Research on the Forecast of Stock Return Direction based on Financial News

Fangge Hu*

School of Journalism, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China

*h13655505300@163.com

Abstract

Quantitative investment strategy is a hot issue in the field of stock investment. In quantitative investment, the goal is to maximize the income based on minimizing the risk. This paper takes A-share listed companies in Shanghai and Shenzhen stock markets as samples to study stocks' quantitative timing investment strategy. The income base of the stock market lies in accurately grasping the trend of stock price. Aiming at ordinary investors who account for most stock market investments, this paper constructs a sentiment index dictionary with financial news as the theme. Then, it compares the prediction accuracy of Logistic, TVF, and F-TVF models. The final empirical results show that the Time-varying factor-weighted nonparametric density function model(F-TVF) with influence factors has a better prediction effect. That is, the sentiment index of economic newspapers is positively correlated with the rate of return.

Keywords

Quantitative Investment; Stock; Financial News; F-TVF Model.

1. Introduction

1.1. The Research Background

At present,the vast majority of investors in China's stock market are an investment collective composed of individual investors who lack consistent theoretical knowledge,but because individual investors can easily make indecisive decisions,the uncertainty of investment is further expanded(Wu Songtao et al.,2017 [1] (7)). With the rapid development of China's economy, online media is becoming increasingly wealthy and powerful. Various stocks, news portals, and forums have entered young people's daily lives in the new era. Investors can transmit information through these platforms and know the financial information of markets and sectors at any time. Media reports have a significant impact on the capital market (Chen Guojin et al.,2019(5) [2]). Compared with institutional investors, China's private investors account for the vast majority, while the stock market has a noticeable herd effect and intense speculation. Therefore, when the financial news rises, investors buy stocks to seize time and increase the stock price. When financial information is sluggish. Investors may reduce the expected psychology and enthusiasm of investment(Fu Kui et al.,2019 [3]).

Therefore, to a certain extent, the impact of online financial news on the stock market is continuous. The text analysis method is to digitize text and transform it into structured data. An excellent analytical approach is to study the relationship between emotion and financial and economic texts through an emotion dictionary. However, at present, there is still no perfect Chinese dictionary of economics and finance. Based on the multi-dictionary and word2vec tool, this paper puts forward an emotion dictionary suitable for stock information analysis and prediction. Because the stock market reflects economic development to a great extent, stock forecasting has always been a hot issue and core issue in financial research. Many former scholars have studied the impact of economic policies and corporate management on stocks. In the era of much-consulting information, the market information reflected behind financial news has a considerable effect on the investment behavior of investors who do not have much professional investment knowledge. The analysis of unstructured economic texts and the construction of sentiment index provides a constant reference value for studying the relationship between financial news and the stock market,which affects the stock market. Nowadays, financial investors are always interested in financial and economic information. For investors, it is of great practical significance to judge the income direction according to financial news. Financial news may also affect investors. Because investors usually get the relevant information through the Internet. Therefore, it is of great significance to dig deep into the influence of economic news on the stock market and forecast the stock market through financial information.

2. Literature Review

The success of a quantitative trading strategy is based on the appropriateness and strict assumptions of investment theory. Under the same conditions, given this theory's universal applicability, more research on quantitative trading strategies also has the characteristics of homogeneity, such as the emergence of machine algorithms and artificial intelligence in recent years. In addition, new algorithms such as genetic algorithms are emerging one after another. In the literature at home and abroad, the primary methods to predict market behavior by building models are divided into the following categories:

