2014

DOES WEB NEWS MEDIA HAVE OPINIONS? EVIDENCE FROM REAL ESTATE MARKET PREDICTION

Daoyuan Sun
Department of Information Systems, School of Computing, National University of Singapore, dysun2010@126.com

Ce Zhang
Renmin University of China, fancyzc@foxmail.com

Wei Xu
Renmin University of China, weixu@ruc.edu.cn

Mei Yun Zuo
Renmin University of China, zuomy@ruc.edu.cn

Junjie Zhou
Henan University of Economics and Law, jjzhou@huel.edu.cn

See next page for additional authors

Follow this and additional works at: http://aisel.aisnet.org/pacis2014

Recommended Citation
Sun, Daoyuan; Zhang, Ce; Xu, Wei; Zuo, Mei Yun; Zhou, Junjie; and Du, Yudie, "DOES WEB NEWS MEDIA HAVE OPINIONS? EVIDENCE FROM REAL ESTATE MARKET PREDICTION" (2014). PACIS 2014 Proceedings. 374.
http://aisel.aisnet.org/pacis2014/374

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Authors
Daoyuan Sun, Ce Zhang, Wei Xu, Mei Yun Zuo, Junjie Zhou, and Yudie Du
Abstract

In this paper, we propose a novel method for real estate price prediction using web new media sentiments by incorporating human searching behavior on the web. By combining online daily news’ sentiments and Google search engine query data, we construct a web news content and online search behavior-based integrated model for real estate prediction. Besides these factors, real estate price time series data are also considered into the model in order to improve the forecasting performance. Furthermore, we make a comparison between the integrated model and the baseline model without search engine query data. Experimental results indicate that the integrated model outperforms the non-integrated model, which suggests that online user searching behavior is of great value in enhancing the prediction performance. These findings imply that the proposed integrated model is effective and feasible for real estate market prediction.

Keywords: Real Estate Prediction, Web News Media, Sentimental Analysis, Search Engine Query Data, Data Mining
INTRODUCTION

Real estate is an important industry which contributes a lot to national GDP and has profound influence on many different fields. Because of its close connection with individuals, enterprises and other industries, research on real estate market prediction can significantly help to the decision making of individuals, enterprises and government. Since the financial crisis in 2008, global real estate market has suffered and changed a lot (Profile of Home Buyers and Sellers 2011), the research on real estate market is becoming popular. The empirical study of real estate bubble in U.S. analysed the abnormal growth rate and predicted the future turning point (Zhou & Sornette 2006). Chinese real estate market also encountered its bottleneck after its rocketing in 2009 (Xie et al. 2011), it senses meaningful to research on the market prediction. Moreover, the real estate industry has changed due to the revolution of information and communication technology (ICT) which led to much more efficient information exchange (Kummerow & Lun, 2005). Thus making the market prediction from the aspect of information and internet technology is natural and meaningful. Also it is an urgent challenge to researchers that proposing new method to make predictions.

In the past years, many methods and models have been put forward for real estate market prediction. The sensitive indicator of real estate market, usually the real estate price index, has been forecasted by many researchers who hope to predict the future trend of real estate market. Besides the real estate price, others such as the stock price, commodity price (Cashin et al. 2002), agriculture price (Yu, et al., 2011) etc. are also valuable indicators for the economy. Various methods of prediction have been proposed for the purpose of examining the future operation and trend of different markets recently. A Neutral Network model was applied to the real estate price prediction by Khalafallah (2008) with only slight forecasting error. Moreover, Lu et al. (2009) used the independent component analysis (ICA) and Support Vector Regression (SVR) to forecast financial time series and got outstanding results. Nevertheless, those methods simply rely on pale statistical data and basically ignore the effects of actual human behaviour of the market participants. As a result, these traditional methods not only cause the waste of valuable information but usually lead to low prediction accuracy, especially when the collected statistical data cannot perfectly reflect the true situation of the market. Since the social and economic systems, such as financial market and real estate market, are directly influenced by the behaviours of participants such as purchasing, investing, advertising and even focusing, human behaviour is the most direct and effective indicator of the operation of real estate market. That is why we introduce the factor of online users’ searching behaviour which related to the real estate market.

