Multimed Tools Appl (2016) 75:1509–1528
DOI 10.1007/s11042-014-2270-1

Predicting movie Box-office revenues by exploiting large-scale social media content

Ting Liu · Xiao Ding · Yiheng Chen · Haochen Chen · Maosheng Guo

Received: 26 December 2013 / Revised: 19 August 2014 / Accepted: 9 September 2014
Published online: 2 October 2014

Abstract Predicting the box-office revenue of a movie before its theatrical release is an important but challenging problem that requires a high level of Artificial Intelligence. Nowadays, social media has shown its predictive power in various domains, which motivates us to exploit social media content to predict box-office revenues. In this study, we employ both linear and non-linear regression models, which are based on the crowd wisdom of social media, especially the posts of users, to predict movie box-office revenues. More specifically, the attention and popularity of the movie, purchase intention of users, and comments of users are automatically mined from social media data. In our model, the use of Linear Regression and Support Vector Regression in predicting the box-office revenue of a movie before its theatrical release is explored. To evaluate the effectiveness of the proposed approach, a cross-validation experiment is conducted. The experimental results show that large-scale social media content is correlated with movie box-office revenues and that the purchase intention of users can lead to more accurate movie box-office revenue predictions. Both the linear and non-linear prediction models have the advantage of predicting movie grosses in our experiments.

Keywords Movie box-office revenue · Social media · Prediction · Purchase intention mining

T. Liu · X. Ding · Y. Chen · H. Chen · M. Guo
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China

T. Liu
e-mail: tliu@ir.hit.edu.cn

Y. Chen
e-mail: yhchen@ir.hit.edu.cn

H. Chen
e-mail: hcchen@ir.hit.edu.cn

M. Guo
e-mail: msguo@ir.hit.edu.cn

X. Ding (✉)
Harbin Nangang district Jiaohua street No. 29 6th floor, Harbin 151000, China
e-mail: xding@ir.hit.edu.cn

1 Introduction

Film is a high-risk cultural industry. Among approximately 103 movies released during the first half of 2012 in China, only 10 obtained a profit. Given that movie box-office revenue is a direct profit of the film industry, it is an important indicator for measuring the success of a movie [29], [32]. The accurate prediction of movie box-office revenues is highly significant for the reduction of market risk, improvement of the management of the film industry, and promotion of the development of a film-related derivative product market [42], [43]. However, predicting movie box-office revenues is a challenging problem, as it is very difficult to discover the essential reason for the volatility of the movie box-office revenue [29]. With the wide and rapid development of the social media platform, the rich social media data provide new opportunities for the prediction of movie box-office revenues. Hence, it inspires us shed some of our obsession for causality in exchange for simple correlations. By letting us identify a suitable proxy for a phenomenon, correlations allow us to capture the present and predict the future. In this task, social media has the following advantages:

- Volumes of data about movies are available on social media. Movies are widely discussed on social media. According to our statistics, at least 10 million user posts talk about movies per week in the Sina microblog.[1] Therefore, sufficient data is available for the analysis.
- Data on movie box-office revenues are easy to obtain. Income from the first week of a movie and its gross income can be acquired from the MTime[2] in China and the Internet Movie Database (IMDB)[3] in the US.
- Social media content and movie box-office revenues have a clear logical correlation [1]. The user who posts a tweet to express his/her purchase intention for a specific movie indicates his/her interest in the movie and his/her likelihood to watch the movie. The first week pre-release data have the strongest correlation with the gross income than those in any other pre-release time periods [43]. After the movie's release, user posts, especially those with positive or negative sentiments, become a kind of electronic Word of Mouth. It can influence other potential customers [19] and further affect the gross income of the movie.

Using large-scale social media content, our approach seeks to predict movie box-office revenues by mining correlation factors from unstructured texts. Most previous studies predicted the movie gross income based on structured IMDB data analysis of specific characteristics [42], [43], [6], e.g., the number of one-week-old theaters, the rating from the Motion Picture Association of America, director, main actors, movie's genre, budget, and so on, but with somewhat limited success. Nevertheless, recent work [2], [12], [31] has shown the power of social media in predicting financial market phenomenon such as stock price movement, product sales, and financial risk. Asur and Huberman [1] indicated that social media content can effectively predict movie box-office revenues. Their work contains two main assumptions. The first assumption is that movies that are most talked about can be the most watched, while the second assumption is that movies with much Word of Mouth will have high gross revenues. However, after intensive analysis, we find that these two assumptions are not always correct. More frequent mention of a movie does not necessarily mean more positive reviews. In addition, many positive reviews cannot automatically translate to more people watching the

[1] http://www.weibo.com/
[2] http://www.mtime.com/
[3] http://www.imdb.com/

said movie in the cinema. Only a few movies actually have both good reviews and high gross revenues. For example, the movie "Painted Skin: The Resurrection" in China did not receive a high review score (5.7 out of 10) in Douban,[4] a famous Chinese movie review website, but its gross income was more than US$ 117.3 million. Therefore, in this paper, we propose a novel approach in mining the purchase intention of users for a particular movie from social media. For example, the tweet "Want to see 3D Atmos Transformers Age Of Extinction" indicates the user's intention to see the movie "Transformers: Age of Extinction". More specifically, we want to determine the number of users who express their intents to watch a specific movie on social media. We have observed that regardless of how much the movie is discussed and how perfect the movie is rated, the factor that is most related to the movie box-office revenue is how many people are willing to see the movie.

In this paper, our contributions are as follows.

- We provide a comprehensive method for predicting movie box-office revenues using social media data as well as a detailed analysis of movie-related social media data.
- We propose a new task of mining the purchase intention of users from social media, which has not been studied in previous movie gross prediction studies.
- Through large-scale analysis, we prove that social media data are capable of helping people build prediction models with better performance.