First, from the indicators and parameters, further, analyze the company's value, and evaluate the fundamentals of the company, to select stocks to establish the best asset portfolio. Second, starting from the mathematical model, showtime series models, such as ARCH, ARIMA, etcto use this model to simulate market conditions. Then, the influence of volatility on value is analyzed to explain the emergence of volatility and use it as a reference for asset investment strategy. Ma Juan et al. (2019) used the improved RSI index to build a transaction model [from the price-volume relationship of financial investment [4]. Research shows that stock price volatility has fixed memory and internal connection. Zhang Xudong et al. (2020), based on Markov chain theory, made full use of the primary data of market and A-shares from 2012 to 2015 and the A-share model based on GARP theory [5]. Cui Wenzhao et al. (2019) used the Piotroski method to construct Fama-French multi-factor model further to better verify the effectiveness of the multi-factor quantitative method in investment through univariate regression [6]. Luo Xin and Zhang Jinlin(2020)show the disorder of the market and the memory characteristics of charts by measuring volatility and amplitude.From the perspective of time domain and frequency domain,the composite deep neural network is used to predict the stock price [7].

3. Theoretical Mechanism.

3.1. Financial News Emotional Dictionary Construction

In the era of big data, financial information is mainly presented in the form of text in various media, and the text format is unstructured data (Jerry Lee et al.,2017[8]). Unstructured data contains valuable information, so it is necessary to obtain confidential information through processing ability. The financial news information investigated in this paper is mainly emotional. A piece of news can be regarded as a combination of several words in an article. News emotion is expressed in language. In this paper, the Harvard Psychological Dictionary, a classic emotional dictionary, is further innovatively constructed (Yin Hairen and Wang Panpan,2015 [9]). In this paper,based on the semi-automatic method and word2vec tool,we determine the emotional trend of vocabulary for better prediction.

3.2. The Theoretical Framework of the Stock Yield Direction Prediction Model

For the directional prediction model of stock return, the traditional methods mainly focus on the linear and parametric models, and this model has potential assumptions. However, the stock market is a complex system, and the data used does not necessarily follow the hypothesis, but the actual distribution function is quite different from the idea, and the prediction effect of the model is significantly reduced (Zhao Qingguo et al.,2020 [10]). For example, the classic logistic model creates an assumption that follows the ascending and logistic distributions for the binary selection problem and the ascending predicted values discussed in the text. To avoid the presupposition of the model, this paper chooses the nonparametric method to build a new model, indicates the rising and falling direction of stock price,and compares it with the Logistic model. Furthermore,in this paper, the emotional index is added to the Time-varying factor weighted density function model to achieve the best prediction effect.

3.2.1. Logistic Model

The logistic model has an actual application in the financial field and plays a critical role in solving the binary problem. As far as the stock price is concerned, the stock price has only two rising and fall (if the price remains unchanged, one is included). In this paper, the logistic model is regarded as a two-class problem to predict the return direction of stocks. The focus of stocks is changeable. Stock gains are as follows.

$$P = P(Y = 1 | X_1, X_2, \cdots X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k)$$

The logistic model is a probability model, p is when $Y = 1$ that is, the probability of stock rising, the result value can be well converted into 0 and 1 by threshold. Because the value of p to be obtained is a constant value in the range of 0~1,the Sigmoid function satisfies these conditions well.

$$Y = \frac{1}{1 + e^{-2}}$$

Therefore, the Logistic model is finally defined as

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}$$

Given the dependent variable x, the parameter β you know, you can estimate the probability p of the stock rising. For parameters $\{\beta_0, \ \beta_1, \ \beta_2, \cdots, \ \beta_k\}$, the maximum likelihood method is used to estimate the model parameters.

Finally, it is concluded that

$$L(\beta_0, \beta_1, \beta_2, \cdots, \beta_k) = \sum_{i=1}^{N} \left[y_i \cdot Z - \log(1 + \exp(Z)) \right]$$

The parameters can be obtained by taking the maximum value from the above formula $\{\beta_0, \ \beta_1, \ \beta_2, \cdots, \ \beta_k\}$ estimated value.

