 doi:10.20944/preprints201812.0340.v1

Article

Sales prediction by integrating heat and sentiments of product dimensions

Xiaozhong Lyu [1,*], Cuiqing Lastname [2,*], Yong Ding [3], Zhao Wang [4] and Yao Liu [5]

[1] School of Management, Hefei University of Technology, Hefei 230009, China; adolflv@mail.hfut.edu.cn
[2] School of Management, Hefei University of Technology, Hefei 230009, China; jiangcuiq2017@163.com
[3] School of Management, Hefei University of Technology, Hefei 230009, China; dingyong@hfut.edu.cn
[4] School of Management, Hefei University of Technology, Hefei 230009, China; xcwangzhao@163.com
[5] School of Management, Hefei University of Technology, Hefei 230009, China; liuyaoemail@foxmail.com

* Correspondence: adolflv@mail.hfut.edu.cn; Tel.: +86-1520-551-5211 (X.L.); jiangcuiq2017@163.com; Tel.: +86-1500-551-9191 (C.J.)

Abstract: The accuracy of sales prediction models based on the big data of online word-of-mouth (eWOM) is still not satisfied. We argue that eWOM contains heat and sentiments of different product dimensions, which can improve the accuracy of these models. In this paper, we propose a dynamic topic analysis (DTA) framework in order to extract heat and sentiments of product dimensions from the big data of eWOM. Finally, we propose an autoregressive-heat-sentiment (ARHS) model, which integrates heat and sentiments of dimensions into the baseline predictive model. The empirical study in movie industry confirms that heat and sentiments of dimensions can improve the accuracy of sales prediction model. ARHS model is better for movie box-office revenue prediction than other models.

Keywords: big data; sales prediction; online word-of-mouth; dynamic topic model; dimension heat; dimension sentiment

1. Introduction

Sales prediction is an important step of product and service management, because it is a foundation for business operations likes promotional marketing. Sales prediction with high accuracy and timeliness can allow firms to reduce the profit losses and improve market performance [1]. With the superiority of big data in online review systems, the frequency of sale prediction become higher than before to acquire more accurate of prevision to support real time decision making. However, the accuracy of these models is still not satisfied. We need extract more predictive information from the high-frequency online big data to improve sales prediction accuracy.

High-frequency big data, such as online word-of-mouth (eWOM) [2] and online search data (OSI) [3], contains timely information and can improve the accuracy of sales prediction [4]. However, the accuracy of sales prediction is still not satisfied for irregular or non-seasonal sales trends [5,6]. EWOM implies detailed information, such as the heat and sentiments of product dimensions, which previous predictive models do not consider. These factors have effects on product sales [7,8]. To improve accuracy of sales prediction, this paper proposes a framework to extract heat and sentiments of product dimensions from eWOM simultaneously and then integrates them into sales prediction model.

We choose movie industry as our research context. We crawl reviews from IMDb.com, online search data from google.com and film-related data from BoxOfficeMojo.com. Finally, we get a big data set including film-related data, Google Trends and 349269 reviews of 122 movies.

In order to extract heat and sentiments of product dimensions, we propose a dynamic topic analysis (DTA) framework in this study, which integrates machine-learning technique and lexicon-based method. It has two major functions. First, DTA captures product dimensions from eWOM without manual annotation. Second, DTA can extracts heat and sentiments of the extracted dimensions simultaneously. After that, we integrate dimension heat and sentiments to construct a new sales prediction model called autoregressive-heat-sentiment (ARHS) model. We get three movie dimensions from online reviews: *star*, *genre* and *plot*. We find that the proposed ARHS model has a better accuracy for predicting movie box-office revenue. Furthermore, ARHS model can predict sales of all kinds of products, which have enough eWOM.

We organize the remainder of this paper as follows. Section 2 describes the related literatures. Section 3 describes our methodology. Section 4 shows the results of our empirical study. The final section summarizes the main conclusions and discusses the implications of our study.

2. Literature Reviews

2.1 EWOM's effect on sales

As argued by [9], eWOM can reduce consumers' uncertainty about products by reduce the information asymmetry between reviewers and potential consumers.

Volume of eWOM represents the amount of information supplied by reviewers, such as the number of online reviews. Previous studies show that volume of eWOM is positively associated with movie box-office revenues [10,11]. Talking heat of product dimensions reflects the amount of information about product dimensions. Therefore, heat of some dimensions can influence product sales differently based the weightage of these dimensions [7]. We demonstrate that dimension heat has predictive power in predicting movie box-office revenues in this paper.

Valence of eWOM can be the average rating on the rating scale (e.g. 1-5) or the binary of positive and negative, which also can be regard as overall sentiment of eWOM. The overall sentiment of eWOM transmits reviewers' emotion to consumers. However, the aggregation process of the overall sentiment may offset the dimension-specific sentiments. This can be one of the reasons for why prior studies found overall sentiment of eWOM has no effect on movie box-office revenues [10,12]. Chen and Xie [13] demonstrate that eWOM provides product-dimension preference information that help consumer find products that match their needs. Potential consumers will have a different attitude to the product after perceiving the sentiments of different dimensions of the product from online reviews [8]. We argue that analyzing eWOM sentiments of different product dimensions can provide new insights for sales prediction and overcome the shortcoming of overall sentiment.

2.2 EWOM-based and GSI-based sales prediction

Online search data is the index (from 1 to 100) of the frequency of the object searched in online search engine, such as Google.com. It has been used for predicting movie box-office revenues [14]. Bughin [15] finds that valence of eWOM influences sales larger than Google Trends. Geva et al. [3] find that adding Google search data to models based on the more commonly used eWOM data significantly improves accuracy of prediction model. They also find that Google search index (GSI) models based on inexpensive Google Trends provide accuracy that is comparable, at least, to that of eWOM-based prediction models. These studies have proved online search data and eWOM both have powerful predictive ability. However, the predictive abilities of heat and sentiments of product dimensions have not researched yet. This research try to improve the prediction accuracy of movie box office revenues by proposing a comprehensive model, which first time integrates heat and sentiments of product dimensions simultaneously.

3. Materials and Methods

Figure 1 shows the framework of our study, which helps researchers to conduct a sales prediction model for products with abundant eWOM. First, we conduct eWOM model and GSI model, which integrate eWOM and Google Trends into autoregressive model, respectively. Then we integrate online search data into eWOM model following the method of [3], and name this baseline model autoregressive-online (ARO) model. Finally, we use DTA to extract the heat and sentiments of different dimensions of movies from eWOM and integrate them into ARO model to see if the new model, ARHS model, has a better prediction accuracy.

