be helpful when designing marketing and advertising campaigns. 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 we ll 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. 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

B. LITERATUR REVIEW

Although Twitter has been very popular as a web service, there has not been considerable published research on it. Huberman and others 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 investigated community structure and isolated different types of user intentions on Twitter. Jansen and others 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 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 use linear regression from text and metadata features to predict earnings for movies. Sharda and Delen have treated the prediction problem as a classification prob lem 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 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.

C. METHOD

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 @user$_{id}$ indicating the intended destination. A retweet is a post originally made by one user that is forwarded by another user. These retweets are a popular means of 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. 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 [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 consist ency, 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. 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. there is a greater percentage of tweets containing urls in the week prior to release than afterwards.

This is consistent with our expectation. 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 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. Next, we investigate the power of social media in predicting real-world outcomes. 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.

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 coeffi-cient 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 all tweets for the 24 movies considered over the week *prior to their release*. 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 predic tive relationship. Notice that this performance was achieved using only one variable (the average tweet rate). To evaluate our predictions, we employed real box-office revenue informati on, extracted from the Box Office Mojo website [5].

The movie T ransylmania 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. The results of the regression experiments are shown in Table 4. Note that, in all cases, we are using only data available prior to the release to predict box-office for the opening weekend.

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 vari-ance 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 8. 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 9 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.

D. CONCLUSION

Specifically, using t he 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 improvi ng 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.

REFERNCES

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. A˚. 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 twit-ter: 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 su ccess 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.