Neural Computing and Applications
https://doi.org/10.1007/s00521-020-05411-7

S.I. : HIGHER LEVEL ARTIFICIAL NEURAL NETWORK BASED INTELLIGENT SYSTEMS

Applying BERT to analyze investor sentiment in stock market

Menggang Li[1,3,4] · Wenrui Li[2,3] · Fang Wang[2,3] · Xiaojun Jia[1,4] · Guangwei Rui[2]

Received: 21 August 2020 / Accepted: 29 September 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract
This paper is an analysis of investor sentiment in the stock market based on the bidirectional encoder representations from transformers (BERT) model. First, we extracted the sentiment value from online information published by stock investor, using the Bert model. Second, these sentiment values were weighted by attention for computing the investor sentiment indicator. Finally, the relationship between investor sentiment and stock yield was analyzed through a two-step cross-sectional regression validation model. The experiments found that investor sentiment in online reviews had a significant impact on stock yield. The experiments show that the Bert model used in this paper can achieve an accuracy of 97.35% for the analysis of investor sentiment, which is better than both LSTM and SVM methods.

Keywords Bert model · Investor sentiment · Online reviews · Cross-sectional regression

1 Introduction

Many studies have shown that investor sentiment is an important factor in the price of stocks and other assets. From this perspective, changes in investor sentiment can predict market trends. Research shows that investor sentiment will be affected by news reports, and even the content, method, and frequency of news reports will affect investor sentiment. Therefore, investor sentiment can reflect investment willingness and expectations of market trends to a certain extent and has a certain ability to predict market yield, fluctuations, and transaction volume. It can be used as business operations, financial institution deposit, and loan decision-making and asset management, as well as new information channels for policy-making departments and regulatory agencies to conduct anticipated management.

However, one of the most challenging problems in time series forecasting is generally considered to be stock market forecasting, due to its noisy and volatile nature. Another unresolved issue in modern socioeconomic and social organization is how to accurately predict stock movements. A lot of relevant research has emerged as an economic science research hotspot, coupled with the attraction of investor interest and high yields. Traditional stock market forecasting methods focus on time series analysis. De Gooijer et al. [1] reviewed papers published in journals administered by the International Association of Forecasters (Journal of Forecasting 1982–1985; International Journal of Forecasting 1985–2005) and found that in these journals more than a third of the published papers focused on time series prediction. In particular, traditional time series analysis methods include autoregressive (AR), moving average (MA), autoregressive and moving average (ARMA), and autoregressive integral moving average (ARIMA) [2]. All of these methods focus on the time series itself and ignore other influences, such as background

✉ Wenrui Li
liwenrui@bjtu.edu.cn

✉ Xiaojun Jia
xjjia@bjtu.edu.cn

[1] National Academy of Economic Security, Beijing Jiaotong University, Beijing, China

[2] School of Economics and Management, Beijing Jiaotong University, Beijing, China

[3] Beijing Laboratory of National Economic Security Early-warning Engineering, Beijing Jiaotong University, Beijing, China

[4] Beijing Center for Industrial Security and Development Research, Beijing Jiaotong University, Beijing, China

Published online: 17 October 2020

information. Specifically, they assume that the preceding and following data are independent and dependent variables, respectively, and aim to obtain a quantitative relationship between them. In addition, these methods often require a number of assumptions and preconceptions, such as the underlying data distribution, the valid ranges of the various parameters and their linkages [3]. However, as a complex system with many influences and uncertainties, the stock market tends to exhibit strong nonlinear characteristics, which makes traditional analytical methods ineffective. Moreover, the amount of information processed for stock market modeling and forecasting is often very large, posing a significant challenge to algorithm design. These characteristics make predictions of stock markets based on traditional methods inadequate [4].

Considering the limitations of the sample and technical aspects, this paper analyzes the sentiment indicators of investor reviews by Bert model to obtain the sentiment tendency of investors, its impact on stock market volatility (mainly yield), and verifies the relationship between investor sentiment and stock yield using two-step cross-sectional regression. To determine the impact of investor sentiment on stock yield, with a view to helping investors capture the sentiment in social media comments and take investment-friendly decisions, as well as informing government regulatory decisions. Experiments have shown that the Bert method can be used to analyze investor sentiment indicators with an accuracy of 96.9%.

2 Literature review

In recent years, with the development of the Internet, big data, and other technologies, people have more diverse access to information and the amount of data is larger, especially unstructured financial "data", through these data, portraying investor sentiment, then studying the impact of investor sentiment on financial markets and asset prices, and thereby understanding the mechanism of market sentiment. It has become one of the hotspots of research in recent years.

2.1 Definition of investor sentiment

On the classic case of the existence test of investor sentiment, close-end fund discount research, Zweig [5] first thought that the change of fund discount was the result of investors' expectations. Lee et al. [6] believe that the change in the discount of the fund reflects the change in the mood of individual microinvestors. The empirical results of Swaminathan [7] show that the change in the discount of closed-end funds can predict the stock returns of small companies better than that of large companies. In the currency market, Alexander Kurov [8] found that monetary policy has a significant impact on investor sentiment, and it mainly depends on the current market conditions. In a bear market, a market that is sensitive to changes in investor sentiment and credit market conditions, monetary policy has a greater impact on it.

2.2 Sentiment measurement and tapping

Most methods of sentiment analysis focus on methods for adapting sentiment resources (e.g., dictionaries) from resource-rich languages (usually English) to other languages with few sentiment resources. For example, Mihalcea produced a subject vocabulary for Romanian by translating an existing English subject vocabulary [9]. They then used the dictionary to create a rule-based classifier of sentence-level subjectivity (e.g., Riloff and Wiebe [10]) that can determine whether sentences in Romanian are subjective or objective.

However, much of the research on multilingual sentiment and subjectivity analysis has focused on building resources to support supervised learning techniques for the desired target language—techniques that require training data to be annotated with appropriate sentiment labels (e.g., document-level or sentence-level positive or negative polarity) [11]. Such data are difficult and costly to obtain and must be obtained separately for each of the languages under consideration. For example, Mihalcea also studied the creation of (sentence-level) subjective annotated Romanian corpora, manually translating a corpus from English and (automatically) projecting the subjective class labels of each English sentence into its Romanian counterpart. With this corpus, they use standard supervised learning methods to obtain a classifier directly from the Romanian text. Their experiments found that the parallel corpus approach worked better than their dictionary translation method described above.

In earlier work, Kim and Hovy [12] conducted a similar study for German and English: They manually translated the target utterance (German or English) into a second language (English or German, respectively) and used an existing affective lexicon in the source language to determine the affective polarity of the target utterance.

More recently, automated machine translation engines have also been employed to obtain the necessary subjective or sentimentally tagged corpus. For the task of classifying sentence-level subjectivity, Banea does so. For example, Banea [13] studied the translation of an English corpus into five different languages, mapping sentence-level labels to translated text. They found that the method worked consistently well regardless of the target language.

Approaches that do not explicitly involve resource adaptation include Wan [14], which usesco-training with

English vs. Chinese features comprising the two independent—views‖ to exploit unlabeled Chinese data and a labeled English corpus and thereby improves Chinese sentiment classification. Another notable approach is the work of Boyd-Graber and Resnik [15], who proposed a generative model—supervising multilingual latent Dirichlet assignments—to jointly model topics that are consistent across languages and adopt them to better predict affective ratings [12].