We only use one week pre-release and post-release data in the experiments in this study. All predictions are out-of-sample predictions. In practice, our approach provides a feasible and more accurate estimation regarding the investment worthiness of some pre-release investors and almost all post-release investors.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 provides the formalized definition of our problem. Section 4 introduces the details of our methodology in extracting information from social media. Section 5 presents the prediction models. Sections 6 and 7 provide the experimental setups and experimental results. Finally, Section 8 concludes the paper and discusses its limitations and further research directions.

2 Related work

2.1 Social media based prediction

The previous decades had brought explosive growth of the social media, especially online social networks [40]. Such networks aggregate the feelings and opinions of diverse groups of people at low cost. Mining large-scale social media content provides us an opportunity to discover rules on social and economic functions, qualitatively and quantitatively analyze user intention, and predict future human-related events.

Social media-based prediction has been widely studied [17], [26], [33]. The main motivation of these studies is the acquisition of large-scale user information (e.g., comments and opinions) from social media at low cost. For example, conventional pre-election polls can be accomplished via telephone surveys. However, performing these surveys is costly. As a newly emerging information platform, web surveys through social media provide an opportunity to accomplish the same task at low cost. Though a very simple method is employed, by using a

[4] http://movie.douban.com/

number of related social media content, a successful prediction result is still obtained. For example, Williams and Gulati [41] successfully predicted the result of the 2008 US presidential election based on the number of Facebook supporters. O'Connor et al. [22] also showed that public sentiment can be helpful for predictions.

In addition to election result prediction, Bollen and Zeng [2] studied the correlation of the large-scale collective emotion of Twitter users with the volatility of the Dow Jones Industrial Average (DJIA). Their experimental results showed that changes in the public mood along specific mood dimensions match shifts in DJIA values three to four days later. Ritterman et al. [25] extracted H1N1 (Swine Flu)-related information from social media and then demonstrated the public belief about the possibility of a pandemic. UzZaman et al. [38] also utilized social media data to help predict game outcomes for the 2010 FIFA World Cup tournament. Bothos et al. [4] combined social media information with prediction markets to derive actionable information and developed a system for predicting Oscar movie awards based on the wisdom of the crowds.

Though social media-based prediction has received considerable attention in recent years, a debate is also ongoing on whether current social media data are effective for predicting real-world outcomes. For example, in the 2001 provincial election in British Columbia, the number of mentions on Internet message boards does not indicate the relative strength of parties [13]. This phenomenon can be explained by the following reasons. First, on Twitter, 4 % of all users are responsible for more than 40 % of the content [37]. Second, social media does not reflect the demographics of the society. In terms of age, statistics in the US in 2000 recorded 36 % were between 18 and 24, 50 % between 25 and 34, and 68 % were over 35 [20]. However, on Twitter, more than 60 % of users are under 24 [35]. Thus, random sampling on social media is biased. Moreover, determining the age of social media users is difficult because the profiles of users are confidential. Therefore, statistically unbiased sampling in terms of age, and similar other attributes, such as region and ethnicity, on social media is impossible. In addition, basic natural language processing (NLP) techniques (e.g., segmentation, POS tagging, and sentiment analysis) does not achieve comparable performance when used on social media texts [20], [11]. One possible explanation is because the vocabulary used in most NLP systems is designed for well-written and standard text rather than for short posts on social media [22], [14].

According to the above analysis, we stress that the topic of social media-based prediction should be carefully selected. Social media is useful for prediction on the following two aspects. On one hand, social media is a data collection platform. For example, if a group of people in particular area post "I have a cold" on Twitter, this area is likely to be experiencing an epidemic. On the other hand, social media is a collective wisdom platform. For example, social media users prefer to post their own predictions (e.g., "I believe Obama can win the election" and "I think Brazil will win the World Cup"). These predictions are untapped collective wisdom on social media. This paper focuses on the prediction of movie box-office revenues, and the reasons behind the predictions are discussed in Section 1. Asur and Huberman [1] were the first to predict movie box-office revenues based on social media. They counted the number of mentions and referred to the positive or negative comments of users regarding a movie to predict movie box-office trends. However, before a movie is released, film reviews are not usually available. Therefore, movie box-office revenues cannot be predicted based on user comments during the first week. This paper carefully analyzes user posts on social media and initially proposes mining the purchase intention of users for movies from social media (i.e., the number of people who express their intention to watch a specific movie) to predict movie box-office revenues. We notice that users tend to express their purchase intention for movies before they are released.

2.2 Movie box-office revenue prediction

A considerable amount of prior research has studied the problem of movie gross prediction from different perspectives [18], [34], [27], [30]. Most previous work [32], [42], [6] had presented forecasts on movie box-office revenues based on IMDB data using regression or stochastic models. However, recent studies have explored the incorporation of other information sources in prediction models. Zhang and Skiena [43] used the combination of IMDB data and news data to predict movie box-office revenues. Joshi et al. [16] used the text from the reviews of film critics from several sources to predict opening weekend revenues and showed that text from reviews can substitute metadata during prediction. Sharda and Delen [29] regarded the prediction problem as a classification problem rather than a problem that involves forecasting the point estimate of box-office receipts and used neural networks to classify movies into categories from "flop" to "blockbuster".

Moreover, substantial interest has been shown in using movie reviews as a domain to test sentiment analysis methods, e.g., [5], [24]. The movie reviews of users become a type of Word of Mouth, which influences other potential customers. Opinion comes in many types: positive, negative, and neutral mixed. Novel techniques in sentiment analysis allowed the aggregate level quantification of positive versus negative mentions with reasonable accuracy. Pang and Lee [23] provided a detailed review in this domain. Mishne and Glance [21] showed that movie sales have some correlation with movie sentiment references, but the researchers neither built prediction models nor showed the value of the correlation because they think the result is not sufficient for accurate modeling. Asur and Huberman [1] demonstrated how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.

In recent years, with the rapid development of social media, big data present on social media have attracted considerable attention. As introduced in Section 2.1, social media-based prediction has its amazing power in many application areas. The emergence of big data on social media allows the prediction method not to explore the causality relationship and instead to focus on discovering and utilizing correlation. Viktor Mayer-Schönberger, in his book "Big Data: A Revolution That Will Transform How We Live, Work, and Think" stressed that by deriving a good phenomenon-related factor, correlation can help us capture the present and predict the future. Correlation is very useful not only because it presents a new perspective but also because all of these perspectives are clear. Therefore, this paper uses a new perspective to study movie box-office revenue prediction. Moreover, mining the collective wisdom on social media is expected to improve the accuracy of prediction.