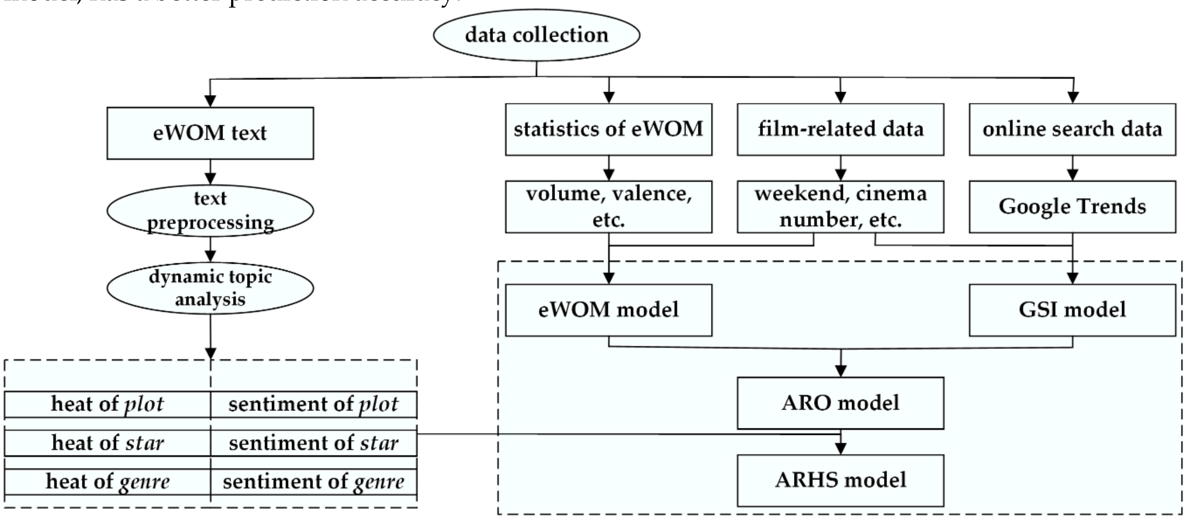

Figure 1. Framework for constructing ARHS model

3.1 Data and Variables

3.1.1 Data collection

EWOM takes many forms, including online reviews [2], blogs [16], microblogs [17], etc. With movies as our research object, we focus on online reviews because statistics suggest that online reviews are more prevalent than other types of eWOM in movie industry [18]. We choose IMDb.com as the resource of online reviews because it is the biggest movie review website in the world. We select movies by the rule in the website. After data filtering, we identify 349269 reviews for 122 movies each released for more than 49 days with at least 100 reviews. We use the threshold, 100 reviews, to guarantee that reviews is enough to train DTM. Our final data set is an online big data set, which contain most of the movie genres shown in Table 1. We chose 49 days as study period in order to achieve a panel data set with enough observations to reach credible experimental results. We divide the data set into two parts based on time to avoid overfitting. The first part is training set for training prediction model. The second part is test set for testing the out-sample performance of the trained model.

Table 1. Category of Movies

Genre	Freq.	MAPP rating	Freq.
comedy	37	R	57
drama	38	PG-13	50
action	13	PG	14
thriller	14	NC-17	1
sci-Fi	10	Total	122
horror	9		
animation	8		
romance	2		
crime	6		
fantasy	5		
adventure	3		
sports	2		
music	2		
documentary	1		
war	1		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

Table 2 lists the statistics of eWOM and film-related variables. First, we measure eWOM valence and volume represented by $v_{t,2}$ and $v_{t,1}$. Valence is the mean of daily reviews' ratings, which reflects the overall sentiment of reviewers on a special movie [19]. Volume is the daily number of reviews [18]. Second, we use variable $v_{t,3}$ to denote the number of days since the movie release to consider the time effect. Third, we set the dummy variable $v_{t,4}$ to one if the day is weekend and zero otherwise to consider the seasonal effect. Fourth, the variable $v_{t,5}$ represents the number of cinemas, which play the films [20]. Finally, we use the Google Trends of movie names as online search data, which ranges from one to 100.

Table 2. Key Variables for each movie: Numerical

Variable	**Description** (for each movie)	**Measure and Data Sources**
Sales	Daily box-office revenue	Dollars (log-transformation); BoxOfficeMojo.com
$v_{t,1}$	Daily number of reviews	Number (log-transformation); IMDb.com
$v_{t,2}$	Daily valence of reviews	Average of daily ratings (1-10); IMDb.com
$v_{t,3}$	Days from initial release	Number (1-49)
$v_{t,4}$	Whether the day is weekend	1= the day is weekend (Fri, Sat and Sun),0 = others
$v_{t,5}$	Daily number of cinemas	Number (log-transformation); BoxOfficeMojo.com
$v_{t,6}$	Daily Google Trends of movie name	Number (0-100); Google.com

3.1.2 Dynamic topic analysis

For 122 movies, we construct the framework DTA by integrating dynamic topic model (DTM) [21], lexicon-based method [22] and Stanford NLP technique [23] to derive the dimension heat and sentiments from online reviews. We obtain 122 daily documents by integrating hundreds of daily reviews of each movie into one document. Finally, the daily documents over 50 days comprise our review corpus. The corpus contains 349269 reviews. Figure 2 shows the structure of the corpus.

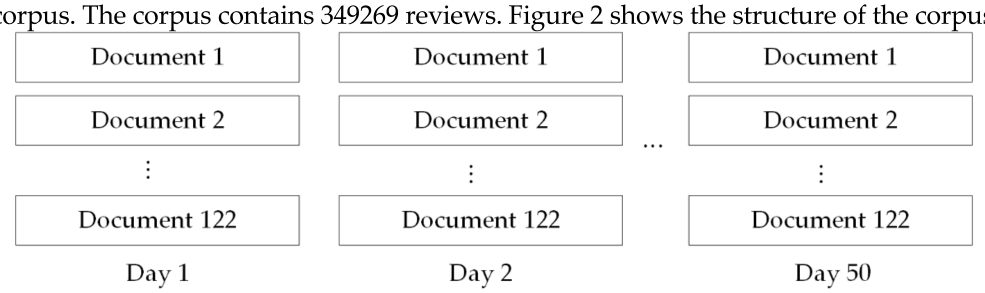

Figure 2. The structure of our review set

We pre-process each document by using the steps used in the study of Guo, Barnes, and Jia (2017). First, we eliminate non-English words and spell errors, such as web sites, punctuations and numbers. Then, we use Stanford NLP package for word text tokenization, part-of-speech tagging, and word stemming. Finally, each document becomes a word-of-bag.