Recently, online users’ behaviour has drawn the attention of researchers. As the world has entered an information era, people began to use the Internet to obtain or spread information of real estate market. A profile by National Association of Realtors (NAR) demonstrates that 90% of the real estate buyers use the Internet to find useful information and 92% sellers use the Internet to publicize trade information (Profile of Home Buyers and Sellers 2011). Apart from that, according to an unofficial survey made by SouFun.com in China, more than 60% of Chinese web users obtain real estate information through the Internet and approximately more than 80% Chinese web users access the website of real estate per month. These survey results suggest that online users’ behaviour is almost a mirror of real world human behaviour in the real estate market. So it can reflect the situation of real estate market to a large extent. Thus it is reasonable and feasible to take online users’ behaviour into consideration when forecasting the real estate market.

Findings in behavioural economics tell us that emotions can affect individual behaviour profoundly. So it can be stated that emotions or sentiment of words in online opinions or articles could serve as an effective indicator of real world and therefore could be used for prediction. Extant research already shows that sentiment in web articles or other forms of web data has great potential in market prediction. Multiple forms of web data have been used to make all kinds of predictions by extracting their sentiment. Das and Chen (2007) made the study of using sentiment of words on certain message boards to discover the relationship between web sentiment and stock returns. Moreover, Bollen and Mao (2011) used twitter mood as the input of Fuzzy Neural Network to prove that public mood and
the Dow Jones Industrial Average close value are closely related. Similarly, online opinion ensemble was also used to predict the financial market (Xu et al. 2012). These examples remind us that sentiment in web data can be integrated into the model of real estate market prediction.

Among various forms of web data, online news articles from reliable sources contain the richest, timely, convincing and valuable information about real estate market in reality. Therefore, the sentiment of these articles could be used to make predictions and is introduced into the proposed prediction model in this paper.

Although news sentiment can reflect part of human behaviours in the real estate market, the sentiment alone is not enough to make the prediction accurate. The content of a news article with few viewers cannot indicate the actual fluctuation the real estate market, however strong its sentiment is. To solve the problem, search engine query data is integrated into our model in order to make more precise prediction. Search engine query data are a record of online users’ searching behaviour. The data is a time series format with two attributes, search key word and search frequency. Since people search things they focus on, these data are a direct reflection of their interest and specific behaviour such as purchasing desires, investing possibilities, bargaining or selling motivations during a certain period of time. (Wu and Brynjolfsson, 2009). Researchers have already made use of search engine query data for prediction. Xu, et al.(2012) used search engine query data combined with machine learning model to predict the unemployment rate and received significant results. Ginsberg, et al. (2009) also applied the search engine query data to detect the seasonal influenza epidemics. Furthermore, Wu and Brynjolfsson (2009) used Google search engine query data to predict the real estate price index and quantity volume which gained profound outputs.

Although previous study has used news sentiment and search engine query data respectively for prediction, no one ever combined the two indicators in a single integrated model. In this paper, a new method of real estate price index prediction is proposed by introducing human behavioural factor into the forecasting model. By combining online daily news’ sentiments and Google search engine query data, we construct an integrated data mining model that has satisfactory forecasting performance. In our model, real estate price index time series data and their lags are also integrated into the model in order to improve the forecasting performance. The prediction is finally completed with a SVR model that has the best performance among several SVR models. Empirical results confirm the effectiveness and feasibility of the proposed method.

The paper is organized as follows. Section 2 gives theoretical background on support vector regressions (SVR). Section 3 shows the research framework and the modelling process in details. Then the empirical analysis is carried out to verify the effectiveness and feasibility of the proposed method and to compare the proposed method with baseline models in Section 4. Finally Section 5 gives conclusions and future work.