3.2.2. Time-varying Density Function Model (TVF Model)

Harvey&Oryshchenko (2012) combined the exponential weighted moving average method with nonparametric kernel density estimation to obtain a Time-varying probability density function. Assuming a time series observation value, it can be obtained by exponential weighted moving average:

$$\hat{f}_i = \frac{1}{h}\sum_{i=1}^{T} H(\frac{y-y_i}{h})w_{ti}, (t = 1,2,\cdots,T)$$

among $H(\bullet)$ and $K(\bullet)$ The corresponding cumulative kernel function, $w_{ti} = (1-w)w^i$ Is the weight factor. Continue to analyze available

$$\hat{F}_{t+1|t}(y) = \omega\hat{F}_{t|t-1}(y) + (1-\omega)V_t(y), (t = 1,2,\cdots,T)$$

Calculate the probability density function $\hat{f}_{t+1|t}(y) = \omega\hat{f}_{t|t-1}(y) + (1-\omega)v_t(y), (t = 1,2,\cdots,T)$ among $V_t(y) = H(\frac{y-y_t}{h}) - \hat{F}_{t|t-1}(y)$, $v_t(y) = K(\frac{y-y_t}{h}) - \hat{f}_{t|t-1}(y)$.

3.2.3. Time-varying Factor Weighted Density Function Prediction Model (F-TVF Model)

Compared with the linear model, the nonparametric Time-varying density function estimates the proper distribution of variables using self-data, which solves missing detection. However, the stock data does not contain all the sentiment mentioned above index and macroeconomic variables. Therefore, in this study, we can consider our data and add influencing factors to density estimation. If x is the influencing factor of y, the influencing factor will be added to the weight of the above model.

$$\widetilde{\omega} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

The model needs to satisfy $0 \le \omega < 1$ .The weight of that nonparametric density function becomes $w_{t,i} = (1\text{-}\widetilde{\omega})\widetilde{\omega}^i$

Its iterative form is as follows: $\hat{f}_{t+1|t}(y) = f_{t|t-1}(y) + (1-\widetilde{\omega})v_t(y), (t = 1,2,\cdots,T)$

And the iterative form of its cumulative distribution function is

$$\hat{F}_t(y) = \widetilde{\omega}F_{t-1}(y) + (1-\widetilde{\omega})\sum_{i=1}^{T} H(\frac{y-y_i}{h}), (t = 1,2,\cdots,T)$$

In the above two formulas, we call it the Time-varying factor weighted density function prediction model, abbreviated as the F-TVF model.

3.3. The Empirical Analysis of Stock Yield Direction Prediction

We created a crawling program through Python to collect news on Sina Finance website. According to its release time, 15:00 is taken as the dividing line, news released before 15:00 is included in the current financial news, news released after 15:00 is included in the next issue, and information released on holidays is postponed and included in the next trading day. Taking Shanghai Composite Index as the research object,the sample period is the daily data from October 1,2019,to October 1,2020,with a total of 271 trading days and a total of 60,498 news data,among which the news text of 2019 is used to construct the financial news emotion dictionary.For the daily yield series of the Shanghai Composite Index, this paper adopts

$y_t = 100 \cdot (P_t / P_{t-1})$ formula calculation y_t :the daily yield, P_t t for Closing price. Firstly, descriptive statistics and normality tests are carried out on the Shanghai Composite Index, converted into the rate of return, as shown in Table 1.

Table 1. Descriptive Statistical Analysis and Normality Test of Shanghai Stock Exchange Index Return Rate

	Mean	Variance	Skewness	Kurtosis	S-W	K-S	ADF
SYL	0.05	0.298	-2.87	1.31	0.000	0.000	0.012

The variance of SYL of the Shanghai Stock Exchange Index is 0. 298.Meanwhile, the S-W and K-S tests are used to test the normality of the Syl series of the Shanghai Stock Exchange Index, and it is found that the P values are all less than 0.05. In addition, the ADF test shows that the yield series of the Shanghai Composite Index is stationary.