However, in recent years, data on sentiment labels in languages other than English are emerging. And the existing monolingual (including English) sentiment classifiers still have much room for improvement, especially at the sentence level [13]. With this in mind, Lu [16] addressed the task of bilingual sentiment analysis: They assumed a certain amount of sentiment label data for each of the studied language pairs and aimed to improve the sentiment classification of both languages simultaneously [14]. Considering label data in each language, they developed a method that exploits a label-free parallel corpus and the intuition that two parallel sentences or documents (i.e., translations of each other) should exhibit the same sentiment—their sentimental labels (e.g., polarity, subjectivity) should be similar. Their solution is a maximum entropy-based EM approach that jointly learns two monolingual sentiment classifiers by treating sentiment labels in unlabeled parallel text as unobserved latent variables, maximizing the regularized joint likelihood of language-specific label data and inferred sentiment labels in parallel text.

2.3 Investor sentiment and stock market

2.3.1 Measurement of investor sentiment

With regard to the measurement of investor sentiment, it can be divided into direct and indirect indicators, depending on the indicator calculation and data source. Direct indicators are constructed to obtain subjective sentimental attitudes of investors through questionnaires or other data sources. Indirect indicators take corresponding objective indicators and data already available in the financial markets and use them as proxy variables for investor sentiment.

In the study, many scholars use questionnaires to construct sentiment indicators. In 2004, Meijin Wang uses the "CCTV watch" data to construct investor sentiment data based on the number of bullish and bearish calls [17]. Based on the number of bullish people, construct investor sentiment data. Robert B Barsky and Eric R. Sims [18] constructed market confidence indicators based on statistical indicators of the proportion of bullish and bearish people, and examined the 12-month and 5-year periods in the time span. In 2012, Barsky and Eric R. Sims is also based on a statistical indicator of the proportion of bullish and bearish calls and examines 12-month as well as 5-year time horizons over the time horizon. In addition to the questionnaire method, the construction of sentiment indicators using unstructured data through text analysis methods such as natural language processing is now becoming a hot topic. Text analysis methods, in turn, can be divided into two categories: sentimental dictionary and machine learning. The lexicon of sentiments is simpler and the machine learning is one of the future directions, which will be increasingly used. Machine learning methods are used to analyze a large number of unstructured text sources and categorize documents by content to obtain useful knowledge and information by combining techniques such as word processing and natural language recognition. The specific methods are: plain Bayesian method, VSM model, LSA model, LSTM model, and so on [19].

Indirect indicators are generally the collation and further analysis of objective data from the securities market. Poniiff et al. found that closed-end fund discounts can be used as a proxy for changes in investor sentiment in 1995. Baker and Wurgler in 2006 construct a BW sentiment index to describe investor sentiment by principal component analysis using six objective market indicators such as turnover rate, stock market yield, number of open accounts, and closed-end fund discount. In 2009, Barber et al. advocate that volume can portray market liquidity, which in turn can be used as a proxy variable to reflect investor sentiment. Zhang Zongxin and Wang Hailiang used principal component analysis to develop an investor sentiment index for China in 2013. The study by Qi Luo and Bill Zhang draws on the momentum effect method as a proxy indicator in 2013. Next year, Liu et al. used monthly new account openings as a proxy for sentiment. In 2014 Chun Wang, on the other hand, uses quarterly net inflows data from open-end equity funds to study investor sentiment [20].

2.3.2 Investor sentiment and stock market returns

In 2007, Tetlock's study of newspaper column text sentiment shows that negative sentiment can be an effective predictor of future changes in the Dow Jones and stock market trading volume, especially over a two-day period. Jiangetal constructs text sentiment about managers using an LM dictionary on data from public company earnings reports and conference calls index in 2016 and empirically found that the sentiment index is significantly and negatively correlated with future stock market yield. In 2018, Meng et al. used the sentiment dictionary to construct a network sentiment index and used the ARMA–GARCH model and Granger causality test to study the interaction between network sentiment and stock yield. It was found that most stock yield can be influenced by network

sentiment in the short run [21]. In 2018, Dongliang Yuan improves the Fama–French three-factor model and investigates the relationship between investor sentiment and stock yield based on the text of microblogs. It is found that Weibo sentiment is positively correlated with stock yield; moreover, the scale effect of China's stock market shows significant performance under pessimistic sentiment. Next year, Yuhui Jiang constructs a regression model of sentiment and SSE returns based on Weibo text to construct sentiment indicators and control variables, such as total market capitalization and lagged one-period returns of the SSE Composite Index; the study finds that sentiment is significantly and positively correlated with SSE yields, especially in an upturned market, where changes in sentiment have a greater impact on yield [22].

2.3.3 Investors and stock market risk

Downes Circle et al. use the number of posts in the Eastern Wealth stock bar as an indication of investor sentiment as individual investor concern in 2018. Through multiple regressions [23], the results show that the higher the number of posts, the higher the risk of future stock price collapse. In addition, when analyst attention increases, the risk of stock price collapse decreases. In 2019, Bing Shen and Xijuan Chen use the mispricing of Tobin's Q-value as an indicator of investor sentiment to study the mechanism of the impact of stock price collapse on A-share listed companies under the condition of controlling shareholder pledge [24]. The results show that sentiment is positively correlated with the risk of a stock price crash, and that the risk of a "flash crash" will be further exacerbated in the case of an equity pledge. Moreover, at the same year Roo-Wei Zhao et al. draw on BW index to construct a sentiment index through principal component analysis and find that investor sentiment exacerbates the risk of share price collapse for mega-cap stocks in an empirical study of Chinese A-shares of different market capitalization sizes [25], while the effect is not significant for stocks of other market capitalization sizes. In 2019, Luo Peng et al. construct a risk prediction model using a Baidu search index and separate statistics for macro [26], banking, nonbanking, and Internet finance types and find that when the frequency of public searches for relevant risk keywords increases, the capital market will be more likely to engage in behaviors such as "catch-up." At the same year, Mengyu Li and Zhihui Li [24] found that market manipulation can affect investor sentiment by constructing a model of market manipulation [27]. Market manipulation leads to optimistic investor sentiment accompanied by an irrational increase in stock price, while the end of the manipulation and the decline in investor sentiment create the risk of stock price collapse [28]. It is also found that this phenomenon is more pronounced in small and mid-cap stocks.

3 Overview

This paper consists of two main modules as shown in Fig. 1, a multilevel sentiment propensity analysis of investors based on the Bert model and a cross-sectional regression analysis model.

Investor multilevel sentiment propensity analysis is the use of the Bert model for sentiment analysis of investor online reviews, the analysis of the resulting investor sentiment into three broad categories—positive, negative, and none, and then refined into eight subcategories—like, happiness, surprise, anger, fear, disgust, sadness, and none.

Cross-sectional regression analysis is the use of cross-sectional one-dimensional regression model to validate the relationship between investor sentiment and stock yield, and the analysis finds that investor sentiment has a more significant influence on stock yield.

In securities investment projects with complex characteristics, the use of the model can provide a scientific basis for securities investment decisions in order to reduce subjectivity and blindness in the securities investment decision-making process.

4 Multilevel sentiment analysis of investor reviews

Through the analysis of the research on investor sentiment in the stock market, it is found that most of research focuses on the single-level investor sentiment analysis of the stock yield in the market. Few experts have studied multilevel sentiment analysis of investors in detail. This paper builds on existing research by breaking down into multiple aspects of multidimensional investor sentiment on stock yield.

