Social Media Based Stock Prediction

Zhi-Hua JIANG\textsuperscript{1,a}, Bo-Yang LIN\textsuperscript{1,b}, Dong-Ning RAO\textsuperscript{2,c,*}

\textsuperscript{1}School of Information Science and Technology, Jinan University, Guangzhou, China
\textsuperscript{2}School of Computer, Guangdong University of Technology, Guangzhou, China
\textsuperscript{a}tjiangzhh@jnu.edu.cn, \textsuperscript{b}2018735@qq.com, \textsuperscript{c}raodn@gdut.edu.cn
\textsuperscript{*}Corresponding author

Keywords: Stock prediction, Nature language processing, Regression.

Abstract. The stock market is one of the most important part of the financial market. However, it is difficult for investors to extract information from data. In the meantime, sentiment analysis is a hot topic in computer science. In this paper, we build an emotion dictionary for finance area. We can calculate the emotional values of finance texts based on the dictionary. Further, with a regression analysis by Fama-French model, we discovered the relationship between the emotional value and stock price by experiments. Simply put, the emotional value and stock’s return rate did not have a significant correlation; the emotional value and stock’s volatility don’t have a significant correlation; but the emotional value and stock’ relative return rates had a positive correlation.

Introduction

We hope the stock price to be predictable. There is a hypothesis, which is called the Efficient Markets Hypothesis (EMH). EMH was proposed in 1970 [1]. EMH states that asset prices fully reflect all available information, it is impossible to "beat the market". However, EMH assumes that a stock price is always driven by unemotional investors [2]. This assumption is not true, at least in some degree. Under most circumstances, the investors do have emotion and may not have all available information. After that, considerable studies suggest that some markets are not fully efficient [3]. These previous studies predict the stock market and beat random selection. Recent years, researchers that aimed at the Chinese stock market [4] appeared.

Therefore, people presented many methods to predict stock prices. Traditionally, people used fundamental analysis and technical analysis. The fundamental analysis [5] is the analysis of a business's financial statements and its competitors and markets. It only focuses on the overall state of the economy, and considers factors. There are many factors, including interest rates, production, earnings, employment, GDP, housing, manufacturing and management. The fundamental analysis is performed on historical and present data. The technical analysis [6] is a methodology for forecasting the direction of prices through the study of past market data, primarily price and volume. There are many tools for technical analysis. However, the efficacy of these tools is disputed by the EMH.

In the meantime, nature language processing (NLP) attracted more and more researchers in computer science. Many techniques are proposed for NLP. Some of these papers focused on stock markets. However, the existing studies are mostly focused on social media [7].

Further, as far as we know, predicting the Chinese stock price based on the posts on the forums has not been studied. The posts on forums are the news released by individual persons. Under some circumstances, it is not reliable, even wrong. In other words, the analysis has bias. However, it can reflect the public opinion.

Therefore, in this paper, we try to predict stock price based on the public opinion through NLP. Specifically, we focus on the posts on the forums. We first collect the posts on forums. Then we process them by NLP techniques. More specifically, we calculate the emotional value of the text. For this purpose, we build a dictionary for the financial domain. At last, we use regression techniques to
find out the relationship between public opinion and the stock price. We found that the emotional value and stock’s return rate don’t have a significant correlation; the emotional value and stock’s volatility don’t have a significant correlation; but the emotional value and stock’ relative return rates had a positive correlation.

The following of this paper is organized as follows. In the next section, we introduce the background of the framework. Then, in the third section, we formalize the problem and present our method. The experiments are presented in the fourth section. At last, this paper is summarized with the conclusion and future work.

Background

We introduce the background knowledge of this paper briefly in this section. The background knowledge includes the financial intelligence and the NLP.

Half century ago, finance studies focused on institutional and legal aspects. The situation was changed during the 1950s. Today, there are new eras of finance, and new tools have been developed. For example, advanced intelligent systems use neural networks, support vector machines, evolutionary computation, fuzzy systems, and so on. Researchers proposed econometric methods for the time-series forecasting problem [8], optimization techniques for the asset allocation problem [9], and multiple criteria decision making methods for corporate performance evaluation and banking problem [10].

