Designing a Social-Broadcasting-Based Business Intelligence System

Huaxia Rui [1] and Andrew Whinston [2]

Abstract

The rise of social media has fundamentally changed the way information is produced, disseminated, and consumed in the digital age, which has profound economic and business effects. Among many different types of social media, social broadcasting networks such as Twitter in the U.S. and "Weibo" in China are particularly interesting from a business perspective. In the case of Twitter, the huge amounts of real-time data with extremely rich text, along with valuable structural information, makes Twitter a great platform to build business intelligence (BI) systems. We propose a framework of social-broadcasting-based BI systems that utilizes real-time information extracted from these data with text mining techniques. To demonstrate this framework, we designed and implemented a Twitter-based BI system that forecasts movie box office revenues during the opening weekend and forecasts daily revenue after 4 weeks. We found that incorporating information from Twitter could reduce the mean absolute percentage error (MAPE) by 44% for the opening weekend and by 36% for total revenue. For daily revenue forecasting, including Twitter information into a baseline model could reduce forecasting errors by 17.5% on average. On the basis of these results, we conclude that social-broadcasting-based BI systems have great potential and should be explored by both researchers and practitioners.

Categories and Subject Descriptors: H.4.2 [**Information Systems**]: Information Systems Applications – *Decision Support*

General Terms: Design, Economics, Management

Additional Keywords and Phrases: Business intelligence, social broadcasting, Twitter, forecasting

[1]The University of Texas at Austin, E-mail: huaxia@utexas.edu

[2]The University of Texas at Austin, E-mail: abw@uts.cc.utexas.edu

1 Introduction

Business intelligence (BI) is defined as the result of "acquisition, interpretation, collation, analysis, and exploitation of information" in the business domain [Davis 2002, p.313]. Although it is common for companies to analyze and understand patterns from their internal data sources (e.g., their operational data), the explosive growth of user-generated content in the past decade has offered companies another perspective to understand their businesses through intelligent use of this new data source. One particularly important type of user-generated content is the real-time stream of people's chatter from social broadcasting networks, such as Twitter in the U.S. and "Weibo" in China.[3] These social broadcasting networks are radically different from traditional social networks such as Facebook. Although individual privacy is an important concern for the latter, public openness is one of the defining features of the former. Users can freely establish asymmetrical ties to (known as *follow*) any other users whose content (e.g., "tweets" on Twitter, which are text-based messages of up to 140 characters) they find interesting. By default, content produced by users is accessible to anyone, and those who subscribe to receive a user's tweets in real-time are called the user's "followers." The openness on Twitter facilitates the dissemination and consumption of content, and probably also leads to more decentralized content generation. People frequently share their information and opinions on a variety of topics openly on Twitter. As a result, huge amounts of real-time data about opinions and activities of a rapidly growing proportion of the population are constantly being generated, disseminated, and eventually consumed.[4] With so much readily available data, the question is, "How can companies make use of it to build their BI systems?"

We propose in the present article a framework of social-broadcasting-based BI systems that companies can use to make better decisions regarding customers, suppliers, and logistics. There are several advantages of such BI systems. Take Twitter, for example. One obvious advantage is that the real-time stream of tweets can be "pushed" to a BI system by the Twitter server and the data are well structured, which makes it convenient to automate data acquisition and integrate acquired data with other business applications. Similarly, reaction or interaction based on the BI output can be "pushed" back to the customers instantly with Twitter. In fact, many companies have realized this convenience and are actively listening to and engaging with their customers through Twitter. However, we believe a systematic solution is more efficient than an ad hoc strategy, and the study of social-broadcasting-based BI is thus important.

Twitter provides a set of application program interfaces (API) that allows companies to obtain valuable structural information on Twitter, including users' social network information and the

[3] We call these sites *social broadcasting networks* because they each are simultaneously a social network and a broadcasting network.

[4] For example, according to Forbes, as of January 2011, there were nearly 200 million registered users on Twitter who posted 110 million tweets per day. For more details, please see *http://www.forbes.com/sites/oliverchiang/2011/01/19/twitter-hits-nearly-200m-users-110m-tweets-per-day-focuses-on-global-expansion/*.

number of times each tweet is retweeted.[5] For example, the number of followers a user (she) has could play an important role in accounting for the level of influence she has because it tells us the number of recipients of each tweet sent by her. The number of followers she has is the size of her immediate audience. The more followers she has, the more people she can reach, which may imply greater influence on the community. Indeed, when we incorporated such structural information along with some other tweet-related variables into a prerelease movie revenue forecasting system, the forecasting error was significantly reduced in terms of mean absolute percentage error (MAPE). In the case of opening weekend revenue forecasting, the forecasting error was reduced by 44%, and in the case of total revenue forecasting, the forecasting error was reduced by 36%.

Another important feature of Twitter is that there are a lot of intention tweets—tweets through which people explicitly or implicitly express their intentions to consume certain goods or services. People on Twitter frequently talk about their plans or intentions to take certain actions—for example, watching a movie or going to a restaurant. This type of content is not common in online forums, in which, discussion on products or services is often based on review. Interestingly, these intention tweets turned out to be very useful information in our demonstration system, in which we used tweets to predict daily movie sales. In our demonstration system, intention tweets were automatically extracted by a text-mining algorithm from the real-time stream of tweets on selected movies. The number of intention tweets was then fed into a time-series model that generated the forecast of the box office revenue of those movies for the next day. We tested the predictive power of the demonstration system using 20 movies. As compared with the benchmark system that did not use any tweet information, the Twitter-based BI system reduced the MAPE by 17.5% on average. For a movie such as *The Blind Side*, the MAPE reduction was 4.47% in absolute value, and for the movie *Wolfman*, including tweet information reduced the forecasting error almost by half.

The potential value of Twitter as a business tool is increasingly being recognized [Rui et al. 2009; Bollen et al. 2010]. How Twitter, or social broadcasting networks in general, can be used to improve business operations is a challenge facing both companies and researchers. In the present article, we make an early attempt to take on the challenge by introducing social-broadcasting-based BI and exploring its potential through the construction of a Twitter-based BI system. Our results suggest that the effectiveness of Twitter data as a source of BI systems may critically rely on how well structural information on Twitter is exploited and how novel text mining techniques can be applied to analyze tweets.

The rest of this article is organized as follows: We review relevant literature from different perspectives in Section 2 and introduce a social-broadcasting-based BI framework in Section 3. In Section 4, we illustrate the framework by constructing a Twitter-based BI system that forecasts movie sales. In Section 5, we conclude our article and point out future research directions.

[5]Retweeting refers to the behavior of re-posting a tweet posted by another user.

2 Related Works

BI, Text Mining, and Sentiment Analysis

There has been a growing interest in the IS community in BI over the past two decades. Early BI systems often worked only with internal data sources, and the data were not delivered in real time. However, a trend has emerged in BI to move toward real-time analysis [Wixom et al. 2008] and to develop web-based BI systems [Chung and Chen 2009; Chung et al. 2005]. More recently, the explosive growth of user-generated content has triggered a new wave of interest in analyzing user-generated content for the purpose of BI. For example, Abbasi and Chen [2008] proposed a framework advocating the development of systems capable of representing the rich array of information types inherent in computer mediated communication (CMC) text. They also developed the CyberGate system based on the framework and provided guidelines regarding feature selection and visualization techniques that CMC text analysis systems should use.

