



Real Estate Confidence Index Based on Real Estate News

Yukun Ma, Bin Xu & Xiaofei Xu

To cite this article: Yukun Ma, Bin Xu & Xiaofei Xu (2018) Real Estate Confidence Index Based on Real Estate News, *Emerging Markets Finance and Trade*, 54:4, 747-760, DOI: [10.1080/1540496X.2016.1232193](https://doi.org/10.1080/1540496X.2016.1232193)

To link to this article: <https://doi.org/10.1080/1540496X.2016.1232193>



Published online: 25 Oct 2017.



Submit your article to this journal [↗](#)



Article views: 394



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

Real Estate Confidence Index Based on Real Estate News

Yukun Ma¹, Bin Xu¹, and Xiaofei Xu²

¹*Department of Computer Science and Technology, Tsinghua University, Beijing, China;* ²*School of Social Sciences, Tsinghua University, Beijing, China*

ABSTRACT: A real estate confidence index (RECI) is used to evaluate real estate industry development, and it has become an effective and powerful measure in China's real estate market (REM). RECI research based on big data is the new trend in finance and economics. In this article, we apply some methods of text classification to research on the construction of RECI. First, the Naïve Bayes algorithm is used to evaluate data and to classify the extent to which this measure describes confidence in the REM. Second, experiments on different perspectives are performed to probe the relationship between variables and the accuracy of the classifier. Third, we use the classifier to predict the weekly news. Ultimately, construction of the RECI based on financial and economic news is achieved by applying the classifier to the time and existence of major financial and economic news.

KEY WORDS: accuracy, naïve Bayesian, RECI, text classification

The real estate index (real estate confidence index [RECI]) is very important to the real estate market (REM) as well as the wider economy. First of all, RECI is a reflection of the residents' desire to buy real estate, which can directly impact the investment demand. In accordance with the principles of macroeconomics, the purchase of real estate and other fixed assets should be attributed to the investment demand. China is a country with special preference toward real estate, in addition to the purchase of property owner occupied and the preference of Chinese residents with the purchase in the form of real estate asset value. In addition, Chinese residents always take the real estate as a gift to the descendants. Therefore, RECI determines the behavior of residents to buy real estate. Once the RECI enhances, residents will buy real estate, which will increase investment demand and the total GDP. Second, RECI indirectly impacts consumer demand. China's real estate prices have been very high. Beijing, Shanghai, and other first tier cities are even higher. Residents' purchases of real estate are mostly in the form of loans. Thus, once the RECI increases, residents will use loans to buy real estate. Then, the level of residents' consumption will decline, which will reduce the consumer demand and will impact the GDP. Third, RECI also affects other aspects of the economy. With the changes of RECI, residents will change their demands on stocks, bonds, foreign exchange, and other financial assets. If the RECI increases, residents will choose to invest in real estate, and then the investment in stocks, bonds, foreign exchange market funds will be reduced, which will affect the development of those types of markets. Because RECI exerts great importance to REM and other economy fields, many researchers have made a lot of exploration in these areas. For example, Shiller (2008) used sentimental data to inform on turning points in the REM. Clayton, Ling, and Naranjo (2009) explored the role of sentiment versus fundamentals in commercial real estate valuation. Recently, Lang and Schaefer (2015) used an economic confidence index in a study of European real estate sector stock returns. In summary, RECI has a great impact on many variables of macroeconomy and then affects the whole economy. It is very clear that RECI is a critical variable in economy analysis.

In China, concern has grown over the real estate industry, and it serves as an indicator of national economic development. Increased public attention to this market has resulted in demand for effective

Address correspondence to Yukun Ma, Department of Computer Science and Technology, Tsinghua University, Beijing, China. E-mail: yukunma@aliyun.com

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/mree.

and efficient evaluation of RECI. In western countries, RECI has become a popular system that reflects and effectively measures the REM. However, the REM in China is far different from those of other countries, despite the use of readymade evaluation indices. Moreover, the healthy development of the REM is important for economic growth, industrial structure, and population quality of life. More importantly, both market and governmental disability are inevitable when addressing China's incomplete REM problems. Consequently, the RECI needs to be more heavily studied, and a new effective evaluation index specific to the Chinese REM is necessary.

Opinions differ greatly with respect to real estate confidence, which seems to be a drawback in RECI construction because of the need to take additional factors into consideration. Otherwise, the index we construct will turn out to be ineffective. For instance, a traditional technical framework for constructing a domestic real estate index and its architecture mainly stem from analysis of the domestic REM and overseas experiences. The confidence indices mainly include efficient supply and demand, latent demand, and latent supply. As for the efficient supply and demand index, 25 factors influencing efficient supply and demand for housing were analyzed via questionnaire and factor analysis. Four main factors were determined from this analysis including population, purchasing power, price, and market. Some relevant index models are established based on these factors. However, the factor we have chosen to construct the RECI may not fully indicate people's opinions. For example, some seemingly unimportant factors that exert great influence may be omitted while others are emphasized. Therefore, direct consideration of people's opinions about real estate may be a valid approach for constructing the RECI. An abundance of news about real estate offers the chance to collect data about confidence in real estate. So, after some surveys and analysis are conducted, we find that people's sentiments about real estate news have close relationships with housing prices and the change of REM. For instance, if most people hold a pessimistic attitude toward the news they read, that may hint housing price rises fast and they never afford, and vice versa. This tendency can be caught in the latter news labeling. In fact the news labeling procedure is part of our survey. Real estate news is useful data and contributes to information-oriented measurement. However, we must answer three questions regarding such information: (1) How to get more people's opinions about real estate? (2) How to deal with the different opinions from different people? (3) How to balance the data's objectivity and human subjectivity? As for the first question, the answer is presented in the data description. The latter two questions need to be probed by experiment and will be answered later.

