Using Social Media Mining Technology to Assist in Price Prediction of Stock Market

Yaojun Wang
Institute of Computing Technology, Chinese Academy of Science; University of Chinese Academy of Sciences
Beijing, China.
e-mail: wangyaojun@ict.ac.cn

Yaoqing Wang
Zhejiang University of Science and Technology
School of Economics and Management
Hang Zhou, China.
e-mail: wangyq1021@163.com

***Abstract*—Price prediction in stock market is considered to be one of the most difficult tasks, because of the price dynamic. Previous study found that stock price volatility in a short term is closely related to the market sentiment; especially for small-cap stocks. This paper used the social media mining technology to quantitative evaluation market segment, and in combination with other factors to predict the stock price trend in short term. Experiment results show that by using social media mining combined with other information, the stock prices prediction model can forecast more accurate.**

Keywords-Stock price prediction; social media; data minin; stock market

I. Introduction

Stock price is a kind of time series in financial domain. The fluctuations of financial time series are dynamically, selectively, and nonlinear, non-stationary and with a lot of noises, which make it difficult to forecast [1]. Effectively predict stock price by using data mining or machine learning techniques in the near future has become one of the most significant issues. However, it is hard to make predictions from the principle of the efficient market hypothesis, since the stock price will follow a random work pattern[2], [3]. In addition, a stationary prediction strategy is also not possible because investors will soon discover such strategies and those successful forecasting rules will lead to self-destruct.

There are many factors that lead to the rise and fall of financial market movement. Predictions of stock market price and its direction are quite difficult. Statistical analysis methods stand a good chance at finding the main factor that impacts the short term stock volatility. While data mining techniques have been profitably to generate high prediction accuracy of stock price movement. A great number of financial analysts and stock market investors' are convinced that they can make profits by employing one of the technical analysis approaches to forecast stock market. Some use time series models expressed by financial theories to forecast a series of stock price data. ANN is usually chosen as a stock prediction tool compared to other methods [4]. However, these approaches cannot work alone because the market value which is always subject to external impact. The stock market is affected by system uncertainties and other unknown factors. To make predictions, several computing techniques must work synergistically rather than exclusively. Recently lots of researchers have found that market segmentation is one the most influential factors affect the direction of stock price movement and is necessary to achieve more accurate outcome.

With the development of the Internet and the popularity of a variety of social media, it becomes more easily for people to gain access to information through crawling data from web site. This web information includes documents, news, blogs, forums, emails, and etc. Among them, the micro blogging, such as Sina Weibo, Twitter and Facebook, features informative content and wide spread speed, has attracted an increase number of users to receive and share information or comments. This rapid growth of available textual instant information has given rise to a research field devoted to knowledge discovery in unstructured data (textual data) --known as social media mining.

Social media mining originates from the relevant field of data mining, which mines patterns from structured data instead of unstructured. It is also related to other fields like information retrieval, web mining, statistics, computational linguistics and natural language processing [5], [6].A valuable application of social media mining is text sentiment analysis, also referred to as opinion mining. This technique is applied to discover the sentiment of a written text and categorize text documents into a set of predefined sentiment categories (e.g. positive or negative sentiment categories). Sentiment analysis is a good standing point when applying data mining to analyze stock comment from the social media. This is because positive market emotion is likely to have positive influence in stock price, while the opposite are true to negative.

This paper focus on the daily sentiment analysis from the information on popular financial social media. The market sentiment will follow some random distribution like the market price which follows a random walk [7-9]. Therefore, it is assumed that if the market sentiment is not separated and has an impact on stock price movement. First, It has to be determined whether it is effective to use the sentiment evaluated from social media to predict stock price or not.

The rest of the paper is organized as follows. Section 2 introduces system design and describes the sub-suctions of the stock prediction system. Section 3 evaluates the approach by applying the method to the test data to get the metrics. Section 4 gives some related works of the subject. Section 5is the conclusion of this paper.

