A Study on Investor Public Opinion Fluctuation Inversion Monitoring Based on Network Information Sentiment Analysis

Lin Lin GUO, Li Wei XIA
School of Management, Northwestern Polytechnical University, Xi’an, Shanxi 710129, China
Email: kelin7777@126.com

Abstract. In this paper, not only was data mining performed for investors’ social networking platforms of micro-blog and Guba, but a method combining pattern matching and machine learning was adopted for sentiment classification and computing. On this basis, an investor network public opinion fluctuation monitoring could be established to monitor and analyze fluctuations and inversion. Furthermore, validity of this model was verified finally.

1. Introduction
Due to the rapid development of internet as well as characteristics such as openness and virtuality, etc. of internet, more and more people prefer to express their own opinions on the internet, which makes the network gradually become a primary site where public opinion topics are produced and propagated. In addition, public opinion forces formed by web reviews also turn into the force that cannot be ignored by both the government and social governors. At present, an increasing number of investors gather together on internet forums to discuss stock market information and make use of network information resources to make investment decisions; as a consequence, intensive fluctuations of asset prices incurred by network public opinions thus formed take place occasionally. Considering this, sentiment analysis is carried out in this paper by collecting investor network information data in China so as to set up a Chinese investor network public opinion fluctuation mode that is further utilized to monitor and predict the status of their public opinions. This is beneficial for supervisors to effectively comprehend network public opinions, make reasonable decisions and relieve risks in advance.

Analytical investigation on network public opinion is an interdisciplinary research focusing on multiple fields including computer networking technology, artificial intelligence technology and data mining, etc. Information mining and analytical study of network public opinions mainly consist of two aspects. One is how to effectively obtain network public opinions and valuable public opinion research objects; the other is to perform network public opinion information analysis, processing, classification, monitoring and early warning. In recent years, researchers have done much in orientation analysis field and studies on orientation analysis are carried out by Hatzivissiloglou and McKeown [1] at the earliest. They use connectives “and” and “but” as cue words to predict sentiment orientation of adjectives, which pioneers in the orientation analysis investigation. As social network rises, orientation under the circumstance of social networking begins to attract the attention from researchers. Bak et al. [2] comprehensively take advantage of theme, emotion, sentiment and lexical patterns, etc. to explore depth and relationship strength among social individuals in the social network, etc. Combined with indices such as the number of followers, the number of comments and orientation, etc., Hui and Gregory [3] quantify sentiment orientation and influence of a certain topic. As for Liu et al. [4], a global emotional...
thesaurus construction method is presented and then applied into opinion mining. Gindl et al. [5] utilizes Bayes classifier to classify unknown words by defining context of subjective words to achieve the purpose of defining orientation. Regarding applied researches on stock market network information sentiment analysis, Ghiassia and Skinnerb [6] adopts network information to analyze investor sentiment and further establish a model interacted with stock price fluctuations to predict stock prices. With the help of text mining tools of google, Bollen and Mao [7] probe into investor sentiment; besides, they also take advantage of Twitter to collect relevant information that is later used as investor sentiment of their study; and, Granger causality is subsequently employed to investigate the relation between sentiment and the rates of securities returns. Comparing with mature stock market abroad, that of China is featured with “being newly-emerging and transition”, which makes it have a stronger mental trait and a significant non-rational behavior as a whole. Therefore, such a study is of certain research and practical significances to monitor the state of Chinese investor network public opinions and take countermeasures in advance.

2. Sentiment Analysis on Stock Market Network Public Opinion Information

Finite At present, text sentiment classification study based on machine learning is a research hot spot of the existing text mining field. Fundamental procedures of this approach can be described as follows. Feature representation sets are extracted for sentiment texts in the first place; then, machine learning algorithm can be used to classify the text into the commendatory or the derogatory according to feature sets selected. Currently, common text classification algorithms based on machine learning consist of the Bayes classifier, the maximum entropy model and the support vector machine, etc..

A method of support vector machine was selected to carry out sentiment analysis on stock market network information in this study. In terms of the support vector machine, it is a machine learning method based on structural risk minimization principle. To be specific, input space can be mapped to the high-dimensional space by virtue of nonlinear transformation defined by the inner-product function to obtain optimal classification face in the latter space. Regarding support vector machine based sentiment classifications, a majority of them are deemed as binary classification problems and training sets of sentiment analysis corpus are divided into those forward and reverse.