2 THEORETICAL BACKGROUND

In this section we presented the background of SVR which is the theoretical foundation of prediction in this paper which are critical important and meaningful.

2.1 Introduction to Support Vector Regression


The SVR is a kind of machine learning method associated with the SVM. There are two type of SVR models, ε – SVR and ε–SVR. Both of the SVR models have gained lots of achievements in solving time series and non-linear regressions issues.

Suppose there are a set of data described as \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \in \mathbb{R}^m \times \mathbb{R}. x_i represent the input vectors which could be time series data. y_i represent the vested actual values of the targets, here i varies from “1” to “n” which is the number of the total input vectors.
For instance, to establish $\varepsilon - SVR$ regression model, the following equations was formulated (Vapnik 1999) to mapping into a high dimensional feature space non-linearly.

$$f(\mathbf{x}) = \mathbf{v} \cdot \phi(\mathbf{x}) + b$$

In equation (1), $\mathbf{v}$ represent weight vector, $b$ is a constant variable and $\phi(\mathbf{x})$ is a mapping function in the feature space $\mathbf{Z}$.

Then the $\varepsilon$-intensive loss function (formula (2)) could be employed in the SV Regression model.

$$L_i := \begin{cases} |f(\mathbf{x}_i) - \mathbf{y}_i| - \varepsilon, & \text{if } f(\mathbf{x}_i) - \mathbf{y}_i \geq \varepsilon \\ 0, & \text{otherwise} \end{cases}$$

The weight vector $\mathbf{v}$ and constant $b$ would be calculated by using the regularized equation (3), where the variable $\mathcal{C}$ is used to specify the trade-off, $\frac{1}{2} \| \omega \|^2$ also used to control the trade-off of the regression model.

$$\mathcal{R}(\mathcal{C}) = \frac{1}{n} \sum_{i=1}^{n} L_i + \frac{1}{2} \| \omega \|^2$$

The former desired weight vector of the regression hyper plan $\mathbf{v}^* = \sum_{j=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}_j)$. Hensley the desirable SVR function (equation (4)) comes out. $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function and $\alpha_i, \alpha_i^*$ are the Lagrangian multipliers variables.

$$f(\mathbf{x}) = f(\mathbf{x}, \alpha, \alpha^*) = \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}_j)$$

$\nu$-SVR is a new variant SVR model which uses $\nu$ to control the number of support vectors effectively. Here we don’t describe the detail of it. The variant regression function is shown in formula (5) and the variables have the same meanings above.

$$f(\mathbf{x}) = f(\mathbf{x}, \alpha, \alpha^*) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b$$

The Table 1 listed common kernel functions which could be used in SVR model. The symbols in Table 1 are used to represent corresponding SVR models in the following sections we proposed.

<table>
<thead>
<tr>
<th>Kernel Functions</th>
<th>Kernel Formula</th>
<th>SVR Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$x^T x_i$</td>
<td>$\varepsilon - SVR$</td>
</tr>
<tr>
<td>Polynomial</td>
<td>$(\gamma x^T x_i + 1)^q$</td>
<td></td>
</tr>
<tr>
<td>Radial basis function</td>
<td>$\text{exp}(-|x_i - x|^2/\sigma^2)$</td>
<td>$\nu - SVR$</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>$\tanh[\gamma x_i^T x + c]$</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Machine Learning Methods and Kernel Functions

The Figure 1 below illustrated the logical structure of SVR and the classification of SVM which shown the basic procedure and core structure of this machine learning tool.
Figure 1. The Logical Structure of SVR and Description of SVM Classification

3 RESEARCH METHODOLOGY

In this section, a new method of real estate price prediction is proposed by combining online daily news sentiments and Google search engine query data that record the frequency of words searched by online users. We construct this integrated data mining model that both addresses real estate news and human searching behaviours. Apart from this, the lags of house price index time series data are also integrated into the model for prediction. Then we make a comparison of predicting the house price index (HPI) data between the integrated model and the non-integrated model without human searching factor. The HPI is an index constructed by China Index Academy (CIA) which is one of the authoritative index of China cities’ real estate price. The framework of the proposed model is shown in Figure 2. More details will be discussed in the following subsections. We take Beijing city as our research object to clarify the research methodology in this paper.
3.1 Data Acquisition and Data processing

Experimental data are acquired and processed in terms of the following steps:

Step 1: Crawl online news articles & real estate price index time series (HPI monthly data).