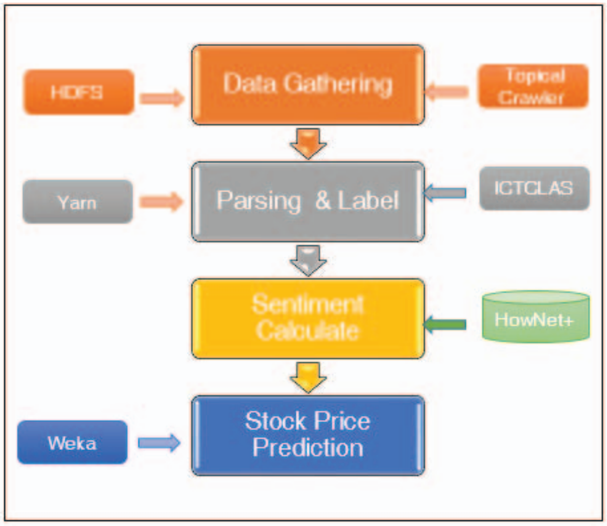

Figure.1. Stock price predicting system

II. Approach

Figure 1 is a diagram of overview of the approach. The first step is gathering stock comment information from social media by virtual of Topical Crawler, then make a preprocessing to the data, afterwards transform the redundant data to a labeled vector and then feed it into the sentiment analysis algorithm to evaluate the segment index of each stock. Next make predictions of the stock price using the SVM model which includes the segment index.

A. Data Fetching

The stock history data is gathered through financial API from the Yahoo and Google. The stock comment data used in this paper was fetched from Chinese popular financial social media by topical crawler. All the data stored with Hadoop distributed file system (HDFS).The websites used to collect data included: Sina Weibo, Tong Hua Shun Network and Dong Fang Cai Fu network.

Sina Weibo, China's largest and most widely used social media sites, is a Chinese equivalent to twitter in the United States. At present, it has more than two hundred million active users. Instantly express their feelings about life through the micro blogging has become an important part of daily life. Tong Hua Shun Network and Dong Fang Cai Fu Network are two famous web site which provide stock information and also have social media platform for registered users to write stock comments. There usually have hundreds of comments for one stock and the contents are mainly in Chinese.

These social media texts contains a huge commercial value, which using the emotional color analysis can help predict the movie box office, conduct monitoring public opinion, to understand the user experience. For implement in stock price predicting, the comments related to the stocks and financials in necessary and enough. To realize the emotional data fetching, a topical web crawler is developed and used to fetch the stock comments after the closing of each trading day.

B. Sentiment Analysis

The first step in data preprocessing is feature extraction, in other words, text parsing, to fetch a vector of words and phrases that describes the stock comment. Using the Chinese segmentation tool ICTCLAS [10] to realize word segmentation. This tool was developed by the colleagues of our institute.

Furthermore, label the segment tags to the word of each vector (each word was labeled with negative or positive) based on the How Net+ sentiment dictionary. How Net [11] is an on-line common-sense knowledge base that can unveil inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. How Net sentiment dictionary explicitly divides the emotion words with negative types and positive types. How Net+ is an improved version of How Net with additional 756 financial sentimental words fetched form manually labeled 1250 financial comment articles.

TABLE I. Sentiment Scores of Sample Data

Date	Close Price	Volume	Amount	Sentiment index
2015-11-16	7.08	11026037	152886112	1.23
2015-11-17	7.14	12472447	177067424	1.21
2015-11-18	7.35	17809164	263919872	0.92
2015-11-19	7.50	11528767	169354048	0.38
2015-11-20	8.06	22994500	363453440	-0.53
2015-11-23	7.95	14395990	229473776	0.24
2014-11-24	7.66	17324408	263894032	-1.32

Finally, calculate the sentiment index and sentiment discrepancy indexes of each stock's labeled word vector using scoring formula as below:

$$SI = ln\frac{1+N_{positive}}{1+N_{negtive}} \quad （1）$$

$$SDI = \left|1 - \left|\frac{N_{positive}-N_{negtive}}{N_{positive}+N_{negtive}}\right|\right| \quad （2）$$

SI is the sentiment index while SDI is sentiment discrepancy index,$N_{positive}$ is the positive word number and $N_{negtive}$ is the negative emotion word numbers. Each stock corresponds to a pair of value SI and SDI. Table I is a sample of calculated sentiment score. The trading data is a data slot from stock 600719.SS.