Ideally, it is impossible for forward samples to overlap with reverse samples and only the maximum margin between forward and reverse samples needs to be found to construct the corresponding classifier. However, as far as sentiment analysis is concerned, there are many cases in which semantic orientation cannot be acquired definitely; as a consequence, a penalty factor denoted by P is introduced. A training set \( M = \{ (x_1, 1), (x_2, 1), \ldots, (x_n, -1) \} \) was defined, where, \( x_i \) referred to emotion features, and category markers 1 and -1 respectively meant that orientation of these features was forward or reverse. As a result, this question resolved itself into obtaining a linear decision function \( f(x) = wx + b \) that could distinguish such two categories of sentiment samples. Furthermore, their margin was maximized under the circumstance that a constraint condition of \( y_i (wx + b) \geq \pm 1 \) has been satisfied. Thus, the corresponding optimization problem turns into:

\[
\min \frac{1}{2} \|w\|^2 + P \sum_{i=1}^l \delta_i \\
\text{s. t. } y_i (wx_i + b) \geq 1 - \delta_i
\]

(1)

(2)

where, \( \frac{1}{2} \|w\|^2 \) is structural risk and \( \delta_i \) is a non-negative relaxation factor; in this case, the following Lagrangian dual problem can be solved and then we obtain another equation below:

\[
\max \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l a_i a_j y_i y_j x_i \cdot x_j
\]

(3)

\[
\text{s. t. } \sum_{i=1}^l a_i y_i = 0, \ 0 \leq a_i \leq C, \ i = 1, \ldots, l
\]

(4)

Where, \( a_i \) stands for Lagrange multiplier. This is a convex quadratic programming problem and its locally optimal solution is identical to its globally optimal solution. If \( a^* = (a_1^*, a_2^*, \ldots, a_n^*) \) is an optimal solution, then, \( w^* = \sum_{i=1}^l a_i^* y_i x_i \).
Considering polysemy of Chinese words, changeability of syntax and differences in linguistic rhetoric, etc., sentiment orientation computing need to take contextual information and rhetoric, etc. of lexical phrases into account; and, such contextual information and rhetoric and others are all factors that have influence on sentiment classification and polarity classification accuracy. Hence, emotion feature extraction was combined together with machine learning in this study to improve the existing machine learning approach; in other words, after word segment has been conducted basically for texts of stock market network information, keyword matching was performed to form keyword sequences to be matched; subsequently, value of sentiment orientation was figured out for each emotion feature; finally, sentiment orientation of these texts were achieved by a machine learning approach. Below, such an approach was utilized to design sentiment orientation computing of stock market network information and the relevant major procedures were presented as follows.

1. Classification. At the training stage, sentiment orientation of comment texts was labeled manually; and, forward and reverse stock market opinions expressed online were marked as 1 and -1 respectively; after word segmentation has been conducted for texts with sentiment orientation labeled, keywords were extracted to construct keyword database. As for the definition of keyword, it will be introduced in the following section. Furthermore, a keyword combination mode that can decide the sentiment classification to the most extent was adopted to constitute a pattern base. At the sorting phase, keywords were also extracted after word segmentation has been carried out for texts to be classified and these keywords were used to generate an emotional pattern according to the established mode. Besides, such an emotional pattern was referred to as a mode feature. Furthermore, pattern generated by keywords was matched with the pattern base. Sentiment orientation values of every pattern were calculated if matched successfully to obtain a feature sequence constituted by patterns. The corresponding eigenvalue represents commendatory or derogatory intensity of these patterns. In the end, machine learning methods such as SVM, etc. were utilized for classification to acquire orientation intensity of the related text.

2. Keyword extraction. Keywords of sentiment classification problems refer to vocabularies that play a decisive role in commendatory and derogatory derivation of statements in the text. After word segmentation, the part of speech was divided into seven categories as sentiment orientation analysis on texts was considered, including the noun, the verb, the adjective, the preposition, the conjunction, the adverb and so on. As far as habits of language use are concerned, roles played by all seven categories in syntactic analysis are rather fixed, which is beneficial for extraction and matching of patterns. Keywords primarily consist of affective evaluation words, privatives and peripheral word sequences of sentiment words. Dependent on impacts of each word on text orientation, these words can be classified into different categories. In this paper, keywords were divided as follows. Category a refers to affective evaluation words with commendatory or derogatory meanings, including some nouns, adjectives, adverbs and verbs. Category b incorporates adverbs of degree, such as quite, very, particularly especially and exceptionally, etc. Category c is the privative such as no and none, etc. Category b is composed of conjunctions signifying transition relations (however, but, nevertheless, etc.). Among such four categories of keywords, only those of category a have commendatory or derogatory senses, while other three categories are all vocabularies necessary for pattern formation.