The construction of the RECI is based on emerging data and assessments. Machine learning is used to generate a classifier to help people make judgments about the real estate news. An illusion we may have is that the assessments from only one person will generate a high-accuracy classifier. However, one person's opinion may not represent those of others'. Different assessments from different people may reduce the discrimination of news differences but may turn out to be valid. Reducing discrimination means reducing the quality of the classifier, which becomes a paradox we must handle. The naïve-Bayes algorithm and support machine vector (SVM) are widely used and have proven to be very effective. In this article, two algorithms are implemented to generate two classifiers. The one with the best performance will be chosen. Additionally, other factors (e.g., titles, news origins, number of labels, and accuracy) in real estate news need to be considered, and their relationship may enhance the quality of the classifier.

The RECI will be issued weekly. Therefore, the classifier computes a probability value each week and the weekly probability value is based on the probability values of each of the 7 days of the week. Thus, the discrete model classifier is transformed into a continuous model classifier and the prediction value can be obtained every week. Ultimately, the RECI trend can be revealed by the continuous weekly prediction value.

In summary, we consider real estate news as big data and some effective machine-learning methods are applied to the dataset so the RECI can reflect the REM. The contributions of this work include the following:

- We formally propose a new perspective on big data to construct the RECI.

- We use real estate news as the medium to reflect people's confidence about the REM in China and apply the machine-learning method of text classification to construct China's RECI.
- We deeply probe the relationship between elements in the news on real estate and accuracy of the classifier.

Related work on RECI and sentimental text classification is summarized first. Then, we describe the elementary theories of the classification algorithms. Detailed experimental data are provided in the following section. After that, we show experimental results applied to our dataset and provides detailed analysis of the results. The conclusions of the article are made, and then suggestions for future work are presented..

Related Work

Research on the RECI appeared earlier in foreign countries and a relative index, called the consumer confidence index (CCI), was proposed. Assessments of confidence in this article are derived from latent or direct household consumers. Consequently, the construction of the CCI has useful implications in constructing the RECI. As research on the CCI has shown, this measure has achieved favorable performance in western developed countries. Moreover, the RECI based on consumer sentiment has helped inform policy for the REM. Since the second quarter of 1999, the Research Centre for Real Estate at Texas A&M University has been constructing the Texas RECI (TRECI), releasing data regularly. The principle of the TRECI is similar to that of the CCI. Specifically, the TRECI chooses the number 100 as the threshold value and a booming market is signified when the index is more than this level. Otherwise, a recession market is said to take place. In the last few years, the construction of news-based sentiment/confidence measures has become more prevalent in the finance literature. In 2007, Tetlock quantitatively measured the interactions between the media and the stock market and found that high media pessimism predicted downward pressure on market prices followed by a reversion to fundamentals and unusually high or low pessimism predicted high market trading volume. In 2011, Engelbers and Parsons compared the behaviors of investors with access to different media coverage of the same information event to solve the problem of disentangling the causal impact of media reporting from the impact of the events being reported. They found that local coverage strongly predicted local trading, after controlling for earnings, investor, and newspaper characteristics. In China, some research on the RECI has been undertaken. For example, the BRE Index was compiled by Hong Kong Polytechnic University for the Hong Kong residential property market. The BREI is similar to that of the CCI and is derived from questionnaires about the REM. The methods applied in the RECI are varied. However, its basic principle is not different from the CCI. Some insignificant differences appear in the content of the index, selected threshold, size of the sample, and computation details. Another example of RECI is based on web GIS and SPSS WebAPP. In this study, a new RECI system is developed by integrating latent supply data, current demand, and GIS and WebAPP technologies. The dissemination of indices is effectively conducted by using online technologies. Generating the RECI according to the people's confidence is undoubtedly feasible. The article drew from the CCI to reconstruct the RECI based on financial and economic big data. Data were gathered via consumer confidence. The advantage of constructing the RECI based on big data mainly lies in the ability to collect real-time data.

In this article, some machine-learning algorithms are involved. Text classification is widely applied to a range of fields, and research in this area is growing. With the development of the Internet and the advent of the big data era, text classification shows great data analytic potential. Since Salton and McGill proposed the bag-of-words model, plenty of algorithms have used this model to undertake classification. Naïve Bayes algorithm is also based on this model, and the algorithm is very simple and effective. The Naïve Bayes algorithm has two kinds of models. One is the Bernoulli model, and the other is the multinomial model. In 1998, McCallum and Nigam showed that the multinomial model performs better. The classifier algorithm in this article is based on the multinomial model. The

classifiers based on Naïve Bayes have been generated via different improved ideas, but the improvement varies based on data in the experiments. Enhancement of the classification ability is necessary in this article. The other algorithm involved in this article is the SVM. SVM has been successfully used in classification and has been proved to be efficient for the analysis of high-dimensionality data. SVM is a profound mathematic theory, and many of its variants have been widely studied. SVM can deal with different data because of kernel function. For classification, many researchers have worked on machine memory reduction and cost-time optimization, which are very useful in practical settings.

As for different text data, a theoretically effective algorithm may perform badly in the experiment. Thus, it is necessary to implement more than one algorithm to make some comparisons. In this article, experiments will be conducted to prove that the Naïve Bayes model is more effective than SVM, despite the theoretical elegance and superiority of the latter.