4.1 Bert model

4.1.1 Bert-related theory

Bidirectional encoder representations from Transformers, or the Bert model, are a pre-trained model that uses the encoder part of the Transformer as the basis of the model, unlike convolutional and recurrent neural networks. Because of the power of the Transformer encoder, Bert models can be added to very deep depths, fully exploiting the properties of deep neural network models and improving model accuracy. The Bert model uses a multi-headed attention mechanism by performing multiple linear transformations of the input vector $X_a \in R_k$ to obtain different linear values and then inputting them into the attention block to compute the attention weights, as shown

in the equation, where h denotes the number of multiple heads in the attention mechanism. The final output of the multi-headed self-attentive mechanism is obtained by combining the output values of the multi-headed attention mechanism and performing another linear transformation. The multi-headed attention mechanism is calculated as follows.

$$\mathrm{Att}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$V_l = \mathrm{Linear}(W_l\,\mathrm{concat}(\mathrm{Att}_1, \mathrm{Att}_2, \ldots, \mathrm{Att}_h) + b_l) \tag{2}$$

where Q, K, V are the word-embedding representations of the text, and $X_a = Q = K = V$. The similarity between words is calculated by point multiplication, then multiplied by a scaling factor $1/\sqrt{d_k}$ to prevent the point multiplication result from being too large to affect the back propagation of the gradient, and finally the softmax computational weight probability is multiplied by V to get the attentional output V_l. The transformer structure as shown in Fig. 2.

After obtaining V_l for the multiple attention mechanism, a new $V_a = V_l + X_a$ is formed by the residual structure. And then, the normalized V_a is output to the forward propagation network, and the output value of a Transformer is obtained again by a single residual structure.

$$V_t = \mathrm{Feed}(W_f V_a + b_f) + V_a \tag{3}$$

where Feed is a linear function. The output value of a Transformer is represented by V_t, and all the calculation processes in the Transformer are represented by Trans. That is, for any vector X_a input to the Transformer, denoted by the computational expression. Bert is composed of multiple Transformers superimposed on each other, and the

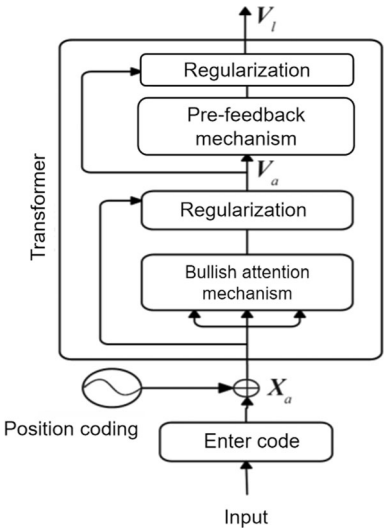

Fig. 2 Transformer structure

output value of Bert in terms of V_b. Bert denotes the computational process in Bert. For any vector, X_a of input Bert is calculated as follows.

$$V_t = \mathrm{Trans}(W_t X_a + b_t) \tag{4}$$

$$V_b = \mathrm{Bert}(W_b X_b + b_b) \tag{5}$$

4.1.2 Input processing of Bert model

In addition to adding position embedding to the token embedding like Transformer, the Bert model also adds a segment embedding to handle problems involving sentence pairs, as shown in Fig. 3.

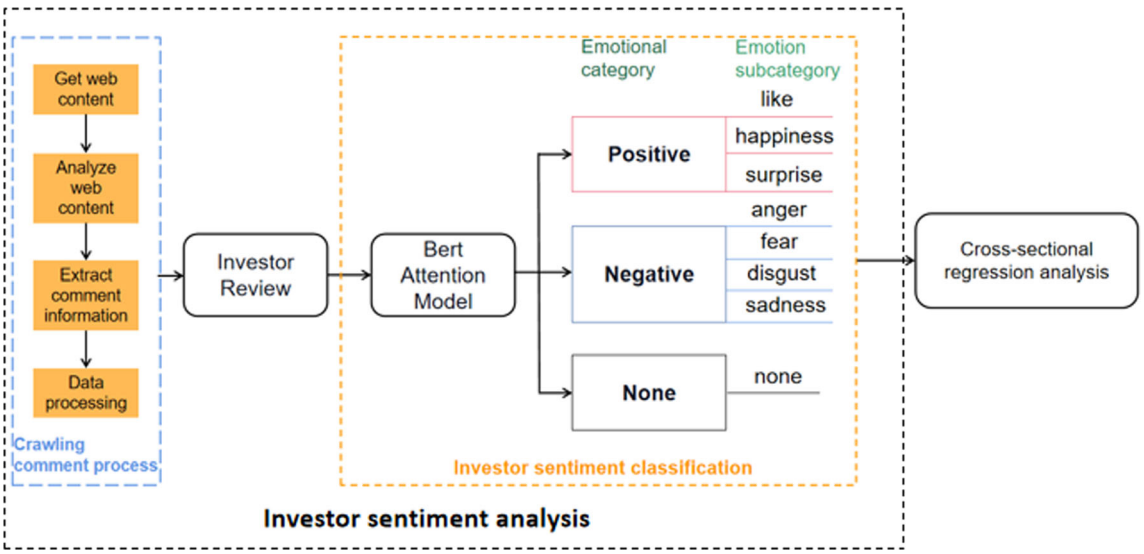

Fig. 1 Overview

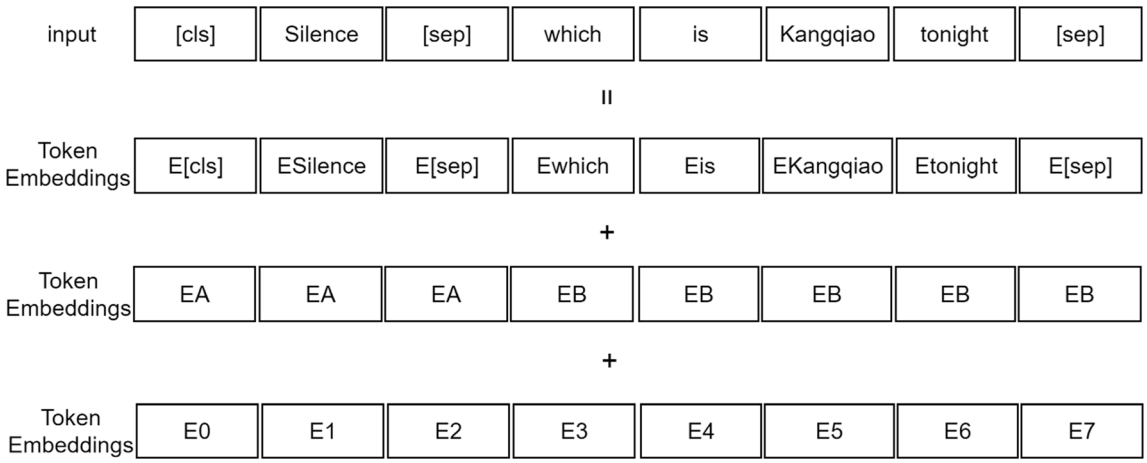

Fig. 3 Bert model input processing

4.1.3 Masked language model

The masked language model is a new prediction task used in the Bert model. Unlike the traditional prediction task for word prediction, this prediction task learns expression word vectors in a process more consistent with human language conventions by fully masking some of the words in the input text, allowing contextual information to be fully considered when predicting the masked words. In the masked language model, the following two steps are performed in order to obtain better training results.