3. Pattern Extraction and Matching. According to characteristics of investor web reviews in China, 8 part-of-speech collocation patterns were put forward in this paper. As shown in Tab. 3-1, n stands for the noun and adj for the adjective; v signifies that the part of speech of a word is verb; sow_{q_i} represents the part of speech of sentiment word q_i and sow_{q_i} \in \{n, adj, v\}. 


Tab 1 Phrase Collocation Patterns

| $f_i$ | $e_i$ | collocational patterns | comment collocation |
|-------|-------|------------------------|---------------------|
| n     | v     | $<n, sowe_q1, v>$      | Noun + Sentiment Word + Verb |
| n     | n     | $<n, sowe_q1, n>$      | Noun + Sentiment Word + Noun |
| n     | adj   | $<n, sowe_q1, adj>$    | Noun + Sentiment Word + Adjective |
| v     | n     | $<v, sowe_q1, n>$      | Verb + Sentiment Word + Noun |
| v     | adj   | $<v, sowe_q1, adj>$    | Verb + Sentiment Word + Adjective |
| adj   | v     | $<adj, sowe_q1, v>$    | Adjective + Sentiment Word + Verb |
| adj   | n     | $<adj, sowe_q1, n>$    | Adjective + Sentiment Word + Noun |
| adj   | adj   | $<adj, sowe_q1, adj>$  | Adjective + Sentiment Word + Adjective |

According to the above table, $f_i$ and $e_i$ are collocational words closest to $q_i$ within the range of one sentence. In addition, different sentences are segmented by punctuations. When no context up to the corresponding requirements could be matched, the affective units were denoted as $v_i = <null, q_i, e_i>$, $v_i = <f_i, q_i, null>$ and $v_i = <null, q_i, null>$. For specific steps, they are described as follows.

Step 1: Sentiment dictionary was utilized to match with the sentiment word $q_i$ in text d.

Step 2: Regarding the sentiment word $q_i$, its position in the relevant sentence is expressed in $\text{position}(q_i)$; in addition, a window with a length of 1 was defined as $\text{Win}_1$, original value of the position as $\text{position}(q_i) = \text{position}(\text{Win}_1)$. Then, in line with the part-of-speech collocation pattern, the window moved forward to a position of $(\text{Win}_1) - 1$, and it should stop moving when a word $f_i$ satisfying the requirements was matched or in the case of $\text{position}(\text{Win}_1) = 0$.

Step 3: Position of $\text{Win}_1$ was initialized to $\text{position}(q_i) = \text{position}(\text{Win}_1)$. Likewise, the window moved reverse to a position of $\text{position}(\text{Win}_2) + 1$; then, it should stop moving when a word $e_i$ satisfying the requirements was matched or in the case of $\text{position}(\text{Win}_2) = l + 1$, where $l$ refers to the length of text A.

Step 4: An affective unit of $v_i = <f_i, q_i, e_i>$ was constructed.

Step 5: Above procedures were repeated up until all sentiment words in text A were matched with.

Step 6: The affective units were adopted as feature items to update representation of the text into $A' = (q_1, q_2, \ldots, q_n)$, where $n$ refers to the number of sentiment words in text A.

By affective unit construction and pattern matching, keywords related to emotional characteristics of the text were extracted. Below, sentiment computing for the text will be conducted.

4. Sentiment Orientation Computing of Text. Sentiment orientation computing of text mainly covers three aspects of keyword orientation computing, pattern orientation computing and comment orientation computing. Among four categories of keywords defined above, only orientation of affective evaluation words (category a) with commendatory or derogatory implications needs to be computed. Firstly, a certain quantity of commendatory and derogatory terms that were most commonly used and had the most intensive sentiment orientation was selected from various fields according to experience as the paradigm word. In terms of words falling into the category a not included in the benchmark lexicon, their comment orientation could be achieved by figuring out the closeness of their semantic relation with two groups of paradigm words. The corresponding computing method is described as follows. For two terms of $k_1$ and $k_2$, if their similarity is denoted as $\text{San}(k_1, k_2)$ and the distance between them as $\text{Dis}(k_1, k_2)$, then,
\[
\text{Sam}(k_1, k_2) = \frac{\partial}{\text{Dis}(k_1, k_2) + \partial}
\]

(5)

Where, \( \partial \) is an adjustable parameter and the value of word distance \( \text{Dis}(k_1, k_2) \) depends on factors such as the depth and area density of a concept hierarchy tree, etc. is a constant. In terms of two nodes of the same path length, their semantic distance can be greater if they are at lower layer of the concept hierarchy. Similarly, for such two nodes at a high-density area of the concept hierarchy tree, their semantic distance should be longer than those located at a low-density area.

