Investor Sentiment Index Based on Large Language Models and Its Predictive Analysis for the Shanghai Composite Index

Xutong Wang

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, 650221, China

wangxutong@stu.ynufe.edu.cn

Abstract. In the realm of financial markets, investor mood often intensifies or lessens the influence of new information on price changes. Recently, progress in Large Language Models like the Bidirectional Encoder Representations from Transformers (BERT) algorithm has shown better predictive abilities than traditional methods in foreseeing market trends, so this study builds an investor sentiment index using the BERT model to test its predictive effectiveness regarding the Shanghai Composite Index, which involves gathering post data from the East Money Stock Forum, after that, the BERT model was fine-tuned for sentiment classification and scoring, thus creating a daily sentiment time series, and the predictive power of the constructed index was assessed using two approaches, conventional econometric regression analysis and a multi-scale forecasting framework combining Complementary Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Long Short-Term Memory(LSTM) network models. The empirical results proved that the BERT-based sentiment index could precisely grasp variations in investor sentiment, improving the explanatory and predictive ability for stock returns and the direction of market movement, demonstrating that Large Language Models are useful for sentiment analysis in financial text, offering a practical tool for investors and regulators to utilize sentiment-related insights.

Keywords: Large Language Models, BERT, Investor Sentiment Index, Shanghai Composite Index, Stock Forecasting.

1. Introduction

In the realm of financial markets, the variations in stock prices are affected by a combination of factors and investors' mood tends to either strengthen or weaken the influence of market information, thus being crucial in deciding price instability; in the context of behavioral finance, mood is the main means through which irrational aspects get into market activities and investors' emotional understandings of market comments, company revelations, and other signs guide their trading decisions, which later on have an effect on the trend of stock prices when the market opens [1].

Traditional sentiment analysis approaches are burdened with multiple drawbacks such as insufficiently dense data presentation, lack of adequate quantitative support, and a weak connection between the measured sentiment figures and real-life investor actions which made it difficult to accurately describe how sentiment affected market results and moreover the scattered nature of sentiment data along with the subtle meanings common in financial conversations was too much for ordinary analytical tools to fully recognize its predictive signs so previous research had not yet successfully clarified the changing relationship between investor sentiment and price trends.

The progress in Large Language Models (LLMs) has opened up fresh paths for analyzing financial texts and measuring investor sentiment, as they are marked by deep semantic understanding and flexible generative abilities and had shown great potential in various financial uses like deciphering financial news, clarifying corporate disclosures, and improving financial question-answering systems, which made it easier to look into the connection between investor sentiment and stock market movements [2-5].

The present research utilized Large Language Models (LLMs) to construct an investor sentiment index aiming to evaluate its predictive value for stock market indices, presented a methodological framework that combined the preprocessing and quantifying of pre-market textual sentiment with econometric regression analyses so as to measure the precision of predictive results, had a new feature

which was the development of a finance-specific semantic lexicon guided by LLMs that improved the accuracy and specificity of sentiment analysis, and established a quantitative connection between sentiment indices and market volatility thereby forming an integrated analytical model that united LLM-driven sentiment quantification, index building, and market prediction.

2. Literature Review

Stock price fluctuations are influenced by various factors and the market information as well as the collective investor sentiment are of great significance, traditional prediction methods like that used by Loughran and McDonald involve making special financial sentiment dictionaries for examining companies' 10-K disclosures in text form and their research found a connection between the frequency of negative words and later decreases in share prices around the announcement times, Boudoukh et al [6, 7]. used the Loughran-McDonald sentiment dictionary to measure sentiment in news content and offered proof that certain news sentiments have a substantial effect on stock price changes, in the same way, Ferguson et al [8]. measured the sentiment of British newspaper articles with the LM lexicon and showed that media sentiment can predict stock returns to some extent, all these studies continuously confirm that investor sentiment serves as an important driver of stock price movements both domestically and internationally [9].