As an important technique for BI, text mining refers to the general process of extracting useful information from electronic text. Closely related to text mining is sentiment analysis (also known as opinion mining), which refers to the particular process of identifying and extracting subjective information. Pang and Lee [2008] gave an overview of the field of sentiment analysis. Recently, researchers have been actively applying text mining and sentiment analysis techniques to BI. For example, Chung and Tseng [2010] developed a new framework for designing BI systems that correlate the textual content and the numerical ratings of online product reviews from Amazon.com. Dang et al. [2010] studied the sentiment classification of online product reviews and proposed a lexicon-enhanced method that combines the machine learning approach and the semantic orientation approach. Their experiments on an Epinion data set and on Blitzer's multi-domain sentiment data set indicate significant improvement of their method over previous methods. Liu et al. [2010] utilized two recent tools of sentiment analysis, OpinionFinder and SentiWordNet, to examine how five text and sentiment measures of online word of mouth (WOM), including WOM volume, valence, subjectivity, number of sentences, and nubmer of valence words, correlate with movie sales measured by five different metrics. They found that WOM volume is the most useful predictor of movie success.

Word of Mouth

Tweets can be viewed as a type of online consumer WOM, which has received a lot of attention recently for its potential effect on product sales. Godes and Mayzlin [2004] pioneered the study of online WOM. They collected WOM information on 44 television shows during the 1999 to 2000 season from the Usenet newsgroup. The WOM information was then used in a panel data model to explain the ratings of those TV shows. They identified the explanatory power of the entropy of conversations across newsgroups but concluded that the volume of conversations did not have any explanatory power. Liu [2006] collected 12,136 WOM messages on 40 movies from Yahoo!Movie and included WOM volume and WOM valence (measured as a percentage of positive/negative WOM)

in a cross-sectional study. He found that most of the explanatory power of WOM information comes from the volume of WOM but not from its valence. Using similar data but with a different empirical model, Duan et al. [2008] also found that box office sales are significantly influenced by the volume of online postings, but not by the ratings of online postings that measure WOM valence. More recently, Chintagunta et al. [2011] measured the impact of national online user reviews on designated market area-level local geographic box office performance of movies, and they suggested that it is the valence that drives box office performance, not the volume. Some researchers have examined the effect of online WOM on the sales of products other than television shows and movies. For example, Chevalier and Mayzlin [2006] studied the effect of WOM on book sales. Sonnier et al. [2011] studied the effect of online communications on the sales of some durable goods from a certain company.

Movie Revenue Forecasting

Because our demonstration system involved the forecasting of movie sales, we will now review some earlier research on this topic. The high economic importance and the significant cultural impact of the movie industry have attracted many researchers to study the ingredients of a successful movie.

One stream of research has focused on how various factors during the production, distribution, and exhibition of a movie can affect its sales. Because these factors are known before the release of a movie, one obvious advantage of this approach is the ability to forecast a movie's financial performance before its release. Early works in this stream of literature used linear regression models and examined factors such as MPAA rating, genre, star actors and directors, releasing schedule, and critical reviews [Litman 1983; Litman and Kohl 1989]. It has been reported that MPAA ratings and subject matters are largely irrelevant, except for science fiction movies. On the other hand, star actors and directors, production cost, and critical rating are positively correlated with movie revenues. Eliashberg and Shugan [1997] found that critical reviews correlate with late and cumulative box office receipts but do not have a significant correlation with early box office receipts. Recently, Eliashberg et al. [2000] proposed and implemented a decision support system called MOVIEMOD for prerelease market evaluation for the motion picture industry. Their model was based on a behavioral representation of the consumer adoption process for movies as a macroflow process and was calibrated in a consumer clinic experiment. The evaluation of MOVIEMOD by the Dutch exhibitor and the distributor showed promise of their system. More recently, Sharda and Delen [2006] and Delen et al. [2007] converted the forecasting problem into a classification problem and explored the use of neural networks in the classification of a movie's financial performance based on variables including MPAA rating, competition, star value, genre, technical effects, sequel, and number of screens. They further developed a web-based decision support system to predict a movie's box office receipts before its initial release.

Another stream of research has focused on forecasting the gross box-office revenues of a new movie based on early box office data. Sawhney and Eliashberg [1996] drew upon a queuing theory

framework and developed a parsimonious model to forecast revenues in Week 4 or later. They reported reasonable accuracy of their model. Dellarocas et al. [2007] proposed a family of diffusion models for a similar purpose, but they emphasized the importance of including online product reviews. In fact, they were able to show that the addition of online product review metrics to a benchmark model that includes prerelease marketing, theater availability, and professional critic reviews substantially increases its forecasting accuracy.

Twitter

Since Twitter became popular in 2007, a year after its creation, it has garnered much attention from researchers in many fields. Asur and Huberman [2010] did a cross-sectional study using 24 movies, and they suggested that social media could be used to predict real-world outcomes. Bollen et al. [2010] analyzed the text content of daily Twitter feeds by Google-Profile of Mood States (GPOMS) and then used a Self-Organizing Fuzzy Neural Network (SOFNN) to predict DJIA closing values. They reported a significant reduction of prediction error.

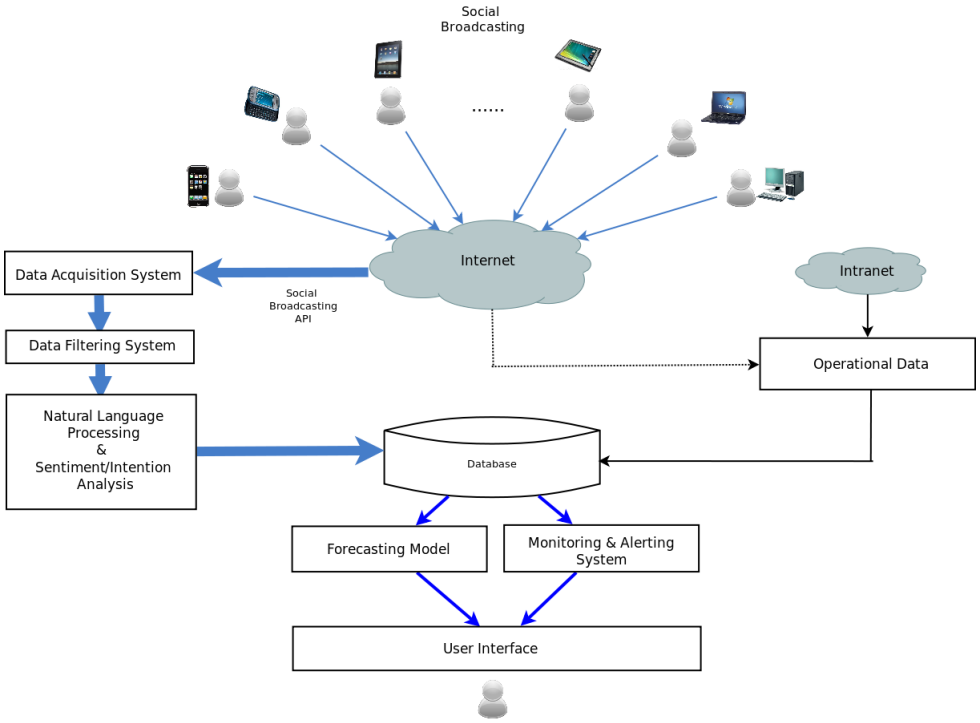
Besides prediction, researchers have used Twitter data to study a variety of interesting phenomena and have related their findings to theories in many different disciplines. Wu et al. [2011] used Twitter data to study several long-standing questions in media communications research regarding the production, flow, and consumption of information. For example, they found considerable support to the classical "two-step flow" theory of communications, and they found that roughly 50% of URLs consumed are generated by just 20,000 users. Romero et al. [2011] analyzed the ways in which hashtags [6] spread on a network defined by the interactions among Twitter users, and their results validated the complex contagion principle from sociology. Rui and Whinston (2011) studied users' tweeting behavior on Twitter, and their results suggested that attention from others is an important driver for content generation. O'Connor et al. [2010] analyzed several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and they found that these surveys were correlated with sentiment word frequencies in contemporaneous Twitter messages. Their results highlighted the potential of text streams as substitutes for and supplements to traditional polling. Shi et al. [2010] studied Twitter users' retweeting behavior, and their results suggested that weak ties are more likely to lead to retweeting.[7]

