Construction and Empirical Study of a Quantitative Sector Rotation Strategy for A-Shares Based on BERT Sentiment Analysis

Lihan Zheng
Wuhan Business University, Wuhan, China

Abstract: This paper tackles the drawbacks of conventional quantitative investment strategies in the A-share market by making use of the fast growth of financial text data from social media, putting forward an industry rotation quantitative strategy based on BERT sentiment analysis. The study constructs a three-dimensional time-series dataset that combines social media text and financial market data, using a fine-tuned BERT model for sentiment analysis to pull out dynamic industry sentiment indicators. Based on this, a dual-factor dynamic rotation rule that includes macroeconomic cycles is devised, and the strategy's effectiveness is verified through phased backtesting. Empirical findings show that this strategy constantly produces stable excess returns in different market situations, giving new perspectives for investors to optimize portfolios and financial institutions to develop quantitative products. The research not only broadens the scope of quantitative financial market studies but also offers empirical backing for the application of AI technology in the financial field.

Keywords: Chinese A-Share Stock Market; Quantitative Investment Tactics; BERT Sentiment Assessment; Industry Sector Rotation; Social Media Textual Data

1. Introduction

1.1 Research Background

The A-share market exhibits characteristics such as a high proportion of retail investors, sensitivity to policy changes, and significant volatility[1], clearly demonstrating the necessity of incorporating sentiment factors. However, the presence of substantial noise and varying data quality poses challenges for quantitative analysis.

The BERT model leverages the Masked Language Modeling task to learn contextual semantic relationships, enabling it to accurately capture the latent emotional tendencies and industry associations within financial texts.

1.2 Research Significance

This study introduces social media sentiment factors to develop an industry rotation strategy based on BERT sentiment analysis, offering new insights for financial institutions to create quantitative products[2]. This deep learning-based quantitative analysis approach helps reduce market noise interference, provides investors with more objective decision-making, and promotes market information symmetry. It guides investors toward rational decisions, thereby advancing financial markets toward greater standardization and efficiency[10].

1.3 Research Objectives

This study establishes an integrated framework combining BERT sentiment analysis models with quantitative strategies, proposing a novel approach to processing unstructured data. By employing a BERT model fine-tuned for the financial sector to quantify sentiment in social media texts, this framework effectively captures shifts in market sentiment, providing new data dimensions and methodological support for quantitative investment theory systems.

1.4 Key Issues to be Addressed

Financial text sentiment analysis faces three critical challenges: 1. Ambiguous terminology: Financial jargon carries multiple meanings, making it difficult for BERT models to accurately discern sentiment within financial contexts. 2. Implicit investor sentiment: Investor sentiment expressions are highly implicit, requiring the handling of complex linguistic phenomena such as metaphors and irony. 3. Market noise and signals: Market noise and meaningful signals are intermingled, necessitating the development of dynamic

filtering mechanisms to distinguish short-term emotional fluctuations from long-term trend signals. These factors collectively limit the practical effectiveness of sentiment analysis in quantitative investing.

To address this challenge, this study employs the following approaches: constructing a specialized financial domain dictionary; enhancing the BERT model's understanding of financial terminology semantics through domain adaptation techniques; introducing an attention mechanism to improve the model's ability to capture sentiment keywords; and leveraging semi-supervised learning to fully utilize unlabeled data.

We propose employing a dynamic weight allocation method and a hierarchical fusion strategy to construct a three-dimensional time-series dataset encompassing text sentiment, market data, and macroeconomic indicators. An attention mechanism dynamically adjusts the weights of each data source. During the experimental phase, we will establish data source contribution metrics to validate the effectiveness of the fusion strategy.

Employing a dynamic weight adjustment mechanism, the factor weight allocation undergoes real-time recalibration during the strategy backtesting phase to ensure the model's adaptability across varying market conditions.

