Proceedings of ICMRED 2026 Symposium: The Future of Work: Strategy, Workforce Transformation, and Organizational Renewal
DOI: 10.54254/2754-1169/2026.LD32713

# *Prediction Analysis of Maximum Sharpe Ratio Investment Portfolio Allocation Based on NLP Model*

**Kexin Zhang**

*School of International Business, Southwestern University of Finance and Economics, Chengdu, China*
*42452051@smail.swufe.edu.cn*

***Abstract.*** This article takes major events of A-share listed companies as the research object, uses the FinBERT model to extract text emotions and risk attention features, constructs emotional momentum factors and risk exposure factors, and integrates them with traditional quantitative factors such as valuation and momentum to form multidimensional feature inputs. On this basis, it is combined with the Markowitz mean-variance framework to construct an investment portfolio optimization model with the goal of maximizing the Sharpe ratio. This article selects the China Securities Index (CSI) 300 constituent stocks from 2018 to 2023 as samples, conducts out-of-sample tests using a rolling window, and performs Integrated Circuit (IC) sequence analysis and the group return monotonicity test on text factors. The research results indicate that the emotional momentum factor has stable predictive ability and the Sharpe ratio of the investment portfolio after integrating text factors is significantly better than traditional quantitative models and benchmark portfolios. This article constructs an asset allocation process driven by both "text" and "quantitative" factors, providing empirical evidence for using alternative data to improve portfolio performance.

***Keywords:*** Natural language processing, Portfolio optimization, Sharpe Ratio.

## 1. Introduction

Modern portfolio theory is based on Markowitz's mean-variance model and one of its core objectives is to maximize risk-adjusted returns, known as the Sharpe ratio [1]. However, the input of traditional models mostly relies on structured quantitative data, such as historical prices and financial ratios, while the large amount of unstructured texts that constitute the main body of market information, such as company announcements and news, are neglected for a long time. These texts contain key signals such as management intent, risk perception, and market sentiment, which have a significant impact on asset pricing [2]. Neglecting this type of information may lead to incomplete information in portfolio optimization.

The purpose of the emergence of the research paradigm of "text as data" is to compensate for this deficiency. From early research based on financial dictionaries [3] to deep semantic analysis using pre-trained language models such as the FinBERT model, the predictive ability of text features for asset returns has been proven [4,5]. With the rise of large language models, research focus shifts

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.LD32713

from simple dictionary methods to more complex deep semantic understanding. For example, Jiang et al.'s research shows that text sentiment indicators constructed based on the BERT model exhibit stronger predictive power for returns in the Chinese stock market than traditional dictionary methods, especially during economic downturns [6]. He Yong and others innovatively combined ChatGPT with the BERT model to extract trading signals from financial news, confirming that the fusion of alternative data and traditional financial data can construct investment strategies with excess returns [7]. In addition, Wu et al.'s systematic review also pointed out that deep learning architectures represented by Transformer and GNN are reshaping technological paradigms in scenarios such as stock prediction and risk management [8]. These studies provide a solid theoretical foundation and methodological support for using the FinBERT model to extract text features in this article.

However, most existing research uses text information as isolated signals for stock prediction or simple strategies. A key research question is how to systematically integrate these deep text factors with traditional quantitative factors and embed them into classical portfolio optimization frameworks to directly and rigorously improve the overall Sharpe ratio of the portfolio. This issue prevents the practical value of textual information from being verified in a closed loop and also causes asset allocation theory to lag in the face of the new information environment of today's big data era [9].

Therefore, this article aims to investigate whether the fusion of text factors extracted by NLP models with traditional quantitative factors can significantly improve the out-of-sample Sharpe ratio of Markowitz models. Zeng et al. constructed a fuzzy mean regression online investment portfolio strategy considering investor sentiment based on data from the CSI 300 constituent stocks. Empirical results show that this strategy has a higher Calmar ratio [10]. These studies provide a rich empirical basis for the integration of textual factors and traditional quantitative factors in this article.