3 A Social-Broadcasting-Based BI Framework

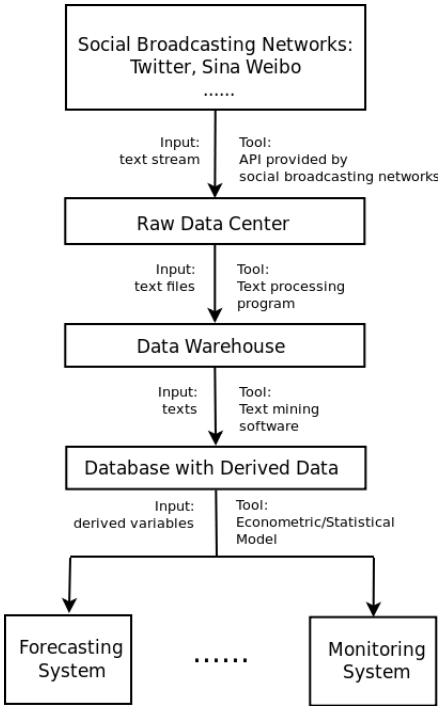
Figure 1 illustrates the basic framework and a process map of the proposed social-broadcasting-based BI system.

[6] hashtags are words or phrases concatenated and prefixed with a hash symbol (#), such as those in "#RealAle is my favorite kind of #beer".

[7] On Twitter, two users are considered strongly tied if they follow each other; they have a weak tie if only one user is following the other.



(a) Conceptual Framework



(b) Process Map

Figure 1: Social-Broadcasting-Based BI

Collecting content and processing it constitute two key components of the system that are very different from a traditional BI system in which data are often available from internal sources. These components of collecting and processing need to be customized according to the specific needs of a company. For example, retailers might want to know customers' plans for shopping. Manufacturers might be interested in customers' comments or complaints about their products so that they can respond quickly.[8] Investors might be interested in people's sentiment toward a certain brand, an industry, or the whole economy, so that they can make better investment decisions. Different objectives imply different data collection procedures and, more importantly, different methods to process the data.

Once the desired information is extracted from social broadcasting networks, it can be fed into a monitoring system for the purpose of monitoring the public sentiment toward the variables of interest, or it can be combined with the company's operational data (e.g., sales data) for the purpose of forecasting the variables of interest.

Data Acquisition

Social broadcasting networks, such as Twitter and Weibo, provide public API for developers to pull data from their servers. For example, both Twitter and Sina Weibo provide a search API, which allows developers to search for content using conditions such as keyword, time stamp, language, and location. In particular, Twitter also provides streaming API for high-volume and repeated queries, which makes it even more convenient for developers to receive real-time tweets that contain a fixed set of keywords.[9] Regardless of whether companies use Twitter search API or streaming API, the challenge is to pick a set of keywords that can yield tweets containing information relevant to their interests.

In addition to tweet information, companies might also need to obtain information regarding Twitter users, such as the social network information of the author of certain tweets. This type of information can also be obtained easily using Twitter API.

Data Processing

Processing textual content is challenging. In the case of Twitter, the first step is often to filter out spam tweets. Such filtering can be done either through supervised classification methods based on the tweet content or by checking the authors of the tweets; the latter is similar to the blocklist approach widely used to filter email spam. Language filtering may also be used to select or exclude text in certain languages if texts in different languages are mixed together. Another important

[8] Dell provided a good example of a manufacturer's successfully using such data. According to The New York Times, Dell noticed people complaining on Twitter about the apostrophe and return keys being too close together on the Dell Mini 9 laptop, and designers fixed the problem in the Dell Mini 10. http://www.nytimes.com/2009/04/14/technology/internet/14twitter.html

[9] In addition to the free API, Twitter also sells limited access to its firehose through products like Halfhose and Decahose, which contain 50% and 10% of all tweets, respectively.

filtering procedure is to filter out irrelevant content. For example, tweets containing "apple" might refer to the company Apple or to the fruit.

After the data are filtered, text mining and sentiment analysis techniques can be used to extract high-quality information from the huge amounts of text. For example, companies may want to estimate the collective sentiment of the population toward a brand and monitor the sentiment over time. They also might want to monitor people's attitudes toward the economy through the analysis of people's chatter about their consumption plans. In the example we present in Section 4, we extracted people's plans of watching certain movies by identifying their intention tweets. For example, a tweet saying "I can't wait to watch Sherlock Holmes!" is an intention tweet about the movie *Sherlock Holmes*.

4 A Twitter-Based BI System

In this section, we present a Twitter-based BI system, which is an instantiation of the proposed social-broadcasting-based BI. The system runs in real-time to forecast a movie's financial performance over the opening weekend and to forecast the daily box office revenue from the start of the fifth week to the end of seventh week.

Data Acquisition

We used Twitter search API [10] to collect tweets related to 55 movies that were widely released between June 2009 and February 2010.[11] For each movie, we collected tweets for 8 weeks: 1 week before the release of the movie, and 7 weeks after the release of the movie.

Data Processing

After the tweets were collected into the system, a simple filtering program was periodically executed to filter out advertising tweets. Although we used several rules to determine whether a tweet was an advertising tweet, the most effective one was simply by checking whether the tweet contained a URL. There were also some irrelevant tweets containing the search keyword but were not about the movies. This was particularly a problem if the movie name was a single word or a commonly used phrase. We first randomly selected 200 tweets for each movie and manually checked whether there were irrelevant tweets. For some movies such as *Ninja Assassin* and *Shutter Island*, there was almost no irrelevant tweet because these two phrases were rarely used on Twitter in contexts other than those movies. However, for some movies such as *Wolfman*, *The Hangover*, and

[10] Twitter streaming API were not available at the time we started the data collection.

[11] We excluded movie titles for which it is difficult to correctly identify tweets that were related to those movies. For example, it is very hard to distinguish tweets talking about the movie 2012 from tweets talking about the year 2012. We also excluded movies with a complicated release schedule. For example, the movie, *The Princess and the Frog*, was first released in New York and Los Angeles on November 25, 2009, and then was widely released on December 11, 2009.

It's Complicated, there were irrelevant tweets. To reduce those irrelevant tweets, we adopted two approaches. First, we used a movie lexicon containing words or phrases such as *movie, cinema, film, theater*, or *ticket* to pick out relevant tweets containing words or phrases in the lexicon. Second, for each movie, we used a customized lexicon for those irrelevant tweets and eliminated tweets containing words or phrases in that lexicon. For example, for the movie *The Hangover*, if a tweet contained the phrase "suffering from a hangover" or the word "drunk," then that tweet was classified as irrelevant. If there were still tweets undetermined after these two procedures, we manually classified them.