To extract product dimensions from large corpus of text data effectively, previous studies have used latent Dirichlet allocation (LDA) model to extract product dimensions from large number of online reviews [24,25]. DTM is more suitable for extracting product dimensions from our structured review set [21], and is an extend method of LDA [26]. DTM can quickly discover a mixture of connected topics from huge number of documents over different time windows that LDA alone cannot achieve.

As a machine learning method, DTM is highly efficient to handle online big data. We use DTM to extract product dimensions, heat of these dimensions, words that represent each dimension and changes of these factors over different time windows. DTM assumes that a review comprises a sequence of N words, $d = (w_1, w_2, \dots, w_N)$, D reviews form a review set, $C_t = [d_1, d_2, \dots, d_D]$, whilst T review sets form a corpus over T time windows, $C = \{C_1, C_2, \dots, C_T\}$. It also assumes that reviewers share K dimensions across the corpus over the T time windows. In each time window, DTM assumes reviewers express their experience about product or service over K dimensions. For instance, a reviewer may comment the movie in the review based on three dimensions with different heat and sentiments: 30% and 4.9 for movie stars, 40% and 3.4 for story plot, and 30% and 2.1 for background music. 30% is the dimension heat of movie stars, which means that a third of the review is about movie stars. 4.9 is the sentiment strength of movie stars, which means that the reviewer has a strong sentiment with movie stars.

The DTM model consists of three hierarchies, and has correlation between different time-windows. Figure 3 shows the probabilistic graphical model of DTM. The circle w is the observable words. Circle Z and ɳ are latent variables. The rectangular boxes represent replications. The outer boxes represent documents, and the inner boxes mean repeatedly generating dimensions and words within a document. α and β are hyper-parameters at the document set level. DTM samples α and β based on the distributions of preceding α and β respectively so that we can extract the same dimensions in document sets of all time windows.

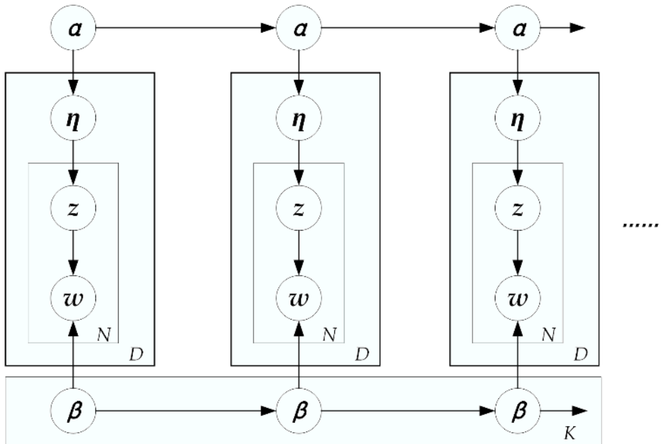

Figure 3. DTM model with plate notation

In DTM modelling, the followed steps formulated the generative process of a review set in time window (day) t:

1. Draw parameter $\beta_t|\beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$
2. Draw parameter $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, a^2 I)$
3. For each document:
   (a) Draw dimension distribution $\eta \sim N(\alpha_t, \delta^2 I)$
   (b) For each word:
      (1) Draw dimension $Z = k \sim Mult(\pi(\eta))$
      (2) Draw word $W_{t,d,n} = w \sim Mult(\pi(\beta_{t,k}))$,

where $\pi(\beta_{k,t})_w = \frac{e^{(\beta_{k,t,w})}}{\sum_w e^{(\beta_{k,t,w})}}$ maps the multinomial natural parameters to the mean parameters,. For a K-dimension model with N words, $\beta_{t,k}$ denotes the N-vector of words distribution for dimension k on day t. The value of DTM parameters that we must set are the first value of parameter α, the first value of parameter β and the number of dimensions K. The first values of parameter α and β are set according to experience: $\alpha = 0.01$ and $\beta = 50/K$. Through comparing the perplexity of DTM and semantics of dimensions when using different value of K, we determine the optimal number of dimensions. Finally, we find three movie dimensions that can represent the review corpus perfectly. The formula of perplexity of DTM for the document set on day t is as follows:

$$perplexity(C_t) = exp\left(-\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\log\sum_{k=1}^{K}p(W_{d,n} = w|Z_{d,n} = k)p(Z_{d,n} = k|d)}{\sum_{d=1}^{D} N_d}\right). \quad (1)$$

C_t is the document set on day t. D is the number of documents on day t. N_d is the number of words in document d. K is the number of dimensions. $p(W_{d,n} = w|Z_{d,n} = k)$ is the heat of word w in dimension k. $p(Z_{d,n} = k|d)$ is the heat of dimension k in document d. DTM learning with Gibbs Sampling can generate the heat of words and dimensions simultaneously. Let $\vartheta_{i,t}$ be the heat of the k^{th} dimension of the i^{th} movie on day t. $\vartheta_{i,t}$ can be calculated as follows:

$$\vartheta_{i,t,k} = \frac{\sum_{d=1}^{D_{i,t}} p(Z=k|t,d,i)}{D_{i,t}}, \quad (2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

where $p(Z = k|t, d, i)$ is the heat of dimension k in document d of movie i. $D_{i,t}$ is the number of documents for movie i on day t. In our research context, $D_{i,t}$ equals one.

We name the three dimensions *plot*, *star* and *genre* following the method in Guo et al. (2017). Table 3 shows the changes of dimension *plot* in different time windows.

