Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 174 (2020) 191–199

Procedia Computer Science

www.elsevier.com/locate/procedia

2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)

Using News to Predict Investor Sentiment: Based on SVM Model

Diya Wang[a], Yixi Zhao[a]

Tianjin Yaohua High School, No.106 Nanjing Road Heping District Tianjin China
JessicaWang0708@163.com
ULC Cambridge International School of Optics Valley, 101 foxconn west road, donghu new technology development zone, jiangxia district, wuhan city, hubei province, China
1366317647@qq.com

These authors are contributed equally to this work

Abstract

In previous studies, people paid more attention to the influence of news on stock price, but neglected the role of investors in stock trading. News influences investors as they make investment decisions. In this article, we look at how news affects investor sentiment. We selected PSY (psychology line), commonly used in stock investment, as an indicator to measure investor sentiment. Based on the SVM model, we establish a model that USES news texts to predict changes in investor sentiment. Our model can achieve high accuracy, which is helpful for guiding the prediction of investor sentiment in the stock market and making correct decisions. After comparing several indicators and models, we found that the investor sentiment prediction model based on SVM has a high accuracy, which can analyze the investor sentiment of the day according to the stock related news and guide investors to make correct investment decisions.

Keywords: Investor Sentiment , PSY, SVM

1. Introduction

Today, social media has become the main way of generating and spreading public opinion. On the other hand, in the stock market, news articles play an essential role on influencing the investors' judgment and confidence in stock value. Research on stock market enables investors to be more successful on Their investments.

Wesley s. Chan[1] believes that news has an impact on stock prices.News can often convey investors' feelings about

10.1016/j.procs.2020.06.074

the market, which can influence their investment decisions.When news about a stock occurs, share prices usually react accordingly.For example, share prices tend to rise when the news is positive;when share prices are negative, they tend to fall.

At present, the research on the stock price of news has been very mature, but this paper pays more attention to how news affects investor sentiment, and thus affects the stock price. Through empirical analysis, scientists find that investor sentiment is a very important factor affecting index returns. Compared with other models, the SVM model considering investor sentiment has a very obvious advantage in the positive direction of winning rate, and thus obtains higher cumulative returns and overall winning rate.

As is known to all, news media often directly or indirectly affect investors' expectations and market operation through the prediction and evaluation of the stock market. First of all, the optimism tendency of media reports and the market index yield have a significant positive impact on emotions, but the density of media reports and the optimistic prediction of the future do not have a significant impact on emotions. Secondly, the market return rate itself presents a certain degree of reversal effect. The optimistic word ratio and optimistic report ratio reported by the media also have a positive impact on the market return rate.

Finally, the bullish bias reported by the media in bull market environments has a more significant impact on sentiment than in bear market environments. Therefore, in order to better improve the capital market management system, it is necessary to effectively supervise the news media to ensure timely, accurate and effective transmission of stock market information [2].

To sum up, it is not hard to see from various experiments that investors' behaviors and emotions are quite important factors worth considering in the prediction of financial time series. The introduction of machine learning into nonlinear modeling will be a powerful analytical tool for investors engaged in quantitative investment.

2. Related work

Wen, Song and Tian proposed a new kernel called semantic and structural kernel, referred to S&S kernel, to forecast the stock's fluctuation based on SVM(Support Vector Machine) relating to not only the contents of the news article but also the information structures among them [1].

Wen, Xiao, He and Gong conducted experiments on the singular spectrum analysis (SSA), and SVM. They found that it is more effective to use combination prediction which combines the decomposition of original index into series with certain economic implications to the SVM [2].

For further research, Nam and Seong also designed a procedure to predict stock price movement based on the financial news considering causality. In their method, they analyzed the relation between the news and the stock, examined the causal relationship between companies, and also calculated the transfer entropy to present the causality and multiple kernel learning to combine features of target firm and causal firms [3].

What's more, Wang and Liu proposed a Chinese text categorization method based on graph model by using a weighting method to select the relevant feature building graph, improving the text representation model, and designing a learning algorithm to classify Chinese text through graphs [4].