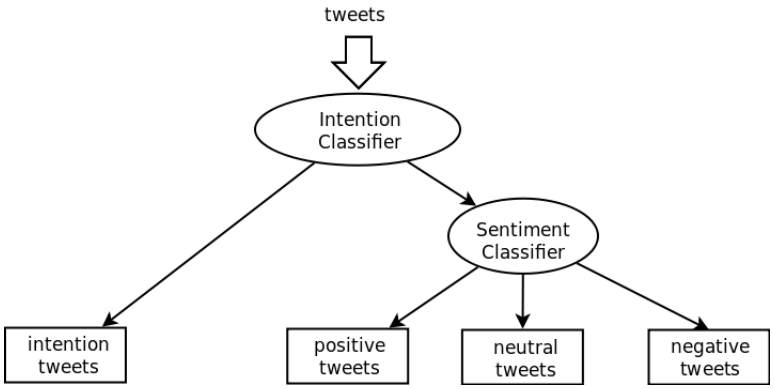


Figure 2: Tweet Classification

After filtering out the advertising tweets and irrelevant tweets, we classified each tweet into one of the four mutually exclusive categories: intention, positive, negative, and neutral. An intention tweet is a tweet in which the author expresses his/her intention to watch a certain movie in the future. A positive tweet is a tweet in which the author expresses positive sentiment toward the movie. Similarly, a negative tweet is a tweet in which the author expresses negative sentiment toward the movie. Neutral tweets are all other tweets that do not belong to any of the aforementioned three categories. Figure 2 illustrates the classification scheme.

We used an intention lexicon to extract features from tweets and then used a support vector machine (SVM) to construct the intention classifier. The intention lexicon was built from the movie tweets in our sample. For the sentiment analysis of tweets, we constructed a Naive Bayesian classifier that drew upon a lexicon of positive words/phrases and negative words/phrases. Naive Bayesian classifiers are often used in the literature for text mining because of their simplicity. Of course, there are more sophisticated classifiers for sentiment analysis in general that might yield higher accuracy. An in-depth study of these methods is beyond the scope of the present article and is left as a future research topic. Both classifiers were trained and tested on a corpus of over 3,000 tweets that were manually labeled. The precisions and recalls for the intention classifier and the sentiment classifier are reported in Table 1 and Table 2, respectively.[12]

[12]Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. For example, a precision of 98% and a recall of 87% for the class of intention tweets mean that 98%

Table 1: Precisions and Recalls of the Intention Classifier

	Precision	Recall
Intention tweets	98%	87%
Non-intention tweets	93%	99%

Table 2: Precisions and Recalls of the Sentiment Classifier

	Precision	Recall
positive	75%	80%
negative	65%	71%
neutral	75%	68%

Prerelease Revenue Forecasting

We will first focus on how Twitter information can help us forecast financial performance of a movie before its initial theatrical release. We measured the financial performance of a movie by its box office revenue per theater. The opening weekend is particularly important to the financial success of a movie. Indeed, it is reported that 25% of total revenue of a movie comes from the first 2 weeks [Litman and Ahn 1998]. Hence, we tried to forecast both the box office revenue per theater during the opening weekend and the box office revenue per theater during a movie's lifetime in theater.

Rather than develop a full-fledged movie revenue forecasting system, our objective was to illustrate how information extracted from tweets could be used to improve the performance of a forecasting model. Hence, we tried to keep our forecasting models as simple and parsimonious as possible. More specifically, Model 1 and Model 3 below are linear forecasting models based on some traditional independent variables, and Model 2 and Model 4 are models based mainly on variables extracted from tweets.[13] By comparing the forecasting errors of these models and studying the predictive power of each variable, we can deepen our understanding of the advantages and limitations of information from Twitter for the purpose of sales forecasting.

- **Model 1**:

$$OpenRPT_i = a_0 + a_1 \cdot Budget_i + a_2 \cdot Sequel_i + a_3 \cdot Scifi_i + a_4 \cdot Comedy_i + a_5 \cdot PG_i + a_6 \cdot R_i \quad (1)$$

of the tweets classified by the program as intention tweets are indeed intention tweets based on human judgement, and 87% of the intention tweets based on human judgement are classified by the program as intention tweets.

[13] We use a and b to represent coefficients from models without Twitter variables (Models 1 & 3) and Greek letters (α and β) to represent coefficients from models with Twitter variables (Models 2 & 4).

- **Model 2**:

$$OpenRPT_i = \alpha_0 + \alpha_1 \cdot Budget_i + \alpha_2 \cdot Sequel_i + \alpha_3 \cdot PreFol3_i + \alpha_4 \cdot PreSenti3_i + \alpha_5 \cdot PreInt3_i \quad (2)$$

- **Model 3**:

$$TotalRPT_i = b_0 + b_1 \cdot Budget_i + b_2 \cdot Sequel_i + b_3 \cdot Scifi_i + b_4 \cdot Comedy_i + b_5 \cdot PG_i + b_6 \cdot R_i \quad (3)$$

- **Model 4**:

$$TotalRPT_i = \beta_0 + \beta_1 \cdot Budget_i + \beta_2 \cdot Sequel_i + \beta_3 \cdot PreFol7_i + \beta_4 \cdot PreSenti7_i + \beta_5 \cdot PreInt7_i \quad (4)$$

The selection of variables for Models 1 and 3 was consistent with that in early literature in which researchers studied the ingredients of financially successful movies. For example, Liu [2006] used similar variables for movie genres and MPAA ratings. For Models 2 and 4, we wanted to use a variable that could proxy the number of people who were aware of the movies, because marketing researchers have long recognized that consumer awareness is the first step in the consumer decision-making process [Lilien et al. 1992]. The most natural variable to consider was the total number of tweets mentioning each movie, which served as the volume of WOM. Indeed, similar measurements have been used in many previous studies [Liu 2006; Duan et al. 2008]. However, we chose $PreFol3$ ($PreFol7$), the total number of followers of those who tweeted about the movie during the 3 (7) days before its release, as a predictor of people who were aware of the movie. The rationale was that the number of people who are informed of a movie might be a better predictor of the opening weekend revenue than the number of people who are informing others of the movie. Another motivation was that the number of followers a user has might be a rough proxy of the user's social influence; thus, $PreFol3$ ($PreFol7$) could be interpreted as the weighted count of the people who were discussing the movie, where the weight was each user's social influence measured by the number of followers each user had.

Table 3: Description of Variables Used in the Prerelease Revenue Forecasting Models

Variable Name	Definition
$OpenRPT_i$	Revenue per theater for movie i during the first weekend (i.e., Friday, Saturday, and Sunday)
$TotalRPT_i$	Revenue per theater for movie i, calculated by dividing the total revenue of a movie by the accumulated total number of theaters over the life time of the movie.
$Budget_i$	The production budget of movie i
$Sequel_i$	1 if movie i is a sequel, 0 otherwise
$Scifi_i$	1 if the genre of movie i is science fiction, 0 otherwise
$Comedy_i$	1 if the genre of movie i is comedy, 0 otherwise
PG_i	1 if movie i is a rated PG in MPAA ratings, 0 otherwise
R_i	1 if movie i is a rated R in MPAA ratings, 0 otherwise
$PreFol3_i$ ($PreFol7_i$) [1]	Total number of followers of the Twitter users who tweeted about movie i during the three-day (seven-day) period before the release of movie i
$PreSenti3_i$ ($PreSenti7_i$)	Sentiment of the Twitter community towards movie i, measured by the ratio of the number of positive tweets to the number of negative tweets during the three-day (seven-day) period before the release of movie i.
$PreInt3_i$ ($PreInt7_i$)	Twitter users' intention to watch movie i during the three-day (seven-day) period before the release of movie i, measured by the number of unique authors of tweets on movie i that were posted during that three-day (seven-day) period and were classified as intention tweets by our classifier.