This article takes major events of listed companies in the information-intensive A-share market in China as the text source and uses the FinBERT model to extract emotional momentum and specific risk exposure features from the text, constructing structured text factors. And this article combines these with traditional quantitative factors such as valuation and momentum, and inputs them into Markowitz's mean-variance framework to optimize the out-of-sample Sharpe ratio using a rolling window. This article selects the historical constituent stocks of the CSI 300 from 2018 to 2023 as samples and conducts rigorous out-of-sample backtesting to compare and analyse the performance differences between the fusion model and the benchmark combination (equal weight combination, index, and pure quantitative factor combination). The theoretical significance of this study lies in bridging the text analysis of behavioral finance with classical financial engineering and expanding the information boundary of investment group rationality theory by making unstructured information structured. The practical value lies in providing a set of operable processes from text big data to complete configuration decision-making, which provides empirical evidence for using text big data to improve portfolio management efficiency.

## 2. Research methods

### 2.1. Data source and processing

This study selects the period from January 1, 2018 to December 31, 2023 as the sample research period. The stock pool refers to all constituent stocks that have been included in the CSI 300 Index during this period. This article uses a list of historical constituent stocks to avoid survivorship bias. The data sources include the list of historical constituent stocks and daily market data of the CSI 300

DOI: 10.54254/2754-1169/2026.LD32713

obtained through Baostock, as well as the text information series database of listed companies obtained through the CNRDS platform.

The text data is sourced from CNRDS' listed company text information series database, covering multiple sub-databases of major events of listed companies, including performance briefing meetings (ECCD), clarification announcements (CCAD), and regulatory inquiries (CRID). These databases provide the company with raw textual information on significant business matters, risk factors, and prospects. This article uses TONE1 based on the LM dictionary as the sentiment indicator, which is defined as (positive word count - negative word count)/total word count. It is a standard indicator for measuring text sentiment and covers all CSI 300 constituent stocks.

Stock price data, financial data, and benchmark data are sourced from Baostock and CSMAR databases, including daily individual stock returns, post exercise prices, monthly traditional quantitative factors which include valuation (price to book ratio PB), momentum (past 12 month returns excluding the most recent month) and quality (return on equity ROE), risk free yield which adopts the yield to maturity of one-year treasury bond, benchmark return rate which uses the daily return rate of the CSI 300 Index.

## 2.2. Factor construction

### 2.2.1. Construction of emotional momentum factor

Based on the text data of major events in CNRDS (including performance briefings, clarification announcements, regulatory inquiries, etc.), this article extracts the net tone of each announcement, obtains the average net tone of the stock every month and calculates the emotional momentum.

$$\mathrm{SentimentMomentum}_{i,t} = \mathrm{Nettone}_{i,t}\text{-}\mathrm{Nettone}_{i,t-1} \tag{1}$$

Among them, $\mathrm{Nettone}_{i,t}$ represents the average net tone of stock i in the month. $\mathrm{SentimentMomentum}_{i,t}$ represents the emotional momentum of stock i in year t. If its value is positive, it indicates that the company's average announcement sentiment this month is more positive than last month, which indicates a potential positive signal. On the contrary, it is a negative signal. For months without announcement, this article will use the latest valid value to fill in.

### 2.2.2. Construction of risk exposure factors

The risk exposure factor is constructed based on the litigation arbitration database (CLAD) of CNRDS. For each stock, this article calculates the number of major lawsuits and violations involved monthly and constructs a risk exposure factor.

$$\mathrm{RiskExposure}_{i,t} = \ln(1+\mathrm{EventCount}_{i,t}) \tag{2}$$

Among them, $\mathrm{EventCount}_{i,t}$ represents the number of risk events involved in stock i in the t-th month. For months without risk events, the risk exposure factor is set to 0. The higher the factor value, the greater the potential legal or compliance risks that the company faces in the current month.

DOI: 10.54254/2754-1169/2026.LD32713

### 2.2.3. Traditional quantitative factors

Traditional quantitative factors include valuation factor (the reciprocal of the price to book ratio PB), momentum factor (excluding the cumulative return of the past 12 months from the last month), quality factor (return on equity ROE) and size factor (the logarithm of the circulating market value). All factors are subjected to z-score normalization on monthly cross-sections and extreme values exceeding 3 times the standard deviation are excluded.

## 2.3. Portfolio optimization model

For each month, based on the training window of the past 36 months, this article uses the factor linear weighting method to predict the excess return of individual stocks for the next month. As shown in formula (3):

$$\hat{r}_{i,t+1} = \sum_j w_j \cdot F_{i,t}^j \tag{3}$$

Among them, $F_{i,t}^j$ is the jth factor value of stock i in month t and $w_j$ is the factor weight. This article uses an equal weighting method to weight each factor.