Among financial intelligent applications, some of them focus on the qualitative prediction of the stock market. Although most of these studies are using traditional macro-economic data, some of them begin to use NLP technologies. Here are some examples. The first kind of these studies used sentiment analysis [11]. There are two kinds of methods for this purpose, generally. One is the lexicon-based approach and machine learning approach. It determines the polarity and the sentiment strength based on the semantic orientation of the words occurring in that text. This method heavily dependent on a sentiment or subjective lexicon. The other kind of methods uses classification algorithms. A classifier is trained on a labeled data set, but it still needs an initial lexicon. The second kind of methods combines quantitative and qualitative features to predict stock prices. People believe that the social media might have a stronger relationship with firm stock market performances than the conventional one. This assumption has been tested by some computer scientist [7]. The third kinds of these methods do text analysis. Because the classifier may misclassify a common word, these studies often have a list of negative words, which will correct it by analyzing the emotion of the financial text [12].

The whole NLP process includes preprocessing and parsing. The preprocessing includes tokenization, stemming, part-of-speech (POS) tagging and lemmatization. There are many popular tools, e.g., the Stanford Name Entity Recognizer (NER) [13]. However, in financial intelligent applications, it is far from enough. To understand the precise meaning, we also need to distinguish conjunctions, adjectives, verbs and adverbs, and so on [14]. Therefore, dictionaries are necessary to help the identification of these words.

Methodology

Problem Specification

In this paper, we try to find out the relationship between the trend of specific stock price and the public opinion about it. To this end, we collect posts on forums to calculate emotion values. Then, we take these values as the quantified public opinion. To find out the relationships, we use regression techniques. More specifically, we are facing a regression problem. In this problem, the inputs are vectors transformed by emotion values, and the outputs are the stock price and trend. The stock price and trends include return rate, relative return rate and volatility.
Overview of the Whole Process

The whole process has 6 components. First, using a spider to get posts on forums, we can get the text we need. Second, these texts need to be transformed into the format we want. Third, a specific dictionary is needed for calculating emotion value. Fourth, with the text from the second step and the dictionary from the third step, we can calculate the emotion value. Fifth, we need a regression model. Sixth, using the emotion value from the fourth step, and the model from the fifth step, we do the regression.

Acquiring Text and Preprocessing

Spider is a widely used technique for web text acquiring, we use it to acquire text in this paper. The preprocessing has three major tasks. First, word segmentation and part of speech (POS) tagging. We can use bidirectional maximum matching algorithm, which is proposed in previous studies [15], to process Chinese word segmentation. Further, following recent researches [16], we use bigram do the POS tagging.

Because the text cannot be used directly for regression, we use a vector space model (VSM). To avoid the curse of the dimension, we need to delete redundant information (words). Therefore, we use the definition of the similarity between a text and a query:

\[
sim(d, q) = \frac{P(R|d)}{P(R|d)}
\]  

In Eq.1, d is the text (document), q is the query. R is the set of text which is relevant, is the set of text which is not relevant. P(R|d) is the probability that d and q is relevant, and P(d) is the probability that d and q is not relevant. In practice, we can use thresholds to decide whether the text is relevant or not.

Building Dictionary and Calculating Emotion Value

The dictionary can be built on a seed set, or an existing dictionary, or manually. In this paper, we use all these methods. First, we use a seed set. The positive seeds are: 上涨, 飘红, 涨停, 利好. And the negative seeds are: 下跌, 飘绿, 跌停, 利空. Then, we can use the synonyms and the antonyms to expand the seed set. Second, we use the emotion word list on HowNet (www.keenage.com) to expand the dictionary. However, many words in the list may have zero influence on the calculation of the emotion value, which would better be erased. Third, because financial domain has its own emotion words, we have to expand the dictionary manually at last. For example, there are specific terms like: 跌停, 利空. Moreover, there are some internet jargons like: 囧. With this dictionary, we can calculate emotion value. The ROST Emotion Analysis Tool (ROST_CM6), is a famous emotion analysis tool. It can be download at http://download.csdn.net/detail/lgxw1027/3872540, and the explanation can be found at http://blog.sciencenet.cn/home.php?mod=space&uid=239936. However, Prof. Yang Shen is no longer working at the Wuhan University, and the official website is not available now.