C. Stock Price Prediction

This paper mainly focuses on the short-term investors' sentiment which is measured by the comments of the performance of the stocks in social media. In addition to investor sentiment, the paper also concerns other four factors to help predict stock price.

The first step uses principal component analysis to find main factors that in fluence the stock price trend. Using the SIand

TABLE II. PREDICTION QUALITY IN THE EXPERIMENT DATASETS

Data sets	SVM		SVM_sentiment	
	Pearson CC.	Max bias	Pearson CC.	Max bias
600719.SS	0.56	5.1%	**0.79**	**3.5%**
Sample Index(20 item small-cap stocks)	0.59	5.8%	**0.76**	**2.3%**

SDI combined with other information of the stock training a model and used in predicting stock price.

They are daily closing price, volume, market index and daily turnover rate. Using all securities' Historical trading data of Chinese stock market with time ranges2009-01-01 to 2015-10-01 as training sets. Using principal component analysis of the training sets and building a liner predicting model as:

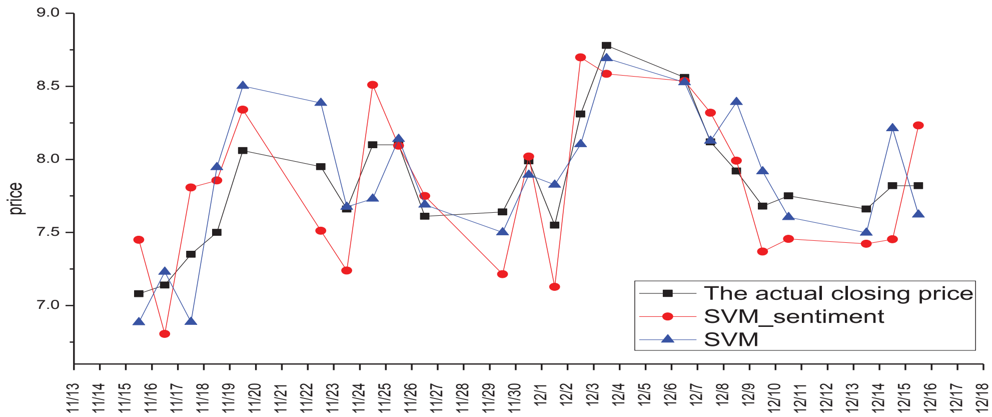

Figure.2. Closing Price Prediction of Stock Code 600719.SS

$$P_i = 0.6238*P_{i-1}+0.0455*V_{i-1}+0.0213*TR_{i-1} + 0.0316*M_{i-1}+0.0423*SDI+\beta \quad (3)$$

Choosing the SVM method to further train the prediction model. The most widely used Radial Basis Kernel was selected as kernel function.

$$K(x_i,x_j)=\exp\left[-\frac{\left\|x_i-x_j\right\|^2}{\sigma^2}\right] \quad (4)$$

III. EXPERIMENT RESULTS

The goal of the experiments is to determine whether our approach is effective in predicting next trading day'sstock price.

A. *Experiment Data Preparations*

The experiment data include20 items small-cap stocks were randomly selected from Chinese stock market. The definition of small cap can vary among brokerages; here we define the small-cap stocks as a company with less than one hundred million of the outstanding shares.

All data are from stock market with the times range from 11-13-2015to 12-18-2015, including 23 trading days. The social media data fetched from Sina Weibo, Tong Hua Shun Network and Dong Fang Cai Fu network including 35 natural days, data size is about10GB.

B. *Stock Price Prediction*

The reason for the experiment choosing the small-cap stocks is based on the statistics that small-cap stocks was more easily affected by market sentiment than large-cap stocks.

In order to evaluate the efficiency of social media mining in price prediction model. We have designed two prediction models, one is "SVM_sentiment" and another is "SVM". The"SVM_sentiment" was an SVM model realized by the formula 3 which includes segment index and "SVM" model used as a control group without segment information (other information is same).

