Abstract

Nowadays, with the fast development of social media, the instant financial news can be quickly spread over the Internet and consequently results in the strong vibration of stock market within a short period of time. To capture the effect of financial news, various classification algorithms are proposed in financial data mining domain. Unfortunately, it has been shown that the positive news might not necessarily drive up the stock price. It is believed that there must exist some underlying events, called anchor events in this paper, which drive the stock market up or down. Therefore, this paper proposed the classification should be performed not only based on the textual features of financial news, but also the semantic distance between the testing set of news and the set of news of anchor events. Accordingly, such semantic distance is measured as deviated expectation calculated based on the distance between two set of extracted topics. The proposed method involves two steps, i.e., calculating whether the testing news is consistent with anchor events or not, and calculating how far the testing news is from the anchor events. We then evaluate our method as well as the state-of-the-art classification algorithms on some real data sets. The promising experimental results have demonstrated that the proposed method is superior to the state-of-the-art classification algorithms in terms of classification accuracy.

Keywords: Stock Price Prediction; Classification; Financial Data Mining, Big Data

1. INTRODUCTION

Nowadays, with the fast development of social media such as Weibo and Twitter, the effect of financial news could be greatly amplified and consequently results in the strong vibration in stock market. It is therefore important to predict whether these breaking news are positive or not. However, aforementioned task is a non-trivial one. Based on the analysis performed on the historical data, there are several difficulties which invalidate most of existing text classifiers. First, breaking news classified as positive sometimes still co-occur with stock price decline, and vice versa. This contradicts with our intuitive assumption that positive news should make stock price go up. Second, most of existing classifiers simply assume that each piece of news is already associated with one or several stocks. However, such kinds of associations are seldom provided in both the
training and testing data set. These difficulties impose a great challenge to most of existing stock mining algorithms.


Moreover, SVM based approaches are naturally proposed due to the excellent ability of SVM in classifying documents. To train a good classifier, a well labeled training data set is a must. To avoid the great effort in labeling financial news, Kaya and Karsligil (2010) have proposed to automatically label financial news based on the volatility of stocks, then using noun-verb pairs as the features of documents to train SVM classifier. At last, the trained SVM is adopted to predict the effect of future news (p.478).

There are also several related approaches which predict the stock risk instead of the direction of stock price. Naïve Bayesian can be combined with KNN algorithm, as proposed by Sanwaliya, Shanker, and Misra (2012), to classify the risk level of each piece of news (p.921); however, their risk level is roughly defined. Pan, Cheng, and Wu (2010) argued that risk is brought by some critical terms instead of news itself (p.271). Accordingly, they have decomposed the news document into a set of keywords, and have filtered out some low-frequency terms as they are assumed not to contribute to the risk prediction. Then, the effect of each term on the set of stocks is modeled using random walk model. The stable status of the model is then used for predicting the risk of testing news. However, the major drawback of this work lies in that, the stock price volatility cannot be simply combined with the calculated term risk. Recently, deep learning is also adapted to this challenging issue (Ding, Zhang, & Liu, 2015, p.2327). In their work, the relationship among entities, corresponding to a specific stock, and their actions on some other stocks are modeled using tensor model. By training the constructed neural network, the short-term network, the medium-term network as well as the long-term network are mixed together to predict the changing direction of stock price. However, aforementioned approaches assume that there exists a strictly positive interrelation between news and stock price. Unfortunately, it has been observed that this is not the case. For instance, the released news on June 29, 2015 are considered as an important good news to the stock market of China, but the index of stock market still falls down. Similar observations are plotted in Figure 1, i.e., both good news and bad news can induce the rise of stock price, and versa vice. This newly observed relationship contradicts the foundations of most of classification based stock price prediction approaches. This reveals that there must exist some hidden factors which play a more important role than classification.