[1] The number of days of monitoring tweets prior to the release seems to matter slightly. For the opening weekend revenue forecasting, results are roughly the same if we choose 2 or 3 days but are slightly worse if we choose $4-7$ days. For the total revenue forecasting, it's exactly the opposite. One possible explanation is that information from dates closer to the release date might be more related to the box office revenue of the opening weekend, but tweet information from a longer period before the release data might be more related to the box office revenue over a longer time after the release.

We used $PreSenti3$ ($PreSenti7$) as an indicator of the general sentiment of the population toward the movie, which was measured as the ratio of the number of positive tweets to the number of negative tweets during the 3 (7) days before the release of a movie. We included the variable $PreInt3$ ($PreInt7$) in order to capture some more direct information about people's movie-going behavior. Of course, our underlying assumption was that the more people who claim on Twitter that they want to watch a particular movie, the more likely that movie will be financially successful.

For consistency, we focused on the 45 movies in our sample that were released on a Friday and for which we had publicly available budget information.[14] All of the variables used in Models 1–4 are defined in Table 3.

Because of the relatively small size of our sample, we adopted a rotating method to calibrate and test our forecasting models. More specifically, each movie in the sample was used once as the movie to be forecast, whereas the rest of the movies were used to estimate the parameters in the models. We then evaluated the forecasting model by averaging the absolute percentage errors of the 45 predictions. Table 4 gives the mean absolute percentage errors of the four models.

Table 4: Mean Absolute Percentage Errors of the Four Forecasting Models

	Model 1	Model 2	Model 3	Model 4
MAPE	64%	36%	39%	25%

From Table 4, we see that including tweet information in the forecasting models significantly reduces forecasting errors. In the case of opening weekend revenue forecasting, the forecasting error is reduced by 44% (from a 64% MAPE to a 36% MAPE), and in the case of total revenue forecasting, the forecasting error is reduced by 36% (from a 39% MAPE to a 25% MAPE). It has been reported in the previous literature that including WOM in the movie revenue forecasting system can reduce forecasting errors by 31% for a movie's opening week and by 23% for the aggregate box office [Liu 2006]. Although the results are not directly comparable because of the differences in the movie samples and in the forecasting models, the results still suggest that information from Twitter could potentially be an important source for BI systems. It should also be noted that our forecasting results were obtained from a completely automated Twitter-based movie revenue forecasting system, which is in contrast with many earlier systems that require significant human involvement.

Finally, we used all movie samples to estimate Model 2 and Model 4 to better understand how the variables we chose for the forecasting models correlated with the financial performance of the movies. The results are reported in Table 5.

We notice that the coefficients of $PreFol3$ and $PreFol7$ are significantly positive, which suggests the potential predictive power of these two tweet variables. To verify our intuition that $PreFol3$ is indeed a better choice than the number of tweets for the forecasting model, we replaced it with the

[14] Movies are typically released on Friday, although some are released on Wednesday or Thursday. Publicly available movie budget information was retrieved from boxofficemojo.com.

Table 5: Estimation Results of the Model 2 and Model 4

	Model 2		Model 4	
Variable	Coefficient	p-value	Coefficient	p-value
$PreFol3$ ($PreFol7$)	1.387e-04	0.0030	3.499e-05	0.0017
$PreSenti3$ ($PreSenti7$)	2.121e+03	0.0244	5.042e+02	0.0238
$PreInt3$ ($PreInt7$)	3.114e-01	0.0830	4.565e-02	0.2769
$Budget$	1.151e+01	0.0008	3.525e+00	0.0000
$Sequel$	5.659e+02	0.2771	1.269e+02	0.3024
Constant	4.592e+00	0.9910	2.734e+01	0.7775
R-squared	0.74		0.76	

number of tweets in Model 2 and repeated the analysis. We find that the average MAPE increases from 36% to 42%, whereas the R-squared of the regression using all movies decreases from 0.74 to 0.68. These facts not only provide strong evidences for the use of $PreFol3$ in the forecasting model but also suggest the potential value of exploiting the underlying social network structure of Twitter. The current approach of using the number of followers as the influence or weight of each tweet is certainly very simple and coarse. Whether a more carefully designed influence measure can improve the forecast remains an open question and is left for future research.

The positive signs of $PreSenti3$ and $PreSenti7$ are consistent with our intuition that public sentiment toward a certain movie does reflect the population's movie-going behavior. It is thus conceivable that by monitoring the public sentiment toward a certain brand or product on Twitter, managers or investors can obtain useful information about future product sales or even stock prices. However, it should be noted that the sentiment of tweets is much less significant than $PreFol3$ and $PreFol7$, which is consistent with the discussion of dominant effect of WOM volume in the previous literature [Liu 2006; Duan et al. 2008].[15] Indeed, it is possible that the sentiment of WOM is generally less significant than the volume of WOM, which probably explains why the valence of WOM is often reported as insignificant in much of the previous literature, whereas volume is reported to be significant. Another explanation is that the awareness effect of tweets is much more significant than the persuasive effect of tweets before the release of a movie because the quality of a movie is revealed to the public mainly after its release.

We also notice that $PreInt3$ is significant at a 10% level in Model 2 whereas $PreInt7$ is insignificant in Model 4. This is not surprising because intention tweets reflect a timely but temporal trend. The population's intention to watch a movie prior to its release might be a reasonable predictor of the movie's opening weekend revenue, but it might not be a good signal of how much revenue the movie can collect over its life span in the theater.

[15]The number of followers of the tweets is highly correlated with the number of tweets. So $PreFol3$ and $PreFol7$ can be viewed as another measure of WOM volume.

Daily Revenue Forecasting

Twitter is well known for its real-time feature. In fact, Twitter has on its homepage the following description of itself: *Instant updates from your friends, industry experts, favourite celebrities, and what's happening around the world.* It is thus interesting to see how Twitter information can predict movie box office revenue at a finer granularity. Because our movie revenue information was collected on a daily basis, we will now turn to the problem of forecasting the box office revenue per theater for each movie on each day, which has never been studied previously in the literature.[16] Timely forecast of product sales helps stores improve their operation management. The purpose of this section is to demonstrate that high-frequency information on Twitter can be integrated into a BI system to forecast daily/weekly/monthly sales of a particular product. To this end, we compared the forecasting errors of two models: one without tweet information and one with tweet information.

There was potentially a large amount of information in those millions of tweets in our sample, and the number of variables we could extract from them was enormous. To avoid an overfitting problem, we tried to keep the prediction model as parsimonious as possible. In fact, we used only one tweet variable, $Intention_t$, which was the total number of intention tweets in the previous five days.[17] We used this particular variable mainly because of its clear interpretation as the number of people who wanted to watch the movie.[18]

We used the following two models to forecast daily revenue per theater. Model 5 is a simple baseline forecasting model in which we used the previous day's revenue per theater to predict the current day's revenue per theater. The two dummy variables, $Weekend_t$ and $Saturday_t$, were included because movie revenue is typically higher on Friday, Saturday, and Sunday than during the rest of the week, and box office revenue on Saturday is usually the highest of all during the week. The variable $Days_t$ denotes the number of days since the movie was released, and was included to account for the decreasing trend of movie revenue. Model 6 is the augmented forecasting model with $Intention_t$ included. All variables related to Models 5 and 6 are summarized in Table 6.