Substantial headway had been made but most traditional forecasting methods still rested on sentiment lexicons or basic statistical systems which were first liked because they were clear and easy to use yet their reliance on surface-level word frequency analysis turned out to be insufficient for spotting subtle or indirect emotional hints in complex textual situations so models depending on static lexicons like the Loughran-McDonald dictionary often failed to meet the strict accuracy demands required by real-world predictive tasks in practical circumstances.

The progress in Large Language Models has brought about a great transformation in the areas of sentiment analysis and financial forecasting by making context-dependent semantic understanding possible, these models can recognize subtle sentiment signs which go beyond simple lexical signs thus improving the precision and fineness in both sentiment detection systems and stock market prediction models, because of this the extraction of detailed sentiment patterns, the evaluation of investors' risk preferences, and the spotting of early warning signs in market discussions have been made easier and these aspects are crucial for showing the mental and emotional bases of trading actions in the field of behavioral finance, incorporating Large Language Models into sentiment analysis has offered a deeper, meaning-centered study of emotional changes clarifying how sentiment alterations affect market trends, looking into the growth and application of Large Language Models for financial predictive analytics is of great theoretical and practical significance as it strengthens the strictness and reliability of investment decision-making procedures [10, 11].

In conclusion, although the significant impact of investor sentiment on stock market activities is widely acknowledged among academics, there are very few studies using Large Language Models (LLMs) in this area, which hinders full comprehension of how LLMs can be beneficial in complex financial situations and also prevents people from thoroughly exploring previous sentiment traits, how they spread, and their overall effects on stock price changes. Based on this, this study utilized web-scraping methods to gather investor sentiment data from East Money Information and then applied LLMs for analyzing the meaning and measuring the sentiment of pre-market text data, thus creating a new stock investor sentiment index, after that, it assessed how well this index could predict the post-market Shanghai Composite Index; the aim of this research was to show how useful LLMs are in detecting investor sentiment and predicting market trends, providing new perspectives for sentiment-based analysis in financial markets.

3. Data Metric Development

3.1. Data Acquisition and Sentiment Quantification

This study constructed an investor sentiment index using the posts collected from the Shanghai Stock Exchange Index discussion board run by East Money Information and text processing was carried out with the Bidirectional Encoder Representations from Transformers (BERT) algorithm, a well-known Large Language Models renowned for its efficiency in dealing with complicated linguistic data and it made it easier to quantify sentiment by turning investors' unstructured, context-dependent language into semantic vectors which could be used as a non-linear predictor of stock price changes and the method consisted of four main steps, getting the data, pre-processing the text, building the lexicon, and optimizing the model.

First, a Python-based web scraping method was utilized for data collection to get metadata like post titles, time stamps, and view counts from the forum, the obtained dataset covered the time span from January 1,2022, to December 31,2024, having around 3.95 million entries, after removing promotional stuff, empty or duplicate data, and other irrelevant things, the dataset was made into a set of 3.45 million relevant records, which averaged roughly 3,150 posts each day during the trading time.

Secondly, establishing a comprehensive sentiment lexicon corpus is crucial for obtaining reliable sentiment quantification analysis results from textual data. By integrating the positive and negative word sets from SnowNLP, BosonNLP, and CNKI Hownet sentiment dictionaries—all built upon the BERT model—and removing duplicates, a complete initial sentiment lexicon corpus was constructed. The SnowNLP stopword list served as the initial stopword database, utilised to filter out post text data incapable of expressing emotional inclination.

Subsequently,in the phase of sentiment quantification, the BERT model was fine-tuned by using sentiment classification datasets to enhance its understanding of the context in textual data, and it produced probability distributions which measured how likely each post was to fall into different sentiment categories, these probabilities were normalized between 0 and 1, with scores lower than 0.5 meaning negative sentiment, those higher than 0.5 showing positive sentiment and a score of 0.5 representing neutral sentiment, after this classification, there were 1,380,345 neutral posts, 1,034,655 negative posts and 1,035,000 positive posts in the dataset.

Because the initial sentiment lexicon was not sufficiently developed, a large amount of text was classified as neutral. Therefore, it is necessary to further train and improve the sentiment lexicon and re-quantify the sentiment of these neutral texts.