Bronselaer and Pasi proposed a new graph-based model by which a decomposition of textual documents is obtained where tokens are automatically parsed and attached to either a vertex or an edge [5].

Not only the contents themselves, sometimes the sentimental words also played an essential role in determining the news and affecting the trend of the certain stocks. In their article, Kim, Jeong, and Ghnani considered about the sentiment words; in their experiment, they applied the sentiment word dictionary to help them recognize and define the ‘emotional state’ expressed in the text [6].

Additionally, Kuo et al. also focused on the framework and presented a compact frame-based facial expression recognition framework, which performed well in their experiments. Compared with other methods, their framework requires less parameters [7].

Many machine learning algorithms required input of a fixed length vector. When it comes to short text, the most commonly used fixed length vector method is bag-of-words. Schumaker and Chen used Bag of Words, Noun Phrases, and Named Entities to approach the textual Analysis of Stock Market Prediction [8].

Despite its popularity, the bag model has two main disadvantages: one is that the bag model ignores word order, while the other is that the bag model ignores the syntax. Thus, when training the model, it causes similar distances among 'powerful', 'strong' and' Paris', whereas in fact 'powerful' should be closer to 'Paris' than to ‘strong’.

3. Mothodology

SVM

In the 1990s, Vapnik **[9]** proposed an efficient classification algorithm for support vector machines, which was subsequently applied in many fields. Over the years, with the continuous improvement of its algorithms, support vector machines have become one of the most useful algorithms in large-scale data classification. [12] The basic idea of SVM learning is to solve for a separate hyperplane that correctly partitions the training data set and has the largest geometric spacing.In the 1990s, Vapnik[11] proposed support vector machine (SVM) as an efficient classification algorithm, which was subsequently applied in many fields.

With the improvement of its algorithm over the years, support vector machine has become one of the most useful algorithms in data classification.**[10]** The basic idea of SVM learning is to solve the separation hyperplane that can correctly divide training data sets and have the largest geometric interval.

Consider a classification problem in n-dimensional spacewith l training samples. The training samples can be defined as:

$$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_l,y_l)\} \qquad (1)$$

where $x_i \in R^n, i=1,\dots,l$ and for a binary classification problem, $y_i \in (-1,1)$. The goal is to identify a new sample x belonging to which class (1 or -1) after training with dataset T. To solve this problem, a decision function f(x) is needed to separate the Rn space into 2 regions:

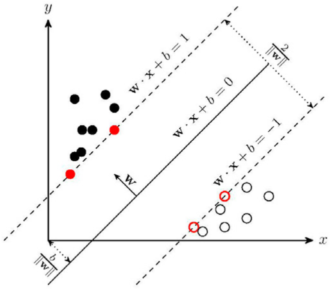
$$f(x)=\operatorname{sgn}(g(x)) \qquad (2)$$

where g(x) is a real function to obtain the value of y for each x. Particularly, for a linear classification problem, g(x) can be a linear function:

$$g(x)=w\cdot x+b \qquad (3)$$

and the corresponding hyperplane is

$$w\cdot x+b=0$$



For nonlinear separation, an appropriate map ∅ is needed to transform an n-dimensional vector x into another m dimensional vector in space R^m. Thus, the maximal softmargin algorithm of the SVM deduces the following primal optimization problem:

$$\begin{aligned} &\min_{w,b,\xi} && \frac{1}{2}\| w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &\text{s.t.} && y_i\left(w\cdot\Phi\left(x_i\right)+b\right)\geq 1-\xi_i \quad (4) \\ & && \zeta_i \geq 0, \quad i=1,\ldots,l \end{aligned}$$

where C is a penalty parameter and ξ_i represent the slack variables. In this paper, we use a nonlinear kernel function, and set $\Phi\left(x_i\right)\cdot\Phi\left(x_j\right) = K\left(x_i, x_j\right)$. A convex quadratic programming problem can be constructed as:

$$\begin{aligned} &\min_{\alpha} && \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right) - \sum_{j=1}^{l}\alpha_j \\ &\text{s.t.} && \sum_{i=1}^{l} y_i\alpha_i = 0 \qquad (5) \\ & && 0\leq \alpha_i \leq C, \quad i=1,\ldots,l \end{aligned}$$