2.3 Problem formulation

The goal of this paper is to study the feasibility of analyzing and predicting movie box-office revenues using large-scale social media content. Given social media text d, we predict the value of a continuous variable v (e.g., movie box-office revenues in this paper). We accomplish this task via a parameterized prediction function f

$$\widehat{v} = f(\mathbf{d}; \mathbf{x}) \tag{1}$$

where $x \in \mathbb{R}^d$ are the parameters. Our approach is to learn a human-interpretable x from a collection of N training examples $\{\langle d_i, v_i \rangle\}_{i=1}^{N}$, where each d_i represents a user post and each $v_i \in \mathbb{R}^d$.

Function (1) shows that the task comprises two problems.

- Information extraction problem. More specifically, we first pre-process the natural language of user posts on social media and then extract correlated factors, such as the attention and popularity of movies, the positive or negative comments of the users and the purchase intention of the users, from the processed texts (Section 4).
- Prediction model construction problem. More specifically, we seek to find the applicable prediction function f to learn from the training data, and then produce accurate movie box-office revenues. This paper adopts linear and non-linear models to address the problem of box-office prediction (Section 5).

As shown in Fig. 1, this paper extracts three categories of information from social media text: purchase intention of users for a movie, attention and popularity of a movie, and positive or negative user comments for a movie. The textual information is represented as features for two prediction models (i.e. Linear Regression model and Support Vector Regression model). Then, we perform a detailed experiment analysis. We introduce each component of the system in detail in the following sections.

3 Information extraction from social media

3.1 Purchase intention mining

3.1.1 Problem statement

Purchase intention can be defined as the intention of an individual to buy a specific product or service. Users tend to explicitly or implicitly express their purchase intention on social media. As shown in Fig. 2, some users post their feelings on social media. Examples include "I CANT WAIT TO SEE CATCHING FIRE" and "really excited to watch the hunger games: catching fire, this weekend", which express user intents to watch the movie "The Hunger Games: Catching Fire". However, although some user posts mention the title of the movie, users do not express their intents to watch that movie in the posts, such as "Everyone's talking about Catching Fire and I'm just like: umm is it Sunday yet". The task of purchase intention mining can be viewed as a binary classification problem. Given a movie title, we first collect tweets

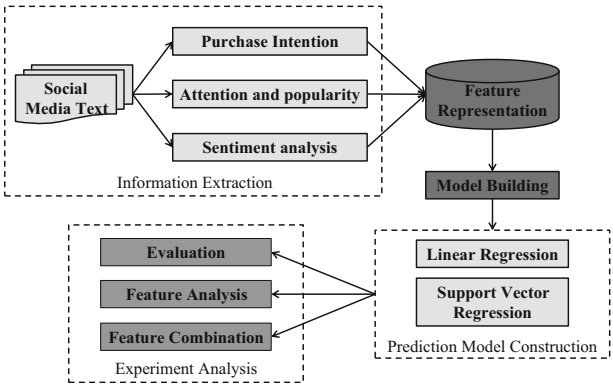


Fig. 1 System Architecture

Fig. 2 Purchase intention examples

that mention the title of the movie from social media content, and then classify those tweets into two categories, namely, containing and not containing users' purchase intention.

3.1.2 Purchase intention mining based on SVM

In machine learning, support vector machine (SVM) is a kernel-based learning algorithm introduced by Boser et al. [3] and Vapnik [39]. SVM was first applied on classification tasks and was later adopted for regression tasks. Predominantly, SVM employs "kernel tricks" for projection of non-linear separable training data onto a high dimensional feature space by preserving dimensions of relatedness in the data. In a classification scenario, SVM then obtains the maximum-margin hyperplane as the decision boundary is pushed by support vectors. Thus, global optimal solutions can be extracted regardless of the sparsity of the training data and become less overfitted. In application scenarios, feature selection is very important for classification performance. This paper selects six feature categories for the task of purchase intention mining. The details are shown in Table 1.

The following sections will introduce each feature in detail:

(1) Bag-of-words feature

First, we remove stop words from all collected social media texts and construct a vocabulary using the information gain approach. The bag-of-words feature is then generated for each user post based on the TF-IDF weighting function.

(2) Mention feature

In microblogs, users can use the "@" symbol followed by the username to remind their friends to see this tweet. An example of a tweet which contains "@" symbol is as follows:

"The Hunger Games -Catching Fire. Can we see the third one now? @wkucab"

By investigating tweet corpora that contain purchase intention for movies, we found that most tweets contain the "@" symbol because when users want to watch a movie, they tend to invite their friends to watch with them. Therefore, the mention feature can help determine whether a tweet contains the purchase intention or not. If a tweet mentions other people, we set its mention feature as "true".

Table 1 Features for purchase intention mining

Feature Symbol	Feature content
B	Bag of words
M	Mention
U	URL
E	Emoticon
L	Length of tweet
T	Trigger word

(3) URL feature

The length of a tweet is limited to 140 characters. Thus, to express more information in a tweet, the user can employ other ways of increasing the quantity of information. For example, the URL of a website or photos can be embedded in a tweet. By closer observation, most regular users do not add URL links in their tweets. However, some spammers usually embed URL links in their tweets such as the following examples:

"Get ready for The Hunger Games: Catching Fire & download the star-studded soundtrack on iTunes today! http://smarturl.it/CatchingFireDlxiT?IQid=gm.twt.src …"

"Seeing @TheHungerGames #CatchingFire this weekend? Get showtimes & tix here: http://goo.gl/a6k6WA pic.twitter.com/7ayiX5EHcl"

As shown in the above examples, advertisement and sales promotion tweets often provide URL links in their content. This phenomenon is useful for our classification problem. If a tweet contains a URL link, we set its URL feature as "true".