[16] The main reason why previous literature did not look at the problem of daily revenue forecasting for movies, as is pointed out by one reviewer, is that for the movie industry, daily forecast does not have much practical value because most movie distribution and promotion decisions are made on a weekly basis. Moreover, predicting the total gross revenues of a movie before its release has more financial significance. Nevertheless, we do daily revenue forecasting for movies here because our purpose is to demonstrate how tweets could be used for predicting product sales and daily or weekly revenue forecasting for products with longer life span could be important for both manufacturers and retailers.

[17] We tried other definitions for $Intention_t$, from 1 day to 6 days. It turned out that 5-day specification gives us the best result, although 4-day specification and 6-day specification produce similar results. One interpretation for this result is that the effect of intention tweets lasts for 5 days on average, that is, when someone says he or she wants to watch a movie on Twitter, he or she is most likely to watch that movie in 5 days after posting that tweet.

[18] Although the total number of tweets on day $t-1$ is highly correlated with the revenue on day t, we think this is mainly because of the fact that $Revenue_t$ is highly correlated with $Revenue_{t-1}$ and the fact that $Revenue_{t-1}$ is highly correlated with the number of tweets on day $t-1$. Hence, it is unlikely that the number of tweets on day $t-1$ has strong predictive power after we incorporate $Revenue_{t-1}$ into the forecasting model. Other types of variable that we could use in the forecasting model are sentiment variables, which reflect people's attitudes toward the movie. Unfortunately, we find that these variables do not change much over time.

Model 5: $$\log(RPT_t) = \beta_0 + \theta \cdot \log(RPT_{t-1}) + \beta_1 \cdot Weekend_t + \beta_2 \cdot Saturday_t + \beta_3 \cdot Days_t + \epsilon_t \quad (5)$$

Model 6: $$\log(RPT_t) = \beta_0 + \theta \cdot \log(RPT_{t-1}) + \beta_1 \cdot Weekend_t + \beta_2 \cdot Saturday_t + \beta_3 \cdot Days_t + \alpha \cdot Intention_t + \epsilon_t \quad (6)$$

Table 6: Description of Variables Used in the Daily Revenue Forecasting Model

Variable	Description and Measure
RPT_t	Revenue per theater on day t
$Weekend_t$	Dummy variable, 1 if day t is Friday, Saturday, or Sunday, 0 otherwise
$Saturday_t$	Dummy variable, 1 if day t is Saturday, 0 otherwise
$Days_t$	Number of days since movie release, 0 if day t is the release day
$Intention_t$	Total number of intention tweets in the past five days

To predict the revenue of a movie on day t, we first used the data of the previous 4 weeks to estimate the coefficients and then predicted RPT_t using the estimated coefficients and the data from day $t-1$. To evaluate the value of new information brought in from tweets, we compared the MAPE of the prediction of the two models for 20 movies, the results of which are summarized in Table 7 and plotted in Figure 3. Figure 4 presents the predicted values of $\log(RPT_t)$ of each model, along with the actual values of $\log(RPT_t)$ of two movies.

On average, including the variable $Intention_t$, a small subset of information from Twitter, reduced the MAPE by 17.5%, from 9.89% to 8.16%. The improvements varied among movies. For some movies (e.g., *The Hangover*, *My Sister's Keeper*), the improvements of Model 5 were marginal, whereas for some other movies (e.g., *The Blind Side*, *Sherlock Holmes*), the improvements were quite significant. There are at least two possible explanations. First, although the improvements for *The Hangover* and *My Sister's Keeper* were very small, the MAPE for these movies were also very small. Hence, it might be difficult to obtain large improvement when the forecasting error of the baseline model is already small. Second, it might be because certain movies are more suitable to forecast with tweet information than others. In fact, even when we forecast daily revenue using only its own lag, the forecasting errors also varied significantly across movies.

Table 7: Forecasting Results of All Movies[1]

movie name	MAPE of Model 5	MAPE of Model 6	Improvement
The Hangover	5.87%	5.64%	0.23%
Land of the Lost	10.25%	8.86%	1.39%
Revenge Of The Fallen	6.17%	4.43%	1.74%
My Sister's Keeper	3.94%	3.73%	0.20%
G.I.Joe [2]	11.99%	9.82%	2.17%
Time Traveler's Wife	8.05%	7.07%	0.98%
Inglourious Basterds	8.12%	6.63%	1.49%
Halloween 2	11.10%	8.23%	2.87%
Couples Retreat	13.04%	11.91%	1.13%
The Blind Side [3]	16.61%	12.13%	4.47%
Ninja Assassin	16.19%	12.71%	3.48%
Invictus	12.22%	12.07%	0.15%
Alvin and the Chipmunks	7.48%	7.46%	0.02%
It's Complicated	6.65%	4.32%	2.33%
Sherlock Holmes	9.25%	5.72%	3.52%
Youth in Revolt	10.99%	9.55%	1.44%
The Book of Eli	9.75%	9.30%	0.45%
Percy Jackson	13.23%	11.82%	1.41%
Wolfman	8.44%	4.48%	3.96%
Shutter Island	8.46%	7.40%	1.06%
Average	9.89%	8.16%	1.73%

[1] For many movies, there are few intention tweets. Considering the noise inherent in the Twitter data and the additional noise introduced by the classification procedures, in this study we used movies that had an average of at least 40 intention tweets per day for the first 7 weeks, which reduced the number of movies to 20. We also tried different thresholds (30 and 50) for the number of intention tweets per day, and the results varied only slightly.

[2] The 31st day after the release of the movie *G.I.Joe* is the Labor Day of 2009, so we set the weekend dummy for that day as 1 to reflect this fact. Controlling this significantly reduced the forecasting error. We thank one of the reviewer for pointing this out. On the other hand, movie sales pattern during the Christmas holiday period is very irregular and we found it very difficult to improve the results by controlling the Christmas holiday effect. To avoid running into data overfitting problem, we did not handle the Christmas holiday effect in our models. This might have resulted in large forecasting errors for movie *The Blind Side* and *Ninja Assassin*.

[3] On the forty-third day of the release of *The Blind Side*, which is January 1, 2010, the film hit $200 million domestically, marking the first time a movie marketed with a sole actress's name above the title (Bullock's) has crossed the $200 million mark. This event triggered a lot of buzz on Twitter. To avoid the bias, we use 42 days rather than 49 days for the movie.