Since these news are used to reflect the actual operation and situation of Beijing real estate market, they should come from reliable sources whose information is true to reality. Thus it is necessary that we crawl the news from authoritative websites at regular intervals. The biggest and historical web station sina.com is a good choice to be the source of news articles. In the meantime, the predicted variable in this paper, the House Price Index, is one of the authoritative index to reflect the real estate market situations. The two data could be gathered by using self-programmed web crawler.

Step 2: Key words query dataset construction & Sentiments extraction.

138 most frequent words in all the crawled articles are recorded and the search frequencies of these words are obtained from Google Trend. The 138 words with their search frequencies form the key words query dataset used in our experiment as a time series input.
Meanwhile, since the news articles contain sentiment words which could reflect public mood and opinion, we can infer the operation of real-estate market by examining the sentiment score of the crawled articles in the same period of time. The sentiment score of each article is calculated using the open-source package ICTCLAS50 and the sentiments dictionary from Department of Chinese, Tsinghua University. There are four types of sentiment score: positive, negative, the addition of positive and negative, the subtraction of positive and negative sentiments.

In this way, we can obtain four sentiment series: $P_t$ (positive sentiment score), $N_t$ (negative sentiment score), $U_t$ ($P_t + N_t$), and $S_t$ ($P_t - N_t$). $P_t$ stands for the positive sentiment score of all the crawled articles at time point t. $N_t$ represents the negative sentiment score which can be calculated by using the sentiments dictionary. $U_t$ and $S_t$ are the addition and subtraction of them.

The equation 6 below described the relationship of variables above where $i$ represent the news article sequence number of a specific time.

$$
P_t = \sum_i P_{ti}, \quad N_t = \sum_i N_{ti}, \quad U_t = \sum_i U_{ti}, \quad S_t = \sum_i S_{ti} \ldots \ldots \ldots (6)$$

The four sentiment series in this step are called original sentiment series.

**Step 3**: News content classification &Weight the sentiment scores.

In this step, we classify the news according to the key words dataset. Each piece of news got a key word by counting the appearance frequencies of each word. Then each news is associated with a specific key word and the query data on a related time. Later the search frequency (query data) of the key word for each article is used as the weight of its original sentiment score to form a new sentiment score which is called weighted sentiment score. Just like in step 2, the weighted sentiment score has four types. And they can respectively be calculated by Formula (7) where the subscript $i$ represent No.$i$ news at time point t. Thus we can acquire four weighted sentiment series that are essential for prediction.

$$W_{sentiments_{it}} = O_{sentiments_{it}} \cdot \frac{SearchingVolume_{it}}{\sum SearchingVolume_{it}} \ldots \ldots \ldots (7)$$

After data processing, we have obtained 9 essential time series: 4 original sentiment series, 4 weighted sentiment series and the HPI time series. With these time series, we can construct a prediction model with the help of SVR. More details of the SVR modelling are to be discussed in the following sections.

### 3.2 Regression Model Construction

Next we could use the time series time to construct the regression model. To continue the three steps in the former subsection, this section can be summarized as the following two steps and Figure 2.

The following steps are below.

**Step 4**: SVR Modelling.

In Table 1 we list four common types of $K(X_t, X)$ and two types of SVR model. For the purpose of enhancing the accuracy of prediction, we try all the eight SVR models in Table 1 that use different SVR types and employ different kernel functions. In each of the eight SVR models, we also integrate different combinations of the four types of sentiment series and the HPI time series and their lags into the regression models. The combinations of the four sentiment series can be seen in the following section in Table2.