Basic Theories

In this section, we provide the basic mathematic theories of the Naïve Bayes model and SVM. However, some preliminaries first need to be presented. Text classification refers to the assignment of a text document to one or more labels. The labels of the training corpus are most often assigned manually, and the purpose of text classification is to train the corpus to generate a classifier and label more documents algorithmically. We consider a document as a large vector, and some important features are selected according to different algorithms. The selected features generate the feature vector, which will be practically used in the classification. Stronger discriminability of the features vector results in higher accuracy of the classifier. However, the phenomena of ambiguity in Chinese text documents are prevalent, presenting new challenges in Chinese text classification. As a result, a pragmatically effective classification algorithm should be chosen.

Terms Definition

In this section, some terminologies in machine learning defined as follows:

- **Corpus:** A corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to train and validate models.
- **Classification:** Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
- **Classifier:** A classifier is a software implementation of a machine-learning algorithm. We use classifier to classify unlabeled data.
- **Training set:** A training set is a set of data used to discover potentially predictive relationships.
- **Test set:** A test set is a set of data used to assess the strength and utility of a predictive relationship. In this article, we use test set to validate model's classification quality.
- **Categories:** Before training a data set, some different categories should be determined. In this article, we have three categories: confidence, no-confidence, and uncertainty.
- **Label:** As a verb, label means we assign one or more categories to a piece of document manually or automatically from the given categories in order to generate training set. As a noun, label is a representation of a document category in data sets. In this article, we first label some amount of news with -1 (no-confidence), 0 (uncertainty), and $+1$ (confidence) as training set, where -1 , 0 , and $+1$ are the news labels, respectively.
- **Accuracy:** The accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's true value, which used to measure the quality of the machine-learning model. In this article, after generating the classifier, we use the classifier to classify labeled test sets and count the correctly classified documents denoted by n_1 . Then, we have $\text{accuracy} = (n_1/n_2)$, where n_2 is the count of documents in test set in a validation experiment.

Generally speaking, if categories of documents distinguish greatly with each other, then we can obtain higher accuracy. So in this article, we define “uncertainty” as a single category so that we can obtain reduced-ambiguity training set, which is based on the fact that people is uncertain of a document and the same to machine.

The Multinomial Naïve-Bayes Model

The classic Bayes theorem is one of the most famous theorems in statistics. As we have presented, a Bayesian classifier gives a conditional probability for every document. The basic formula is as follows:

$$p(y|x) = \frac{p(x|y) \times p(y)}{p(x)} \quad (1)$$

where y stands for the category of a document and x is the document as we have mentioned. Now we consider

$$x = \{a_1, a_2, \dots, a_m\} \quad (2)$$

$$C = \{y_1, y_2, \dots, y_n\} \quad (3)$$

as the unspecified document and category set, respectively. That means, we don't know the category of the document x , and our task is assign a category to x . Meanwhile, we predefine a category set C , where $a_t (t = 1, 2, \dots, m)$ stands for a word in the document x , and $y_i (i = 1, 2, \dots, n)$ represents one category in set C . Then, we compute all the conditional probabilities for the unspecified item and choose the category with maximum probability as the final category.

$$P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\} \quad (4)$$

A basic assumption in the Naïve Bayes model is that each feature in the features vector is independent from all others. So, according to the Bayesian theorem,

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} (i = 1, 2, \dots, n) \quad (5)$$

We can see that the denominator in the right of the theorem is a constant if the training data is determined. Thus, we just need to maximize the numerator.

$$P(x|y_i)P(y_i) = P(y_i) \prod_{j=1}^m P(a_j|y_i) \quad (6)$$

The multinomial model with Laplace's rule of succession can be presented as follows:

$$p(a_t|y_i) = \frac{1 + \sum_{x_j \in C} n(a_t, x_j)}{m + \sum_{t=1}^m \sum_{x_j \in C} n(a_t, x_j)} \quad (7)$$

where $n(a_t, x_j)$ is the number of occurrences of word a_t in document x_j . The probability $p(a_t|y_i)$ is always vanishingly small and would be truncated in computation, so rewriting formula (7) by taking logarithms is the routine, which can prevent losses in precision of the classifier.

Support Machine Vector

The SVM algorithm considers all the data as the vector space. The purpose of classification is to find the decision surface that can optimally separate one category from the other. Usually, the SVM is used to solve two-category pattern recognition problems, and multi-category classification is established on the basic two-category category model. Let $C = \{+1, -1\}^m$ be the label for the document d , where $+1$ stands for positive category and -1 stands for the negative. When given a training set of labeled documents $(x_i, y_i), i = 1, 2, \dots, m$, where $x_i \in R^n$ and $y_i \in C$, SVM solves the optimization problem as follows:

$$\begin{aligned} \min_{w, b, \xi} & \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i \\ \text{s.t.} & y_i(w^T \Phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \quad (8)$$

To simplify the text classification problem, we rely on the two-category model. The multi-category classifier can be implemented by training N one-versus-rest classifiers or training $N(N-1)/2$ one-versus-one classifiers for the N categories case. Here, we assume that the input data point x is linear separable data, and the decision surface can then be written as

$$\langle w, x \rangle - b = 0 \quad (9)$$

where b is a constant that needs to be learned from the training set. For an arbitrary document feature vector $x_i \in R^n$, if $\langle w, x \rangle - b \geq +1$, the document will be assigned as a positive category. Otherwise, the document will be assigned as negative category. If the input data point x is nonlinear, the function Φ in (1) should be shifted as the kernel function as follows:

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \quad (10)$$

The input data points are mapped into a higher dimensional space by kernel function. The decision surface then becomes a hyper-plane that maximizes the distance between two categories. The maximum distance indicates the classifier's strong classification and generalization ability.