where α_i are Lagrangian multipliers. After obtaining the solution a $\alpha^* = \left(\alpha_1^*,\ldots,\alpha_l^*\right)^T$, the optimal separating hyperplane
can be given by:

$$g(x) = \sum_{i=1}^{l} y_i\alpha_i^* K\left(x_i, x\right) + b^*$$

$$b^* = y_i - \sum_{i=1}^{l} y_i\alpha_i^* K\left(x_i, x_j\right) \qquad (6)$$

Then, a new sample is classified as 1 or -1 according to the decision function of formula(2).

4. Model

First we need to get the news text data and stock data we need. We obtained the news data of the medical industry through the web crawler. We first cleaned the data and then vectorized them. For stock data, we obtained PSY and the opening and closing prices of corresponding companies on the day of news from the financial database, and then cleaned the data. After processing the data, we put the news data and PSY into the SVM model, and used the news data to predict PSY and output the prediction accuracy. This model can change the inputs and methods to make multifaceted comparisons. We designed three comparative experiments on the model prototype to test how to obtain more accurate investor sentiment prediction.

5. Experiment

5.1 Data

5.1.1 Data collecting:
First we need enough news text data to support the subsequent models. We used web crawlers to crawl a lot of news texts on major financial news media. In order to avoid the impact of news from different industries, we only obtained data from the medical industry. These news texts from December 3, 2013 to March 31, 2017, a total of 19908 stock news texts and 10914 industry news texts. The crawled news text contains the date, the corresponding company, the specific content, and some text symbols. Then we obtained the stock information of the company on the day of each news from the Wind database. To measure the impact of news on consumers, we

chose a Psychological line (PSY), a commonly used indicator in equity investment. PSY is an emotional indicator for investigating the psychological fluctuations of investors in the stock market. It has certain reference significance for the research and judgment of the short-term trend of the stock market. The researchers found that on the one hand, people's psychological expectations are directly proportional to the market's level, that is, the market is rising, psychological expectations are rising, the market is falling, and psychological expectations are falling. On the other hand, when people's psychological expectations are close to or reach In extreme times, rebellious psychology begins to work and may eventually lead to a reversal of the psychologically expected direction.

$PSY(n)=A/n\times100$

A is the number of cycles in which the stock price rises during the n-cycle. n is the sampling parameter of PSY, which can be divided into clock, day, week and month.

In our model, let's make a hypothesis: if PSY > 50, then the stock price has an upward trend, that is, the news has a positive impact on investors, at this time we put a "+1" label on PSY; if PSY <50, it is determined that the stock price has a downward trend, that is, the news has a negative impact on investors. At this time, we put a "-1" label on PSY.

5.1.2 Data cleaning

After we have obtained the news text data and the stock data, we need to clean the data and have reached the requirements for modeling. For news text data, we first need to segment it. In this article we use Python's jieba participle (requires a reference here). Then remove some symbols and garbled. After that, we need to vectorize the news text. We used the doc2vector model () to complete the vectorization of news text.

For stock data, we cleaned it and label it. For industry news, we didn't classify it, but put it into the universum-support vector machine (U-SVM) model. U-SVM uses 3-class classification approach to solve the 2-class classification Problem. [11] Next, the vectorized news data and stock data are put into the SVM model.

6. Results

In the experiment, we use the model to predict the accuracy rate, which is the accuracy of the prediction of investor sentiment in the model. In our model, news data is used as input and output is the result of prediction.

To measure the impact of industry news on investor sentiment, we also built the U-SVM model as a comparison with the SVM model.