The first step is the selection of the masking area. In Bert, the model randomly selects 15% of the words as masking areas. This ensures the validity of the training. The second step is the choice of the masking method. In Bert, in order to solve the problem that the masked words will not be perceived by the downstream task, after the masked language model has completely masked the words, resulting in inconsistent text between the preprocessing and fine-tuning phases of the model, the following three masking methods will be chosen for the masked words during the training process.

1. Masked words are replaced by special notation [mask] 80% of the time during training, as shown in Fig. 4.
2. And 10% of the time in training is replaced by a random word from the dictionary, as shown in Fig. 5.
3. The masking word will remain the same for the remaining 10% of the time, as shown in Fig. 6.

4.2 Sentiment analysis with Bert model

A Bert model differs from traditional classification models such as convolutional or cyclic models. The model will be pre-trained in a large corpus environment before classification. The pre-trained Bert model can be used directly for classification, but is less accurate and requires further pre-training on a task-specific dataset, at which point the model is retrained on the target data again using the target data from the original pre-training, such as a masking language model or next sentence prediction task, so that its own parameters apply to the target data. A transfer learning session was conducted in the target area. Finally, the fine-tuning of the classification effect is performed on the two pre-trained models that means fine-tuning the training and outputting classification results. The Bert model classification process is shown in Fig. 7.

Thanks to the excellent word vector expression and model parameters after the pre-training phase, the model can easily capture the deep abstract features of the statements, so the fine-tuning phase of the model is mostly just training the parameters of the output layer and slightly adjusting the pre-trained model, which ensures the

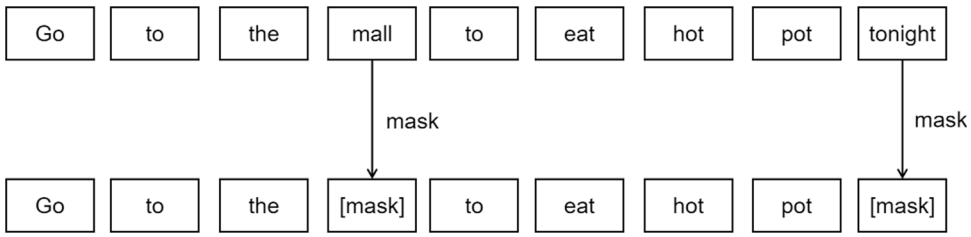

Fig. 4 Masking of the Bert model

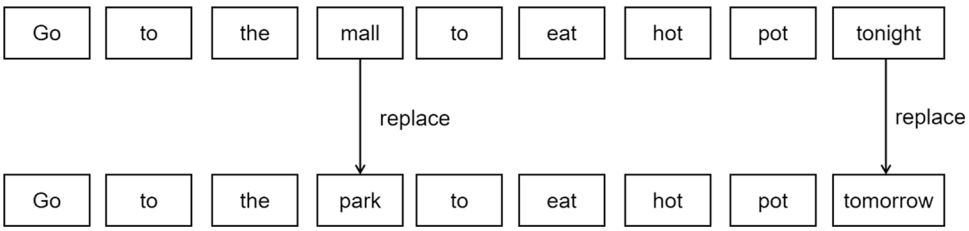

Fig. 5 Replacement of the Bert model

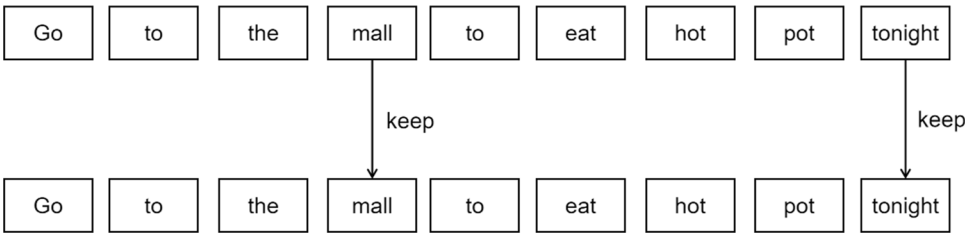

Fig. 6 Invariance of the Bert model

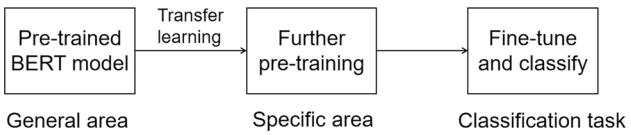

Fig. 7 Bert model training process

convergence speed and classification accuracy of the fine-tuning phase.

The sentiments that people often convey are multilayered, in terms of broad categories, are generally divided into three types of sentiments: positive, negative, and none. In order to describe sentiments in more detail, the sentiments are further divided into eight subcategories: like, happiness, surprise, anger, fear, disgust, sadness, and none. However, the existing research is devoted to the fine-grained sentiment classification problem of identifying small categories of sentiments, usually treats subcategory sentiment analysis directly as a multicategorization problem, with the categories considered to be equally independent of each other. However, the various subcategories of positive sentiments usually have similar sentimental content. At the same time, the subcategories of negative sentiments usually have similar sentimental content. The classification process is illustrated in Fig. 8.

In the classification process, the broad categories of sentiments are first calculated, as shown in the following equation.

$$p(c = \text{CLASSES}|T_n) = \frac{\exp(w_c Y + b)}{\sum_c \exp(w_c Y + b)} \tag{6}$$

where $p(c = \text{CLASSES}|T_n)$ denotes the probability that the nth text category is a sentiment of CLASSES. The category of sentiment is used to calculate the category loss, as shown in following equation.

$$L_C = -\sum\nolimits_C y_n \log(p(c = \text{Classes}|T_n)) \tag{7}$$

Further, calculate the classification probability of the corresponding subclass as shown in the following equation.

$$p(c = \text{classes}|T_n) = \frac{\exp(w_c Y + b)}{\sum_c \exp(w_c Y + b)} \tag{8}$$

where $p(c = \text{classes}|T_n)$ denotes the probability that the nth text belongs to the classes under the CLASSES major class sentiment. Next, the categorical loss is calculated using the subcategory sentiment results, as shown in the following equation.

$$L_{\text{classes}} = -\sum\nolimits_{\text{classes}} y_n \log(p(c = \text{classes}|T_n)) \tag{9}$$

Finally, a combination of categories and subcategories losses is obtained for the final multilevel affective classification loss function, as shown in the following equation.

$$L_{\text{MC}} = \alpha L_{\text{CLASSES}} + (1 - \alpha) L_{\text{classes}} \tag{10}$$

where α is the weighting factor used to balance the major and minor categories of losses. By optimizing multiple levels of sentiment loss, it is possible to perform fine-grained classification of subcategories of sentiment while integrating information from categories of sentiment. This will improve the ability of the sentiment classification network to distinguish between different broad categories and enable more accurate sentiment analysis of text as shown in Table 1.