The newly trained lexicon based on the BERT model utilised the SnowNLP library to quantify the sentiment of neutral post texts, yielding the following results, 138,035 negative texts, 138,035 positive texts, and 1,104,275 neutral texts. The final sentiment classification of forum posts yielded, 1,172,690 negative posts, 1,173,035 positive posts, and 1,104,275 neutral posts.

3.2. Design of the Investor Sentiment Index

This study developed an investor sentiment index through integrating an evaluation of both how frequently people posted and how widespread positive market sentiment was, as had been initially put forward by Antweiler and Frank [12].

$$EI_i = \ln\left(\frac{1+P_i}{1+N_i}\right) \tag{1}$$

Where P_i denotes the number of positive posts on day i, and N_i denotes the number of negative posts on day i. When the number of positive posts exceeds the number of negative posts on a given day, the EI_i value is greater than 0, indicating heightened investor sentiment and optimism towards the stock market; conversely, the EI_i value is less than 0. Where P_i denotes the number of positive posts on day i, and N_i denotes the number of negative posts on day i. When the number of positive posts exceeds the number of negative posts on a given day, the EI_i value is greater than zero,

indicating heightened investor sentiment and optimism towards the stock market; conversely, the EI_i value is less than zero.

3.3. Variable Selection

In order to explore how investor sentiment impacts stock-price volatility, this study selected different elements from the Shanghai Composite Index which were then utilized as dependent variables within several empirical setups; more precisely, the daily returns of the Shanghai Composite Index were computed by taking the logarithmic difference of the closing prices at the close of successive trading days, resulting in a continuously-varying dependent measure ideal for regression analysis and predictive modeling and this measure grasped the market's intra-day price swings and was used as the regressand in ordinary least squares (OLS) regression models, making it easier to specify the model and verify the predictive results.

To evaluate how investor sentiment influenced the probability of market appreciation, a two-way directional indicator was created which sorted daily market returns into positive and negative groups with positive returns marked 'upward' (represented as 1) and negative ones labeled 'non-upward' (represented as 0), and then this binary indicator was put into a Logit regression model to measure the impact of investor sentiment on the chance of the stock market making gains.

In this study, the main independent variable was an investor sentiment index which had been created using the BERT Large Language Models, first, an initial sentiment polarity analysis was carried out on forum posts from the Shanghai Stock Exchange Index discussion platform where each post was classified as positive, negative, or neutral and the corresponding sentiment probabilities were calculated, then these separate sentiment categories were combined daily by comparing the numbers of positive, negative, and neutral posts resulting in a daily investor sentiment measure and the final index ranged from -1 to 1 with positive figures indicating that optimism was predominant and negative ones showing that pessimism was dominant.

To deal with the problem of simultaneity bias, the lagged daily sentiment index which stands for the value of the previous day was added as an independent variable in the regression models and also to cover the long-term course of sentiment trends, daily sentiment indices were combined by averaging or month-weighted ways to form a monthly investor sentiment indicator which was subsequently used as a regressor to assess the influence of intermediate-term investor sentiment on changes in stock prices.

3.4. Data Analysis

Empirical modeling was utilized for evaluating the predictive efficacy of the investor sentiment index and time-series data which included the closing figures of the Shanghai Composite Index were obtained from the Juquan Quant platform covering the period from January 1, 2022, to December 31, 2024, so as to match the time frame of the sentiment index sample.