Data Description

The construction of RECI is based on real estate news, so we need a variety of different dimensional data to explore relationships between these variables. The news we apply in the experiments is derived from the website <http://dichan.sina.com.cn/>. A crawler program has been built to collect all the news in the website since July 31, 2014. A crawled news document includes the title, crawl time, release date, source, and content. Some elements in news document may exert certain influence on the RECI, so consideration of this influence should be taken.

First, we requested some professionals to label a certain amount of news. The news was assigned as confidence, no-confidence, or uncertainty. Each piece of news was labeled by five professionals. Uncertainty may consist of ambiguity and independence. Information from a piece of news is objective, so we can obtain consensus for this data. If opinions vary, then the label of uncertainty is attached. After labeling some news, we built a specialized website for more people to label news. This news was used to evaluate how serious participants were in the labeling process. In the online-labeling process, a piece of news was labeled by three people. After the labeling process, more news with different assessments could be collected.

The labeled news is preprocessed as the training set for the classifier, giving us a number of data with different labels to benefit our research. News could be classified into five classes based on label number. Moreover, the news could be classified into official and unofficial. These classifications are embodied in the latter experiments.

Experiments and Analysis

In this section, all of the experiments were conducted in the same hardware and software environments. The memory size of the experimental machine is 4.0 GB and the CPU is 2.0 GHz. The program execution environment is JRE1.7.0 and the development tool is JDK1.7.0.

First, we digitized the news labels by assigning +1, -1, and 0 to replace confidence, no-confidence, and uncertainty, respectively. The news labeled by five people was chosen as the first training set to determine the choice of classifiers. In this experiment, the sum of five labels was the final label of a piece of news (If $\text{sum} > 0$, label = +1; if $\text{sum} = 0$, label = 0; if $\text{sum} < 0$, label = -1). Then, after the data cleaning and data preprocessing, we established two text classifiers through the Naïve Bayes model and SVM based on 300 pieces of news. Table 1 illustrates the accuracies of the two classifiers through cross validation.

The table illustrates that SVM is not always effective in the field-specified texts, especially in the Chinese sentimental text classification. In our experiment, the prediction values always trend to the category holding the largest quantity in the training set. This tendency implies that the selected feature vectors are ineffective. In the implementation of SVM, feature vectors are selected by the TF-IDF algorithm, and this algorithm selects the terms by their frequencies. However, some important terms that can provide evidence of people's judgment may have low frequencies. Therefore, SVM fails to separate one category from another. The next experiments are based on the Naïve Bayes algorithm. Table 1 also shows that the accuracy of the Naïve Bayes algorithm is not high, and work is needed to improve the classifier in the following sections.

The Label Number Choice for a Document

We hope to obtain the same single label for a document so that the classifier can better separate data points. However, the final single label of a document must truly depict different people's opinions. How to balance the objective information acquisition in a document and subjective opinions from different people is the concern in this section. In the experiment, documents labeled by five people were chosen as the training set, and five classifiers were implemented for five labels. Then, different documents with different labels were combined as the new train sets. As a result, we obtain $C_5^1, C_5^2, C_5^3, C_5^4, C_5^5$ (where C_n^k is the k -combination of n in mathematics) classifiers. We also choose the sum as the final label if one document contains more than one label. We choose accuracy and variance to demonstrate the separation ability of the classifiers and generalization ability of the different opinions. Higher accuracy of a classifier indicates better separation ability. Smaller variance indicates stronger generalization ability.

Table 2 shows that the training set with just one person's label may provide greater accuracy, but it is based on just one person's opinion. Other people may have the opposite views that the classifier does not embody. A moderately large difference among different classifiers for different people is evident, and the variance of this situation is the largest. With an increase in label number, variances decrease gradually, but the means remain unchanged. This finding indicates that the classifier with more persons' labels possesses better generalization. To summarize different opinions, the combination

Table 1. The cross validation results of two classifiers.

Classifier number	Accuracy	
	Naïve Bayes (%)	SVM (%)
1	56.67	33.33
2	56.67	33.33
3	61.13	33.33
4	52.33	33.33
5	50.00	33.33

Table 2. The accuracies and variances of different classifiers generated by different label numbers.

Number	Accuracy										Mean	Variance
C_5^1	0.67	0.70	0.53	0.43	0.56						0.58	0.01197
C_5^2	0.70	0.60	0.50	0.47	0.67	0.57	0.43	0.43	0.53	0.57	0.55	0.00869
C_5^3	0.57	0.60	0.67	0.60	0.53	0.53	0.53	0.53	0.50	0.60	0.57	0.00265
C_5^4	0.53	0.50	0.57	0.60	0.60						0.56	0.00195
C_5^5	0.57										0.57	0.00000

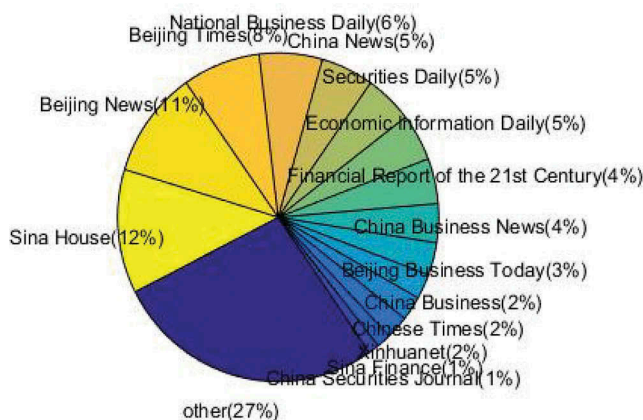
with more labels is the classifier of choice if data are available. Therefore, documents with more labels contribute to the construction of RECI.