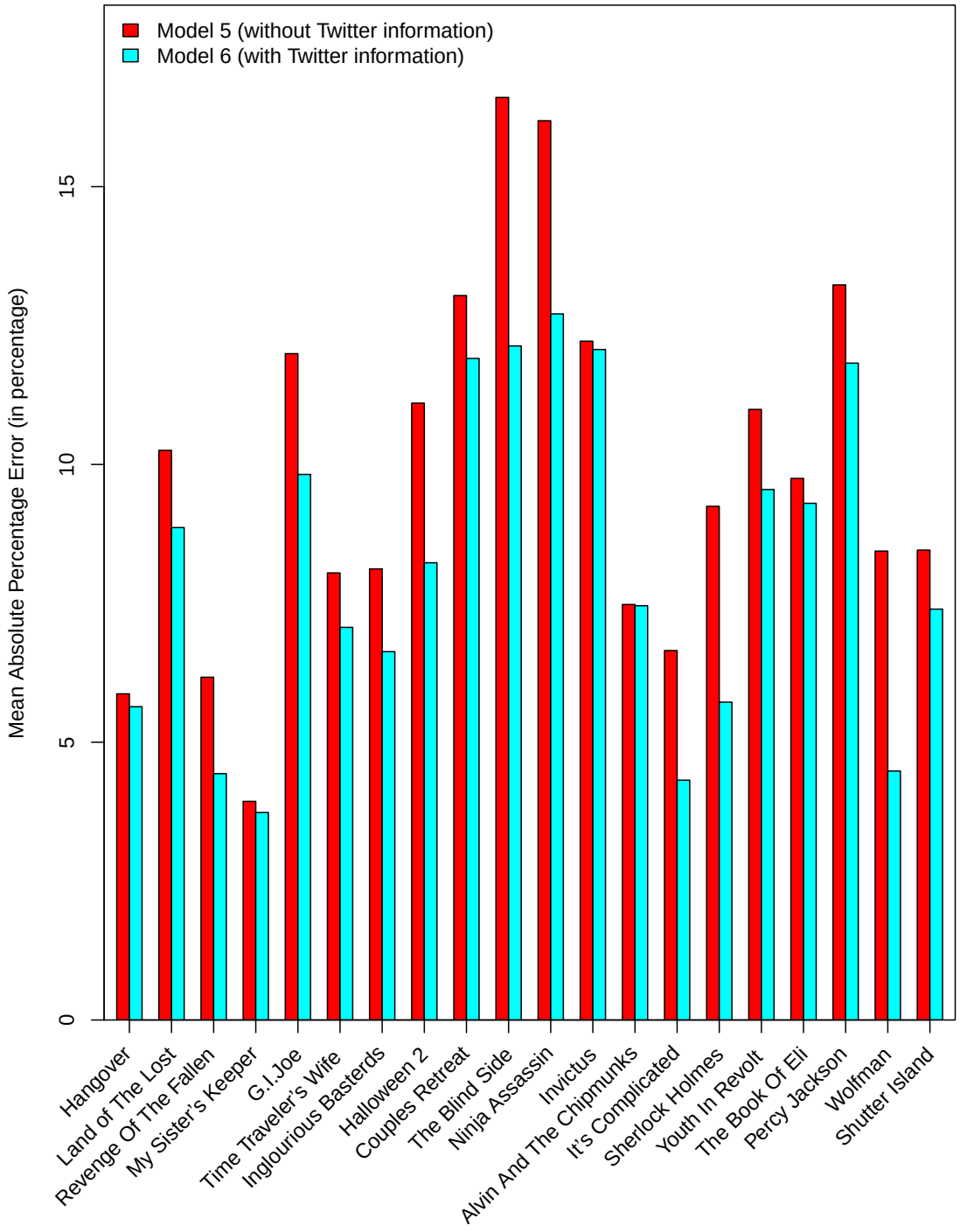


Figure 3: Bar chart comparing the MAPE of Model 5 and Model 6 for each of the 20 movies

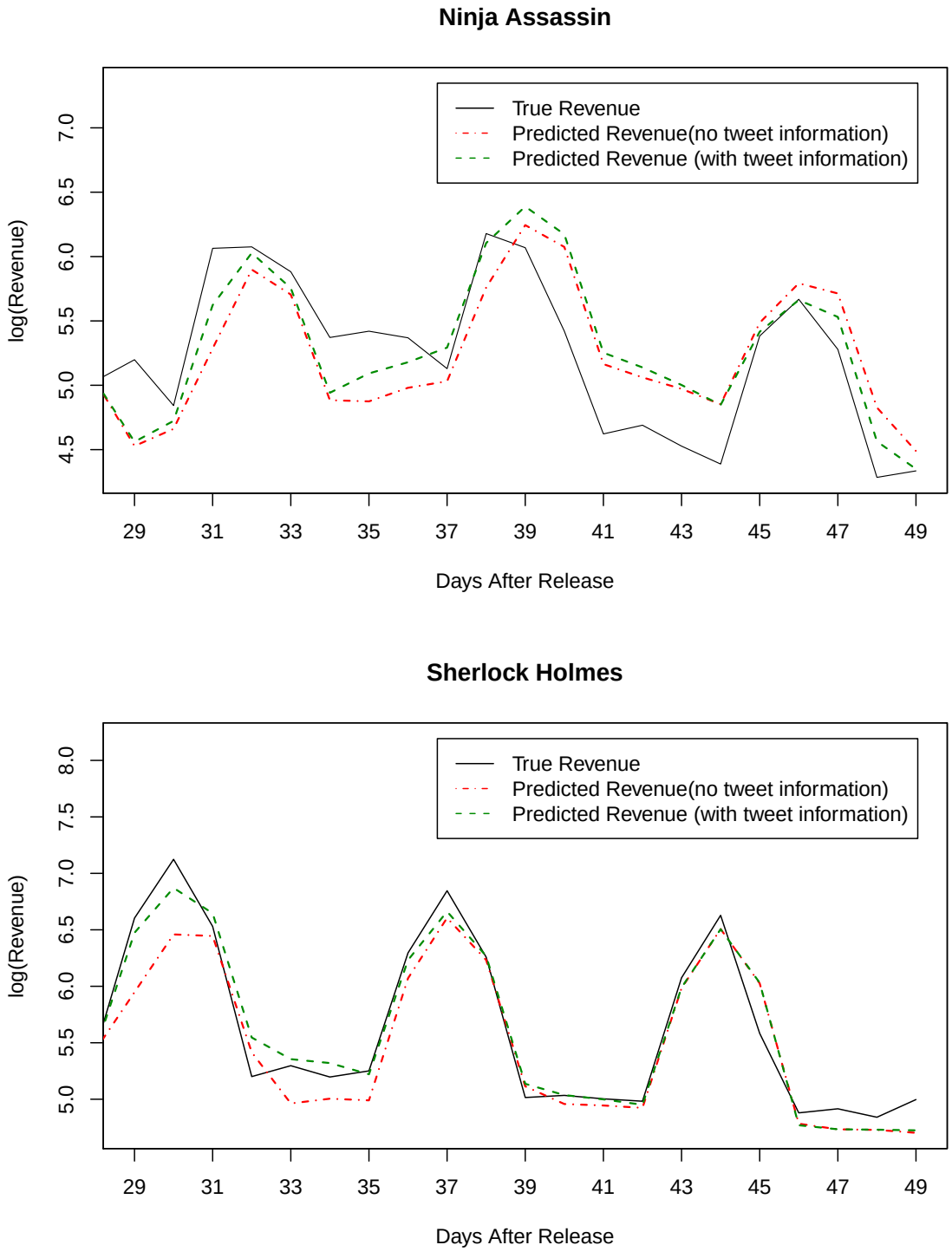


Figure 4: Predicted and actual revenues of two movies: *Ninja Assassin* and *Sherlock Holmes*. Clearly, incorporating information from intention tweets can often adjust the forecast in the right direction, which is why the average MAPE is smaller in Model 6 compared with that of Model 5.

5 Conclusion

As social networks such as Facebook are digitizing the private side of our social life, social broadcasting networks such as Twitter are transforming the public side of our social life. The voices unleashed by social broadcasting networks has profound social, economic, and political impact on today's world. A very natural question from a business perspective is how companies can make use of the huge amounts of real-time information generated and consumed by millions of people through those social broadcasting networks. In the case of Twitter, combining the real-time stream of tweets with structural information about users or tweets makes Twitter a great platform to build BI applications. Indeed, some start-up companies are already trying to grab the business opportunities in this field. For example, using its proprietary system, Fizziology monitors social media buzz from Facebook, Twitter, and blogs to generate meaningful business insights.[19] Although some researchers are starting to use user-generated content (e.g., online product reviews) as a source of BI, to the best of our knowledge, there is no systematic study on using social broadcasting networks to build BI systems. The goal of the present article was to fill the gap by proposing and implementing a framework for social-broadcasting-based BI systems.