Table 3. The change of words and their weightages of dimension *plot*

plot	*weight*	*plot*	*weight*	*plot*	*weight*
story	0.9%	*plot*	0.5%	*plot*	0.5%
plot	0.4%	*story*	0.4%	*story*	0.4%
book	0.4%	*book*	0.3%	*book*	0.4%
horror	0.3%	*horror*	0.3%	*horror*	0.3%
dark	0.2%	*dark*	0.3%	*dark*	0.2%
original	0.3%	*original*	0.2%	*original*	0.2%
scary	0.2%	*scary*	0.2%	*scary*	0.2%
real	0.2%	*maze*	0.2%	*maze*	0.2%
pretty	0.2%	*pretty*	0.2%	*pretty*	0.2%
action	0.2%	*love*	0.2%	*house*	0.2%

Dimensions heat is the proportion that reviewers talk about the product dimension in eWOM. For example, heat of dimension *plot* denotes the proportion that consumers talk about the *plot*-related information in reviews. Figure 4 shows the heat of three movie dimensions changes over the 50 days

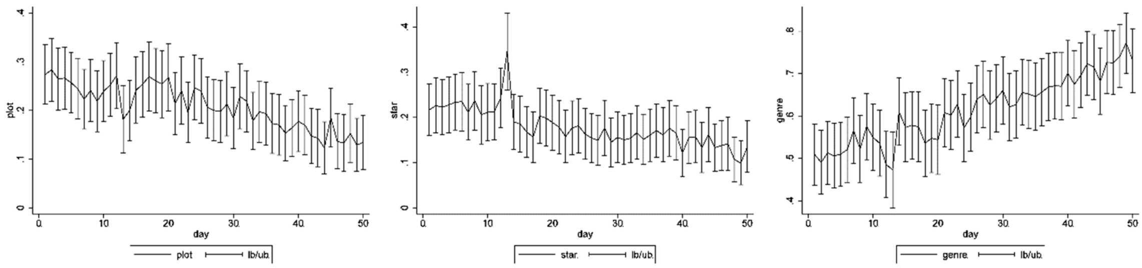

Figure 4. Average heat of the three dimensions over 122 movies

Then, we use the sentiment lexicon and syntax relation to calculate dimension sentiments. Lexicon-based method by using a public-recognized sentiment lexicon is more objective and suitable for big data sentiment analysis than machine-learning-based method that needs expert annotations. Because expert annotation has a high time cost, and it exists deviations between humans. Most studies about dimension sentiment analysis divide the dimensions into positive or negative class [27]. Sentiment analysis methods are different according to different application requires. Our study calculate the sentiment strength of each dimension that can forecast movie box-office revenues. We extract the syntactic relations between dimension words and sentiment words in the daily review sentences by using the Stanford NLP package. We obtain the sentiments of dimension words based on the extracted relations. Table 4 shows the main sentiment mining rules used in our framework.

Table 4. The main rules for sentiment mining of dimension words

Syntax relations	**Examples**	**word sentiments**
Nominal subject	The *plot* is *boring*.	*Plot*: 3.0
Adjectival modifier	She is a *good actor*.	*Actor*: 3.8612
Direct object	I *enjoy 3D*.	*3D*: 3.9782
Open clausal complement	I think the actor *enjoys acting*.	*Acting*: 3.9782
Adverb modifier	Tom *performed earnestly*.	*Perform*: 3.5
Relative clause modifier	I saw the *actor* who people *dislike*.	*Actor*: 3.5417

Finally, we calculate average daily sentiment strength of dimensions for each movie. Let $s_{i,n,d}$ be the sentiment value reflect to the n^{th} dimension word appear at the i^{th} time in document d for one movie. Then, the sentiment of the k^{th} dimension for one movie on t^{th} day can be formulated as follows:

$$\theta_{t,k} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{D}\sum_{d=1}^{D}\frac{1}{I}\sum_{i=1}^{I} s_{i,n,d}. \tag{3}$$

Intuitively, $\theta_{t,k}$ represents the average strength of sentiment of k^{th} dimension. Figure 5 shows the average sentiments of dimension *plot* of 122 movies.

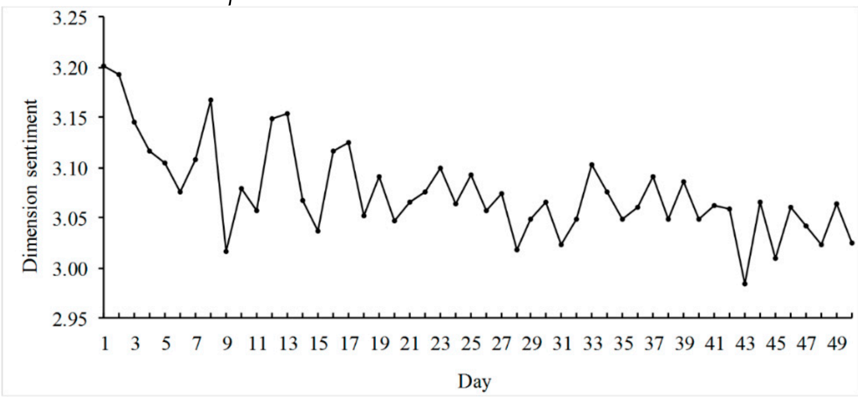

Figure 5. Average sentiments of dimension *plot* over 122 movies

In Table 5, we describe the key variables of dimensions.

Table 5. Key Variables for each movie: Dimensions

Variable	Description	Measures
$\vartheta_{t,1}$	The heat of dimension *plot* on day *t*	Probabilistic
$\vartheta_{t,2}$	The heat of dimension *star* on day *t*	Probabilistic
$\vartheta_{t,3}$	The heat of dimension *genre* on day *t*	Probabilistic
$\theta_{t,1}$	The sentiment of dimension *plot* on day *t*	Numerical value
$\theta_{t,2}$	The sentiment of dimension *star* on day *t*	Numerical value
$\theta_{t,3}$	The sentiment of dimension *genre* on day *t*	Numerical value

3.1.3 Descriptive analysis

Table 6 shows the summary statistics of variables. We can see the sales, volume ($v_{t,1}$), theatres ($v_{t,5}$) are right-skewed distribution and the skewness of volume and sales is very large. That means very few movies have high box-office revenues or high customer attentions, and most movies have low box-office revenues or low customer attention. The distributions of valence ($v_{t,2}$) are relatively evenly distributed. Dimension heat ($\vartheta_{t,i}$) is between zero and one. The median of dimension sentiment ($\theta_{t,i}$) is three.