5 Cross-sectional regression validation

5.1 Cross-sectional data

Cross-sectional data are columns of data consisting of the same statistical indicators in different statistical units at the same time. The difference with time series data is that the

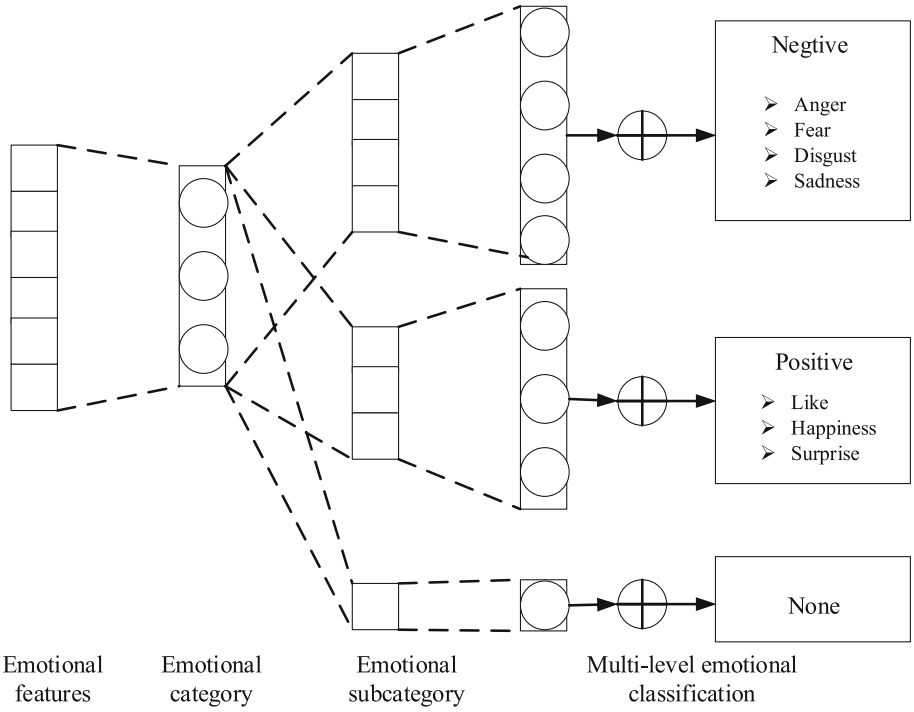

Fig. 8 Investor reviews multilevel sentimental categories

data are arranged in a different standard, as the time series data are arranged in chronological order and the cross-sectional data are arranged by statistical units. Thus, cross-sectional data do not require that the subject and its extent be the same, but that the time of the statistics be the same. This means that the data must be of the same cross section at the same time. In exactly the same way as the time data, the statistical caliber and calculation methods (including the calculation of value quantities) of the cross-sectional data should be comparable.

In analyzing the cross-sectional data, two main issues should be addressed. One is the problem of heteroskedasticity, which inherently varies across individuals or geographic areas because the data are collected from a sample of individuals or geographic areas at a given time. The second is the consistency of the data, which mainly includes whether the sample size of the variables is consistent, whether the sampling period of the samples is consistent, and whether the statistical standards of the data are consistent (Table 1).

As shown in Table 2, there are two cross-sectional datasets, these data columns consisting of the same statistical indicators for different statistical units at the same time.

5.2 Two-step cross-sectional regression validation

In 1973, Fama and MacBeth [29] proposed the Fama–MacBeth Regression in order to test the CAPM. Fama–MacBeth is a two-step cross-sectional regression test method. It very cleverly excludes the effect of residual correlation in cross section on standard errors and is widely used in the industry.

The first step in a Fama–MacBeth regression is to obtain the exposure of individual stock yield to the factor through a time series regression. In general, assuming the factor is not portfolio yield, the first step is to perform a time series regression to determine the β_i:

$$R_{it} = a_i + \beta_i' f_t + \varepsilon_{it}, \quad t = 1, 2, \ldots, T, \forall i \tag{11}$$

In the second step of the cross-sectional regression, Fama-MachBeth performed a cross-sectional regression at each time t. This is the biggest difference between Fama-MachBeth and the cross-sectional regression:

$$R_{it} = a_i + \beta_i' f_t + \varepsilon_{it}, \quad t = 1, 2, \ldots, T, \forall i \tag{12}$$

where the regressors on the right-hand side of the regression equation are the factor yield f_t and the variable on the left-hand side is R_{it}. The coefficients obtained from the regression are the exposure of individual stock i to the factor β_i, the intercept a_i, and the random residuals ε_{it}. In a general cross-sectional regression, we first take the mean value of $R_{it}, t = 1, 2, \ldots, T$ on the time series to obtain the average return of an individual stock $E[R_i]$. Regression was then done on the cross section using $E[R_i]$ and $R_{it}, t = 1, 2, \ldots, T$, so only one cross-sectional regression was done. The Fama–MacBeth cross-sectional regression does an independent cross-sectional regression on each t (if there is $T = 500$ periods, this means 500 cross-sectional

Table 1 Example of investor sentiment analysis

Content of reviews	Sentimental classification	Polarities	Readings
I ate two boards in a row, my heart hurts, I can't even eat well, can anyone tell me when I can get my money back?	Fear	− 1	50
I was once told by a big financial expert that the current stock market is such that if you try to get out of the car to take a piss, you'll be far away from the market	Disgust	− 1	68
I've been going through a lot of bad times lately, and today I finally have something that makes me happy, haha!	Happiness	1	4
Seeing that there is still a bias in people's interpretations of the group	None	0	12
Can't find a reason to go down, so start buying	Like	1	7

Table 2 Examples of cross-sectional data

	SSE (%)	SZSE (%)	GEM (%)	Roger's rate of return (%)
2017	6.56	8.47	− 10.68	22.67
2018	− 24.59	− 34.42	− 28.65	− 12.83

regressions) and then takes the mean of the parameters from these t cross-sectional regressions as the estimate of the regression.

$$\widehat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\lambda}_t \tag{13}$$

$$\widehat{\alpha_i} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\alpha_{it}} \tag{14}$$

The ingenuity of the above method is that it treats the regression results of the T period as T-independent samples. The standard errors parameter depicts how the sample statistics vary across samples. In traditional cross-sectional regressions, we perform only one regression to obtain one sample estimate of the sum. In traditional cross-sectional regressions, we run only one regression to obtain one sample estimate of λ and α. In the Fama–MacBeth cross-sectional regression, we treat the T-period sample points independently and obtain T estimates of λ and α samples. The standard errors for λ and α can then be easily and correctly determined.

$$\sigma^2(\widehat{\lambda}) = \frac{1}{T^2}\sum_{t=1}^{T}(\widehat{\lambda_t} - \widehat{\lambda})^2 \tag{15}$$

$$\sigma^2(\widehat{\alpha_i}) = \frac{1}{T^2}\sum_{t=1}^{T}(\widehat{\alpha_{it}} - \widehat{\alpha_i})^2 \tag{16}$$

In addition, with T α estimates, it is easy to derive the covariance matrix of residuals and thereby test whether the mispricing of individual stocks is jointly zero. From the above description, it is easy to see that the Fama–MacBeth cross-sectional regression differs from the traditional cross-sectional regression in that.

The Fama–MacBeth cross-sectional regression is done with R_{it} and β_i on different t's, and the results of the regression λ_t and α_{it} are then averaged over the time series to obtain $\lambda = E[\lambda_t]$ and $\alpha i = E[\alpha_{it}]$. Conventional cross-sectional regressions are performed by taking the mean of R_{it} in the time series to obtain $E[R_i]$ and then performing another cross-sectional regression to obtain λ and α directly. In short, Fama-MacBeth regression first performs regression and then averaging, and traditional cross-sectional regression performs averaging first and then regression. When the regressor in the cross-sectional regression, that is, β_i, is constant over all T periods, Fama–MacBeth has a significant advantage in dealing with the cross-sectional correlation of the residuals. In Fama and MacBeth (1973), a rolling window is used in solving β_i by time series regression. If we estimate β_i at once using all the sample data, then they take the same value for all t periods.