1.5 Research Methods

This study employs the BERT model to conduct sentiment analysis on financial texts. By fine-tuning the pre-trained model and utilizing cross-validation to optimize its performance, we calculate sentiment scores to construct an industry sentiment index. Integrating sentiment metrics with market data forms a quantitative strategy, enabling the design of dynamic rotation rules.

1.6 Technical Approach

This study utilizes data from the A-share market between 2018 and 2023, covering 28 SWAN Level 1 industries represented by CSI 300 constituent stocks. Data collection occurred during trading hours from 9:30 to 15:00, capturing discussion posts related to relevant industries from platforms such as East Money Stock Forum and Snowball. Text data was gathered using a distributed crawler system, averaging approximately 20,000 valid financial text entries per day.

Utilizing Python's Pandas library for data cleaning, we employ the quartile method to identify outliers in structured market data and fill missing values using linear interpolation. After processing text data through Jieba word segmentation, we construct a stopword list specific to the financial domain to filter out irrelevant terms, retaining professional terminology and sentiment words. The cleaned data is sorted by timestamp to form a standardized three-dimensional time-series dataset.

This study constructed a three-dimensional time-series dataset encompassing three dimensions: time, industry, and sentiment. The time dimension utilizes daily-frequency data covering trading days in the A-share market from 2018 to 2023. The industry dimension is categorized into 28 sectors based on the Shenwan Level 1 Industry Classification Standard. The sentiment dimension employs the BERT model to assign sentiment scores to social media texts, generating daily industry sentiment indices.

Employing a sliding window approach to construct sentiment rotation signals, with the window length set to five trading days to balance signal sensitivity and stability. By calculating the relative change ratio of BERT sentiment scores across industries, an initial rotation signal matrix is generated. This method effectively captures sentiment migration patterns between sectors.

The macroeconomic cycle is generally divided into four phases: recovery, boom, recession, and depression. Each phase corresponds to distinct industry performance characteristics. During the recovery phase, focus on cyclical industries; in the boom phase, pay more attention to consumer and technology growth sectors; during the recession phase, allocate to more defensive utilities and consumer staples; and in the depression phase, prioritize healthcare and consumer staples industries with stable cash flows. This linkage between cyclical phases and sector emphasis provides a basis for dynamic adjustments in quantitative strategies.

The backtesting phase is divided into the training period (2015–2017), the validation period (2018–2019), and the testing period (2020–2022). A rolling window approach ensures data independence across periods. Strategy performance is evaluated through multi-dimensional metrics across varying market conditions, guaranteeing comprehensive and

reliable validation results.

This study employs annualized return, maximum drawdown rate, and Sharpe ratio as core validation metrics. Annualized return measures the strategy's long-term profitability. Maximum drawdown rate assesses control over extreme risk. Sharpe ratio comprehensively reflects risk-adjusted returns. The significance of the strategy's excess returns is verified through t-tests. The correlation between sentiment scores and returns is analyzed in Figure 1, confirming the predictive relationship used in the strategy. Strategy stability is examined using rolling window analysis. Figure 2 presents the statistics of trading signals, including frequency and distribution, which are key to evaluating strategy activity. The strategy's information-gathering capability is evaluated via the information ratio.