![Figure 1. The newly observed relationship between stock price and news](image-url)
The utility of a bad news could be positive, and then its effect on stock price is still positive. The utility is the hidden factor which can be modeled as the deviation of users’ expectation. The reason is straightforward. To imagine that for a bear-market, there exist some “black swan” events which trigger the poor performance of the stock market. If a piece of news, no matter how good it is, is not related to these “black-swan” events, then it will not play a positive effect on the stock price. In this paper, this is called expectation deviation by us. These “black-swan” events are treated as anchor events. Only the future good news is related to these anchor events, then they can be classified as positive, and negative otherwise. Therefore, in the proposed approach, we first determine whether the breaking news is related to the anchor events or not. To judge the breaking news, the latent dirichlet allocation (LDA) will be performed on both the news set of anchor events and the testing news set (Hoffman, Blei, & Bach, 2010, p.856, Qing Li, Yuanzhu Chen, Li Ling Jiang, Ping Li & Hsinchun Chen, 2016, p.1, H. Tahersima, M. Tahersima, M. Fesharaki & N. Hamed, 2011, p.123, Mahajan A., L. Dey & S. M. Haque, 2008, p.423, Mittermayer M.-A, 2004). If the topic extracted can match with each other, then we consider the testing set of news is related to anchor events. Then, a binary classification task will be performed only on the related news set. At last, the positive related news set is considered to be able to drive up stock price.

The remaining of our paper is organized as follows. Section 2 formulates the problem and the proposed approach is illustrated in Section 3. Experimental evaluation is performed in Section 4 and we conclude the paper in Section 5.

2. PROBLEM FORMULATION

Let $ORI\_NEWS$ denote the set of news containing anchor events, $T$ denote a piece of news to be classified. To classify whether a piece of news is positive or not, we need to build a bag of words illustrating the semantic tendency of the news. Conventionally, semantic analysis is performed using certain natural language processing techniques. In this paper, we build such bag of words using terms with explicit preference such as “crisis”, “increase” and “decrease” to avoid the computation complexity. For the semantic analysis of news, users can adopt any technique to further improve the model performance. $C$ denote the class label of the news, it takes value 1 as positive and 0 as negative. If $C=1$, then the corresponding stock price is assume to go up. Note that $C$ is classified based on the proposed utility. And utility is defined in this paper as the deviated expectation of the news. To measure expectation, the topic of the news about anchor events is assumed as expectation, i.e., anchor topic. The topic extracted for testing set of news is considered as general news topic. The difference between news topics and anchor topic is considered as deviated expectation. If the overlap between news topic and anchor topic is greater than a predefined threshold, then we consider it consistent with expectation. Then, this testing set of news is to be classified. If $C=1$, then stock price rises.

3. PROPOSED DEVIATED EXPECTATION BASED NEWS CLASSIFICATION METHOD

3.1 FILTERING NEWS SET AND PREPARATION

As aforementioned, we assume that the association relationship between news and stock is unknown. Therefore, a set of news for each day will be extracted instead of one specific piece of breaking news. First, we manually identify several events which can greatly affect the stock market. Then, this set of news could be quite huge, e.g., containing several hundreds of news per day. However, only a small part of such news is important news. If directly extract topics out of this set of news, the important topics might be overwhelmed by a lot of trivial topics. To avoid this, we first adopt K-means algorithm to perform clustering on this set of news. As the breaking news is generally more important than the rest news, it will appear several times within a short period time. And the rest news is trivial and its semantic meanings are more diverse. By clustering this set of news, important news having repeated content and thus is more likely to be grouped within one or several clusters. However, the rest trivial news is less likely to be grouped together. By empirically setting a small M, only the first M clusters out of K clusters will be kept for the following experiments and the rest clusters will be removed. The K-means clustering algorithm will not be illustrated in this paper.
3.2 Deviated Expectation Calculation