(4) Emoticon feature

Emoticons are widely used on social media. Users prefer to use emoticons to express their emotions. If a user posts a tweet with both a positive emoticon and the movie title, the tweet indicates that the user may want to watch that movie. We will set the emoticon feature as "true" when we detect tweets containing emoticons.

Moreover, if we can classify emoticons into different categories according to their emotion, better classification performance can be achieved. This classification can be handled in future work.

(5) Length of the tweet feature

Statistically, tweets that contain purchase intention are found to be not too lengthy because users tend to use concise language to express their intents, such as "I CANT WAIT TO SEE CATCHING FIRE TOMORROW" and "I'm going to go watch catching fire tonight". However, the length of advertisements and news tweets is longer.

We set the length threshold as 30 characters. If the length of the tweet is longer than the threshold, we set this feature as "true".

(6) Trigger word feature

Social media users usually use some specific words to express their purchase intention for movies, such as the following sentences:

"I want to watch catching fire"

"I'm ready to go see Catching Fire today"

In the above examples, "be ready to go see" and "want to watch" express the purchase intention of users for movies. We manually collect these words in a word list and name these words as trigger words. If a tweet contains trigger words, we set its trigger word feature as "true". In this paper, we carefully select 42 words as trigger words.

3.1.3 Attention and popularity

We are interested in studying the generation of attention and popularity for movies on social media and the effects of this attention on the real-world performance of the movies considered. To use a quantifiable measure on the tweets, we define the post-rate as the number of tweets that refer to a particular movie.

$$Post - rate = \frac{|N_{total}|}{|Time\,window\,size|} \tag{2}$$

The generating rate can be estimated differently according to the size of the time windows, such as hourly, daily, or weekly. The higher the posts generating rate of a movie, the more people are interested in it, and the topic is more attractive. Previous studies showed that the daily generating rate before release is a better predictor for movie box-office revenues [1]. Hence, we set the size of the time window as one day in this paper.

Notably, some movies that were released during the period considered were not used in this study because correctly identifying tweets that are relevant to those movies is difficult. For example, for the movie Starry Night, segregating tweets that discuss on the movie from those referring to the famous painting by Vincent van Gogh is very difficult. We have ensured that the data we have used are disambiguated and clean by manually choosing appropriate keywords.

3.1.4 Sentiment analysis

In this section, we investigate the importance of sentiments in predicting movie box-office revenues. The attention can effectively predict opening week box-office revenues for movies. However, prior to the release of the movie, movie review data are available. We consider the problem of utilizing the sentiment analysis techniques for forecasting movie grosses.

Sentiment analysis is a well-studied problem in the NLP community, with different classifiers and language models employed in earlier studies [23], [8]. Sentiment analysis is commonly viewed as a classification problem where a given text is labeled as Positive, Negative, or Neutral. In this study, we construct a sentiment analysis classifier based on a sentiment lexicon (a list of positive and negative sentiment words, e.g., “like” and “hate”). The sentiment lexicon is obtained from the Harbin Institute of Technology in China.

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

$$Sent - rate = \frac{N_{positive} - N_{negative}}{N_{total}} \tag{3}$$

$N_{positive}$, $N_{negative}$, and N_{total} are the number of positive tweets, negative tweets, and total tweets, respectively. The sentiments index is proven to have a strong correlation with the financial market and is useful in the prediction of real-world outcomes [2], [22].

4 Prediction model

4.1 Linear regression model

We use linear regression (LR) to directly predict movie box-office revenues denoted as v based on features x extracted from the movie metadata and the text of the social media. That is, given

an input feature vector $x \in \mathbb{R}^d$, we predict output $\hat{v} \in \mathbb{R}$ using a linear model $\hat{v} = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}$. To learn values for the parameters $\boldsymbol{\theta}=\langle\beta_0,\boldsymbol{\beta}\rangle$, the standard approach is to minimize the sum of the squared errors for a training set containing n pairs $\langle \boldsymbol{x}_i, v_i\rangle$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $v_i \in \mathbb{R}$ for $1 \le i \le n$.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}=\langle\beta_0,\boldsymbol{\beta}\rangle}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}\left(v_i - \left(\beta_0 + \boldsymbol{x}_i{}^T\boldsymbol{\beta}\right)\right)^2 + \lambda P(\boldsymbol{\beta}) \tag{4}$$

A penalty term $P(\boldsymbol{\beta})$ is included in the objective for regularization. Classical solutions use L_1 and L_2 norms, known respectively as ridge and lasso regression. Recently, a mixture of the two has been introduced and called the elastic net [44].

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{n}\left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right) \tag{5}$$

where $\alpha \in (0,1)$ determines the trade-off between L_1 and L_2 regularization. For our experiments, we use the elastic net and specifically, the *glmnet* package, which contains an implementation of an efficient coordinate ascent procedure for training [10].

4.2 Support vector regression model

Support vector regression (SVR) [9] is a well-known method for training a regression model. SVR is trained by solving the following optimization problem:

$$\min_{x\in\mathbb{R}^d}\frac{1}{2}\|\mathbf{x}\|^2 + \frac{C}{N}\sum_{i=1}^{N}\max(0, |v_i - f(d_i;\mathbf{x})| - \varepsilon) \tag{6}$$

where C is a regularization constant and ε controls the training error. Given the embedding h of tweets in $\mathbb{R}^d$, ε defines a "slab" (region between two parallel hyperplanes, sometimes called the "ε -tube") in $\mathbb{R}^{d+1}$ through which each $\langle h(d_i), f(d_i;x)\rangle$ must pass to have zero loss. The training algorithm obtains parameters X that define a function f minimizing the (regularized) empirical risk.