As seen above, the major advantage of the Fama–MacBeth regression is that it eliminates the effect of residual cross-sectional correlation on the standard error. The residual yield of stocks is highly cross-sectionally correlated, so this correction is critical for accurate calculation of standard errors, but it also has shortcomings.

First, the Fama–MacBeth regression does not help with the temporal correlation of residuals. In 2009, Petersen analyzed different regression techniques for panel data that ignored the temporal or cross-sectional correlation of residuals, leading to inaccurate standard errors (underestimating their true values). Second, as mentioned above, the β_i used in cross-sectional regressions is not known, but is an estimate from time series (generated regressors) and therefore subject to error. The Fama–MacBeth regression

cannot help with this and requires Shanken correction, but now we have a powerful tool like GMM that can easily handle the various correlations of the residuals. But do not forget that the Fama–MacBeth regression was developed almost 10 years before GMM. In the absence of GMM or other advanced methods, the Fama–MacBeth regression has been widely recognized by the academic community and has a profound influence on the development of cross-sectional regressions by cleverly eliminating the influence of residual cross-sectional correlation.

In general, the Fama–MacBeth regression is a cross-sectional regression where the factors can be portfolio yield or other indicators. Like a normal cross-sectional regression, the first step is to obtain the exposure of the investment to the factor β_i through a time series regression. Once β_i is obtained, a cross-sectional regression is run using individual yield R_{it} and β_i on a cross-sectional basis for each period (T periods in total), yielding a factor λ_t and an individual residual α_{it} for that period. After obtaining T estimates from T cross-sectional regressions, they are averaged to obtain the mean factor yield $\lambda = E[\lambda_t]$ and the mean individual residuals $\alpha_i = E[\alpha_{it}]$. The Fama–MacBeth regression rules out the effect of residual cross-sectional correlation on standard errors, but does not help with temporal correlation.

This paper uses a two-step cross-sectional regression to validate the relationship between investor sentiment and stock yield, thus demonstrating that investor sentiment has a very significant effect on stock yield.

The first is the choice and definition of variables. We subdivide the investor sentiment in the review information into eight indicators of like, happiness, surprise, anger, fear, disgust, sadness, and none to study the stock yield and then get the empirical results of the impact of investor sentiment on the stock yield of the company based on the online reviews. In this paper, the monthly stock yield from September 2018 to July 2020 is used as the experimental sample, and the data are obtained from the Guotaian database; the sentiment analysis of the Oriental Fortune online reviews from September 27, 2018, to July 20, 2020. See Table 3 for the interpretation of the corresponding variables (Tables 4, 5).

Next cross-sectional regression analysis is carried out, according to the findings of existing studies found that: investor sentiment has some explanatory power for stock market yield, in order to analyze the relationship between the two in depth and detail, this paper studies the relationship between investor sentiment and company stock yield, and the paper analyzes the impact of investor sentiment on stock yield in China based on the Oriental Wealth online reviews. In this section, the investor sentiment indicators based on the Eastern Wealth online reviews, which are selected as the explanatory variables, and the yields are used as the explained variables to construct the cross-sectional one-dimensional regression model. The cross-sectional regression analysis model takes into account the differences between different stocks and is a regression analysis of data consisting of the same statistical indicators in different statistical units at the same time, and the cross-sectional one-dimensional regression analysis focuses on the degree of explanation of the individual explanatory variables for the explained variables. Through research and combing of the existing literature and the analysis of the relevant theories, this section selects eight indicators as the explanatory variables, namely like, happiness, surprise, anger, fear, disgust, sadness, and none, and selects the stock yield from the Guotaian database as the explanatory variable to construct the monthly investor sentiment indicator-stock yield cross-sectional one-dimensional regression model, as follows.

$$P_t = \alpha + \beta \times Q_m + \varepsilon. \tag{17}$$

6 Experiments and results

6.1 Data sources

In this paper, we choose to obtain stock reviews in the Eastern Stock Exchange, as shown in the figure, with a total of 30 W articles, covering the period from September 27, 2018, to July 20, 2020, and divide the processed data into long and short texts according to byte size, using different sentiment analysis models for texts of different lengths (Fig. 9).

Bert layer, using Google's pre-trained English model "BERT-Base, Uncased," the model uses a 12-layer Transformer, the hidden size is 768, the parameter of multi-

Table 3 Variable definition

Variable symbol	Variable name	Variable meaning
R_t	Closing price	Closing price at time t
P_t	Yield rate	Yield at time t $P_t = R_t/R_{t-1} - 1$
Q_m	Monthly sentiment indicators	Arithmetic average of the sentiment value of all reviews per month

Table 4 Two-step cross-sectional regression validation

Fama–MacBeth (1973) Two-step procedure				Number of obs = 4859 Num.time periods = 187 $F(1190) = 5.74$ Prob > $F = 0.0173$ avg.R-squared = 0.1456	
stockreturn	Coef.	Fama–MacBeth SE	t	|t|	(95% Conf.Interval)
bzps	.2,048,637	.0831456	2.56	0.016	.0435698 .3850610
_cons	.0361589	.027687	0.95	0.420	− .0310265 .0865781

Table 5 Categories

	Difficult	Easy
Positive	Positive difficult	Positive easy
Negative	Negative difficult	Negative easy

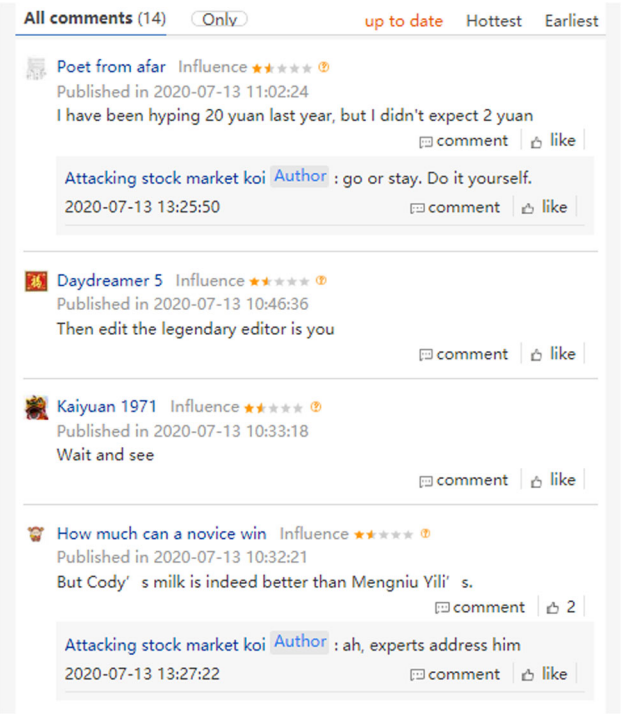

Fig. 9 Investor review

headed attention is 12, the total parameter size of the model is 110 MB, the optimizer uses Adam, the learning rate is set to 2E − 5, the maximum sentence length is set to 128, the training set and the test set's batch_size is 16 and 8, respectively, and the training epoch is set to 3. There are three metrics to evaluate the effectiveness of the experiment in this paper, which are accuracy, recall, and F Score. Accuracy is the proportion of positive reviews and news reviews that are correctly predicted by the model to the total number of samples. Recall is the proportion of news reviews that are correctly categorized among all samples that are actually news reviews, which indicates how many positive cases of the true sample are correctly predicted from the perspective of the original sample. And F score is an evaluation index that combines accuracy and recall.