Table 6. Summary statistics of key variables

Variable	Mean	Median	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis
Sales	1039207	263875	35167017	10	2204483.7	5.351	46.855
$v_{t,1}$	22.11526	11	506	0	38.754737	3.956	26.541
$v_{t,2}$	3.801627	4	10	0	3.6083277	0.162	1.410
$v_{t,5}$	1483.412	1195	4324	1	1264.8718	0.343	1.634
$v_{t,6}$	33.49281	28	100	2	21.922873	1.101	3.815
$\vartheta_{t,1}$	0.269317	0.009709	0.99999565	1.93E-06	0.4063216	1.082	2.284
$\vartheta_{t,2}$	0.135738	0.009709	0.99999565	2.16E-06	0.3078812	2.184	5.992
$\vartheta_{t,3}$	0.594945	0.95943	0.99999488	1.43E-06	0.4544953	-0.426	1.250
$\theta_{t,1}$	3.073407	3	4.83333	0.130435	0.3685301	-3.347	28.987
$\theta_{t,2}$	3.094447	3	4.90476	0.130435	0.3519009	-2.837	28.437
$\theta_{t,3}$	3.086102	3	4.60417	0.130435	0.299457	-3.282	35.261

Figure 6 (a) shows the relationship between Google Trends and box-office revenues of movie *Gravit*. Figure 6 (b) shows the relationship between eWOM volume and box-office revenues of movie *Gravity*. We can see that they eWOM and GSI both have high correlations with movie box-office revenues.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

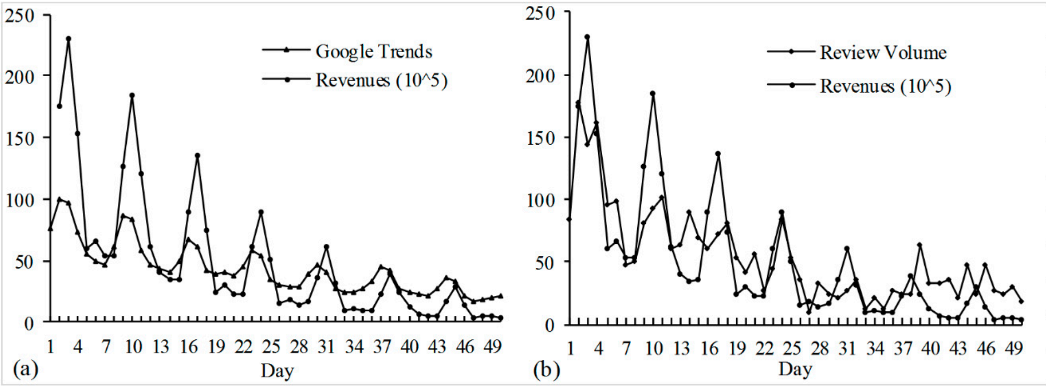

Figure 6. The relationship between online information and box-office revenues of movie *Gravity*. (a) The relationship between Google Trends and box-office revenues. (b) The relationship between review volume and box-office revenues.

3.2 Predictive Model

To forecast movie box-office revenues, we construct the proposed approach based on autoregressive model, because regressive model is the most efficient predictive model [5]. We also need to address some methodological concerns. First, we take a log-transform to some skewed variables to make them similar with normal distribution. Second, we use the variance inflation factor (VIF) to assess multivariate multicollinearity. The VIF values are lower than the threshold five, so multicollinearity was not a serious issue [28]. We use the first 40-days data to train predictive model and the last 9-days data to test the trained model.

3.2.1 Autoregressive Model

We start with an autoregressive (AR) model as our base model to forecast movie box-office revenues. We use AR model with parameter p to model the relationship between preceding box-office revenues and current box-office revenue as follows:

$$\log(Sales_t) = \alpha + \sum_{i=1}^{p} \varphi_i \log(Sales_{t-i}) + \epsilon_t, \quad (4)$$

where $\varphi_1, \varphi_2, \ldots, \varphi_p$, are the parameters to be estimated, α is the effect of combination of time-invariant variables, such as production budgets and genres of movies, and ϵ_t is an error term.

3.2.2 Autoregressive-online Model

Besides preceding box-office revenues, online information, such as Google Trends and eWOM volume, might greatly influence box-office revenues. According to the discussion in above sections, we propose a predictive model by integrating online information into AR model. This model include all variables of previous GSI models and eWOM models. Our ARO model is similar to the model proposed in [3], and can be formulated as follows:

$$\log(Sales_t) = \alpha + \sum_{i=1}^{p} \varphi_i \log(Sales_{t-i}) + \sum_{i=0}^{q} \sum_{j=1}^{J} \rho_{i,j} v_{t-i,j} + \epsilon_t, \quad (5)$$

where $v_{t,j}$ represents the j^{th} online information variable on day t. We determine p and q by comparing the model accuracy when using different p and q. φ_i and $\rho_{i,j}$ are parameters that need estimations. Parameter q specifies the lags of preceding days of online information variables. J indicates the number of these variables.

3.2.3 Autoregressive-heat-sentiment Model

According to previous studies, heat and sentiments of product dimensions are very important to sales [7,8]; thus, it is desirable to integrate the heat and sentiments of movie dimensions into predictive

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

model to achieve better accuracy. In this section, we extend ARO model to ARHS model. We formulate ARHS model as follows:

$$\log(Sales_t) = \alpha + \sum_{i=1}^{p} \varphi_i \log(Sales_{t-i}) + \sum_{i=0}^{q}\sum_{j=1}^{J} \rho_{i,j} v_{t-i,j} + \sum_{i=0}^{\gamma}\sum_{k=0}^{K} \omega_{i,k}\vartheta_{t-i,k} + \sum_{i=0}^{\delta}\sum_{k=0}^{K} \mu_{i,k}\theta_{t-i,k} + \epsilon_t, \tag{6}$$

where p, q, γ and δ are user-defined parameter, ϵ_t is an error term, and $\varphi_i, \rho_{i,j}, \omega_{i,k}$ and $\mu_{i,k}$ are parameters that need estimations. $\vartheta_{t,k}$ and $\theta_{t,k}$ are the heat and sentiments of the k^{th} dimension at time t, which are obtained by using DTA. p, q, γ and δ specify how far the model "looks back" into the history, whereas J and K specify how many related variables that we would like to consider. J and K are fitted as we discussed in section 3.1. We use least square method to train all the models.

4. Results

In this section, we compare ARHS model with AR model, eWOM-based model, GSI-based model and ARO model to validate its effectiveness.