The Influence of News Sources

We classify the training set into official and unofficial as mentioned in the data description. Intuitions provide some information that varies from official to unofficial, which has some influence on the confidence index. To research this problem, we first obtained statistical information concerning news sources, which is illustrated in the pie chart (Figure 1). Proportions of official versus unofficial news are demonstrated in the other pie chart.

From the first pie chart, it is apparent that news sources include almost all of China's important real estate websites, which represents the real estate industry. This strong representativeness means that RECI based on real estate news is reliable and credible. Figure 2 shows that official and unofficial news are almost equally divided. This indicates that official and unofficial voices are embodied in our data. Next, we used official and unofficial news as training sets to generate two classifiers, respectively. Additionally, we thoroughly investigated the relationship between different training sets and their classifiers' accuracy. After a 10-fold cross validation, our results are shown in Table 3.

Table 3 illustrates that official news leads to relatively worse classifier accuracy. This may indicate that official news contains more ambiguous terms, leading to confusion in our classifier. Meanwhile, unofficial news contributes to higher accuracy, demonstrating that such news provides more information in our classifier. We can improve the predict-ability of our classifier by merging official news and unofficial news. The construction of the RECI both based on official and unofficial news should be

**Figure 1.** News sources.

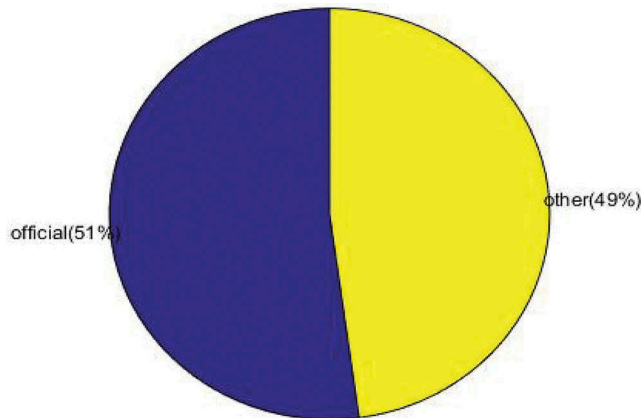


Figure 2. The proportion of official new.

Table 3. The accuracies of official and unofficial training sets.

Category	Accuracy (%)
Official	48
Unofficial	58

considered because of the objectiveness of this variable. Official news illustrates the government's attitude toward the real estate industry to some degree. Taking this fact into account helps to enhance the validity of the RECI. If different weights are attached to official and unofficial news, the objectiveness of the RECI is impaired, despite better quality of the classifier. As a consequence, we set identical weights on the two data.

Other perspectives to improve the quality of the classifier will be considered in the following sections. Title weight, different clusters of training sets, and different weights for different labeled training sets will be the concerned variables in the next experiments, respectively.

The Relationship Between Title Weight and Accuracy

News titles are very important and may provide distinctive information about its content. However, the weight of the title we need in the classifier remains unknown. Therefore, the experiment conducted in this section explores the relationship between title weight and accuracy of the classifier. Figure 3 shows the experimental results of different title weights.

Classifier accuracy is best when the title weight is between 0.1 and 0.3. If the title is not considered, accuracy is worse. However, the accuracy gets lower as title weight increases. If we just consider the title and omit the content, accuracy is the worst. Hence, in our construction of RECI, a weight of 0.2 is chosen in addition to the following experiments.

The Relationship Between the Different Clusters of Training Sets and Accuracy

As mentioned in the beginning of Section 5.1, we chose the sum of different labels as the final label for a document. Therefore, if a document was labeled by five people, then final labels are between -11 and 11 . The experiments conducted above are based on the principle: if $\text{sum} > 0$, $\text{label} = +1$; if $\text{sum} = 0$, $\text{label} = 0$; if $\text{sum} < 0$, $\text{label} = -1$. However, in this section, we try different combinations of the final label as evidence that a document belongs to the "confident cluster," "no-confidence cluster,"

or “uncertain cluster.” For example, if $\text{sum} > 1$, $\text{label} = +1$; if $-2 < \text{sum} < 2$, $\text{label} = 0$; if $\text{sum} < -1$, $\text{label} = -1$. Moreover, this case is represented by $[-5, -2]$, $[-1, 1]$, and $[2, 5]$. Based on the preliminaries, we conducted experiments to examine the relationship between different clusters of training sets and accuracy. The experimental results are summarized as follows (Table 4 and Figure 4).

Table 4 provides concrete experimental results and Figure 4 depicts the classifier trend under different clusters of training sets. The table and figure show that the case of $[-5, -1]$, $[0]$, and $[1, 5]$ has the best accuracy when compared with other situations. This result implies that when some other heterogeneous data are added into the training set of an “uncertain cluster,” the likelihood of confusion increases in the classifier. This implication leads to a bad classification result.

The Relationship Between Different Clusters' Weights and Accuracies

In the last section, we simply add all of the documents with different final labels as the final training set. This was done based on the principle that if $\text{sum} > 0$, the training set is in the confidence category, if $\text{sum} = 0$, then uncertainty; if $\text{sum} < 0$, then no-confidence. If a document is all labeled $+1, 0, -1$ by

Table 4. The accuracies of different training sets.