6.2 Results

6.2.1 Multilevel sentiment analysis of investors

Adam is an SGD-based first-order optimization algorithm, which differs from SGD in that SGD does not change the learning rate during the training process, while Adam dynamically changes the learning rate by computing the first- and second-order moments of the gradient, which is an adaptive learning rate optimization algorithm that combines the advantages of AdaGrad and RMSProp. The Adam optimization algorithm was used to train the neural network, and the experimental results show that Adam's algorithm has the excellent results, as shown in Fig. 10, and the loss on both the training and test sets can be reduced to about 0.05 after using Adam's algorithm.

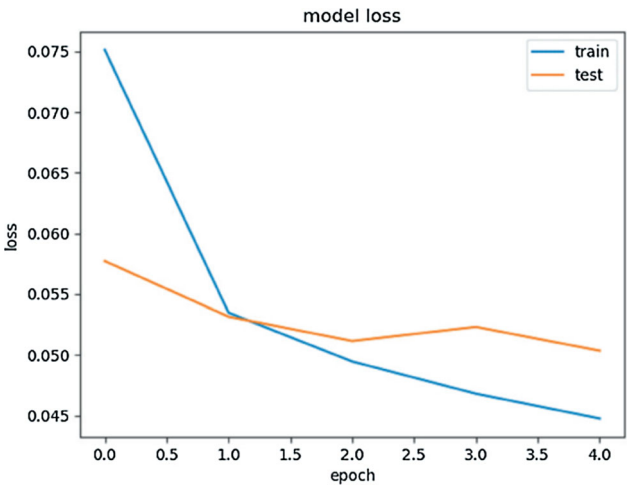

Fig. 10 Training set and test set losses

In the experiments, we set the batch size to 128, and the algorithm is close to convergence and achieves over 98% accuracy on the test set after 3 epochs, as shown in Fig. 11.

The performance of the paper's proposed method is compared to other machine learning methods for sentiment recognition in datasets.

The sentiment analysis of investor reviews is used to obtain the sentiment tendency of investors. From Fig. 12, it can be seen that the accuracy, recall, and F-value of the Bert + Attention method are significantly better than the SVM and LSTM + Attention methods, so using Bert + Attention to predict investor sentiment will produce the results that are closest to the real situation.

The resulting investor sentiment indicators are combined with the previously collected trends in stock yields, obtain the comparative relationship shown in Fig. 13.

From the graph, we again verify that the Bert model is more accurate than LSTM and SVM, and we can see that investor sentiment and stock yield are positively correlated, and when investor sentiment is more positive, the stock market also trends upward; when investor sentiment is more negative, the stock market trends downward. This conclusion will be further verified in Sect. 6.3.2. (The trend direction of the fold of the stock yield amount and the stock yield is the same, just in different units.)

6.2.2 Cross-sectional regression validation

The two-step cross-sectional regression validation method can be used to analyze whether investor sentiment indicators have a significant impact on stock yield, and the model regression results are shown in the following table.

Thus, the final estimate of the model is:

$$R_t = 0.03 + 0.21 \times Q_m$$

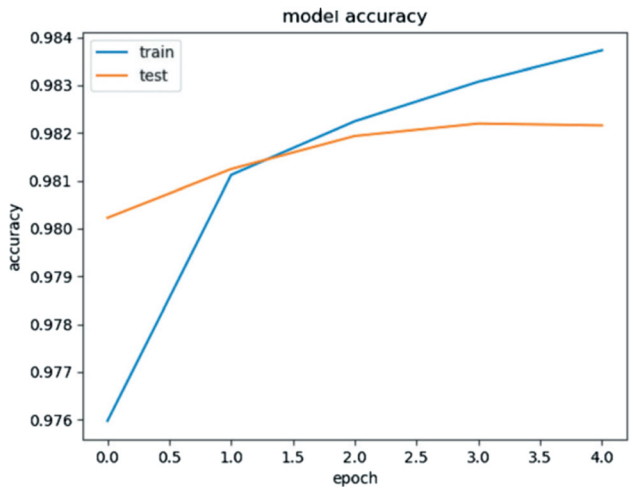

Fig. 11 Training set and test set accuracy

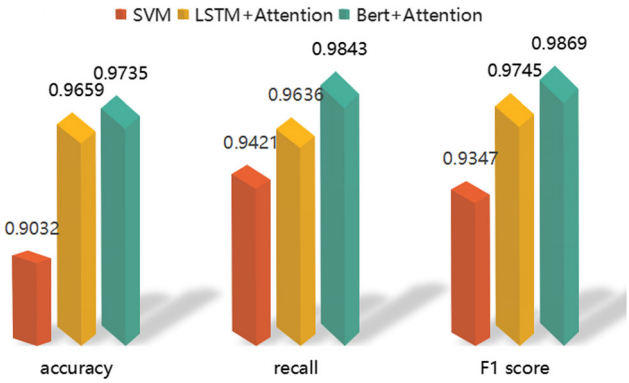

Fig. 12 Comparison of different methods

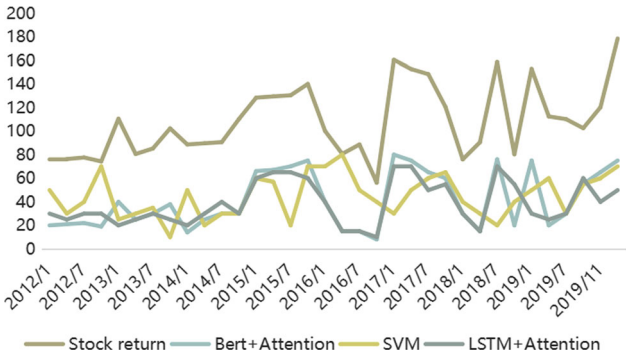

Fig. 13 SVM, LSTM, and Bert sort out investor sentiment versus stock yields

The above regression results show that: the p value of the model is 0.0173, indicating that the model is significant at the 95% confidence level; the p-value of the intercept term is 0.420 and the intercept term is not significant; the coefficient of the sentiment term is 0.20 and the p-value of the sentiment term is 0.016 and the sentiment term is significant, indicating that investor sentiment has a relatively strong explanatory power on the stock returns. And when investor sentiment is more positive, the returns of stocks trend up, and when investor sentiment is more negative, the returns of trend down.

6.3 Experimental analysis

From the experimental results, the application of neural networks to predict stock market investor sentiment indices has many advantages, such as less stringent requirements for data distribution, nonlinear data processing, strong robustness, and dynamics, which make it a hotspot in the early warning research of stock market investor sentiment indices. However, at the same time, artificial neural networks also have some shortcomings, which are mainly manifested by the imbalance in the number of difficult and easy samples.