Unlike traditional BI systems in which information is pulled from internal data sources, for a Twitter-based BI system, data acquisition and processing are very challenging. For data acquisition, companies need to carefully select a set of keywords that best serves their business purposes. On the other hand, the technique of collecting data is becoming less and less difficult as Twitter keeps refining its API structure. For data processing, the challenge is to understand the information hidden in the huge amounts of text. Developments in text mining in the past two decades have provided us with a variety of useful tools to analyze text. Among these tools, the techniques of sentiment analysis are particularly useful. However, the current state-of-the-art in this field is still far from mature. Moreover, even within a tweet of only 140 characters at most, information is multidimensional. For example, as we have illustrated in the example of forecasting daily box office revenue, people's intentions expressed in the tweets are an important source of information, and it is not yet even studied in the field of sentiment analysis. Despite all of these challenges, we have shown in our movie revenue forecasting examples that there is rich and valuable information on Twitter that can be incorporated into BI systems and that might be used for a variety of business purposes, including improving business forecasting and monitoring certain variables of interest.

Note that although we have used Twitter throughout the article to illustrate the BI framework and to construct a BI system accordingly, the framework and methods proposed presently are not limited to Twitter or to the forecasting of movie sales. Rather, practitioners can easily adapt our framework and methods for the design of other social-broadcasting-based BI systems.

As the popularity of social broadcasting networks keeps growing, and with more research in the field of text mining being done, there will be more and more opportunities for social-broadcasting-

[19] *http://fizziolo.gy/*

based BI systems in the business world. We hope that our work can serve as a first step in this direction and that this article can inspire more researchers in the IS field to start looking at the business implications of social broadcasting networks.

References

ABBASI, A. AND CHEN, H. 2008. CyberGate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly* 32, 4, 811-837.

ASUR, S. AND HUBERMAN, B.A. 2010. Predicting the future with social media. In *Proceedings of the ACM International Conference on Web Intelligence.*

BOLLEN, J., MAO, H. AND ZENG, X. 2010. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1, 1-8.

CHEVALIER, J., AND MAYZLIN, D. 2006. The effect of word of mouth on sales: online book reviews. *Journal of Marketing Research* 43, 3, 345-354.

CHINTAGUNTA, P. K., GOPINATH, S., AND VENKATARAMAN, S. The effect of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets. *Marketing Science*, forthcoming

CHUNG, W., CHEN, H., AND NUNAMAKER, J.F. 2005. A visual framework for knowledge discovery on the web: an empirical study of business intelligence exploration. *Journal of Management Information Systems* 21, 4, 57-84.

CHUNG, W., AND CHEN, H. 2009. Web-based business intelligence systems: a review and case studies. In *Handbooks in Information Systems: Business Computing*, G. Adomavicius and A. Gupta (Ed.), 373-396, Emerald Group Publishing, Bradford, England.

CHUNG, W. AND TSENG, T. 2010. Extracting business intelligence from online product reviews: an experiment of automatic rule-induction. In *Proceedings of the 31st International Conference on Information Systems.*

DANG, Y., ZHANG, Y., AND CHEN, H. 2010. A lexicon-enhanced method for sentiment classification: an experiment on online product reviews. *IEEE Intelligent Systems* 25, 4, 46-53.

DAVIES, P.H.J. 2002. Intelligence, information technology, and information warfare. *Annual Review of Information Science and Technology* 36, 313-352.

DELEN, D., SHARDA, R., AND KUMAR, P. 2007. Movie forecast Guru: a web-based DSS for Hollywood managers. *Decision Support Systems* 43, 4, 1151-1170.

DELLAROCAS, C., ZHANG, X., AND AWAD, N. F. 2007. Exploring the value of online product reviews in forecasting sales: the case of motion pictures. *Journal of Interactive Marketing* 21, 4, 23-45.

DUAN, W., GU, B., AND WHINSTON, A.B. 2008. Do online reviews matter? An investigation of panel data. *Decision Support Systems* 45, 4, 1007-1016.

ELIASHBERG, J. AND SHUGAN, S.M. 1997. Film critics: influencers or predictors? *Journal of Marketing* 61, 2, 38-78.

ELIASHBERG, J., JONKER, J., SAWHNEY, M.S., AND WIERENGA, B. 2000. MOVIEMOD: an implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science* 19, 3, 226-243.

GODES, D., AND MAYZLIN, D. 2004. Using online conversations to study word-of-mouth communication. *Marketing Science* 23, 4, 545-560.

LILIEN, G.L., KOTLER, P., AND MOORTHY, S.K. 1992. *Marketing models.* Prentice Hall, NJ.

LITMAN, B.R. 1983. Predicting success of theatrical movies: an empirical study. *Journal of Popular Culture*, 16, 4, 159-175.

LITMAN, B.R. AND KOHL, L.S. 1989. Predicting success of motion pictures: the '80s experience. *Journal of Media Economics* 2, 2, 35-50.

LITMAN, B.R. AND AHN, H. 1998. Predicting financial success of motion pictures. In B. R. Litman (Ed.), *The Motion Picture Mega-Industry.* Allyn & Bacon, MA.

LIU, Y. 2006. Word of mouth for movies: its dynamics and impact on box office revenue. *Journal of Marketing* 70, 3, 74-89.

LIU, Y., CHEN, Y., LUSCH, R. F., CHEN, H., ZIMBRA, D., AND ZENG S. 2010. User-generated content on social media: predicting market success with online word-of-mouth. *IEEE Intelligent Systems* 25, 1, 75-78.

O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B.R., AND SMITH, N.A. 2010. From tweets to polls: linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*

PANG, B. AND LEE, L. 2008. *Opinion mining and sentiment analysis.* Now Publishers Inc, MA.

ROMERO, D.M., MEEDER, B., AND KLEINBERG, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th ACM International World Wide Web Conference*

RUI, H., WHINSTON, A. B. AND WINKLER, E. 2009. Follow the tweets. *Wall Street Journal*, November 30, 2009.

RUI, H. AND WHINSTON, A.B., 2011. Information or attention? an empirical study of user contribution on Twitter. Forthcoming at *Information Systems and e-Business Management*

SHARDA, R., AND DELEN, D. 2006. Predicting box-office success of motion pictures with neural networks. *Expert System with Applications* 30, 2, 243-254.

SHI, Z., RUI, H., AND WHINSTON, A.B. 2010. Information sharing in social broadcast: evidences from Twitter. In *Proceedings of the 21st Workshop on Information Systems and Economics*

SAWHNEY, M.S AND ELIASHBERG, J. 1996. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science* 15, 2, 113-131.

SONNIER, G. P., MCALISTER, L., AND RUTZ, O. J. 2011. A dynamic model of the effect of online communications on firm sales. *Marketing Science* 30, 4, 702-716.

WIXOM, B.H., WATSON, H.J., REYNOLDS, A.M., AND HOFFER J.A. 2008. Continental airlines continues to soar with business intelligence. *Information Systems Management* 25, 2, 102-112.

WU, S., HOFMAN, J.M., MASON, W.A., AND WATTS, D.J. 2011. Who says what to whom on Twitter. In *Proceedings of the 20th ACM International World Wide Web Conference.*