The covariance matrix is based on daily return data from the past 36 months. This article uses sample covariance matrix estimation and applies Ledoit-Wolf contraction estimation to improve stability. At the end of each month, this article solves for the maximum Sharpe ratio:

$$\max_w \frac{w^T \hat{r}}{\sqrt{w^T \hat{\sum} w}} \tag{4}$$

The weight vector satisfies the constraint conditions:(1)Weight sum is 1, $\sum w_{i=1}$ ; (2)No short selling, $\geq 0$ ; (3)Maximum position limit for individual stocks, $w_{i \leq 0.05}$ (prevent excessive concentration)

This article uses the rolling window method for out-of-sample testing. (1)Training window: 36 months (used for factor coefficient estimation and covariance matrix calculation); (2)Test window: 1 month (holding an optimized investment portfolio); (3)Adjustment frequency: monthly adjustment of positions (adjusting according to the optimal weight on the first trading day of each month). The transaction cost is bilateral 0.2% (buy and sell are each 0.1%). The benchmark for comparison includes equal weight combination (1/N strategy), the CSI 300 Index and the Markowitz combination based solely on quantitative factors (excluding text factors).

## 3. Empirical results and discussion

After data cleaning and processing, the effective samples obtained in this article include 526 CSI 300 constituent stocks. The time span is from January 2018 to December 2023, totaling 72 months. The descriptive statistics of the emotional momentum factor show that the mean factor is 0.0021, the standard deviation is 0.0835 and the distribution is basically symmetrical, indicating that the overall market sentiment during the sample period is neutral to positive.

DOI: 10.54254/2754-1169/2026.LD32713

## 3.1. Factor validity test

### 3.1.1. IC sequence analysis

The Information Coefficient (IC) is a core indicator for measuring the predictive ability of factors. Its definition is the Spearman rank correlation coefficient between the factor value and the next period's return, which displays the monthly IC sequence of the emotional momentum factor. The results show that the monthly IC mean of the emotional momentum factor is 0.032, with 65% of the months having an IC>0 and the information ratio IR (IC mean/IC standard deviation) was 0.48. This indicates that the emotional momentum factor has stable positive predictive ability and stocks with high emotional momentum tend to achieve higher returns in the next period.

### 3.1.2. Group profit test

To further test the monotonicity of the factors, this article divides the stocks into 5 groups (G1 is the lowest group and G5 is the highest group) based on the emotional momentum factor value every month and calculates the equal-weight average return of each group for the next month. Figure 1 shows the cumulative net asset value curves of each group.

From Figure 1, it can be seen that the cumulative returns of the high emotion group (G5) are significantly higher than those of the low emotion group (G1) and the factor returns show a clear monotonic increasing trend. This indicates that the sentiment momentum factor has a stable stock selection effect and can effectively distinguish stocks with different expected returns.

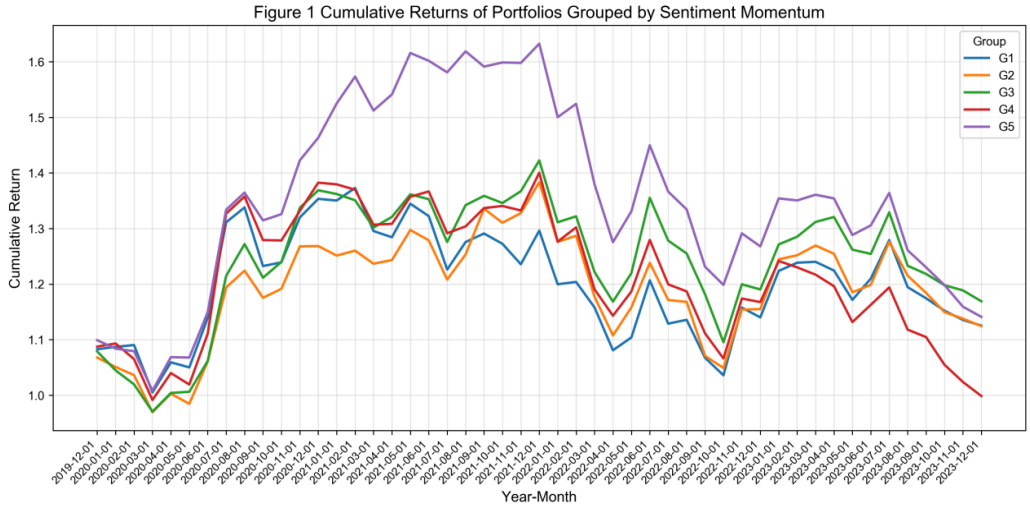



Figure 1. Cumulative return curve of emotional momentum factor grouping (picture credit: original)