**Extracting News Topics**

We assume the topic of news as the expectation. To calculate the topic, it is natural to choose LDA algorithm, written as

\[
p(w_t, \theta_t, z_t, \varphi | \alpha, \beta) = p(\theta_t | \alpha) \prod_{k=1}^{K} p(\varphi_k | \beta) \prod_{n=1}^{N_t} p(z_{tn} | \theta_t) p(w_{tn} | z_{tn}, \varphi_k) \tag{1}
\]

where \( T \) denote the set of news, \( K \) denote the set of topic, \( \alpha \) denote parameter of the topic distribution over \( T \), \( \theta_t \) denote topic distribution in news \( t \), \( z_{tn} \) denote the topic of the \( n \)-th term in news \( t \), \( w_{tn} \) denote the \( n \)-th term in news, \( \beta \) denote the parameter of term distribution in the \( k \)-th topic, \( N_t \) denote the total number of terms in news \( t \). We have \( \theta_k \sim \text{dirichlet}(\alpha) \), and \( \varphi_k \sim \text{dirichlet}(\beta) \). For any topic, we have \( z_{tn} \sim p(z_{tn} | \theta_t) \), and for any term, we have \( w_{tn} \sim p(w_{tn} | z_{tn}, \varphi_k) \).

To resolve the LDA formulation, we adopted Gipps sampling method.

**Calculating Deviated Expectation**

If topics extracted from anchor events are consistent with topics extracted from the testing set of news, then they are considered as consistent expectation. To determine whether the two topics are consistent or not, we proposed to calculate their similarity based on cosine similarity metric, written as

\[
sim(T, ORI\_NEWS) = \frac{T \cdot ORI\_NEWS}{\| T \| \cdot \| ORI\_NEWS \|} \tag{2}
\]

where \( ORI\_NEWS \) is the filtered set of news about anchor events, \( T \) is the calculated topics of the testing set of news, \( \| \cdot \| \) denote the number of that term. We then define a threshold, if the similarity is greater than this threshold, then the two topics are considered as consistent, the corresponding formulation is written as

\[
l(T) = \begin{cases} 
1, & \text{sim}(T, ORI\_NEWS) > \text{threshold} \\
-1, & \text{otherwise}
\end{cases} \tag{3}
\]

where \( l(T) \) is the indicator function, if \( l(T) = 1 \), then the topics of testing set of news is consistent with the expectation extracted from anchor events.

3.3 Deviated Expectation Based Classification

After acquiring the consistent topics of testing set of news, we need to measure the distance of each piece of testing news from the news set of anchor events, and this distance is called deviated expectation. It is impossible to directly calculate such deviation. However, the consistent news is thought to be related to anchor events. We can sum up the tendency extent of each term in the news, and calculate the difference between their summed tendencies as the deviated expectation, given as

\[
F(T) = f(T) - f(ORI\_NEWS) \tag{4}
\]

where \( f(ORI\_NEWS) \) is the degree of positive or negative tendency of the news set of anchor events, \( f(T) \) is the degree of positive or negative tendency of the testing news set. \( f(T) \) can be calculated as

\[
f(T) = \frac{\sum_{i=1}^{n} w_i d_+ - \sum_{i=1}^{n} w_i d_-}{|d_+| + |d_-|} \tag{5}
\]

where \( |d_+| \) denote the number of positive terms in news set \( T \), \( |d_-| \) denote the number of negative terms in \( T \), \( d_+ \) and \( d_- \) denote the \( i \)-th positive term and its weight, \( d_- \) and \( d_- \) denote the \( j \)-th negative term and its weight, respectively. Similarly, for the news set of anchor events, we can calculate its tendency as

\[
f(ORI\_NEWS) = \frac{1}{|ORI\_NEWS|} \sum_{k=1}^{N_{ORI\_NEWS}} \left[ \sum_{i=1}^{n} w_i d_+ - \sum_{i=1}^{n} w_i d_- \right] \tag{6}
\]

The notation is the same as Equation 3.