We use the *mean absolute percentage error (MAPE)* to measure the performance of predictive models in this paper.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|Pred_i - True_i|}{True_i} \times 100\%, \tag{7}$$

where n is the number of predictions that made on the testing data, $Pred_i$ is the predicted box-office revenue, and $True_i$ represents the true value of the box office revenue. In statistic, *MAPE* is a suitable measure of accuracy for the time-series-value predictions. We can compare the error of fitted time series because it is a percentage error. All the *MAPE* results reported herein are mean value of independent runs of 122 movies on different days. This metric is robust to compare the performance of sales prediction models [3,29].

4.1 Parameter determination for ARHS model

In ARHS model, Parameters p, q, γ and δ provide the flexibility to fine tune the model to optimal performance. We now study how the choices of these parameter values affect the prediction accuracy.

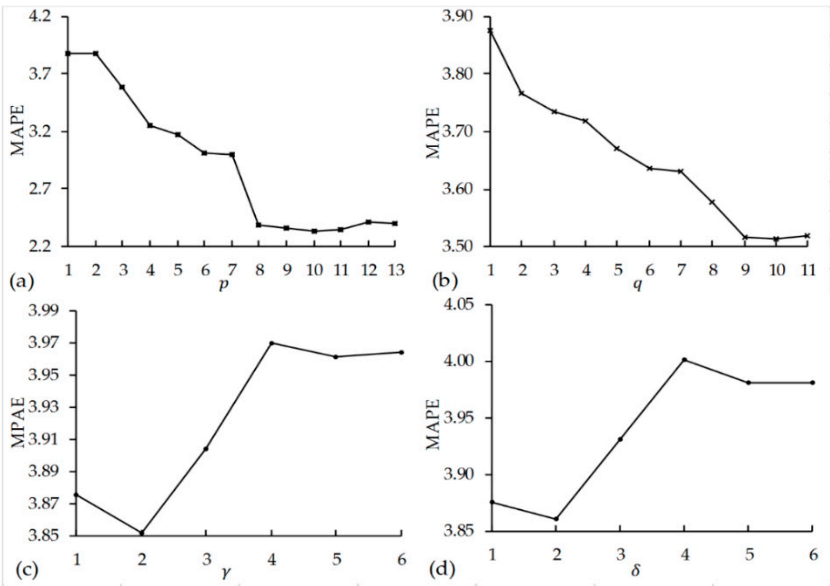

Figure 7. The effects of parameters on the prediction accuracy. (a) Effects of p. (b) Effects of q. (c) Effects of γ. (d) Effects of δ.

doi:10.20944/preprints201812.0340.v1

First, we vary p with fixed values of parameters q, γ and δ ($q = \gamma = \delta = 1$) to study how preceding box-office revenues affect the prediction accuracy of ARHS model. As shown in Figure 7a, the model achieves its best prediction accuracy when $p = 10$. The change of accuracy is minor after $p = 8$. The accuracy even goes down after $p = 11$. These findings suggest that p should be large enough to factor in all significant influences of preceding box-office revenues, but should not be too large to let irrelevant preceding box-office revenues reduce prediction accuracy.

Then, with fixed value of p, γ and δ values ($p = \gamma = \delta = 1$), we then vary the value of q from one to 11 to study its effect on prediction accuracy. Figure 7b shows that the model also achieves its best performance when $q = 10$. However, the accuracy is basically the same after $q = 9$, which means that numerical online information will affect box-office revenues in the next nine days. From the above results, we can suggest that the predictive power of numerical online information for box-office revenues last a little longer than preceding box-office revenues.

By using fixed values of p, q and δ ($p = q = \delta = 1$), we vary γ from one to six to study the prediction accuracy of ARHS model. As shown in Figure 7c, the ARHS model achieves the best prediction accuracy at $\gamma = 2$, which implies that the effect of dimension heat captured from the text of eWOM lasts two days.

We also vary δ from one to six, with fixed p, q and γ ($p = q = \gamma = 1$). As shown in Figure 7d, ARHS model achieves the highest accuracy at $\delta = 2$, which implies that the effects of dimension sentiments on box-office revenues also last two days.

From the results above, we can conclude that product dimension information captured from online comments has a shorter effect on box-office revenues than numerical online information. We think the reason is that consumers only look through the text of eWOM posted in recent days, but glance over the numerical information of eWOM posted in a longer period before they decide to see a movie.

4.2 Comparison with other prediction models

To verify the superiority of ARHS model, we compare its performance with other models' performance.

First, we compare ARHS model ($q = 10$, $\delta = \gamma = 2$) against AR model. As shown in Figure 8, ARHS model constantly outperforms AR model as p ranges from one to ten. We can see that ARHS model has a much better accuracy when p is small, which implies that eWOM of movie can supply more predictive power when we know little about preceding box-office revenues. When p equals four our proposed sales prediction model improve the MAPE of AR model at 27.65%. When the lags of sales equals eight, the improvement of MAPE is the smallest. However, it has a 2.69% improvement. These improvements suggest that ARHS model has a better accuracy.

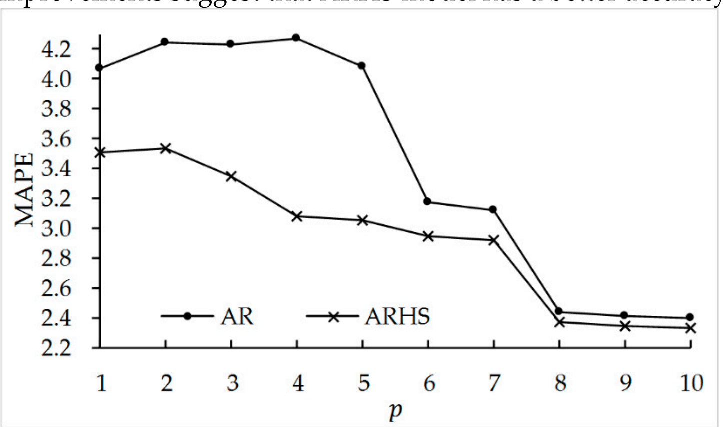

Figure 8. Comparison with autoregressive prediction model

Then, we conduct experiments to compare ARHS ($\delta = q = \gamma = 1$) model against eWOM model, GSI model [14] and ARO model [3]. Our study and previous studies prove that these models are better than AR model. As shown in Figure 9a, eWOM model and GSI model nearly have the same performance in accuracy. As shown in Figure 9b, 9c, and 9d, we can see that ARHS model always

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

outperforms eWOM model, GSI model and ARO model during p ranges from one to six. Thus, ARHS model is the best among these models. The effects of eWOM text on box-office revenues decrease over time, and our test is at the end of movies' release time. Therefore, comparing with eWOM model, GSI model and ARO model, ARHS model improves the MAPE not very high. We argue that the accuracy improvement of ARHS model will be higher at earlier periods of movie release. Because of the high gross of movies, even very little improvement in forecasting accuracy might result in a difference in millions of dollars. ARHS model is meaningful to movie marketers and theatre managers.