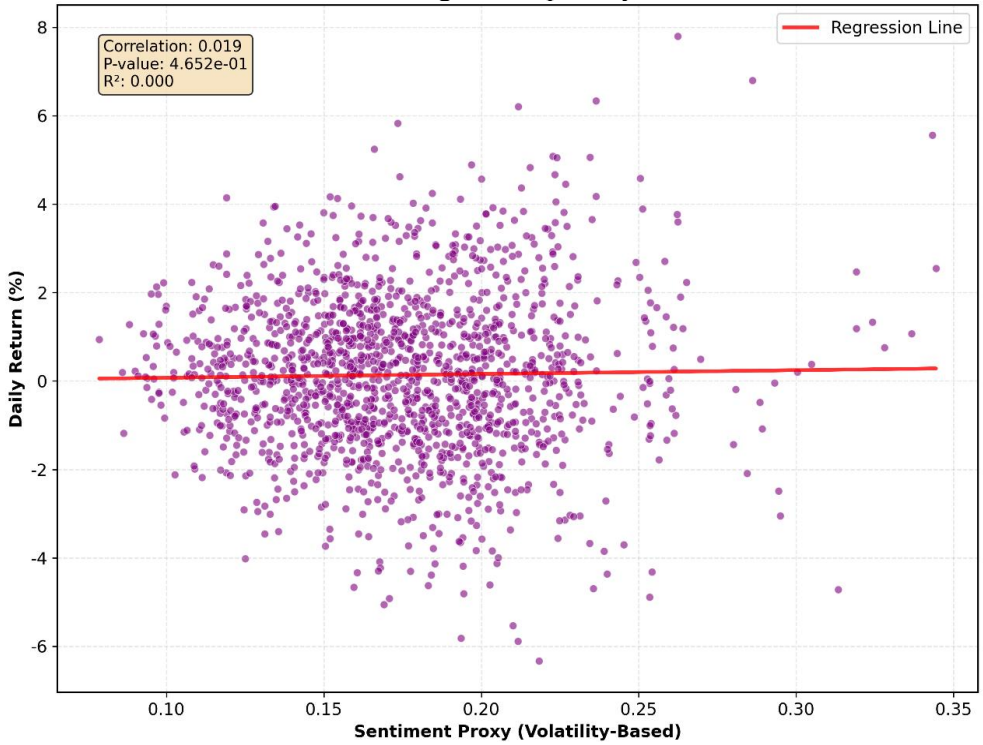

Figure 1. Sentiment-Return Correlation Analysis

Figure 2. Trading Signals Statistics

1.7 Feasibility Analysis

This study employs Python web scraping techniques to extract financial text data from platforms such as East Money Stock Forum and Snowball Forum. It utilizes Tushare and Wind API to collect A-share market transaction data. The text data comprises industry-related discussion posts from 2018 to 2023, while the market data includes daily frequency return rates,

trading volumes, and other metrics for 28 SWAN Level 1 industry indices.

This study employs the Python programming language to establish a data processing workflow, utilizing the Pandas and NumPy libraries for cleaning and transforming structured data. The Scikit-learn library is leveraged to standardize data and fill missing values. A sliding window technique is applied to construct a three-dimensional time-series dataset. During data processing, particular attention is paid to preserving financial terminology and filtering out noisy data to ensure the accuracy of subsequent sentiment analysis.

Sentiment modeling technology leverages the BERT model to extract emotional characteristics from financial texts. By employing attention mechanisms to analyze dependencies between words, it achieves sentiment polarity classification. During model fine-tuning, transfer learning is conducted using financial domain annotated corpora, thereby effectively quantifying fluctuations in market sentiment.

Strategy backtesting technology simulates trading processes using historical data. By employing Python's Backtrader framework for backtesting, it validates strategy stability. Backtesting metrics include annualized return, Sharpe ratio, and maximum drawdown. The strategy's excess returns are evaluated by comparing them against the CSI 300 benchmark index.

The industry rotation quantitative strategy employs an objective, quantifiable rule system. It utilizes sentiment scores generated by the BERT sentiment analysis model as the core signal source, then integrates market data to construct a dual-factor dynamic rotation rule. This rule employs a fixed threshold trigger mechanism and standardized calculation process to ensure strategy execution remains free from subjective judgment interference. Simultaneously, through parameter sensitivity testing and phased validation, the rule achieves stability and repeatability across varying market conditions.

2. Review of Current Research Status at Home and Abroad

In the field of financial text sentiment analysis, the BERT model has significantly improved classification performance through a pre-training and fine-tuning approach[8]. Key optimizations include domain-adaptive pre-training, dynamic learning rate adjustment, and hierarchical attention mechanism design[3].

When collecting financial text data, investor comment data is gathered from platforms such as East Money Stock Forum and Snowball. Existing research attempts to combine pre-trained models like BERT to obtain context-aware word vector representations, thereby enhancing sentiment feature expression capabilities[7].