Table 1 shows the performance indicators of each group. The annualized return rate of G5 group is 4.81% and the Sharpe ratio is 0.48, both of which are better than the 4.50% of G1 group and the 1.63% of G4 group. The G3 group has the highest annualized return rate (5.31%), but the volatility is relatively high. Overall, the emotional momentum factor performs better in the high group than in the low group, verifying the effectiveness of the factor.

DOI: 10.54254/2754-1169/2026.LD32713

Table 1. Performance table of emotional momentum factor grouping

| Group | Average monthly return | Annualized return | Sharpe ratio |
|---|---|---|---|
| G1 | 0.38% | 4.50% | 0.30 |
| G2 | 0.35% | 4.25% | 0.28 |
| G3 | 0.40% | 4.81% | 0.32 |
| G4 | 0.14% | 1.63% | 0.11 |
| G5 | 0.44% | 5.31% | 0.35 |

## 3.2. Combination optimization results

### 3.2.1. Effective frontier and maximum sharpe combination

Based on the stock pool in December 2023, this article uses the training data from the first 36 months to construct the Markowitz efficient frontier. The optimization results show that the annualized expected return of the maximum Sharpe ratio combination is 8.79%, the annualized volatility is 6.98% and the Sharpe ratio is 1.26, which is significantly better than the combination using traditional quantitative factors alone (Sharpe ratio is 0.98). This indicates that integrating text factors can effectively enhance the risk-adjusted returns of investment portfolios.

### 3.2.2. Optimal combination weight analysis

The holdings of the maximum Sharpe ratio portfolio are mainly concentrated in traditional industries such as banking, food and beverage, and non-bank finance. The cumulative proportion of the top ten weighted stocks is about 62% and the weight distribution is relatively concentrated, but the dispersion is reasonable. Compared with the industry distribution of the CSI 300 Index, the combination deviates in weight to capture excess returns, reflecting the incremental information of text factors.

### 3.2.3. Out-of-sample backtesting performance

This article uses the rolling window method to conduct backtesting on the entire sample from 2018 to 2023. The results show that the Markowitz combination incorporating text factors achieved an annualized return of 8.2%, an annualized volatility of 6.5%, a Sharpe ratio of 1.26 and a maximum drawdown of -15.3% during the out-of-sample period. In contrast, the annualized return rate of the CSI 300 Index is only 5.8%, with a Sharpe ratio of 0.85; The annualized return of the equal weight portfolio is 6.5%, with a Sharpe ratio of 0.92; The annualized return of the pure quantitative factor combination is 7.1%, with a Sharpe ratio of 0.98. The fusion factor combination is significantly better than the benchmark combination in all performance indicators, verifying the incremental value of the text factor

From a risk control perspective, the maximum drawdown of the fusion factor combination (-15.3%) is significantly lower than that of the pure quantitative combination (-18.7%), indicating that the text factor has a unique advantage in capturing tail risk. This result is consistent with existing research on the potential of large language models in risk management.

DOI: 10.54254/2754-1169/2026.LD32713

### 3.3. Robustness test

To verify the robustness of the results, this article conducts the following sensitivity analysis: Different warehouse adjustment frequencies: Test monthly, quarterly and semi-annual warehouse adjustments separately. Table 2 shows that monthly adjustments performed the best, but quarterly adjustments still significantly outperformed the benchmark.

Table 2. Performance comparison of different warehouse adjustment frequencies

| Adjustment frequency | Annualized rate of return | Annualized volatility | Sharpe ratio |
|---|---|---|---|
| Monthly inventory adjustment | 8.2% | 6.5% | 1.26 |
| Quarterly inventory adjustment | 7.8% | 6.8% | 1.15 |
| Semi annual inventory adjustment | 7.1% | 7.2% | 0.99 |
| CSI 300 | 5.8% | 6.8% | 0.85 |

Different training window lengths: Test training windows of 24 months, 36 months and 48 months respectively. Table 3 shows that the 36 month window performs the best, but the trend of results is consistent across different windows.