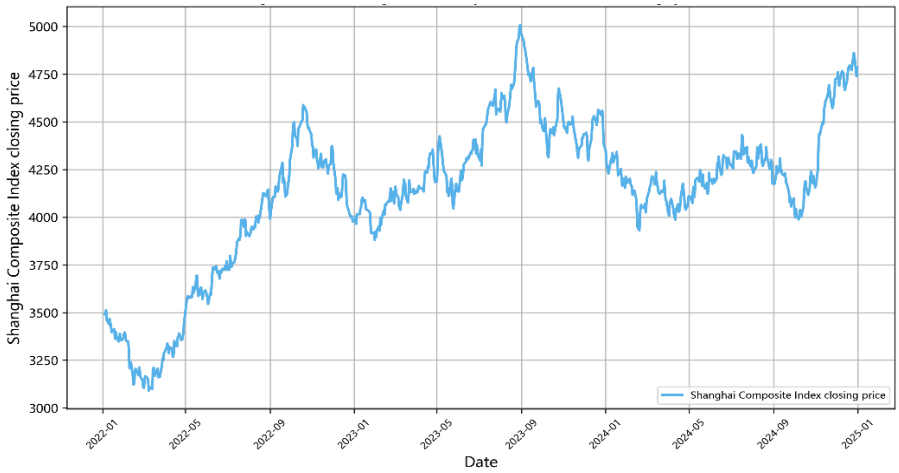

Figure 1. Shanghai composite index closing price

The graph in Figure 1 showed that during the sample period, the Shanghai Composite Index went through three distinct stages, a downward trend, a time of recovery, and later on fluctuations, which together covered the main market trends in the examined time frame, making the empirical results more solid and dependable.

This study evaluated the predictive capacity of the BERT-based investor sentiment index by using two different methodological approaches, a quantitative regression analysis model and a multi-temporal scale predictive modeling framework.

3.4.1. Path Analysis in Econometric Regression

First, Ordinary Least Squares (OLS) is employed to examine the explanatory power of LSI_{t-1} over the return rate r_t Subsequently, a Logit model is utilised to test the predictive capability of LSI_{t-1} over the directional indicator D_t thereby assessing the validity of the sentiment index.

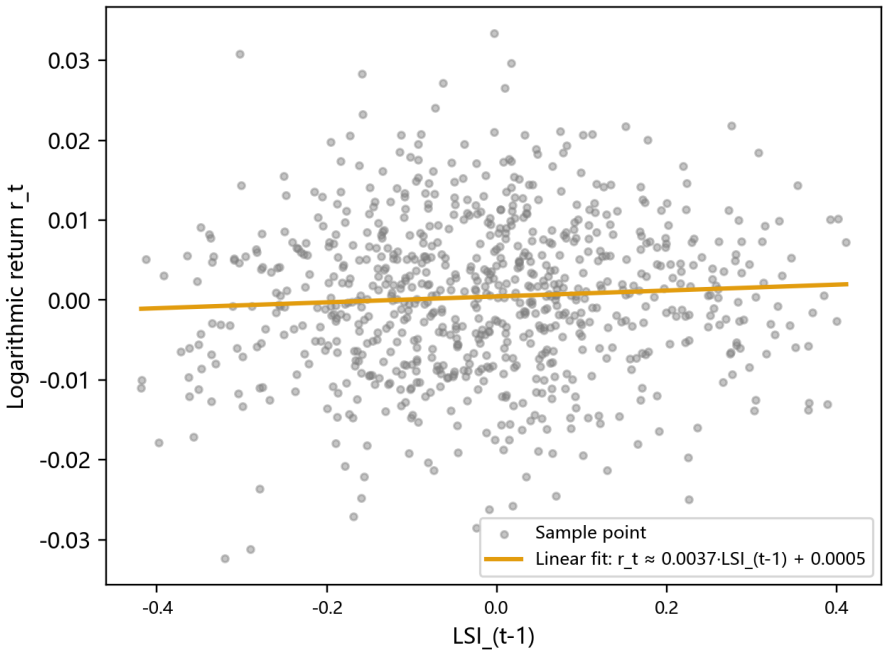

Figure 2. Relationship between r_t and lagged mood LSI_{t-1}

As shown in Figure 2, the regression line exhibits a slight upward slope, indicating a directional relationship where greater optimism correlates with higher returns the following day. However, the daily frequency noise is substantial, limiting the explanatory power of LSI alone. This is why additional control variables such as momentum fluctuations are incorporated.