Number	Training set by the label sum	Accuracy (%)
1	$[-5, -1]$, $[0]$, $[1, 5]$	63
2	$[-5, -2]$, $[-1, 1]$, $[2, 5]$	60
3	$[-5, -3]$, $[-2, 2]$, $[3, 5]$	55
4	$[-5, -4]$, $[-3, 3]$, $[4, 5]$	50
5	$[-5]$, $[-4, 4]$, $[5]$	43

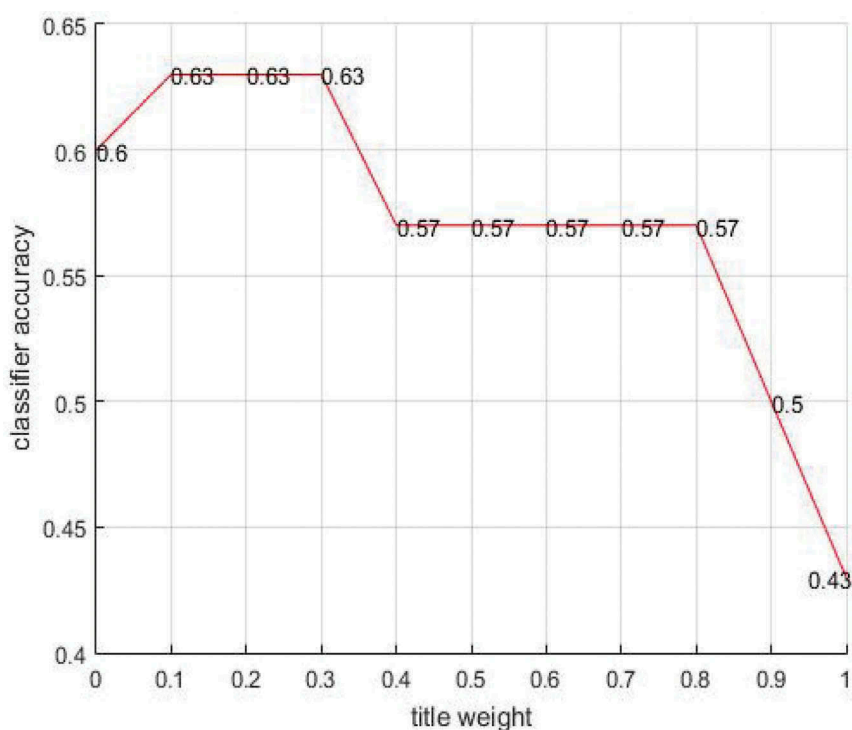


Figure 3. The relationship between title weight and accuracy.

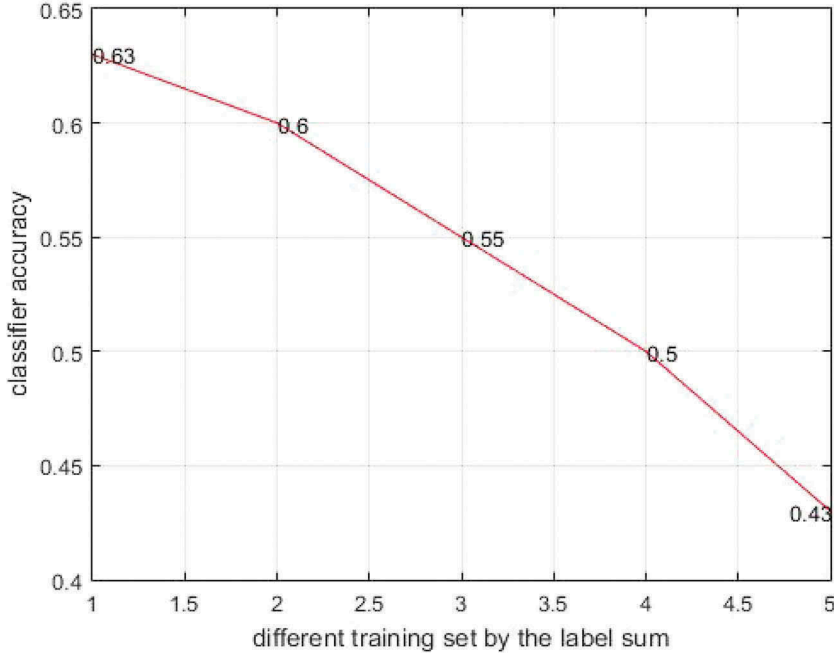


Figure 4. Cursive description of Table 4.

five people, then the document appears to possess strong objectiveness of confidence, uncertainty, and no-confidence. Specifically, sum = 11 and sum = 1 are not the same situation and we must address their differences. The approach of simply adding all the documents to construct the final training set may not be appropriate.

If we consider the documents with one sum value as a vector subspace when five people label a news source, the sum value is between -5 and $+5$ and we will obtain 11 vector subspaces. Now, we deal with the confidence training set and the no-confidence training set in the same approach. Some symbols are as follows:

$$VS_i (i = 1, 2, 3, 4, 5) \quad (11)$$

$$\sum_{i=1}^5 \lambda_i = 1 (1 > \lambda_i > 0), \quad i = 1, 2, 3, 4, 5 \quad (12)$$

where (11) represents the confidence training set vector subspaces whose label sums are 5, 4, 3, 2, 1, respectively. We add the combination factors (12) into the vector subspaces, providing the following convex combination of training sets:

$$\sum_{i=1}^5 \lambda_i VS_i \quad (13)$$

Our purpose is to determine how the value of λ_i influences classifier accuracy. A larger absolute sum value should result in a larger value of λ_i . We conducted our experiments to clearly find out the actual values of λ_i .

The experimental results in Table 5 contradict our early conjecture. The accuracy improvement is inversely related to the different clusters' training set weight. Results reveal that all documents with

Table 5. The crossing validation accuracies of different clusters' weights.