With the deviated expectation calculated according to Equation 4-6, the classification of the testing news is proposed as

\[
C(EVENT\_NEWS) = \begin{cases} 
+, & \sum_{i=1}^{N} l(EVENT\_NEWS_i) \cdot F(EVENT\_NEWS_i) > 0 \\
-, & \text{otherwise}
\end{cases} \tag{7}
\]

where \( C \) is the class label, \( F() \) is calculated using Equation 4, and \( n \) is the number of consistent news with anchor events in testing news set, which is defined as

\[
n = \sum_{i=1}^{N} l(EVENT\_NEWS_i) \tag{8}
\]

where \( l() \) is calculated using Equation 3 measuring whether the testing news is consistent with anchor events or not. \( F() \) is the deviation which can be
negative or positive. Therefore only if both \(f()\) and \(F()\) are negative or positive, this testing news will drive up the price of the corresponding stocks.

4. Experiments

To evaluate the performance of the proposed approach, we have carefully performed rigorous experiments on extracted news data sets and stock price set of Chinese stock market. The state-of-the-art text classification algorithms are also implemented by us on these data sets for performance comparison, which are KNN, native Bayesian and SVM.

To prepare the data sets, we have crawled news from Sina7x24, Sina Financial and Phoenix Financial, from March 2014 to August 2016. In this collected data set, we have over 200,000 news data. The statistics of this data set is reported in Table 1. The corresponding stock price is also collected for the same period of time. The collected set of news for each day is treated as the daily set of news. This daily set of news may contain news of abnormal events. These abnormal events are considered as anchor events. To automatically find these anchor events, we utilize the co-current vibration of stock price and news. Therefore, we set a threshold, e.g., 5%. If the market index is greater than this threshold, then the daily set of previous day is considered to trigger this abnormal price vibration. Note that previous day stands for the period after 15:00 of previous day till 9:30 of current day. News collected in this period is considered to be able to affect the open price of current day. For this news set, we filtered out 9 abnormal sub set of anchor events. However, for the extraction of testing set of news, we screen out abnormal events after each anchor event using the same method. Accordingly, we have 29 sub testing sets of news. The number of filtered anchor events and testing set of news is reported in Table 2.

<table>
<thead>
<tr>
<th>Anchor news</th>
<th>Number of testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnchorNews 1</td>
<td>3</td>
</tr>
<tr>
<td>AnchorNews 2</td>
<td>3</td>
</tr>
<tr>
<td>AnchorNews 3</td>
<td>3</td>
</tr>
<tr>
<td>AnchorNews 4</td>
<td>4</td>
</tr>
<tr>
<td>AnchorNews 5</td>
<td>2</td>
</tr>
<tr>
<td>AnchorNews 6</td>
<td>5</td>
</tr>
<tr>
<td>AnchorNews 7</td>
<td>3</td>
</tr>
<tr>
<td>AnchorNews 8</td>
<td>3</td>
</tr>
<tr>
<td>AnchorNews 9</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2. Extracted sub set of anchor and testing news

criterion adopted in the experiments is the accuracy as we already filtered out both anchor events and testing news set, and thus the general recall will not be used for evaluation.

4.1 Effect of Number of Topics

We first evaluate how the number of topics of LDA affects the expectation extracted for news. In this experiment, we set the number of topics as 3, 6 and 8, respectively. Then, the topics extracted by LDA algorithm is reported in Figure 2, Figure 3 and Figure 4, respectively. The x-axis in the figures represents different topics, while y-axis represents the topic distribution in the news. From figure 2, it is well noticed that all topics are contained in each piece of news, which indicates that the 3-topic clustering cannot discriminate the importance of different news. For the 6-topic clustering results, as shown in Figure 3, although each news contains one different topic, it still contains the rest topics. Therefore, both 3 and 6 topics clustering results are not satisfying. When the number of topics is set to 8, we can observe from the Figure 4 that, each piece of news only contains few different topics and each contained topic is mainly distributed in that news rather than the rest news. Moreover, topics contained in different news have less overlapping part. This indicates that when the number of topics equals to 8, the corresponding clustering results can well separate different news which facilitates to filter out the trivial news. In the rest of experiments, the number of topics is fixed to 8.