3.4.2. Multi-Timescale Prediction Pathways

The Composite Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) algorithm was employed to break down both the price and sentiment index time series into Intrinsic Mode Functions (IMFs) that functioned within specific frequency ranges, after that Long-Short Term Memory (LSTM) networks were separately designed for each IMF, in the experimental setup a control group was given only price data while the experimental group incorporated lagged sentiment variables that corresponded to the frequency-specific IMFs, once the predictive results for each IMF were obtained the forecasts were combined to reconstruct the main time series which were then compared with actual data points to assess the effectiveness of the LSTM models.

The model's predictive ability was evaluated by means of a full range of metrics including root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), directional precision and economic effectiveness measures.

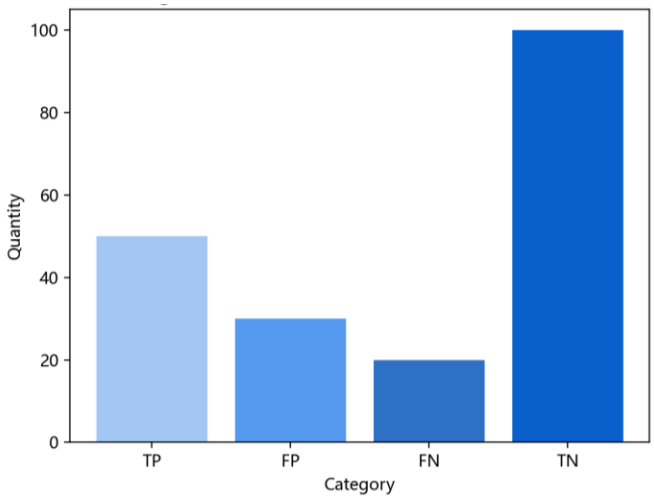

Figure 3. Direction prediction confusion count

The graphic depiction in Figure 3 shows that the model's predictive results have a far larger quantity of true positives (TP) and true negatives (TN) than false positives (FP) and false negatives (FN), which indicates that the model has a great deal of precision in predicting market trends, particularly in foreseeing downturns, and most of the market fluctuations, whether they went up or down, were precisely forecasted with very few wrong classifications, leading to less frequent occurrences of false positives and false negatives, so the model's function demonstrated its ability to match the real-world market behaviors in the vast majority of cases.

It was observed that the number of true negatives was far greater than that of true positives, indicating that the predictive model tended to be cautious when predicting market downturns perhaps showing a tendency towards market stability, and although having such a conservative inclination, the model worked well being able to distinguish effectively between the times when the market went up and when it went down.

4. Empirical Research Findings

4.1. Analysis of the Relationship Between Investor Sentiment Index and Market Volatility

Examining Figure 4 shows a strong connection between mood swings and market fluctuation, which was especially evident when the market was unstable, and it could be noticed that when the market had more turmoil, the mood index (the orange line) and the Shanghai Composite Index (the blue line) moved together and both had similar ups and downs, implying that under market pressure, changes in investors' feelings were closely linked to price changes and suggesting that mood changes might have a great impact on market directions.

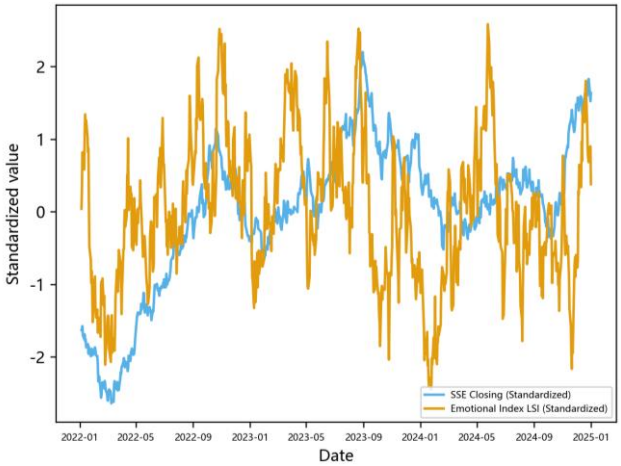

Figure 4. Standardized Shanghai composite closing and sentiment index

4.2. Benchmark Results

Table 1. Benchmark regression model results: OLS and Logit Regression analysis.