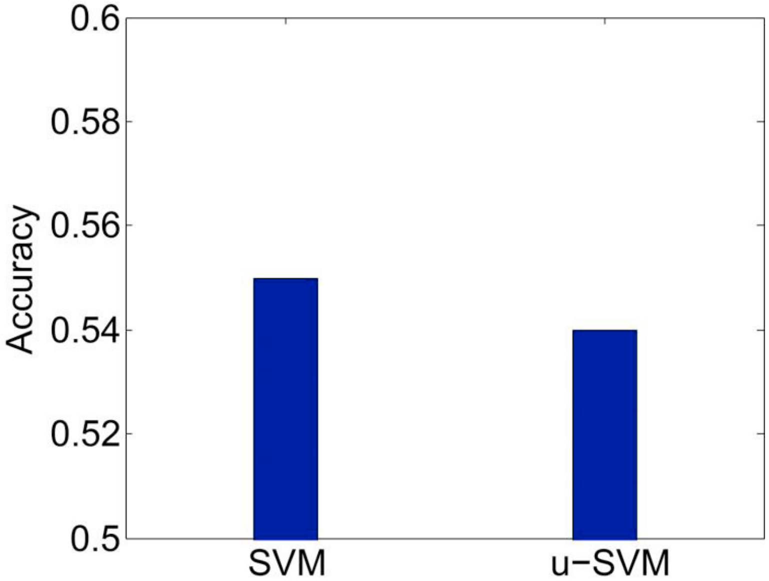


Fig.1. Accuracy among different models

Fig.1. illustrates that the SVM model has a good effect on the prediction of investor sentiment. We found that without adding industry news, the accuracy rate was higher than after joining, the former reached 55%, and the latter only 54%. The reason for this situation is that the emergence of a large number of industry news has led to over-fitting.

At the same time, in order to measure, how different amounts of news data will affect the prediction results, we also designed a comparison experiment based on the SVM model to change the amount of data.

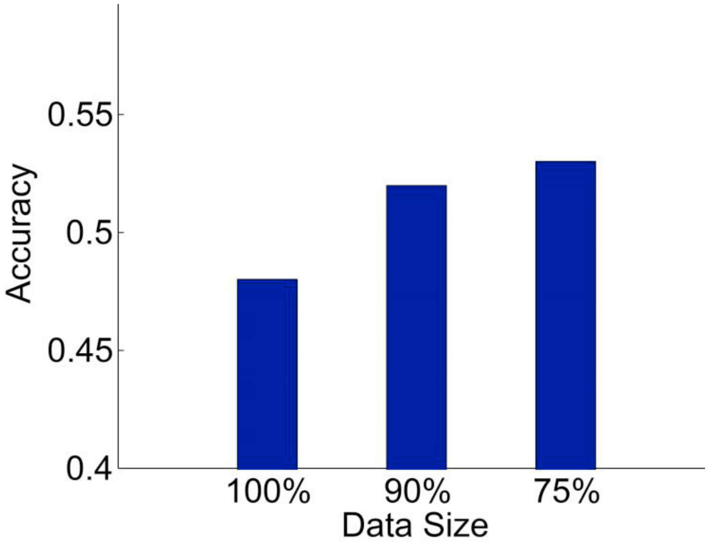


Fig.2.Accuracy among data size

Fig.2. illustrates that changing the amount of news data based on the SVM model leads to a lower accuracy of model prediction. When the amount of data is 100%, it reaches 55%; when the amount of data is 90%, the accuracy rate reaches 52%; when the amount of data is 50%, the accuracy rate reaches 53%.

Finally, in order to observe how news affects the stock price changes of the day, we also use ratio as an indicator to test.

Ratio=（close price/open price）*100%

When ratio>1, it means that the stock price is rising that day, we will have a “1” label for this ratio; when ratio<1, it means that the day is down, we give this ratio a “-1” label.