Regression

According to Fama [1][2], we have:

\[
RE_{i,t} - RE_{f,t} = \alpha + \beta_m (RE_{m,t} - RE_{f,t}) + \beta_S SMB_t + \beta_H HML_t + \epsilon_{i,t}
\]  

Where \(RE_{i,t}\) is the risk-free rate of return at time \(t\); \(RE_{m,t}\) is the market return at time \(t\); \(RE_{i,t}\) is the return of asset \(i\) at time \(t\); \(SMB_t\) is the simulated portfolio returns of market value factor at time \(t\); \(HML_t\) is the simulated portfolio returns of book value factor at time \(t\).

Base on this we can build our regression model. First, the return rate:
\[ RE_{i,t} = \alpha + \beta_0 BDI_{i,t} + \beta_m RE_{m,t} + \beta_s Size_{i,t} + \beta_h PB_{i,t} + \epsilon_{i,t} \]  

Here we use stock size and PB (price to book) replace SMB and HML. The BDI parameter is the effect of public opinion. Second, the relative return rate:

\[ RR_{i,t} = \alpha + \beta_0 BDI_{i,t} + \lambda_m RE_{m,t} + \lambda_s Size_{i,t} + \lambda_h PB_{i,t} + \epsilon_{i,t} \]  

Third, the volatility:

\[ \text{Sigma}_{i,t} = \alpha + \beta_0 BDI_{i,t} + \gamma_1 MSigma_{i,t} + \gamma_2 Size_{i,t} + \gamma_3 PB_{i,t} + \gamma_4 TOR_{i,t} + \epsilon_{i,t} \]  

Here the MSigma^2_{m,t} is the market volatility, and the TOR_{i,t} is the stock turnover rate.

**Experiments**

**Acquiring Text and Preprocessing**

We collect data from eastmoney.com. There are 10 stocks, and the posts are posted in April 2014. See Tab.1 for details.

<table>
<thead>
<tr>
<th>Stock Name</th>
<th>Code</th>
<th># of Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>宝钢股份</td>
<td>600019</td>
<td>2819</td>
</tr>
<tr>
<td>大立科技</td>
<td>002214</td>
<td>2807</td>
</tr>
<tr>
<td>宇通客车</td>
<td>600066</td>
<td>2660</td>
</tr>
<tr>
<td>广日股份</td>
<td>600894</td>
<td>2511</td>
</tr>
<tr>
<td>海信电器</td>
<td>002415</td>
<td>2644</td>
</tr>
<tr>
<td>京东方</td>
<td>600075</td>
<td>2802</td>
</tr>
<tr>
<td>海信家电</td>
<td>600606</td>
<td>2641</td>
</tr>
<tr>
<td>金风科技</td>
<td>002202</td>
<td>2632</td>
</tr>
<tr>
<td>南山铝业</td>
<td>600219</td>
<td>2613</td>
</tr>
<tr>
<td>松芝股份</td>
<td>002454</td>
<td>2505</td>
</tr>
</tbody>
</table>

| Code | 600019 | 002214 | 600066 | 600894 | 002415 | 600075 | 600606 | 002202 | 600219 | 002454 |
| # of Messages | 2819 | 2807 | 2660 | 2511 | 2644 | 2802 | 2641 | 2632 | 2613 | 2505 |

**Building Dictionary and Calculating Emotion Value**

We used ROST_CM6 to calculate emotion value. However, we use the methods in Section III build our own dictionary. The final dictionary is illustrated in Tab.2.