As figure 2and table II shows, on the 600719.SS dataset, the Pearson correlation coefficient count between "SVM_sentiment" predicted price and the actual price is 0.79.On the contrast, "SVM" showed a rate of correlation of 0.56. Sample index includes 20 item stocks were constructed using the arithmetic mean method. We also test these two models on the sample index with the same time range as 600719.SS.The experiment on real data has shown that the SVM model which containing the segment indexes predicted price closer to the actual closing price with small bias.

IV. RELATED WORKS

A. Stock Price Prediction Using FinancialNews

There have been many researches aimed at identifying the relationship or predicting stock market movements using news to analyze. A Korea researcher named Yoosin Kim designed a method consists of the NLP categorizing and extracting the sentiments and opinions expressed by the writers. But the effect it not satisfactory [12]. A New York researcher RoberP. Using a synthesis of linguistic, financial and statistical techniques to create a prediction system called AZF in Text [13].It is known that trading return performed well in the top 10 quantitative mutual funds of 2005. By looking back upon testing data sets from other years, it can be seen that the return rate is generally below 15%.

B. Price Trend Prediction Using Outlier Data Mining Algorithm

Baylor University's Zhao, Lei [14] proposed an outlier mining algorithm to detect anomalies on the basis of volume sequence of high frequency tick-by tick data of stock market. Using anomalies on distributions of trading volume to predict upward trends of stock prices. This approach predicts the stock trend effectively and can make profits on the Chinese stock market in a long-term usage. The situation is the same as "Financial News" methods that the features use for predicting stock price are too deficient to make accurate prediction.

V. CONCLUSION

In this paper, starting from the efficient market hypothesis we fetch the stock comments information from social media and then preprocessing the data to emotion vectors. By calculating the segment value of each stock's emotion vector, we found that the segment value is very sensitivity to the stock price movement. In addition, through testing our social media mining algorithm we discovered that the SVM model contains segment index has higher prediction accuracy than SVM model not combined segment index.

REFERENCES

[1] TSAY R S. Analysis of financial time series [M]. John Wiley & Sons, 2005.

[2] TIMMERMANN A, GRANGER C W. Efficient market hypothesis and forecasting [J]. International Journal of Forecasting, 2004, 20(1): 15-27.

[3] MALKIEL B G. The efficient market hypothesis and its critics [J]. Journal of economic perspectives, 2003, 59-82.

[4] ZHANG G, PATUWO B E, HU M Y. Forecasting with artificial neural networks:: The state of the art [J]. International journal of forecasting, 1998, 14(1): 35-62.

[5] TANG L, LIU H. Community detection and mining in social media [J]. Synthesis Lectures on Data Mining and Knowledge Discovery, 2010, 2(1): 1-137.

[6] CORLEY C D, COOK D J, MIKLER A R, et al. Text and structural data mining of influenza mentions in web and social media [J]. International journal of environmental research and public health, 2010, 7(2): 596-615.

[7] DAS S, CHEN M. Yahoo! for Amazon: Extracting market sentiment from stock message boards; proceedings of the Proceedings of the Asia Pacific finance association annual conference (APFA), F, 2001 [C]. Bangkok, Thailand.

[8] BODURTHA J N, KIM D-S, LEE C M. Closed-end country funds and US market sentiment [J]. Review of Financial Studies, 1995, 8(3): 879-918.

[9] EICHENGREEN B, MODY A: National Bureau of Economic Research, 1998.

[10] ZHANG H-P, YU H-K, XIONG D-Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS; proceedings of the Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17, F, 2003 [C]. Association for Computational Linguistics.

[11] ZHU Y-L, MIN J, ZHOU Y-Q, et al. Semantic orientation computing based on HowNet [J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.

[12] KIM Y, JEONG S R, GHANI I. Text opinion mining to analyze news for stock market prediction [J]. Int J Advance Soft Comput Appl, 2014, 6(1):

[13] SCHUMAKER R P, CHEN H. A discrete stock price prediction engine based on financial news [J]. Computer, 2010, 1): 51-6.

[14] ZHAO L, WANG L. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm; proceedings of the Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on, F, 2015 [C]. IEEE.