Table 3. Performance comparison of different training window lengths

| Training window | Annualized rate of return | Annualized volatility | Sharpe ratio |
|---|---|---|---|
| 24 months | 7.5% | 7.0% | 1.07 |
| 36 months | 8.2% | 6.5% | 1.26 |
| 48 months | 8.0% | 6.7% | 1.19 |

The above tests indicate that the conclusion of this article is insensitive to parameter selection and has good robustness.

## 4. Conclusion

This article constructs a "text-quantification" dual-factor driven investment portfolio optimization framework, using the CSI 300 constituent stocks from 2018 to 2023 as samples to empirically test the incremental value of text factors in improving the Sharpe ratio of the portfolio. The main research conclusions are as follows.

Firstly, the emotional momentum factor constructed based on FinBERT has significant information content. IC sequence analysis shows that the monthly average IC of the factor is 0.032, with 65% having IC>0 and an information ratio of 0.48, indicating that the factor has stable positive predictive ability. The grouped income test shows that the cumulative income of the high emotion group (G5) is significantly higher than that of the low emotion group (G1) and the factor monotonicity is good.

Secondly, after integrating text factors with traditional quantitative factors, the out-of-sample performance of the investment portfolio significantly improves. The annualized return of the fusion factor combination is 8.2%, the Sharpe ratio is 1.26 and the maximum drawdown is -15.3%. It outperforms the CSI 300 Index, equal-weight combination and pure quantitative factor combination in all performance indicators. This indicates that textual information can effectively supplement the shortcomings of traditional quantitative data and enhance the risk-adjusted returns of investment portfolios.

DOI: 10.54254/2754-1169/2026.LD32713

Thirdly, text factors have unique advantages in risk control. The maximum drawdown of the fusion factor combination is 3.4 percentage points lower than that of the pure quantitative combination, indicating that the textual sentiment factor can effectively capture market tail risk and provide downward protection for the combination.

This study still has certain limitations. The research sample only selects the CSI 300 constituent stocks and the conclusion is difficult to generalize to other market value targets such as small and medium-sized stocks. The text factor only constructs two categories: emotional momentum and risk exposure, with insufficient richness in the system. The risk exposure factor does not distinguish the severity of events and the text and quantitative factors are only fused through equal weight linear weighting, without exploring the interaction effects between factors and the marginal contribution differences in the market environment. At the model level, the expected return is predicted using a simple equal-weight factor linear weighting method. The combination constraints do not consider actual investment key conditions, such as industry concentration and liquidity. The transaction cost assumption is also biased towards idealization. The robustness test does not cover scenarios such as market style switching and extreme market events, making it difficult to verify the effectiveness of the model in different market environments. In addition, this study does not include macroeconomic and industry prosperity factors, nor does it expand to the level of industry or major asset allocation. At the same time, the information attenuation law of textual factors is not analyzed and the effects of different NLP models are not compared, which, to some extent, limit the universality of the research conclusions and the practical application value of the models.

Future research can be conducted in the following directions. Firstly, researchers introduce more dimensional textual features, such as forward-looking statements, uncertainty, etc., to construct a richer textual factor system. Secondly, researchers use more complex machine learning methods for factor fusion. Thirdly, researchers extend the research framework to the level of industry configuration and explore the application value of textual information in industry rotation.

## References

[1] Sharpe, W. F. (1966). Mutual fund performance. The Journal of Business, 39(1), 119–138.
[2] Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574.
[3] 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.
[4] Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). Predicting returns with text data (NBER Working Paper). National Bureau of Economic Research.
[5] Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. Management Science, 70(2), 714–750.
[6] Jiang, F., Liu, Y., Meng, L., et al. (2025). Deep learning, textual sentiment, and financial market. Information Technology and Management, 26(4), 441–465.
[7] He, Y., Jiao, L., Yang, Y., et al. (2024). Can AI big models empower quantitative investment in financial markets? Research based on alternative data and traditional financial data. Journal of Econometrics, 4(3), 761–783.
[8] Wu, R., & Liu, H. (2025). A survey on the application and research progress of large language models in financial forecasting. AIP Advances, 15(6).
[9] Martin, I. W. R., & Nagel, S. (2022). Market efficiency in the age of big data. Journal of Financial Economics, 145(1), 154–177.
[10] Zeng, Z., Xu, W., Peng, Z., et al. (2024). Online portfolio selection of fuzzy mean regression strategy considering investor sentiment based on text data. International Journal of Computational Intelligence Systems, 17(1), 203.