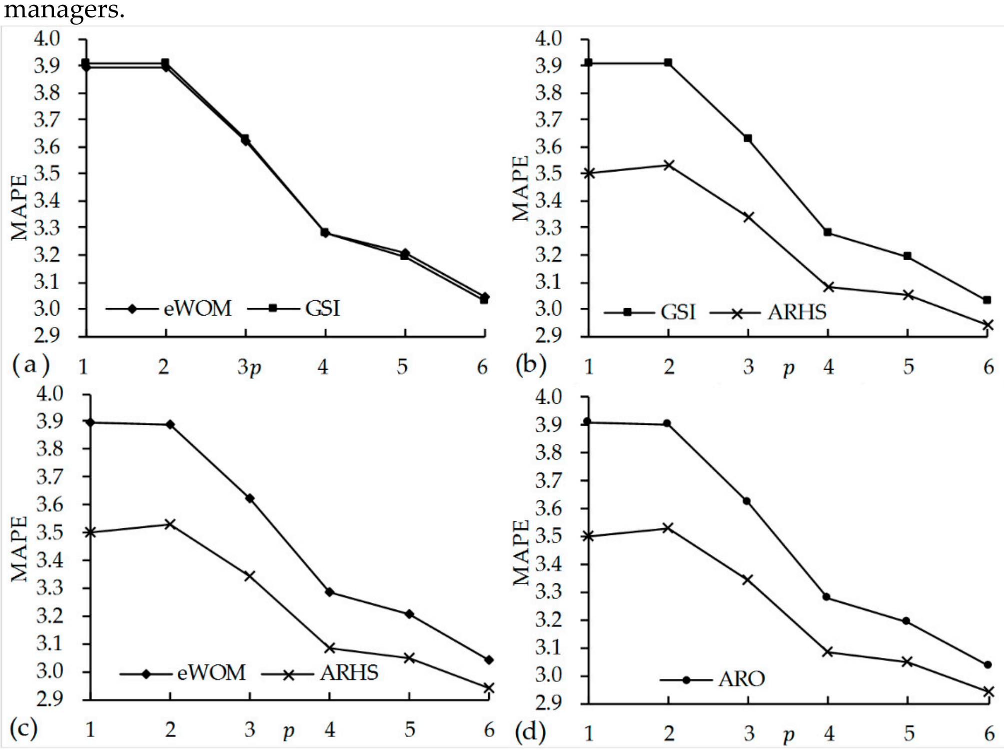

Figure 9. Comparisons of model accuracy. (a) Comparison of eWOM model and GSI model. (b) Comparison of GSI model and ARHS model. (c) Comparison of eWOM model and ARHS model. (d) Comparison of ARO model and ARHS model.

4.3 Time robustness

In order to verify the time robustness of ARHS model, we compare its accuracy in different predictive period. We use the first 20-days, 30-days and 40-days data as training data and the following 9-days data as testing data respectively. Figure 10 shows the results. The prediction accuracy increases during $0 < p < 8$ increases and nearly does not change after $p \geq 8$. The prediction accuracy of 21-29th days is always better than the accuracy of 31-39th days, and the prediction accuracy of 31-39th days is better than that of 41-49th days. That means ARHS model's prediction performance is better in the initial stage of movie released, and the earlier the better. Therefore, we can conclude that heat and sentiments of dimensions have greater predictive power in the early days of movie release.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

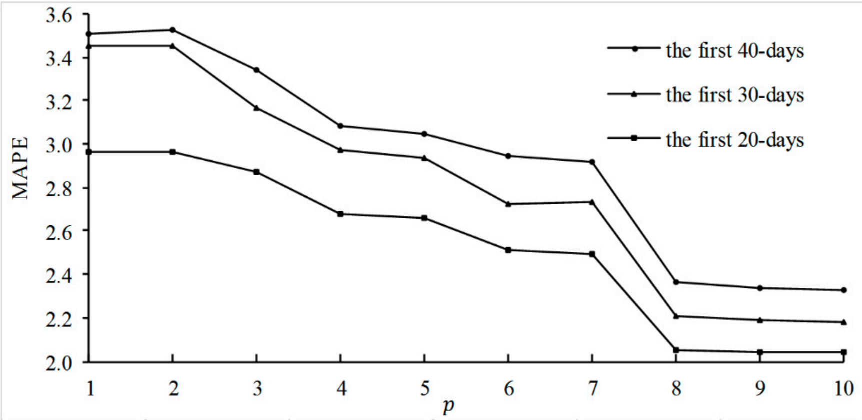

Figure 10. Comparison of different prediction intervals.

5. Conclusion and Discussion

The predictive model for movie box-office revenue is still not satisfied. Previous research demonstrate that eWOM text implies heat and sentiments product dimensions that influence product sales [7,8]. Thus, we propose a method called dynamic topic analysis (DTA) to extract the heat and sentiments of product dimension from eWOM. From the results of DTA, we obtain heat and sentiments of three movie dimensions: *plot*, *genre* and *star*. Then, to improve accuracy, we propose ARHS model by integrating dimension heat and sentiments into the predictive model of movie box-office revenues. By comparing performance with other predictive models, ARHS model is better. We also find that ARHS model performs much better in the early stage of product release.

Our paper has some contributions to managerial implications. First, marketers can use DTA to extract the heat and sentiments of product dimensions from eWOM. This information has great predictive power. Therefore, they can make different marketing strategies at different time according to the different predictive power of these dimension information. Second, the optimal parameters of ARHS model suggest that the predictive power of numerical aspects of eWOM lasts longer than that of eWOM text: heat and sentiments of product dimensions. Therefore, managers should pay attention to numerical aspects of eWOM over a long period and only pay attention to the text of new eWOM. Third, theaters can adjust the projection room number for different movies according to predicted daily box-office revenues.