Different values of $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$	Crossing validation accuracy (%)
(1.0, 0.0, 0.0, 0.0, 0.0)	42
(0.9, 0.1, 0.0, 0.0, 0.0)	47
(0.7, 0.2, 0.1, 0.0, 0.0)	55
(0.4, 0.3, 0.2, 0.1, 0.0)	57
(0.3, 0.3, 0.2, 0.1, 0.1)	60
(0.2, 0.2, 0.2, 0.2, 0.2)	63

different label sums need to be considered. Therefore, we divide the training sets into several categories by the sum values (i.e., we increase the categories of the training sets and keep statistics on the predict result). If a document is labeled by five people, the sum values are between -5 and $+5$. Specifically, our training sets are divided into 11 categories, and the predictive values (pvs) are between -5 and $+5$. If $-1 \leq pv \leq +1$, then the final predictive result is defined as 0. If $pv > +1$, then the final predictive result is defined as $+1$. If $pv < -1$, then the final predict result is defined as -1 . This case is represented as $[-5, -2]$, $[-1, +1]$, and $[+2, +5]$ and all cases are summarized as follows:

Table 6 illustrates that classifier accuracies are not more improved when we divide the training sets into more parts. However, the accuracy sometimes increases because of the poor distribution of documents of different training set categories and lack of robustness. For example, when the number of labeled-0 documents is small, the accuracy may be 90%. However, this percentage may not apply to practical situations. According to our different processing approaches, accuracies are between 50% and 70%. Results from Table 6 suggest that we may transform the discrete classification model into a continuous one so that we can obtain a probability for every prediction document. In this way, we can obtain more detailed information from the classifier. In the next section, this idea will be demonstrated systematically.

The Weekly RECI

Real estate news factors have been considered in previous research, although raising classifier accuracy is a difficult procedure. Therefore, continued research is still needed on this topic. A primary aim of this study was to describe trends in the real estate industry. The classifier just provides a benchmark for consumer sentiment. Additionally, the index system should be assessed over time.

Here, we use the classifier to predict the news week by week. Every week, we can obtain the probabilities of confident and no-confident documents. Based on this probability, we can compile our index value to depict the probabilities continuously.

Table 6. Statistics on the results with the 11 training sets categories.

Processing approaches			Accuracy (%)
No-confidence	Uncertainty	Confidence	
$[-5, -1]$	$[0]$	$[+1, +5]$	50
$[-5, -2]$	$[-1, +1]$	$[+2, +5]$	60
$[-5, -3]$	$[-2, +2]$	$[+3, +5]$	63
$[-5, -4]$	$[-3, +3]$	$[+4, +5]$	77
$[-5]$	$[-4, +4]$	$[+5]$	65

Conclusions

This article proposed the conception of an RECI based on real estate news, and a systematic approach was taken to address this topic. We applied the machine-learning method to the construction of the RECI and developed a confidence index based on real estate news. Moreover, we analyzed the relationship between data and classifier accuracy. Results suggest that, it is difficult to derive some important sentimental feature terms with the TF-IDF algorithm, despite its efficiency in other text classifications. Second, a greater number of participants are needed to increase confidence and objectivity of results about the RECI. Third, news titles provide vital information that urge people to offer similar viewpoints. Therefore, we gave them a common weight to support our classification. Another important issue is the relation between different training sets and classifier quality. Enhancing the accuracy to some degree was found to be beneficial. However, some relationships are complex and require further research.

The ultimate purpose of our research was to construct an RECI based on real estate news and to issue the index weekly. We applied the trained classifier to the weekly news so that we could obtain probabilities of different categories. According to these probabilities, we could construct our RECI by merging the probability values. The construction of an RECI needs to be adjusted to improve its validity.

Future Work

The construction of an RECI is a systematical task and requires more intensive and deliberate consideration and adjustment. This article proposes an RECI based on big finance and economic data. This conception requires more practice validation and technical support. Simultaneously, classifier quality requires improvement. This improvement is a self-adjusted procedure. Sentimental text classification is complex because the classifier cannot separate the text correctly, especially in particular fields. For example, the news is objective to some degree, though people's opinions vary over time. Although we have submitted a type of category called uncertainty, the news is difficult to deal with. Specifically, the classifier faces difficulties finding important feature terms that supports people's opinions. However, we know that some artificially specified key words attach great importance to correctly classifying the text. Therefore, we may manually choose some feature terms to enhance the classifier. Furthermore, crawling many news comments from the Internet can assist and amend existing RECI models. This strategy requires long-term data collection and analysis procedures. We intend to explore more effective algorithms for performing sentimental text classification. Constructing realm-relevant sentimental feature terms may also help enhance classifier quality. Although we now have some complete Chinese sentimental lexicons, they may not fit real estate field. Therefore, we need to establish real estate sentimental lexicons.

On the one hand, the RECI is constructed based on the weekly news, and the most important meaning of the index is generating a time series, and effectively indicating the development tendency of REM. Now, we have more than 70 index records and further research on the records is one of our future work. On the other hand, we have attempted to label terms effectively. Future work is still needed to raise the quality of labeling and make it more professional and systematic. Labeling quality exerts great influence on classifier accuracy. Therefore, we look forward to generating our realm-specified label criterion and standard realm-specified open corpus.

Funding

This work is supported by Beijing Philosophy and Social Science Fund under grant number 15ZDA18 and China Postdoctoral Science Foundation under grant number 2014M560068. It is also supported by the Institute for Innovation and Development, Tsinghua University, and National Laboratory of information science and technology, Tsinghua University.