Let h be a function from the tweets into some vector-space representation $\subseteq\mathbb{R}^d$. In SVR, the function f takes the following form:

$$f(\mathbf{d};\mathbf{x}) = h(\mathbf{d})^T\mathbf{x} = \sum_{i=1}^{N}\alpha_i K(\mathbf{d}, d_i) \tag{7}$$

where Eq. (7) re-parameterizes f in terms of a kernel function K with "dual" weights α_i (i=1…N). K can be seen as a similarity function between two tweets. During testing, a new example is compared with a subset of the training examples (those with $\alpha_i \neq 0$). With SVR, this set is typically sparse. With the linear kernel, the primal and dual weights relate linearly.

$$\mathbf{x} = \sum_{i=1}^{N}\alpha_i h(d_i) \tag{8}$$

Full details of SVR and its implementation, which are described in the study of Scholkopf and Smola [28], are not provided in this paper. SVMlight [15] is a free, available implementation of SVR training that we use in our experiments.

5 Experiment setup

5.1 Dataset

Two kinds of movie data are used in this paper: movie-specific variables and movie-related tweets data. Movie-specific variables from November 2011 to January 2012 are collected from the popular movie website Wangpiao[5] in China. The data include the first week and gross income of 57 movies. Social media data are obtained from Sina microblog, which contain 1.1 billion text contents from November 2011 to January 2012. To ensure that we obtain all tweets that refer to a particular movie, we use keywords present in the movie title as search arguments. Consequently, we collected 5 million tweets.

Given that no public corpus for the task of purchase intention mining is available, this paper manually constructs an annotated dataset. To justify the effectiveness of our method, we carefully conduct user studies into the corpus. For each tweet in the data, two annotators are asked to label whether the tweet contains the purchase intention for a specific movie or not. The agreement between our two annotators, measured using Cohen's kappa coefficient [7], is substantial (kappa=0.85). We ask the third annotator to adjudicate the classified data on which the former two annotators disagreed upon. The annotated dataset contains 2,300 tweets, where 1,600 are used as the training set and the remaining data are used as test data.

5.2 Evaluation measure

We adopt traditional *Precision, Recall*, and *F-Measure* to evaluate our approach of purchase intention mining. The evaluation functions are as follows:

$$Precision = \frac{|correct\ tweets|}{|tweets\ identified\ by\ our\ approach|} \tag{9}$$

$$Recall = \frac{|correct\ tweets|}{|the\ whole\ tweets|} \tag{10}$$

$$F - Measure = \frac{2 \cdot Precision\ Recall}{Precision + Recall} \tag{11}$$

We use the coefficient of determination (adjusted R^2) and Relative Absolute Error (RAE) to evaluate the regression models. The use of adjusted R^2 is an attempt to consider the phenomenon of the automatically and spuriously increasing R^2 as extra explanatory variables are added to the model. Theil [36] had modified R^2 as a factor that adjusts for the number of explanatory terms in a model relative to the number of data points. Unlike previous R^2, the adjusted R^2 increases at the inclusion of a new explanator but only if the new explanator improves the R^2 more than as expected in the absence of any explanatory value being added by the new explanator. The adjusted R^2 is defined as $R^2_{adj} = 1 - \frac{n-1}{n-k-1}\left(1-R^2\right)$, where k is the total number of regressors in the linear model (not counting the constant term), and n is the sample size. The relative absolute error is very similar to the relative squared error because the relative

[5] http://www.wangpiao.com/

Table 2 Experimental result of feature selection

Feature selection	Precision	Recall	F-measure
B	0.40	0.65	0.50
B+M	0.43	0.65	0.52
B+M+U	0.48	0.63	0.54
B+M+U+E	0.52	0.67	0.59
B+M+U+E+L	0.55	0.70	0.62
B+M+U+E+L+T	0.72	0.80	0.76

absolute error is also relative to a simple predictor, which is simply the average of the actual values. Mathematically, the relative absolute error (RAE) is defined as $RAE = \sum_{i=1}^{n} |\widehat{y}_i - y_i| / \sum_{i=1}^{n} |y_i - \overline{y}_i|$, where $\widehat{y}_i$ is the estimation value, y_i is the actual value, and $\overline{y_i}$ is the predicted value.

5.3 Baseline

By carefully studying previous research, we find that only Asur and Huberman from HP laboratories [1] have attempted to predict movie box-office revenues based on social media. Specifically, by using the rate of chatter from almost 3 million tweets from the popular site Twitter, Asur and Huberman constructed a LR model to predict box-office revenues prior to the release of the movie. Given that no public test corpus for movie box-office revenues prediction is available, this study adopts the approach of Asur and Huberman and used the LR model to predict and extract the following features: the number of tweets that refer to a particular movie per hour (*tweet-rate*), the number of theaters in which the movies are released (*thcnt*) for first week income prediction, the ratio of positive to negative tweets (*PNratio*), tweet-rate, and thcnt for the gross income prediction. We carry out the experiment on the dataset introduced in Section 6.1 and use the approach by Asur and Huberman as our baseline system.

6 Experimental results and analysis

The goal of this paper is to predict the first week income and gross income of the movie box-office. The experiments adopt the 5-fold cross validation and all of the extracted data are from the one-week pre-release and post-release of the movie.

6.1 Experimental results of purchase intention mining

To verify the effectiveness of feature selection for purchase intention mining, we use the incremental approach to constantly add features into the experiments. First, we obtain the initial experimental result by using bag-of-words feature (B), and then add mention (M), URL (U),

Table 3 Comparison with the baseline system

	First week		Gross	
	R^2_{adj}	RAE	R^2_{adj}	RAE
Baseline	0.89	0.41	0.64	0.78
Our approach	0.94	0.34	0.67	0.73

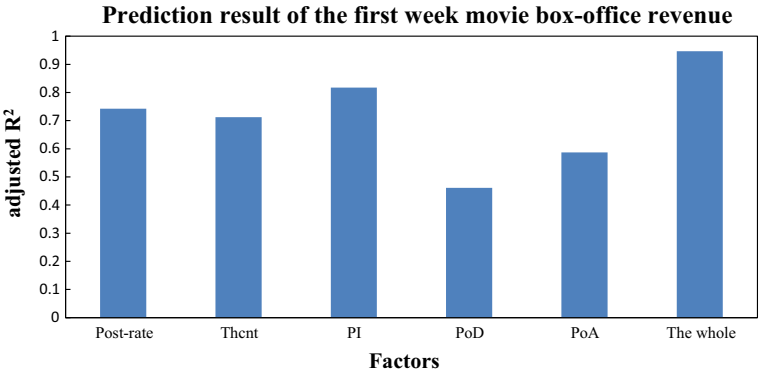


Fig. 3 Analysis of factors (first week income)

emoticon (E), length of text (L), and trigger word (T) features one after the other. The experimental results are shown in Table 2.