The best way to solve the problem of imbalance between the number of difficult and easy samples is to introduce Focal Loss, a single-stage target detector usually generates up to 100 k candidates, and only a few of them are positive samples, which is very unbalanced. The formula for loss-cross-entropy, which is commonly used in the calculation of classification, is as follows.

$$\mathrm{CE} = \begin{cases} -\log(p), & \text{if} \quad y = 1 \\ -\log(1-p), & \text{if} \quad y = 0 \end{cases} \tag{18}$$

But this does not solve the whole problem. In total, the sample can be divided into the following four categories, based on positive, negative, difficult, and easy.

Although α balances the positive and negative samples, it does not help with the imbalance of the easy and difficult samples. In fact, a large number of candidate targets in target detection are easy samples.

The losses for these samples are low, but due to the extremely unbalanced number of samples, the number of easy samples is relatively too large and ultimately dominates the total loss. Easy samples (i.e., those with high confidence) have very little improvement on the model, and the model should focus primarily on those that are difficult to divide, where Focal Loss plays a key role. A simple idea: reduce the losses of the high confidence (p) samples a bit more.

$$\mathrm{FL} = \left\{ \begin{array}{ll} -(1-p)^{\gamma}\log(p), & \text{if} \quad y = 1 \\ -p^{\gamma}\log(1-p), & \text{if} \quad y = 0 \end{array} \right\} \tag{19}$$

When r takes 2, if $p = 0.968$, the $(1 - \mathrm{p})2 = 0.001$ loss attenuates by a factor of 1000 if $p = 0.968$.

The final form of Focal Loss addresses the imbalance of difficult samples and positive and negative samples, respectively, and using the two together yield the following final formula.

$$\mathrm{FL} = \left\{ \begin{array}{ll} -\partial(1-p)^{\gamma}\log(p), & \text{if} \quad y = 1 \\ -(1-\partial)p^{\gamma}\log(1-p), & \text{if} \quad y = 0 \end{array} \right\} \tag{20}$$

In future studies, if Focal Loss can be applied to the loss function, the problem of imbalance in the number of difficult samples can be greatly improved, thus making sentiment prediction more accurate.

7 Conclusion

This paper analyzes online reviews of investors based on the BERT model, obtains sentiment tendencies of investors, and illustrates the impact of investor sentiment on stock yield using a cross-sectional one-dimensional regression model. The study shows that the model is used to analyze the sentimental tendencies of investors, which can provide a scientific basis for stock investment decisions in order to reduce subjectivity and blindness in the stock investment decision-making. Although our experiments successfully confirmed the relationship between investors and stock returns, there are still shortcomings: For example, the difficulty and easy sample problem is mentioned in Sect. 6.3, but this is a common occurrence in the field of artificial intelligence, we hope to make breakthroughs in subsequent research.

Acknowledgments This research is supported by the R&D Program of Beijing Municipal Education commission (Grant No. KJZD20191000401). This research is also supported by the Program of the Co-Construction with Beijing Municipal Commission of Education of China (Grant Nos. B20H100020, B19H100010) and funded by the Key Project of Beijing Social Science Foundation Research Base (Grant No. 19JDYJA001).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. De Gooijer JG, Hyndman RJ (2005) 25 years of IIF time series forecasting: a selective review. Soc Sci Electron Publ 22(3):443–473
2. Farina L, Rinaldi S (2000) Positive linear systems. Theory and applications. J Vet Med Sci 63(9):945–948
3. Fisher KL, Statman M (2000) Investor sentiment and stock returns. Financ Anal J 56(2):16–23
4. Jin Z, Yang Y, Liu Y (2019) Stock closing price prediction based on sentiment analysis and LSTM. Neural Comput Appl 32:9713–9729. https://doi.org/10.1007/s00521-019-04504-2
5. Zweig ME (1973) An investor expectations stock price predictive model using closed-end fund premiums. J Finance 28:67–78
6. Lee CMC, Shleifer A, Thaler RH (1991) Investor sentiment and the closed-end fund puzzle. J Finance 46:75–109
7. Swaminathan B (1996) Time-varying expected small firm returns and closed-end fund discounts. Rev Financ Stud 9:845–887
8. Kurov A (2010) Investor sentiment and the stock market's reaction to monetary policy. J Banking Finance 34:139–149. https://doi.org/10.1016/j.jbankfin.2009.07.010
9. Yueqin L, Yong H, Chao Y (2020) Investor sentiment and stock price: empirical evidence from Chinese SEOs. Econ Model. https://doi.org/10.1016/j.econmod.2020.02.012
10. Riloff E, Wiebe J, Wilson T (2003) Learning subjective nouns using extraction pattern bootstrapping. In: Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003, Morristown, NJ, USA. Association for Computational Linguistics, pp 25–32
11. Zhang D (2017) High-speed train control system big data analysis based on the fuzzy rdf model and uncertain reasoning. Int J Comput Commun Control 12(4):577–591
12. Kim SM, Hovy E (2006) Automatic identification of pro and con reasons in online reviews. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp 483–490
13. Banea C, Mihalcea R, Wiebe J (2010) Multilingual subjectivity: are more languages better?.In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp 28–36

14. Wan X (2009) Co-training for cross-lingual sentiment classification. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. Association for Computational Linguistics, Suntec, Singapore, pp 235–243
15. Boyd-Graber J, Resnik P (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp 45–55
16. Lu B (2010) Identifying opinion holders and targets with dependency parser in Chinese news texts.In Proceedings of the NAACL HLT 2010 student research workshop pp 46–51
17. Mihalcea R, Banea C, Wiebe J (2007) Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th annual meeting of the association of computational linguistics, Prague, Czech Republic. Stroudsburg, PA: Association for Computational Linguistics, pp 976–983
18. Barsky RB, Sims ER (2012) Information, animal spirits, and the meaning of innovations in consumer confidence. Am Econ Rev Am Econ Assoc 102(4):343–177
19. Schmeling M (2009) Investor sentiment and stock returns: some international evidence. J Empir Finance 16(3):394–408
20. Chung SL, Hung CH, Yeh CY (2012) When does investor sentiment predict stock returns? J Empir Finance 19(2):217–240
21. Zhiqing M, Guojie Z, Yunwen Z (2018) Research on the relationship between internet investor sentiment and stock market price——based on analysis of text mining technology. Price Theory Practice 8:127–130. https://doi.org/10.19851/j.cnki.cn11-1010/f.2018.08.031
22. Brown GW, Cliff MT (2004) Investor sentiment and the near-term stock market. J Empir Finance 11(1):1–27
23. Bin L, Tan C, Cardie C, Tsou BK (2011) Joint bilingual sentiment classification with unlabeled parallel corpora. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for Computational Linguistics, Stroudsburg, pp 320–330
24. Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns. J Finance 61(4):1645–1680
25. So WG, Kim HK (2020) The influence of drama viewing on online purchasing intention: an empirical study. J SystManag Sci 10(2):69–81
26. Horvat D, Wydra S, Lerch CM (2018) Modelling and simulating the dynamics of the european demand for bio-based plastics. Int J Simul Model 17(3):419–430
27. Moon KS, Kim H (2019) Performance of deep learning in prediction of stock market volatility. Econ Comput Econ Cybern Stud Res 53(2):77–92. https://doi.org/10.24818/18423264/53.2.19.05
28. He Z, He L, Wen F (2019) Risk compensation and market returns: the role of investor sentiment in the stock market. Emerg Mark Finance Trade 55(3):704–718
29. Fama EF, MacBeth JD (1973) Risk, return, and equilibrium: empirical tests. J Political Econ 113(3):607–636

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.