<table>
<thead>
<tr>
<th>Web site</th>
<th>Sina7x24</th>
<th>Phoenix Financial</th>
<th>Sina Financial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of news</td>
<td>181815</td>
<td>9827</td>
<td>96078</td>
</tr>
</tbody>
</table>

Table 1. Number of news collected from different web sites

For stock price prediction, if the close price of current day is higher than that of previous day, then we consider it as a positive one. The evaluation
Figure 2 Clustering results with 3 topics

Figure 3 Clustering results with 6 topics
4.2 EVALUATION ON THE PROPOSED METHOD

For the testing set of news, although we have filtered part of trivial news as mentioned in Section 4.1, the testing set of news still contains some inconsistent news with the anchor events. To determine whether a piece of news can match with the corresponding anchor event or not, we calculate the consistency of each piece of news with the supposed anchors using Formulation 3. The corresponding results are listed in Table 3. From this table, it can be seen that only a small fraction of the testing set of news, as listed in the second column, can match with their corresponding anchor events. Only the matched news will be assigned with a positive sign which will be used for the prediction of stock price.

We then calculate the deviated expectation for each piece of matched news using Formulation 4-6.

The corresponding results are reported in Table 4. In this table, the second column is the percentage of news whose deviated expectation is positive, whereas the last column is the percentage of news whose deviated expectation is negative. It is obvious that not every news is positively related to the anchor event.

<table>
<thead>
<tr>
<th>Testing Set of News</th>
<th>Percentage of matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing set 1</td>
<td>33.3%</td>
</tr>
<tr>
<td>Testing set 2</td>
<td>54.5%</td>
</tr>
<tr>
<td>Testing set 3</td>
<td>26.6%</td>
</tr>
<tr>
<td>Testing set 4</td>
<td>22.2%</td>
</tr>
</tbody>
</table>

*Table 3. The consistency calculation results*

<table>
<thead>
<tr>
<th>Testing news set</th>
<th>positive deviated expectation</th>
<th>negative deviated expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing set 1</td>
<td>16.6%</td>
<td>83.3%</td>
</tr>
<tr>
<td>Testing set 2</td>
<td>36.3%</td>
<td>63.6%</td>
</tr>
<tr>
<td>Testing set 3</td>
<td>53.3%</td>
<td>46.6%</td>
</tr>
<tr>
<td>Testing set 4</td>
<td>50.0%</td>
<td>50.0%</td>
</tr>
</tbody>
</table>

*Table 4. Deviated expectation between anchor and testing news*
<table>
<thead>
<tr>
<th>Testing news set</th>
<th>Deviated Expectation</th>
<th>Stock Price Direction</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing set 1</td>
<td>negative</td>
<td>Down</td>
<td>Yes</td>
</tr>
<tr>
<td>Testing set 2</td>
<td>negative</td>
<td>Down</td>
<td>Yes</td>
</tr>
<tr>
<td>Testing set 3</td>
<td>negative</td>
<td>Down</td>
<td>Yes</td>
</tr>
<tr>
<td>Testing set 4</td>
<td>positive</td>
<td>Up</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 5. Classification results of the proposed method

At last, we perform the proposed classification method and the results are given in Table 5. From Table 5, it can be seen that our proposed method work well on these four testing sets. Then, we implemented the state-of-the-art classification methods directly on these data sets and compared with ours. The comparison results are reported in Table 6 and Figure 5, respectively. Table 6 reports the classification accuracy of various classification algorithms as well as ours on the testing sets of news corresponding to each set of anchor news. It is well observed that the proposed method is superior to the state-of-the-art classifiers like SVM, KNN and NB. The average accuracy of our method is much better than that of these three classification algorithms. The reason is that the proposed method is performed by not only the content of news but also how close the news is from the abnormal events. However, the conventional classifiers rely more on the content of the news. It is also noticed that SVM is superior to both KNN and NB which is consistent with the common knowledge that SVM is one of the most powerful classifiers. Figure 5 gives a more intuitive way to evaluate all these classification algorithms. The purple bar is the classification results of our method and it is obvious that the accuracy of our method is much higher than that of other classifiers.