Table 2 shows that the performance of classification is improved by the continuous addition of features into the classifier. The experimental result verifies the effectiveness of our features for the task of purchase intention mining. By a closer investigation, we find that the trigger word feature (T) contributes to the maximum improvement in performance because if a tweet is very short and if it simultaneously contains purchase intention trigger words (e.g., "want to watch") and the movie title, the tweet may express the purchase intention of the user for the movie.

6.2 Prediction results

Table 3 lists our experimental results that are compared with those of the baseline system. Our approach achieves better performance on the test dataset (bigger R^2_{adj} value and smaller RAE value).

Note that our approach, as shown in Table 3, is obtained by using the LR model (similar to that of the baseline) but with the following differences. We first propose mining the purchase intention of users for movies on social media to predict movie box-office revenues. Alternatively, the baseline system only uses the popularity of the movie to predict movie box-office revenues. The experimental result shows that our proposed new feature can significantly improve the performance of the prediction model. This finding is consistent with our assumption that when more people want to watch a movie, the higher revenues gained by the movie box-office. The popularity of the movie, however, cannot directly reflect the number of people who want to watch the movie.

This paper also investigates the effectiveness of each factor in predicting movie box-office revenues. We use post-rate, the number of theaters in which the movies are released (thcnt), purchase intention (PI), sentiment analysis (SA), the popularity of the director (PoD),[6] and the popularity of the main actors (PoA).[7] The experimental results are shown in Figs. 3, 4, 5, and 6.

Figures 3 and 4 indicate that when we use each feature to predict movie box-office revenues, purchase intention can achieve the best performance. Fig. 3 demonstrates that purchase intention (PI) achieves better performance than post-rate, and Fig. 4 indicates that

[6] The baidu trends of the director
[7] The baidu trends of the main actors

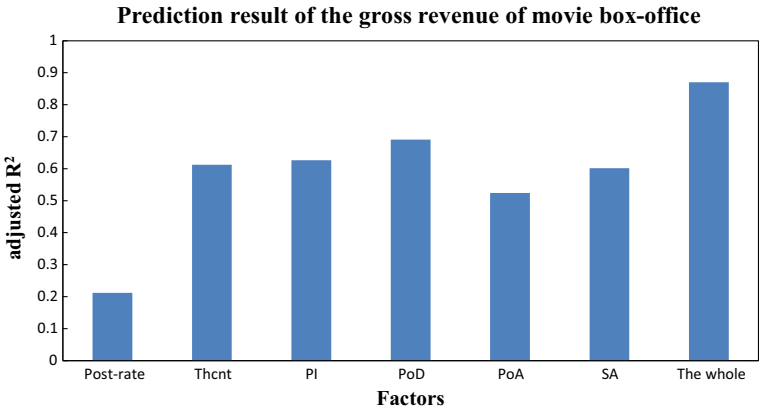


Fig. 4 Analysis of factors (gross income)

purchase intention outperforms sentiment analysis. The experimental results verify our assumption that purchase intention is a better indicator for predicting movie box-office revenues than the popularity of the movie and the sentiment analysis of the movie. Furthermore, if only the popularity of the director and of the main actors on social media is used, a good experiment result cannot be obtained. Having many superstars in a movie does not necessarily mean that the movie will receive more revenues. Users tend to be more rational when they choose films to watch. In addition, although each factor does not lead to an ideal performance, integration of these factors achieves the best performance. These factors therefore reinforce each other and can reflect the movie box-office trends from different perspectives. For example, post-rate and purchase intention can reflect the will of users to watch the movie. Thcnt includes the expectation of experts regarding movie box-office revenues. The popularity of the director and of the main actors affects the ability of the movie to attract audiences. When we combine these factors, we can obtain a better performance.

In addition to studying each individual factor, this paper also considers a combination of different factors. Experimental results are shown in Figs. 5 and 6.

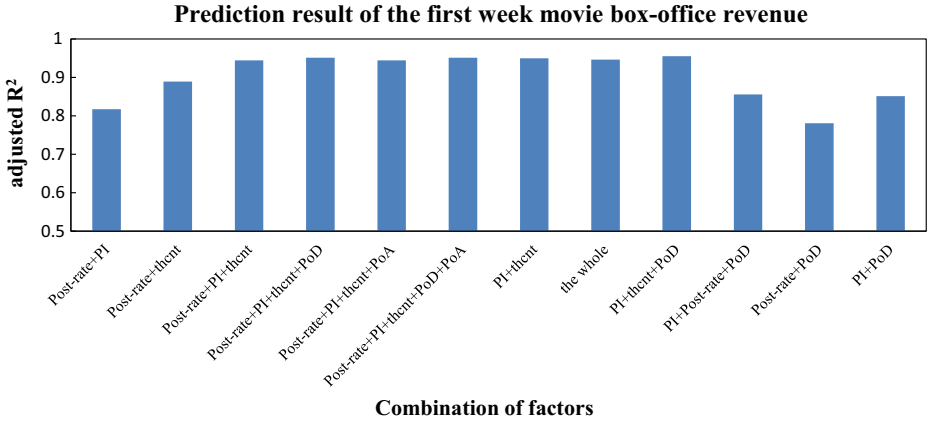


Fig. 5 Analysis of the combination of factors (first week income)