In the experiment, the 10 fold cross-validation method is employed to measure the performances of different models aforementioned. Then the model that outperforms the others will be selected for prediction and comparison.

The whole process of SVR modelling is demonstrated in Figure 2 with multiple time series inputs.
3.3 Prediction and Comparison

The selected model in step 4 is applied to predict the HPI. Finally, the integrated model with weighted sentiment series is compared with the model with original sentiment series which the difference of the two series is the influence of online users’ searching behaviours.

The predicted HPI could be obtained by the regression function below, the variable $i$ is the index of month, the $F_i$ represent the predicted house price index in month $i$, and $n$ is the number of months:

$$
epsilon - SVR : \quad F_i = f(X_i, \alpha_j, \alpha_j^*) = \sum_{j,k=1}^{m} (\alpha_j - \alpha_j^*) K(X_{ij}, X_{ik}) \quad \ldots \ldots \ldots \ldots (8)$$

$$\nu - SVR : \quad F_i = f(X_i, \alpha_j, \alpha_j^*) = \sum_{j=1}^{m}(\alpha_j - \alpha_j^*) K(X_{ij}, X_i) + b \quad \ldots \ldots \ldots (9)$$

Root Relative Squared Error (RMSE) is used as the criterion of model evaluation. The lower its prediction RMSE is, the better the model performs. RMSE is calculated by the equation (10):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(F_i - H_i)^2}{n}} \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (10)$$

4 EMPIRICAL ANALYSIS

4.1 Data Source and Description

The crawled news articles cover the time span from July, 2010 to March, 2013 from the “Sina Real Estate News” web station (site: http://bj.house.sina.com.cn/#). Besides, the sentiment series are formed by the open-source program package ICTCLAS50 developed by the Institute of Computing Technology, Chinese Academy of Science (ICT-CAS) accompany with the sentiments words dictionary constructed by the Department of Chinese Language, Tsinghua University. And the key words search frequency from the Google search engine can be downloaded as CSV file from Google Trends (site: http://www.google.com/trends/#).

In addition, the House Price Index (HPI) is published by China Index Academy (CIA), the professional real estate research and statistics organization, every month. The HPI is one of the most
authoritative real estate indexes in China thus it is used as the indicator of real estate market in our research model.

4.2 Empirical Results

According to the Proposed Method section mentioned before, we have experiments with multiple regression models and combinations of different time series in Table 2. The performances of these models are listed in Appendix Table 1 and Figure 3 below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Attributes ID</th>
<th>Attributes ID</th>
<th>Attributes ID</th>
<th>Attributes ID</th>
<th>Attributes ID</th>
<th>Attributes ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_i$</td>
<td>2</td>
<td>$N_i$</td>
<td>3</td>
<td>$U_i$</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>$p_i,U_i$</td>
<td>7</td>
<td>$p_i,S_i$</td>
<td>8</td>
<td>$N_i,U_i$</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>$p_i,N_i,U_i$</td>
<td>12</td>
<td>$p_i,N_i,S_i$</td>
<td>13</td>
<td>$p_i,U_i,S_i$</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 2. Combinations of Attributes

It is shown in Figure 3 that models with linear kernel function basically have better performance than others. It also implies that the $\varepsilon - SVR$ linear model outperforms the $\nu - SVR$ linear model. Besides, the advantages of the models using weighted sentiment series can also be known easily. Firstly, in a general view, the models using weighted sentiment time series have better performances when comparing the average RMSE in Table 3 which mean the weighted sentiment series have a higher prediction accuracy.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\varepsilon - SVR$</th>
<th>$\varepsilon - SVR$</th>
<th>$\varepsilon - SVR$</th>
<th>$\varepsilon - SVR$</th>
<th>$\nu - SVR$</th>
<th>$\nu - SVR$</th>
<th>$\nu - SVR$</th>
<th>$\nu - SVR$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Poly</td>
<td>RBF</td>
<td>Sigmoid</td>
<td>Linear</td>
<td>Poly</td>
<td>RBF</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>Original Series</td>
<td>371.07</td>
<td>705.42</td>
<td>702.04</td>
<td>700.49</td>
<td>1704.09</td>
<td>11839.06</td>
<td>719.92</td>
<td>714.86</td>
</tr>
<tr>
<td>Weighted Series</td>
<td>220.33</td>
<td>267.27</td>
<td>688.73</td>
<td>686.76</td>
<td>1884.68</td>
<td>11218.43</td>
<td>703.83</td>
<td>668.94</td>
</tr>
</tbody>
</table>