To explain the correlation between emotional index and stock return sequence, the model is established as follows.

$$y_t = \alpha + \sum_{i=1}^{n_1} \beta_i SI_{t-1} + \sum_{i=1}^{n_1} \gamma_i y_{t-i}$$

Among y_t the yield of Shanghai Composite Index SI_{t-i} indicates lag i Emotional index of the period, y_{t-i} shows a larg i e yield of the period. A standard is used to determine the optimal lag order. In order to avoid pseudo-regression, the sentiment index was tested by the ADF unit root test. The results show that the first-order difference of sentiment index is stable, and the sentiment index of the first-order difference is used as the input variable. The regression results are shown in the table. The results show that the sentiment index is significant and positive at 5%, indicating that the sentiment index is positively correlated with the rate of return. Therefore, when the sentiment index of financial news is high, it may promote investors' positive investment sentiment, leading to the rising stock price.

Table 2. Regression result of sentiment index to stock return prediction

	Cons	y_{t-1}	y_{t-2}	SI_t	SI_{t-1}	SI_{t-2}
y_t	0.021	-0.031	0.035	0.968***	0.601**	0.412
t	1.202	-0.624	0.589	3.827	1.698	1.241

To compare with the linear model, this paper uses three forecasting models: Logistic model, TVF model, and F-TVF model to predict the rising and falling direction of the Shanghai Stock Exchange Index. This paper evaluates in two ways: accuracy and scoring rules. ADF unit root test is performed on the impact index factor to ensure its stability.Out-of-sample prediction results are shown in the table.

Table 3. Out of sample prediction results

	L	Logistic	TVF	F-TVF
PanelA	0.300	0.639	0.593	0.683
	0.350	0.572	0.589	0.636
	0.400	0.572	0.553	0.572
	0.450	0.556	0.572	0.569
	0.500	0.557	0.561	0.545
Pamela	Logarithmic score	-0.752	-0.729	-0.721
	Secondary scoring	0.461	0.468	0.482

According to the table, from the accuracy analysis, we can draw the following conclusions:

(1) the prediction accuracy of 1) the Logistic model ranges from 55.6%to 63.9%, and that of the model ranges from 55.3%to 59.3%. The prediction accuracy of the F-TVF model is 54.5%to 68.6%. The accuracy of all models is more significant than 0.5, which indicates that these models have predictive ability. For each model, the maximum accuracy appears near the threshold of 0.3-0.4. (2) after improving the TVF model and adding sentiment index, the accuracy of the F-TVF model is higher than that of the TVF model at every threshold level. The accuracy of the F-TVF model is higher when the threshold changes from 0.3 to 0.5. The accuracy of the F-TVF model is greater than that of the logistic model, which indicates that the sentiment index significantly improves the model's prediction ability. (3) From the analysis of scoring rules, it can be seen that whether it is a logarithmic representation or quadratic representation. The model's score is the lowest in the logic model, while the score of the F-TVF model is the highest. This shows that the sentiment index model based on F-TVF is superior to logistic and TVF models in terms of probability. In a word, financial news has a certain influence on stock returns. By adding financial news sentiment index,the prediction accuracy of the model can be significantly improved.With Logistic and TVF models, the Logistic model has higher precision than the TVF model,and the TVF model has a higher score than the Logistic model. After adding an emotional index, the accuracy and score of the prediction model are improved. This is because of the importance of the sentiment index in predicting the direction of stock returns; the model established in this paper can achieve higher accuracy and higher scores. To draw the above conclusion more clearly, different models are compared horizontally, and the prediction order of each model is listed in this table.

Table 4. Order of out-of-sample prediction results

	L	Logistic	TVF	F-TVF
PanelA	0.300	1	2	1
	0.350	1	2	3
	0.400	2	3	1
	0.450	3	1	2
	0.500	3	2	1
Pamela	Logarithmic score	3	2	1
	Secondary scoring	3	2	1

3.4. Simulation Results of Trading Strategy

Based on the forecasting model, the forecasting results are simulated and traded. The annualized rate of return and the Sharp rate are used as evaluation indicators to evaluate the model's forecasting ability from the economic point of view.