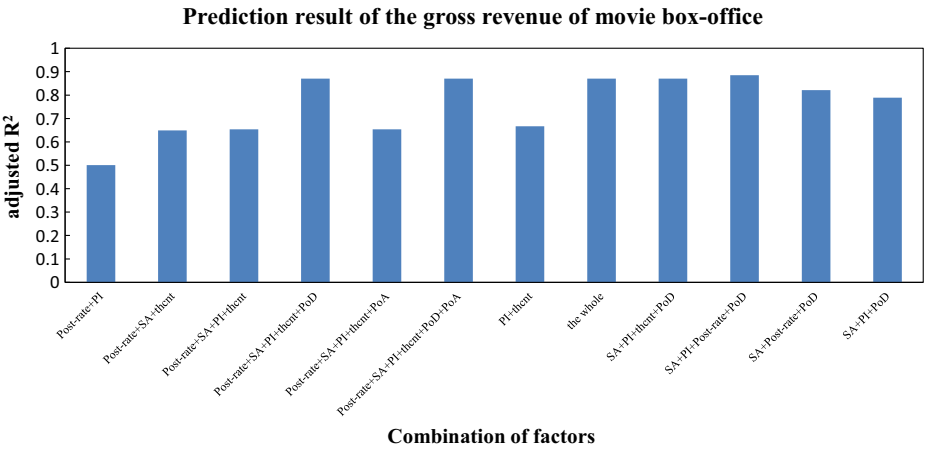


Fig. 6 Analysis of the combination of factors (gross income)

Figures 5 and 6 illustrate that the combination of purchase intention (PI), sentiment analysis (SA), the number of theaters in which the movies are released (thcnt), and the popularity of the director (PoD) achieves the best performance. These four factors support four different perspectives in predicting movie box-office revenues. The combination of post-rate and purchase intention achieves the worst performance because these two factors are similar and the higher the intent to purchase, the more number of times is the movie mentioned on social media. Moreover, Figs. 5 and 6 also show that we cannot achieve the best performance when all factors are used together. We should therefore be careful in selecting factors. Combining some factors may not provide a positive effect, but rather decreases the performance.

6.3 Experiments of different prediction models

In addition to the LR model, this paper also adopts the SVR model with rational basis function (RBF) and linear kernels. The experimental results are shown in Table 4.

Table 4 shows that both the LR and SVR (linear kernel) models achieve better performance than the SVR (RBF kernel) model in the first week movie box-office prediction and worse performance in the gross income prediction. We analyze that the one-week pre-release data have the strongest linear correlation with the first week income, hence the LR and SVR (linear kernel) models can achieve better performance. However, with new influential factors, as well as some unanticipated events, the data do not have a strong linear correlation with gross

Table 4 Experimental result of different prediction models

	First week		Gross	
	R^2_{adj}	RAE	R^2_{adj}	RAE
LR model	0.94	0.34	0.67	0.73
SVR model (linear kernel)	0.95	0.28	0.70	0.71
SVR model (RBF kernel)	0.87	0.55	0.74	0.69

income. Thus, the LR model is less powerful than the SVR (RBF kernel) model. The combination of the linear and non-linear prediction models will be examined in future work.

7 Conclusions and future work

In this paper, we have shown how social media can be utilized to forecast future outcomes. Specifically, using more than 5 million tweets collected from the Sina microblog, we constructed LR and SVR models to predict the box-office revenues of movies prior to their release. We then showed that our results outperformed those of the baseline systems. A strong correlation was found between the purchase intention for movies and movie box-office revenues.

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

Acknowledgment Ting Liu: model building, experiment design, paper writing
Xiao Ding: model building, experiment design, paper writing
Yiheng Chen: model building, experiment design
Haochen Chen: data collection, experiment design
Maosheng Guo: data collection

References

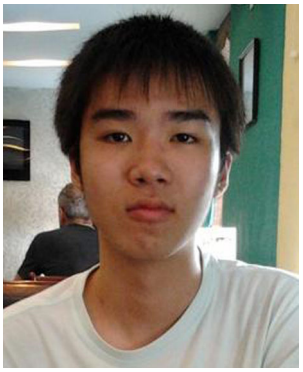
1. Asur S, Huberman BA (2010) Predicting the future with social media [C]//Web intelligence and intelligent agent technology (WI-IAT), 2010. IEEE/WIC/ACM international conference on IEEE 1:492–499
2. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market [J]. J Comput Sci 2(1):1–8
3. Boser B E, Guyon I M, Vapnik V N. (1992) A training algorithm for optimal margin classifiers [C]// Proceedings of *the fifth annual workshop on Computational learning theory*. ACM, 144–152
4. Bothos E., Apostolou D., Mentzas G. (2010) Using Social Media to Predict Future Events with Agent-Based Markets. *IEEE Intelligent Systems*, vol. PP, no. 99.
5. Chaovalit P, Zhou L. (2005) Movie review mining: A comparison between supervised and unsupervised classification approaches [C]//System Sciences, 2005. HICSS'05. Proceedings of *the 38th Annual Hawaii International Conference on. IEEE* 112c-112c
6. Chen A (2002) Forecasting gross revenues at the movie box office [J]. University of Washington, Seattle
7. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
8. Ding X, Liu B, Yu P S. (2008) A holistic lexicon-based approach to opinion mining [C]//Proceedings of *the 2008 International Conference on Web Search and Data Mining*. ACM, 231–240.
9. Drucker H, Burges CJC, Kaufman L et al (1997) Support vector regression machines. J Adv neural inf Process Syst 9:155–161
10. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent [J]. J Stat Softw 33(1):1
11. Gayo-Avello D, Metaxas P T, Mustafaraj E. (2011). Limits of electoral predictions using twitter [C]//*ICWSM*.
12. Gruhl D, Guha R, Kumar R, et al. (2005) The predictive power of online chatter [C]//Proceedings of *the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM, 78–87
13. Jansen H J, Koop R. (2006) Pundits, ideologues, and the ranters: The British Columbia election online [J]. *Canadian Journal of Communication*, 30 (4)
14. Jansen BJ, Zhang M, Sobel K et al (2009) Twitter power: tweets as electronic word of mouth [J]. J Am Soc Inf Sci Technol 60(11):2169–2188