Based on a method that has been achieved to compute the similarity between vocabularies, the measure of semantic orientation of any word \( K \) could be obtained and denoted as \( kp \) or \( kn \). While the former represents the paradigm word group of positive emotion, the latter stands for that of negative emotion. Furthermore, the number of words in such two word groups was expressed in \( s \) and \( t \) separately. In this case, orientation of word \( k \) can be written into an equation below.

\[
\text{Propensity}(k) = \frac{1}{s} \sum_{s=1}^{\sigma} \text{Sam}(k, kp_s) - \frac{1}{t} \sum_{t=1}^{\tau} \text{Sam}(k, kn_t)
\]

(6)

If \( \text{Propensity}(k) \) is positive, it means that the corresponding word has a commendatory sense; otherwise, it signifies that the word is derogatory.

5. Pattern Orientation Computing. After successful pattern matching, emotion features constituted by patterns were obtained. Next, values of orientation corresponding to all features were computed. While adverbs of degree had a very significant influence on sentiment orientation, privatives could change sentiment orientation of commentary words. Considering this, sentiment orientation computing methods should be respectively defined for 8 patterns given in Tab. 1.

6. Commentary Text Orientation Computing. As orientation of words was figured out to obtain values of pattern feature orientation, each text could form a pattern sequence vector. Subsequently, orientation of the pattern sequence vector was classified by a learning machine approach to acquire the orientation of the relevant commentary text.

3. Experimental Results and Analysis

The according to those discussed above, data source was firstly defined by virtue of preferences of Chinese investors. It turned out that microblog and Guba were two major most interactive socialized media dominated by text information and featured with being crowded interactively, a great amount of information and quick updating, etc. Therefore, information data collected from microblog and Guba were principally adopted and word segmentation and key word extraction were also performed for comments made by investors on microblog and Guaba platforms, so as to construct a pattern sequence vector. In order to verify it, network data from August 19 to 20 of 2015 were selected as the test object.

3.1. Data Processing

First, word segmentation and keyword extraction were carried out for comments, which was followed by pattern matching; after feature word groups have been matched with successfully, the pattern sequence vector could be established. In this study, 324 words were manually selected from Hownet Sentiment Vocabulary as the sentiment dictionary of stock comments; among them, 146 were positive sentiment words, while the remaining 178 were negative. Positive words of the maximum sentiment orientation values included happy, delighted, nice, good, excited, bright, like, worthy and expectation, etc.; by contrast, negative words of the largest sentiment orientation values consisted of harsh, bad, garbage, detestable, dark, horror, sick, blackguardly, disappointment, troublesome, not well and tricked, etc. Furthermore, all these words would be used as forward and reverse paradigm words to calculate orientation values of other keywords and then achieve orientation values of the pattern. In Hownet, many phrases with “not” were directly deemed as reverse sentiment words, such as “not happy” and “not well”, etc. Besides, they served as evaluative words during pattern matching in a direct manner.

3.2. Experimental Results

In the process of classifier and feature selection, a method combining statistical learning based on pattern matching come up with in this paper utilized 75% of network data as a training set of this experiment,
while the remaining 25% was network information adopted as its test set. Additionally, all features were selected and the SVM classification approach was made use of. Regarding the SVM classifiers of Matlab7.0 and SVM toolbox, they were both modified before the experiment during which Sigmoid function was utilized as the kernel function, where $a=1$ and $b=1$. Performance evaluation capacity indices of FA, FR, RA and RR employed respectively referred to forward accuracy, forward recall, reverse accuracy and reverse recall. As for experimental results, they are presented in Tab. 2.

| SVM Classifier | FA   | FR  | RA   | RR   |
|----------------|------|-----|------|------|
|                | 83.29% | 81.15% | 82.78% | 79.56% |

The above table signifies that the accuracy up to 83.29% is rather high. As a result, pattern matching was combined together with machine learning in this paper to calculate orientation of key words and patterns by virtue of pattern matching. In addition, relations among features of network data and orientation of words were taken into account as well before an approach of machine learning was adopted. In this way, analysis capabilities of such a combination method could be improved naturally.