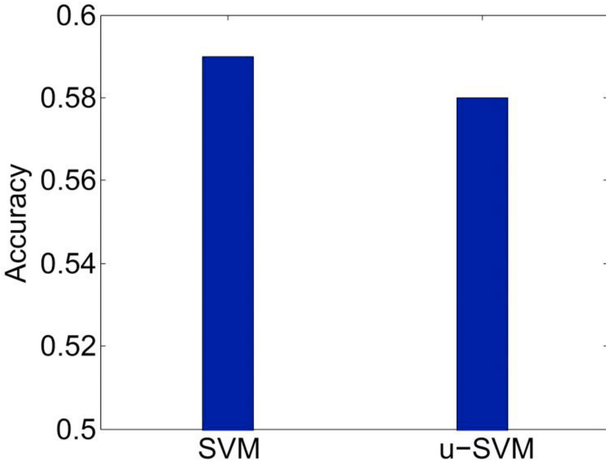


Fig.3. Accuracy among different indicators

Fig.3. shows that the accuracy rate does not change much when the predictive index is changed based on the SVM model. When PSY is selected as the index, the accuracy rate reaches 59%. When ratio is selected as the index, the accuracy rate reaches 58%. Although these two indicators are very different, PSY is used to indicate the change in investor sentiment, and ratio is used to reflect the stock price's rise and fall on the day. This shows that the SVM model has a very good effect in predicting stock information and investor sentiment.

7. Conclusion

Through our experimental research, we found that the investor's emotions are influenced by the news text, and using the SVM model to predict can achieve a higher accuracy. The influence of investor sentiment on stock prices is crucial, so the prediction of emotions can effectively guide investment behavior. Our research also found that adding too much industry news when forecasting will reduce the accuracy of the forecast, so we believe that in order to avoid the impact of macro information on accuracy, we should make predictions based on individual stock news

Acknowledgements

The completion of the thesis is attributed to many people's support and encouragement. Without of the help of many people,it could never be completed, thus here we would like to express my sincere gratitude towards them.
First and foremost, we owe our heartfelt thanks to my distinguished and cordial supervisor,Professor Yingjie Tian, who influenced us with his insightful ideas and meaningful inspirations, guided us with practical academic advice and feasible instructions, and enlightened us while we was confused during the writing procedure. His thought-provoking comments and patiently encouragements are indispensable for my accomplishment of this thesis.
Ultimately, thanks go to our parents who are our mentor and guardian from the very beginning in primary school. Without their refined education and care,we could never grow up in such a joyous and cozy environment nor have the courage to confront any obstacles on our way to success.

References

[1] YIN Hai-yuan. ″A Study on Effect of Media Reports on Investor Sentiment:Evidence from China′s Stock Market.″ Journal of Xiamen University(Arts & Social Sciences) (2016)

[2] Wesley S. Chan. ″Stock Price Reaction to News and No-News: Drift and Reversal After Headlines.″ Journal of Financial Economics (2003): 223-260

[3] Long, Wen, Song, Linqiu, Tian, Yingjie. ″A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity.″ Expert Systems with Applications (2018)

[4] Fenghua WEN, Jihong XIAO, Zhifang, HE. ″Stock Price Prediction Based on SSA and SVM.″ Procedia Computer Science (2014) : 625-631

[5] Nam KiHwan, Seong, NohYoon. ″Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market.″ Decision Support Systems (2018)

[6] Kim, Yoosin, Jeong, Seung Ryul, Ghani, Imran. ″Text Opinion Mining to Analyze News for Stock Market Prediction.″ International Journal of Advances in Soft Computing and Its Applications (2014)

[7] Robert P. Schumaker, Hsinchun Chen. ″Textual analysis of stock market prediction using breaking financial news: The AZFin text system.″ ACM Transactions on Information Systems (TOIS) (2009)

[8] Hsin-Min Lu, Nina WanHsin Huang, Zhu Zhang. ″Identifying Firm-Specific Risk Statements in News Articles.″ Pacific-Asia Workshop on Intelligence and Security Informatics (2009)

[9] Fuhr, Norbert. ″Probabilistic Models in Information Retrieval.″ Computer Journal

[10] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang. ″Forecasting stock market movement direction with support vector machine.″

[11] Yune, HongJune, Kim, HanJoon, Chang, JaeYoung. ″An Efficient Search Method of Product Review using Opinion Mining Techniques.″

[12] Deng N, Tian Y, Zhang C. ″Optimization based theory, algorithms, and extensions.″ CRC Press, New York (2012)

[13] Wen Long, Ye-ran Tang, Ying-jie Tian ″Investor sentiment identification based on the universum SVM.″