From the experiments, we can conclude that classification should be performed based on not only the features of text, but also the utility of text. The utility can be defined case by case. In our situation, the utility is defined as how far the testing news is semantically from the anchor events, i.e., deviated expectation. From the promising results, we have validated the effectiveness of the proposed method when compared with the state-of-the-art classification algorithms.

5. CONCLUSION

With the fast development of social media, breaking news is now playing a critical role in affecting stock market. Due to the poor performance of conventional feature based classification algorithms, we proposed to classify not only based on the content of news, but also the semantic distance between testing news and anchor events. We then model such semantic distance as deviated expectation. The deviated expectation is first represented by topics of the news and the corresponding formula is proposed to calculate the similarity between two extracted topic sets. We evaluate the proposed method as well as the state-of-the-art classification algorithms on some real data sets. The promising results demonstrate that the proposed method outperforms the state-of-the-art classification algorithms in terms of accuracy.

6. ACKNOWLEDGMENT

This work is supported in part by National Nature Science Foundation of China under Grant No. 61372013, National Key Technology Support Program of China under Grant No. 2014BAI05B06, and Shenzhen Science and Technology Program under Grant No.JCYJ20160330163900579 and No.JSGG20150512145714247.
<table>
<thead>
<tr>
<th>Anchor News Set</th>
<th>SVM</th>
<th>KNN</th>
<th>NB</th>
<th>Our Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.356</td>
<td>0.333</td>
<td>0.333</td>
<td>0.683</td>
</tr>
<tr>
<td>2</td>
<td>0.467</td>
<td>0.333</td>
<td>0.333</td>
<td>0.633</td>
</tr>
<tr>
<td>3</td>
<td>0.317</td>
<td>0.333</td>
<td>0.333</td>
<td>0.750</td>
</tr>
<tr>
<td>4</td>
<td>0.425</td>
<td>0.250</td>
<td>0.250</td>
<td>0.625</td>
</tr>
<tr>
<td>5</td>
<td>0.500</td>
<td>0.483</td>
<td>0.500</td>
<td>0.825</td>
</tr>
<tr>
<td>6</td>
<td>0.400</td>
<td>0.200</td>
<td>0.507</td>
<td>0.740</td>
</tr>
<tr>
<td>7</td>
<td>0.344</td>
<td>0.333</td>
<td>0.333</td>
<td>0.800</td>
</tr>
<tr>
<td>8</td>
<td>0.478</td>
<td>0.333</td>
<td>0.333</td>
<td>0.630</td>
</tr>
<tr>
<td>9</td>
<td>0.478</td>
<td>0.333</td>
<td>0.344</td>
<td>0.767</td>
</tr>
<tr>
<td>Average accuracy</td>
<td>0.418</td>
<td>0.325</td>
<td>0.363</td>
<td>0.727</td>
</tr>
</tbody>
</table>

Table 6. Comparison results of various classification methods

Figure 5. The plotted results of classification
7. REFERENCES


Authors

Sishi Ruan is currently a master student at Harbin Institute of Technology Shenzhen Graduate School. Her research interests include financial data mining.

Jennifer Y. M. Lai is an assistant professor at University of Macau. She received her PhD in management from The University of Hong Kong. She is in the field of business and management. She has published over ten top journal papers.

CHEN Xiaoyun is an associate professor at University of Macau. She received her PhD in marketing from The University of Hong Kong. She is in the field of business and management. She has published over ten top journal papers.

Xiaofeng Zhang is currently an associate professor at Harbin Institute of Technology Shenzhen Graduate School. His research interests include data mining, machine learning and big data analytics. He has published more than 30 referred papers.