Table 3. Average RMSE of Each Constructed Models

After we selected the best prediction time series and model, we find the second finding. The Figure 3 reveals that the HPI prediction using weighted sentiment series and the $\varepsilon - SVR$ type with linear kernel function has more stable and accurate effects. It is clearly that the RMSE of the weighted sentiments time series are strictly lower than the original ones and what’s more the majority of the weighted sentiments time series are stable and humble.

Figure 3. The RMSEs of the Selected Best Regression Models
These two findings successfully support the hypothesis that human behaviour, online users’ searching act in, do reflect the real estate market and it is a useful factor and tool for predictions. The forecasting performance will be remarkably improved when actual human searching behaviour is taken into account.

5 CONCLUSIONS AND FUTURE WORK

This paper presents a novel integrated model for real estate market prediction. In the proposed method, online news articles about real estate market are crawled and four types of sentiment series are generated. Then online users’ searching action which is recorded by search engine query data is integrated into the prediction by adding weights to the original sentiment series. After that, different types of SVR models are employed to do the prediction. Then the one that is both stable and has low RMSE is selected for eventual prediction. Eventually, we make a comparison between models with original sentiment series and models with weighted sentiment series. Empirical results reveal that the prediction performance of weighted sentiments series surpasses the original sentiments series. This supports the hypothesis that online users’ behaviour has great value in real-estate market prediction research.

Nevertheless, this paper still leaves questions and potential opportunities which need to be answered and further studied. Firstly, how to improve the quality of the data should be thought over. The prediction will have low accuracy if the experimental data comes from unreliable sources. Moreover, other regression models besides the SVR model should be tried and explored in order to achieve better forecasting performance. For instance, the Neutral Networks and other machine learning models may also fit our model and should be considered. Furthermore, the selection of key words and the weighting process should also be paid more attention because they are crucial in the model. What’s more, the proposed method and idea can also be applied to other research areas such as stock index prediction, e-commerce consumption forecasting as well as other significant areas.

Appendix

<table>
<thead>
<tr>
<th>Model</th>
<th>(\varepsilon - SVR)</th>
<th>(\nu - SVR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Poly</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 O</td>
<td>250.96</td>
<td>648.14</td>
</tr>
<tr>
<td>W</td>
<td>252.66</td>
<td>179.98</td>
</tr>
<tr>
<td>2 O</td>
<td>312.71</td>
<td>261.11</td>
</tr>
<tr>
<td>W</td>
<td>205.63</td>
<td>192.40</td>
</tr>
<tr>
<td>3 O</td>
<td>254.00</td>
<td>628.34</td>
</tr>
<tr>
<td>W</td>
<td>163.60</td>
<td>197.27</td>
</tr>
<tr>
<td>4 O</td>
<td>238.05</td>
<td>658.33</td>
</tr>
<tr>
<td>W</td>
<td>203.98</td>
<td>577.58</td>
</tr>
<tr>
<td>5 O</td>
<td>331.88</td>
<td>656.19</td>
</tr>
<tr>
<td>W</td>
<td>338.75</td>
<td>311.33</td>
</tr>
<tr>
<td>6 O</td>
<td>483.50</td>
<td>768.75</td>
</tr>
<tr>
<td>W</td>
<td>215.73</td>
<td>351.58</td>
</tr>
<tr>
<td>7 O</td>
<td>432.94</td>
<td>772.18</td>
</tr>
</tbody>
</table>
Appendix Table 1  The Experimental Results of SV Regression