Table 5. Trading strategy simulation results

	Logistic		TVF		F-TVF	
L	SR	RC(%)	SR	RC(%)	SR	RC(%)
0.300	0.062	12.313	0.061	12.418	0.113	22.718
0.350	0.024	4.675	0.052	7.984	0.121	15.893
0.400	0.011	3.976	0.043	6.789	0.124	16.912
0.450	0.013	3.452	0.072	9.563	0.097	14.511
0.500	0.003	2.115	0.039	4.797	0.088	15.689

Note: SR is Sharp ratio; RC is the annualized rate of return

It can be seen from the table that under different thresholds l, the sharp rate of the Logistic model fluctuates from 0.003 to 0.062, and the annualized income fluctuates from 2.115%to 12.313%; The Sharp rate of the TVF model fluctuates from 0.039 to 0.072, and the annualized income fluctuates from 4.797%to 12.418%. The sharp rate of the F-TVF model fluctuates from 0.088 to 0.113, and the annualized income fluctuates from 14.511%to 22.718%.

From the horizontal perspective, when L is the same, the trading strategy of the TVF model is better than the Logistic model, and that of the F-TVF model is better than the TVF model.

To show that the sentiment index can improve the forecasting ability of the model, this paper selects two other stock indexes and 30 stocks for empirical analysis to avoid overlap. Firstly, this paper tests the robustness of the model, replacing the Shanghai Stock Exchange Index and the Fund Index.The A-share index reflects China's financial market, and the stock price has hardly changed. Financial news also demonstrates the company's stock and stock situation in China. The fund index reflects the sweeping changes of the fund market. It can examine the impact of financial and economic news on specific industries and markets instead of reading evidence through the fund index. Robustness test results are as follows. The results from table and table are consistent with the above conclusions. In addition to the research of stock price index,this paper also tests the robustness of 100 stocks,estimates and evaluates 100 samples with the above method,and obtains the accuracy and evaluation of the three models under different L values. The stock sample period is from October 2019 to October 2020.Then,two pairs of models are tested.The test hypothesis is:TVF model is not as effective as the logistic model.The prediction effect of the F-TVF model is not as good as that of the TVF model.The forecasting effect of the F-TVF model is not as good as that of the logistic model.As shown in the stock information table of 100 stocks,the robustness test results are shown in the table.

Table 6. Forecast results of A-share index

	L	Logistic	TVF	F-TVF
PanelA	0.300	0.614	0.598	0.693
	0.350	0.569	0.583	0.642
	0.400	0.575	0.565	0.571
	0.450	0.551	0.571	0.564
	0.500	0.554	0.569	0.553
Pamela	Logarithmic score	-0.741	-0.722	-0.713
	Secondary scoring	0.464	0.472	0.491

Table 7. Forecast results of fund index

	L	Logistic	TVF	F-TVF
PanelA	0.300	0.623	0.587	0.672
	0.350	0.565	0.578	0.613
	0.400	0.569	0.557	0.568
	0.450	0.561	0.567	0.574
	0.500	0.562	0.568	0.551
Pamela	Logarithmic score	-0.762	-0.747	-0.702
	Secondary scoring	0.483	0.471	0.446

Table 8. Accuracy test results

L	Logistic	TVF	F-TVF
0.3	4.124** (0.021)	3.481** (0.021)	1.962** (0.021)
0.35	4.001** (0.000)	3.672** (0.002)	1.891** (0.000)
0.4	3.809** (0.001)	3.141** (0.000)	1.892** (0.002)
0.45	4.022** (0.000)	3.671** (0.001)	2.012** (0.000)
0.5	3.897** (0.001)	2.891** (0.000)	1.896** (0.001)

We can see that the testing effect is significant from accuracy or score from the table and table. The results show that the F-TVF model is robust under different samples. Therefore, it can be concluded that the F-TVF model is better than the logistic model and TVF model, and the F-TVF model based on emotion index has a better prediction effect than the logistic model and TVF model.