<table>
<thead>
<tr>
<th>Word</th>
<th>Value</th>
<th>Word</th>
<th>Value</th>
<th>Word</th>
<th>Value</th>
<th>Word</th>
<th>Value</th>
<th>Word</th>
<th>Value</th>
<th>Word</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>创新高</td>
<td>20</td>
<td>创新低</td>
<td>-20</td>
<td>超卖</td>
<td>13</td>
<td>超买</td>
<td>-13</td>
<td>大阳</td>
<td>15</td>
<td>大跌</td>
<td>-15</td>
</tr>
<tr>
<td>涨停</td>
<td>20</td>
<td>跌停</td>
<td>-20</td>
<td>超跌</td>
<td>13</td>
<td>超涨</td>
<td>-13</td>
<td>预增</td>
<td>15</td>
<td>预亏</td>
<td>-15</td>
</tr>
<tr>
<td>涨停板</td>
<td>20</td>
<td>跌停板</td>
<td>-20</td>
<td>获利盘</td>
<td>13</td>
<td>套牢盘</td>
<td>-13</td>
<td>抄底</td>
<td>13</td>
<td>逃顶</td>
<td>-13</td>
</tr>
<tr>
<td>除权</td>
<td>20</td>
<td>基盘</td>
<td>-20</td>
<td>进货</td>
<td>13</td>
<td>出货</td>
<td>-13</td>
<td>利多</td>
<td>10</td>
<td>利空</td>
<td>-10</td>
</tr>
<tr>
<td>除息</td>
<td>20</td>
<td>加息</td>
<td>-20</td>
<td>获利</td>
<td>13</td>
<td>亏损</td>
<td>-13</td>
<td>护盘</td>
<td>10</td>
<td>破盘</td>
<td>-10</td>
</tr>
<tr>
<td>牛市</td>
<td>20</td>
<td>熊市</td>
<td>-15</td>
<td>补涨</td>
<td>13</td>
<td>补跌</td>
<td>-13</td>
<td>小阳</td>
<td>10</td>
<td>小阴</td>
<td>-10</td>
</tr>
<tr>
<td>大涨</td>
<td>15</td>
<td>大阴</td>
<td>-15</td>
<td>中阳</td>
<td>13</td>
<td>中阴</td>
<td>-13</td>
<td>支撑</td>
<td>10</td>
<td>压力位</td>
<td>-10</td>
</tr>
</tbody>
</table>

The emotion value is calculated day by day. For example, the result for Apr. 20th is listed in Tab.3.

<table>
<thead>
<tr>
<th>Stock Name</th>
<th>Emotion Value</th>
<th>Stock Name</th>
<th>Emotion Value</th>
<th>Stock Name</th>
<th>Emotion Value</th>
<th>Stock Name</th>
<th>Emotion Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>松芝股份</td>
<td>4.06</td>
<td>宝钢股份</td>
<td>4.4</td>
<td>海信电器</td>
<td>5.33</td>
<td>京东方</td>
<td>11.19</td>
</tr>
<tr>
<td>南山铝业</td>
<td>3.25</td>
<td>金风科技</td>
<td>2.33</td>
<td>广日股份</td>
<td>6.78</td>
<td>宇通客车</td>
<td>5.36</td>
</tr>
<tr>
<td>宇通客车</td>
<td>5.36</td>
<td>大立科技</td>
<td>1.13</td>
<td>海信电器</td>
<td>5.72</td>
<td>京东方</td>
<td>11.19</td>
</tr>
</tbody>
</table>
Regression
The regression is done with SAS. Using the models we build in Section III, subsection E, we can get our results.

Table 4. Regression results.

<table>
<thead>
<tr>
<th>Model #</th>
<th>Adjusted r-squared</th>
<th>Statistical significance</th>
<th>Corelation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.215</td>
<td>0.000</td>
<td>0.219</td>
</tr>
<tr>
<td>4</td>
<td>0.154</td>
<td>0.006</td>
<td>0.002</td>
</tr>
<tr>
<td>5</td>
<td>0.096</td>
<td>0.126</td>
<td>0.550</td>
</tr>
</tbody>
</table>

The results show that the emotional value and stock’s relative return rates had a positive correlation. However, the emotional value and stock’s return rate do not have a significant correlation. Further, the emotional value and stock’s volatility don’t have a significant correlation.

Summary
In this paper, we try to predict stock price based on the posts on forums through NLP.

Acknowledgement
This paper is supported by the project of the Fundamental Research Funds for the Central Universities (Grant No. 21615438) and the projects of the Natural Science Foundation of Guangdong Province, China (Grant No. 2016A030313084 and No.2016A030313700 and No.2014A030313374) and the Science and Technology Planning Project of Guangdong Province, China (Grant No. 2015B010128007).

References