<table>
<thead>
<tr>
<th>O</th>
<th>W</th>
<th>201.32</th>
<th>197.80</th>
<th>687.90</th>
<th>684.08</th>
<th>377.43</th>
<th>NaN</th>
<th>705.27</th>
<th>205.27</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>W</td>
<td>445.84</td>
<td>669.67</td>
<td>694.16</td>
<td>685.51</td>
<td>487.21</td>
<td>NaN</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>245.37</td>
<td>185.21</td>
<td>683.51</td>
<td>683.54</td>
<td>223.24</td>
<td>NaN</td>
<td>705.27</td>
<td>705.22</td>
</tr>
<tr>
<td>9</td>
<td>W</td>
<td>305.05</td>
<td>686.50</td>
<td>683.54</td>
<td>688.28</td>
<td>381.30</td>
<td>NaN</td>
<td>690.77</td>
<td>702.86</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>202.01</td>
<td>267.25</td>
<td>683.54</td>
<td>683.54</td>
<td>8004.52</td>
<td>NaN</td>
<td>700.68</td>
<td>705.22</td>
</tr>
<tr>
<td>1</td>
<td>W</td>
<td>296.94</td>
<td>685.29</td>
<td>697.67</td>
<td>683.54</td>
<td>401.16</td>
<td>NaN</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>212.11</td>
<td>192.40</td>
<td>683.54</td>
<td>683.54</td>
<td>226.33</td>
<td>NaN</td>
<td>705.27</td>
<td>684.98</td>
</tr>
<tr>
<td>0</td>
<td>W</td>
<td>364.12</td>
<td>878.06</td>
<td>683.54</td>
<td>706.01</td>
<td>432.29</td>
<td>NaN</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>207.84</td>
<td>577.58</td>
<td>684.21</td>
<td>697.82</td>
<td>227.37</td>
<td>NaN</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td>1</td>
<td>W</td>
<td>427.29</td>
<td>756.80</td>
<td>688.96</td>
<td>688.89</td>
<td>321.41</td>
<td>23797.82</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>225.00</td>
<td>197.80</td>
<td>683.54</td>
<td>683.54</td>
<td>219.63</td>
<td>NaN</td>
<td>705.22</td>
<td>705.27</td>
</tr>
<tr>
<td>2</td>
<td>W</td>
<td>359.67</td>
<td>775.22</td>
<td>683.54</td>
<td>683.54</td>
<td>431.22</td>
<td>NaN</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>204.36</td>
<td>197.80</td>
<td>684.11</td>
<td>688.89</td>
<td>210.25</td>
<td>NaN</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td>3</td>
<td>W</td>
<td>302.36</td>
<td>858.66</td>
<td>743.30</td>
<td>726.56</td>
<td>404.87</td>
<td>NaN</td>
<td>718.64</td>
<td>755.06</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>210.16</td>
<td>197.80</td>
<td>695.68</td>
<td>700.38</td>
<td>212.21</td>
<td>23797.82</td>
<td>705.27</td>
<td>705.27</td>
</tr>
<tr>
<td>4</td>
<td>W</td>
<td>760.73</td>
<td>878.08</td>
<td>743.56</td>
<td>743.61</td>
<td>NaN</td>
<td>NaN</td>
<td>705.30</td>
<td>705.22</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>216.45</td>
<td>185.21</td>
<td>697.82</td>
<td>686.43</td>
<td>392.08</td>
<td>23797.82</td>
<td>705.30</td>
<td>705.22</td>
</tr>
</tbody>
</table>

Note: O represent the Original Sentiments Series, W represent the Weighted Sentiments Series. NaN means this model couldn’t find a proper regression solution in the experiment.

Acknowledgement

This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No. 71001103, 91224008, 71273265, 91324015), National Social Science Foundation of China Major Program (Grant No. 13&ZD184), Beijing Social Science Fund (No. 13JGB035), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No. Z131101000413058), and Program for Excellent Talents in Beijing.

References