References

- Bailey, M. J., R. F. Muth, and H. O. Nourse. 1963. A Regression method for real estate price index construction. *Journal of the American Statistical Association* 58 (304):933–42. doi:10.1080/01621459.1963.10480679.
- Bram, J., and S. Ludvigson. 1998. Does consumer confidence forecast household Expenditure? A sentiment index horse race. *Federal Reserve Bank of New York Economic Policy Review* (6):59–78.
- Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2:121–67. doi:10.1023/A:1009715923555.
- Calhoun, C. A. 1996. OFHEO house price indexes: HPI technical description. *Office of Federal Housing Enterprise Oversight*, March: 1–15.
- Campbell, C., N. Cristianini, and A. Smola. 2000. Query learning with large margin classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*.
- Chakrabarty, S., M. Chopin, and A. Darrat. 1998. Predicting future buyer behavior with consumers' confidence and sentiment indexes. *Marketing Letters* 9 (4):349–60. doi:10.1023/A:1008089400754.
- Clapp, J. M. 1990. A methodology for constructing vacant land price indexes. *Real Estate Economics* 18 (3):274–293. doi:10.1111/1540-6229.00522.
- Clayton, J., D. Ling, and A. Naranjo. 2009. Commercial real estate valuation: Fundamentals versus investor sentiment. *The Journal of Real Estate Finance and Economics* 38:5–37. doi:10.1007/s11146-008-9130-6.
- Cotsomitis, J. A., and A. C. C. Kwan. 2006. Can consumer confidence forecast household spending? Evidence from the European Commission Business and Consumer Surveys. *Southern Economic Journal* 72 (3):597–610. doi:10.2307/2011835.
- Cristianini, N., and J. Shawe-Taylor. 2000. *An introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press.
- Curtin, R. T. 1982. Indicators of consumer behavior: The University of Michigan Surveys of Consumers. *Public Opinion Quarterly* 46 (3):340–352. doi:10.1086/268731.
- Dumais, S. T., J. Platt, D. Heckerman, and M. Sahami. 1998a. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*. ACM Press.
- Dumais, S. T., J. Platt, D. Heckerman, and M. Sahami. 1998b. Inductive learning algorithms and representations for text categorization. In *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management*, Bethesda, US, 148–55.
- Engelberg, J., and C. Parsons. 2011. The causal impact of media in financial markets. *The Journal of Finance* 66:67–97. doi:10.1111/j.1540-6261.2010.01626.x.
- Fleming, M. C., and J. G. Nellis. 2005. *The Halifax house price index — Technical details*. 1984. Halifax Building Society. HBOS. "Halifax House Price Index", 1–3. Halifax Group Press.
- Fu, Y. M. 2003. Estimating the lagging error in real estate price indices. *Real Estate Economics* 31 (1):75–98. doi:10.1111/reec.2003.31.issue-1.
- Gatzlaff, D. H., and D. C. Ling. 1994. Measuring changes in local house prices: An empirical investigation of alternative methodologies. *Journal of Urban Economics* 35 (2):221–44. doi:10.1006/juec.1994.1014.
- Gelper, S., A. Lemmens, and C. Croux. 2007. Consumer sentiment and consumer spending: Decomposing the granger causal relationship in the time domain. *Applied Economics* 39 (1):1–11. doi:10.1080/00036840500427791.
- Goetzmann, W. 1992. The accuracy of real estate indexes: Repeat sales estimators. *The Journal of Real Estate Finance and Economics* 5 (1):5–53. doi:10.1007/BF00153997.
- Guo, H., H. Li, Q. Shen, Y. Wang, and Y. Li. 2007. Real estate confidence index based on Web GIS and SPSS WebAPP. *Project Management* 25:171–77. doi:10.1016/j.ijproman.2006.09.015.
- Hui, E. C. M., K. H. Yu, and D. K. H. Ho. 2007. Office price index lagging in Singapore and Hong Kong. *International Journal of Strategic Property Management* 11 (1):33–46.
- Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398 in *Lecture Notes in Computer Science*, 137–42. Heidelberg, DE: Springer Verlag.
- Knight, J. R., J. Dombrow, and C. F. Sirmans. 1995. A varying parameters approach to constructing house price indexes. *Real Estate Economics* 23 (2):187–205. doi:10.1111/reec.1995.23.issue-2.
- Kwan, A. C. C., and J. A. Cotsomitis. 2006. The usefulness of consumer confidence in forecasting household spending in Canada: A national and regional analysis. *Economic Inquiry* 44 (1):185–97. doi:10.1093/ei/cbi064.
- Lang, S., and W. Schaefer. 2015. Examining the sentiment-return relationship in European Real Estate stock markets. *Journal of European Real Estate Research* 8:24–45. doi:10.1108/JERER-10-2014-0036.
- Mark, J. H., and M. A. Goldberg. 1984. Alternative housing price indexes: An evaluation. *Real Estate Economics* 12 (1):30–49. doi:10.1111/1540-6229.00309.
- Maurer, R., M. Pitzer, and S. Sebastian. 2004. Hedonic price indices for the Paris housing market. *Allgemeines Statistisches Archiv* 88 (3):303–26. doi:10.1007/s101820400173.
- OFHEO. 2004. OFHEO house price indexes. *Office of Federal Housing Enterprise Oversight*, December 1:1–75.
- Shiller, R. 2008. Historic turning points in real estate. *Eastern Economic Journal* 34:1–13. doi:10.1057/palgrave.eej.9050001.
- Tetlock, P. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance* 62:1139–68. doi:10.1111/j.1540-6261.2007.01232.x.
- Thibodeau, T. G. 1989. Housing price indexes from the 1974–1983 SIMSA annual housing surveys. *Real Estate Economics* 17 (1):100–17. doi:10.1111/reec.1989.17.issue-1.
- Vicente, R. R., C. M. D. Toledo, V. B. P. Leite, and N. Caticha. 2006. Underlying dynamics of typical fluctuations of an emerging market price index: The Heston model from minutes to months. *Physica A: Statistical Mechanics and its Applications* 361 (1):272–88. doi:10.1016/j.physa.2005.06.095.