Model	N	$\beta\ LSI_{t-1}$	P-value	Other
OLS:r_t	781	0.00355	0.073	$R^2 = 0.0087$, Containing constants, r_{t-1}
Logit:D_t	781	0.802	0.068	Containing constants, r_{t-1}

The data presented in Table 1 indicated that the benchmark model's regression analysis showed a substantial influence of the lagged sentiment index (LSI) on market returns as well as directional forecasts and even though the p-values obtained from the ordinary least squares (OLS) and logistic regressions were close to the traditional thresholds of statistical significance, the results emphasized the significance of taking into account the predictive value of the sentiment index when analyzing the market.

The regression analysis using the ordinary least squares (OLS) approach showed that the sentiment index had a statistically noteworthy positive influence on market returns, with a p-value of 0.073 indicating this clearly, and this result implied that changes in investors' sentiment might considerably affect the volatility of market returns, thus offering useful insights for making predictive market models more effective and accurate.

In the Logit regression analysis, the sentiment index had a coefficient of 0.802 indicating that it was significant as a predictive element for the market direction particularly during periods when there was strong sentiment fluctuation, and even though the p-value was 0.068 which barely failed to reach the traditional standards for statistical significance, the index still maintained great explanatory strength about the market's directional changes showing that sentiment analysis was useful in financial prediction systems.

5. Conclusion

This study utilized the framework of a Large Language Model, namely the BERT model, to construct an investor sentiment index and meticulously assessed its predictive abilities regarding the Shanghai Composite Index, and the findings indicated that Large Language Models were capable of drawing out sentiment clues from complex financial data, breaking through the semantic restrictions of traditional lexical analysis approaches, and an empirical study demonstrated that the sentiment index generated by BERT had substantial explanatory strength for the returns and directional changes of the Shanghai Composite Index, which was shown by both OLS and Logit regression analyses, also in a multi-variable time-series forecasting context, the incorporation of sentiment measures enhanced both the predictive precision and the success rates of directional forecasts, emphasizing the anticipatory and predictive importance of investor sentiment in the dynamics of the equity market.

This research methodically broadened the academic path in the area of affective finance, producing empirical measures that could be used as reference standards by market players and regulatory bodies, but it wasn't free of limitations as the dataset was confined to the East Money Stock Forum which by nature narrowed the scope of the sample, the BERT model employed mainly dealt with textual meaning leaving out the inclusion of multimodal data like visual elements and interaction figures, and the system for classifying sentiment in the analysis was rather basic being unable to fully grasp the nuances of psychological feelings, so future academic efforts could benefit from looking into the combination of multimodal data integration methods, the creation of more advanced sentiment expression models, and the verification of results across different markets and situations.

References

[1] Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), 1259–1294.

[2] Arcas, B. A. (2022). Do Large Language Models understand us? Daedalus, 151(2), 183–197.

[3] Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2013). Which news moves stock prices? A textual analysis (No. w18725). National Bureau of Economic Research.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[5] Ferguson, N. J., Philip, D., Lam, H., & Guo, J. M. (2015). Media content and stock returns: The predictive power of press. Multinational Finance Journal, 19(1), 1–31.

[6] Guo, Z. (2024). A study on the impact of multimodal news sentiment on the stock market (Master's thesis, Southwestern University of Finance and Economics).

[7] Kirtac, K., & Germano, G. (2024). Sentiment trading with Large Language Models. Finance Research Letters, 62, 105227.

[8] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

[9] Moreno, A., & Ordieres-Meré, J. (2025). Predicting stock price trends using language models to extract the sentiment from analyst reports: Evidence from IBEX 35-listed companies. Economics Letters, 112404.

[10] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2025). A comprehensive overview of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 16(5), 1–72.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[12] Zhang, C., Wu, X., Deng, H., & Zhang, H. (2022). A time-varying study of Chinese investor sentiment, stock market liquidity and volatility: Based on deep learning BERT model and TVP-VAR model.