Our research has some theoretical implications. First, DTA provides a framework for researchers to extract the heat and sentiment of multidimensional constructs in many social studies. Second, dimension heat and sentiments indeed improve the accuracy of prediction model. Researchers can use them to predict sales of other products, outcome of election and price of stock. Third, we propose the superior ARHS model for movie box-office revenue prediction.

This paper also has some limitations. We only predict daily box-office revenues to demonstrate the predictive power of dimension heat and sentiments. To improve our theory, we shall predict weekly or monthly sales for different products in future research. Additionally, we only use one type of eWOM in this paper. We should use multi-type of eWOM in future research to have a more robust result.

Author Contributions: Conceptualization, X.L. and C. J.; methodology, X.L.; software, X. L.; validation, X.L., Y.D. and Z. W.; formal analysis, X.L. and Y.L.; investigation, X.L.; resources, X.L.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L., C.J.; visualization, X.L.; supervision, X.L., Z.W., Y.L.; project administration, C.J., Y.D.; funding acquisition, C.J..

Funding: This research was funded by National Natural Science Foundation of China (NSFC), key grant number: 71731005 and grant number: 71571059.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

References

1. Amornpetchkul T, Duenyas I, Şahin Ö. Mechanisms to Induce Buyer Forecasting: Do Suppliers Always Benefit from Better Forecasting? Prod Oper Manag. 2015;24: 1724–1749. doi:10.1111/poms.12355
2. Hu N, Koh NS, Reddy SK. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. Decis Support Syst. 2014;57: 42–53. doi:10.1016/j.dss.2013.07.009
3. Geva T, Oestreicher-Singer G, Efron N, Shimshoni Y. Using forum and search data for sales prediction of high-involvement products. MIS Q. 2017;41: 65–82. doi:10.2139/ssrn.2294609
4. Chern CC, Wei CP, Shen FY, Fan YN. A sales forecasting model for consumer products based on the influence of online word-of-mouth. Inf Syst E-bus Manag. 2015;13: 445–473. doi:10.1007/s10257-014-0265-0
5. Ghiassi M, Lio D, Moon B. Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Syst Appl. 2015;42: 3176–3193. doi:10.1016/j.eswa.2014.11.022
6. See-To EWK, Ngai EWT. Customer reviews for demand distribution and sales nowcasting: a big data approach. Ann Oper Res. 2018;270: 415–431. doi:10.1007/s10479-016-2296-z
7. Li X, Wu C, Mai F. The effect of online reviews on product sales: A joint sentiment-topic analysis. Inf Manag. 2018; doi:10.1016/j.im.2018.04.007
8. Liang TP, Li X, Yang CT, Wang M. What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. Int J Electron Commer. 2015;20: 236–260. Available: https://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=1&sid=d018e9be-cc42-4deb-b342-6c4ac98938c9%40pdc-v-sessmgr01
9. Siering M, Muntermann J, Rajagopalan B. Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. Decis Support Syst. 2018;108: 1–12. doi:10.1016/j.dss.2018.01.004
10. Liu，Yong. Word of mouth for movies: Its dynamics and impact on box office revenue. J Mark. 2006;70: 74–89. doi:10.1509/jmkg.70.3.74
11. Duan W, Gu B, Whinston AB. Do online reviews matter?—An empirical investigation of panel data. Decis Support Syst. 2008;45: 1007–1016. doi:10.1016/j.dss.2008.04.001
12. Zhang Z, Li X, Chen Y. Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews. ACM Trans Manag Inf Syst. 2012;3: 1–23. doi:10.1145/2151163.2151168
13. Chen Y, Xie J. Online consumer review: Word-of-mouth as a new element of marketing communication mix. Manage Sci. 2008;54: 477–491. doi:10.1287/mnsc.1070.0810
14. Lee C, Jung M. Predicting movie incomes using search engine query data. International Conference on Artificial Intelligence and Pattern Recognition. 2014. pp. 45–49.
15. Bughin J. Google searches and twitter mood: nowcasting telecom sales performance. NETNOMICS Econ Res Electron Netw. 2015;16: 87–105. doi:10.1007/s11066-015-9096-5
16. Ha SH, Bae SY, Son LK. Impact of online consumer reviews on product sales: Quantitative analysis of the source effect. Appl Math Inf Sci. 2015;9: 373–387. doi:10.12785/amis/092L12
17. Dijkman R, Ipeirotis P. Using Twitter to predict sales: a case study. arXiv Prepr arXiv 150304599. 2015; 1–14. Available: http://arxiv.org/abs/1503.04599
18. Duan W, Gu B, Whinston AB. The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. J Retail. 2008;84: 233–242. doi:10.1016/j.jretai.2008.04.005
19. Chevalier JA, Mayzlin D. The effect of word of mouth on sales: Online book reviews. J Mark Res. 2006;43: 345–354.
20. Wang F, Liu X, Fang E. User reviews variance, critic reviews variance, and product sales: An exploration of customer breadth and depth effects. J Retail. 2015;91: 372–389.
21. Blei DM, Lafferty JD. Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning. 2006. pp. 113–120. doi:10.1145/1143844.1143859
22. Guerini M, Gatti L, Turchi M. Sentiment analysis: How to derive prior polarities from SentiWordNet. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. pp. 1259–1269. Available: http://arxiv.org/abs/1309.5843
23. Socher R, Perelygin A, Wu J, Chuang J, Manning C, Ng A, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. pp. 1631–1642. doi:10.1371/journal.pone.0073791
24. Guo Y, Barnes SJ, Jia Q. Mining meaning from online ratings and reviews : Tourist satisfaction analysis using latent Dirichlet allocation. Tour Manag. 2017;59: 467–483.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2018 doi:10.20944/preprints201812.0340.v1

25. Tirunillai S, Tellis GJ. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. J Mark Res. 2014;51: 463–479. doi:10.1509/jmr.12.0106
26. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3: 993–1022.
27. Schouten K, Frasincar F. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering. 2016. pp. 813–830. doi:10.1109/TKDE.2015.2485209
28. Kleinbaum D, Kupper L, Nizam A, Rosenberg E. Applied regresion analysis and other multivariable methods. Nelson Educ. 2013;
29. Yu X, Liu Y, Huang X, An A. Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Trans Knowl Data Eng. 2012;24: 720–734. doi:10.1109/TKDE.2010.269