15. Joachims T (1999) Making large scale SVM learning practical [J]
16. Joshi M, Das D, Gimpel K, et al. (2010) Movie reviews and revenues: An experiment in text regression [C]//Human Language Technologies: The 2010 Annual Conference of *the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 293–296
17. Jungherr A, Jürgens P, Schoen H (2012) Why the pirate party won the German election of 2009 or the trouble with predictions: a response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: what 140 characters reveal about political sentiment". J Soc Sci Comput Rev 30(2):229–234
18. Litman BR, Kohl LS (1989) Predicting financial success of motion pictures: The'80s experience [J]. J Media Eco 2(2):35–50
19. Liviu L, Mihaela T (2011) Predicting product performance with social media. J Nforma Educ 15(2):46–56
20. Metaxas P T, Mustafaraj E, Gayo-Avello D. (2011) How (not) to predict elections [C]//Privacy, security, risk and trust (PASSAT), 2011 *IEEE third international conference on and 2011 I.E. third international conference on social computing* (*SocialCom*). IEEE, 165–171
21. Mishne G, Glance N S. (2006) Predicting Movie Sales from Blogger Sentiment [C]//*AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 155–158.
22. O'Connor B, Balasubramanyan R, Routledge BR et al (2010) From tweets to polls: linking text sentiment to public opinion time series. J ICWSM 11:122–129
23. Pang B, Lee L (2008) Opinion mining and sentiment analysis [J]. Found trends Inf Retr 2(1–2):1–135
24. Pang B, Lee L, Vaithyanathan S. (2002) Thumbs up?: sentiment classification using machine learning techniques [C]//Proceedings of *the ACL-02 conference on Empirical methods in natural language processing*-Volume 10. Association for Computational Linguistics 79–86
25. Ritterman J, Osborne M, Klein E. (2009) Using prediction markets and Twitter to predict a swine flu pandemic [C]//*1st international workshop on mining social media*. 9
26. Sakaki T, Okazaki M, Matsuo Y. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors [C]//Proceedings of *the 19th international conference on World Wide Web*. ACM, 851–860
27. Sawhney MS, Eliashberg J (1996) A parsimonious model for forecasting gross box-office revenues of motion pictures [J]. Mark Sci 15(2):113–131
28. Schölkopf B, Smola A J. (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond [M]. *MIT press*
29. Sharda R, Delen D (2006) Predicting box-office success of motion pictures with neural networks [J]. Expert Syst Appl 30(2):243–254
30. Sharda R, Meany E. (2000) Forecasting gate receipts using neural network and rough sets [C]//Proceedings of *the International DSI Conference*. : 1–5
31. Si J., Mukherjee A., Liu B., Li Q., Li H., Deng X. (2008). Exploiting Topic based Twitter Sentiment for Stock Prediction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL-2013), pp. 24–29
32. Simonoff JS, Sparrow IR (2000) Predicting movie grosses: winners and losers, blockbusters and sleepers [J]. Chance 13(3):15–24
33. Skoric M, Poor N, Achananuparp P, et al. (2012) Tweets and votes: A study of the 2011 singapore general election [C]//*System Science* (*HICSS*), 2012 45th Hawaii International Conference on. IEEE, 2583–2591
34. Sochay S (1994) Predicting the performance of motion pictures [J]. J Media Eco 7(4):1–20
35. Sysomos Inc, "An In-Depth Look Inside the Twitter World ". http://www.sysomos.com/insidetwitter/. [Accessed Feb 3, 2012].
36. Theil H (1961) Economic forecasts and policy [J]
37. Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment[J]. *ICWSM*, 2010, 10: 178–185
38. UzZaman N, Blanco R, Matthews M. (2012) TwitterPaul: Extracting and Aggregating Twitter Predictions [J]. arXiv preprint arXiv:1211.6496
39. Vapnik V. (2000) The nature of statistical learning theory [M]. springer
40. Wikipedia, "social media". http://en.wikipedia.org/wiki/Social_media
41. Williams C, Gulati G. (2008) What is a social network worth? Facebook and vote share in the 2008 presidential primaries[C]. American Political Science Association
42. Zhang L, Luo J, Yang S (2009) Forecasting box office revenue of movies with BP neural network [J]. Expert Syst Appl 36(3):6580–6587
43. Zhang W, Skiena S. (2009) Improving movie gross prediction through news analysis [C]//Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society 301–304
44. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net [J]. J R Stat Soc Ser B (Stat Methodol) 67(2):301–320

Ting Liu received his B.E., M.E. and Ph. D. in Computer Science and Technology, from Harbin Institute of Technology, Harbin, China, in 1993, 1995 and 1998, respectively. Currently, he is a Professor with the School of Harbin Institute of Technology, Harbin, China. His current research interests include natural language processing, social computing, text mining, information retrieval and machine translation, etc.

Xiao Ding received his B.E. and M.E. in Computer Science and Technology, from Harbin Institute of Technology, Harbin, China, in 2009 and 2011, respectively. Currently, he is a Ph. D. candidate in Harbin Institute of Technology, Harbin, China. His current research interests include natural language processing, social computing and text mining.

Yiheng Chen received his B.E. in computer application from Northeast Forestry University, Harbin, China, in 2002, M.E. and Ph. D. in Computer Science and Technology, from Harbin Institute of Technology, Harbin, China, in 2004 and 2010, respectively. Currently, he is a Lecture with the School of Harbin Institute of Technology, Harbin, China. His current research interests include information retrieval, social computing and text clustering.



Haochen Chen received his B.E. in Computer Science and Technology, from Harbin Institute of Technology, Harbin, China, in 2012. Currently, he is a Master student in Harbin Institute of Technology, Harbin, China. His current research interests include natural language processing and social computing.

Maosheng Guo received his B.E. in Computer Science and Technology, from Shandong University, Shandong, China, in 2012. Currently, he is a Master student in Harbin Institute of Technology, Harbin, China. His current research interests include natural language processing and social computing.