4. Conclusion and Enlightenment

In the field of finance, the prediction of stock return has always been a hot topic. As a barometer of the economy, the stock market is not always smooth, and its fluctuation characteristics are likely to cause considerable losses to investors. In China, due to the relatively weak ability of individuals to take risks, it is of great significance for investors to accurately predict the stock market. As for the direction of stock profits, investors are the most crucial point. With the rapid development of China's economy, the profitability of network media is becoming stronger and stronger. Once a major event occurs, it will spread rapidly on the network for the first time. No matter what quality investors, this news will affect investors. This paper uses a nonparametric method to predict the direction of stock returns from the perspective of financial news. In this paper, the logistic model, TVF model, and F-TVF model based on emotion index are used to estimate the model. Some conclusions are drawn. First, the sentiment index of economic newspapers is related to the yield and positively impacts the outcome. This may be the current private investor-centered Chinese investor, but individual investors are vulnerable to financial news. When the emotion of economic news is strong, it promotes personal investment and leads to the rise of stock price. Secondly, by adding an emotional index to the model, it is found that the F-TVF model based on emotion index can significantly improve the prediction accuracy and evaluation of the traditional logistic model and TVF model and improve the prediction ability of the model. Thirdly, because the model's accuracy is different under different thresholds, the evaluation criteria are the same as the evaluation criteria.

The F-TVF model based on logarithmic and quadratic evaluation rules is better than the logistic model and the model; the accuracy and evaluation of the model are optimized. This shows the effectiveness and necessity of the F-TVF model. Although they are institutional investors and individual investors, it is still impossible to obtain information on Shanghai amount on the Internet.Even if you can read and analyze every piece of information manually, the cost of time is enormous. Text drilling technology solves this problem well. In this paper, through many financial news analysis feelings, to make a prediction decision on the stock market, which is also very important for investors and academic researchers. Finally, considering the richness of the corpus,this paper constructs a dictionary with more characteristics of the financial field.

References

[1] Wu Songtao, he Jianmin, Li Shouwei. Stock risk and its contagion based on multi-attribute herding behavior [J]. Journal of Beijing University of Technology (SOCIAL SCIENCE EDITION), 2017,19 (01): 64-72.

[2] Chen guojin, Zhang Runze, Xie peilin, Zhao xiangqin. Informed trading, information uncertainty, and stock risk premium [J]. Journal of management sciences in China, 2019,22(04):53-74.

[3] Fu Kui, Liu Yujie, Chen Meili. Stock price prediction based on financial news emotional tendency value [J]. Journal of Beijing University of Posts and Telecommunications (Social Science Edition), 2019, 21(01):87-100.

[4] Ma Juan, Wang Lu, Zuo Liming. Research on Stock Price Forecast of GEM Based on Grey System and Neural Network [J]. Contemporary Financial Research, 2019(02):87-97.

[5] Zhang Xudong, Huang Yufang, Du Jiahao, Miao Yongwei. Stock price prediction based on discrete hidden Markov model [J]. Journal of the Zhejiang University of Technology, 2020,48(02):148-153 + 211.

[6] Cui wenzhe, Li boys, Yu Desheng. empirical analysis of stock price prediction based on the GARCH model and BP neural network model [J]. Journal of Tianjin regular university (natural science edition), 2019,39(05):30-34.

[7] Luo Xin, Zhang Jinlin. Stock price prediction based on multi-time scale composite deep neural network [J]. Wuhan Finance, 2020(09):32-40.

[8] Jerry Lee, Zhu Shiwei, Wei Moji, Yu Junfeng, Li Xintian. Analysis of emotional tendentiousness of news texts based on dictionaries and rules [J]. Shandong Science, 2017,30(01):115-121.

[9] Yin Hairen, Wang Panpan. Media reports, market returns, and investor sentiment based on news analysis [J]. Soft Science, 2015,29(07):136-139+144.

[10] Zhao Qingguo, Kong Xiangyue, Liu Liming, Yang Longqian. Construction of Time Series Weighted Mean Model for Short-term Stock Price Forecast [J]. Journal of Shenyang University of Aeronautics and Astronautics, 2020,37(04):81-89.