After network information data sentiment scoring, a comprehensive index is set by summing up various sentiment values in a certain period of time. To be specific, such an index of network sentiment was denoted by $Y_t$, where $t$ refers to the period of time; that is, $Y_t = \sum x_t^a$, where $a \in$ the quantity of network sentiment information of \{buy, hold, sell\}. Hence, investor public opinion index within such a period of time could be expressed in $ISI_t = \ln\frac{1+Y_t^{buy}}{1+Y_t^{sell}}$.

3.3. Investor Network Public Opinion Monitoring

Considering that the purpose of this paper is to build a Chinese investor network public opinion fluctuation monitoring model, such fluctuations were firstly defined as $\Delta ISI_t = ISI_t - ISI_{t-1}$ after the network public opinion index has been established and the corresponding fluctuating values were standardized. By using 0.5 as the unit scale, investor network public opinions were divided into four states such as extremely pessimistic, extremely optimistic, generally optimistic and generally pessimistic. On this basis, the monitoring model for Chinese investor network public opinions was expressed into,

$\mid \Delta ISI_t \mid \in [0, 0.5]$, where fluctuation of one unit is referred to as a general fluctuation state out of the range of early warning

$\mid \Delta ISI_t \mid \in [0.5, 1]$, where fluctuation of two units is referred to as a drastic fluctuation state entering the range of early warning

$\mid \Delta ISI_t \mid \in [1, 2]$, where fluctuation of two units and above is referred to as a reverse state entering the range of high alert

In consistency with this model, $\Delta ISI_t$ on August 19 and 20 of 2015 were calculated to be 0.675 and 1.221, which meant that mood fluctuations were rather large and there existed a potential tendency of reversal so that it enters a state on high alert. Since then, intensive fluctuations occurred in the Chinese stock market and the fluctuation range of indices went beyond 25% within 5 days. Consequently, severe fluctuations of public opinions imposed a profound influence on the society and it was declared by the government of China that the Chinese market entered a stationary fluctuation phase. Thus, the Chinese investor network public opinion monitoring model constructed in this study was verified to be valid to a certain degree.

4. Conclusion

Network information mining and keyword extraction were carried out specific to two network platforms (microblog and Guba) crowded with the largest number of Chinese investors. On this basis, pattern matching was further conducted to extract affective evaluation words of Chinese investors to form the
corresponding emotion feature sequence in the case of successful matching; then, sentiment orientation values of all features were figured out. Finally, investor sentiment orientation of the text was acquired by a machine learning method. Based on this, a Chinese investor network public opinion fluctuation monitoring model was established to define and supervise fluctuations and reversal of network public opinions expressed by Chinese investors. This is not only beneficial to comprehend diverse states of Chinese investor public opinions in the process of their changes and reasonably predict developing trends of network public opinions, but also conducive for the government to know about stability of investors and further take proactive measures to avoid the occurrence of unexpected group events that may threaten financial system safety.

References
[1] V Hatzivassiloglou, KR Mckeown. Predicting the semantic orientation of adjectives[C]. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. 1997:174-181.
[2] JY Bak, CY Lin, A Oh. Self-disclosure topic model for classifying and analyzing Twitter conversations[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:1986-1996
[3] P Hui, M Gregory. Quantifying sentiment and influence in blogspaces[C]. Proceedings of the First Workshop on Social Media Analytics. ACM, 2010: 53-61.
[4] X Ding, B Liu, P S Yu. A holistic lexicon-based approach to opinion mining[C]. Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008: 231-240.
[5] S Gindl, A Weichselbraun, A Scharl. Cross-Domain Contextualisation of Sentiment Lexicons[C]. Proceedings of the 2010 conference ECAI, 2010:771-776.
[6] M Ghiassia, J Skinnerb, D Zimbra. Twitter brand sentiment analysis: A hybrid system using n -gram analysis and dynamic artificial neural network[J]. Expert Systems with Applications, 2013, 40 (16) :6266-6282.
[7] J Bollen, H Mao, A Pepe. Determining the public mood state by analysis of microblogging posts[J]. Alife, 2010 :56-65