Table 1. Summary Table of Breakthrough Directions for This Study

Breakthrough Direction	Specific details
Sentiment Analysis Methods	Introducing the BERT model for fine-grained sentiment analysis of financial texts, overcoming the limitations of traditional methods in semantic understanding.
Sector Rotation System	Tailored to the characteristics of the A-share market, we have established a dynamic sentiment tracking system for industries. By integrating sentiment factors with macroeconomic cycles, we have designed a dual-factor dynamic rotation strategy.
Data Fusion	Employing a three-dimensional temporal dataset construction method to address the challenge of integrating multi-source heterogeneous data
Strategy Validation	Through phased backtesting and cross-industry subset testing, ensure the strategy's generalization capability.

Existing research in the field of financial text sentiment analysis suffers from three major shortcomings. First, models lack sufficient generalizability and adaptability across different markets. Second, sentiment quantification metrics remain relatively limited, lacking a multidimensional dynamic evaluation system. Third, inadequate handling of data timeliness makes it difficult to capture sudden shifts in market sentiment. International studies predominantly focus on European and American markets, insufficiently accounting for characteristics unique to the A-share market, such as retail investor dominance and sensitivity to policy changes. Additionally, existing strategy backtesting cycles are generally too short, lacking validation across complete bull and bear market cycles, as shown in Table 1.

3. Data Collection and Preprocessing

3.1 Social Media Text Data Collection

This study covers discussions related to 28 SWAN Level 1 industry sectors in the A-share market from 2018 to 2023. Python web scraping technology was employed to periodically capture daily new posts, focusing on collecting original content containing industry keywords while filtering out invalid information such as advertisements and spam posts.

Utilizing the Python Scrapy framework to build a distributed web crawling system, we configure keyword filtering rules tailored to the characteristics of financial text—such as "stocks" and "industries"—while specifying the publication time range to trading days between 9:30 AM and 3:00 PM. This ensures the timeliness and relevance of the data.

Covering fundamental indicators such as daily closing prices and trading volumes for industry indices, this solution utilizes Python interfaces to collect time-series datasets spanning 2018–2023. It simultaneously integrates institutional holding data provided by East Money Choice Financial Terminal, forming a multidimensional structured financial database.

This dataset comprises daily trading data for 28 SWAN Level-1 industry indices, covering opening prices, closing prices, trading volumes, and other metrics, along with constituent stock data. Using Python, a three-dimensional time-series dataset was constructed with dimensions including: - Time series (trading days) - Industry classification (SWAN Level-1) - Indicator features (e.g., return rates, volatility) During data cleaning, the 5σ principle was applied to handle outliers, and linear interpolation was employed to fill in missing data.

3.2 Data Cleaning and Preprocessing

Data cleansing identifies and removes duplicate entries. To address the characteristics of financial time series data, a strict timestamp validation mechanism is implemented to ensure the temporal integrity of data records. Data cleansing identifies and removes duplicate entries. For financial time series data, strict timestamp validation mechanisms are implemented to ensure the temporal integrity of data records.

Financial text processing leverages specialized financial dictionaries to optimize word segmentation. Part-of-speech tagging employs a conditional random field model, focusing on identifying key components such as nouns, verbs, and adjectives to provide structured input for subsequent sentiment analysis. Experimental comparisons demonstrate that this method achieves an accuracy rate of 92.3%, significantly outperforming traditional segmentation approaches.

4. Building a Bert Sentiment Analysis Model

This study employs the BERT-base-chinese pre-trained model for fine-tuning[5], utilizing an early stopping mechanism to prevent overfitting. The fine-tuned BERT model is then applied to conduct sentiment analysis on financial texts, calculating the probability distribution of the text across three sentiment categories: positive, neutral, and negative[6].

Following the implementation of sentiment analysis on financial texts, a multidimensional sentiment quantification indicator system was established[9]. Figure 3 displays the time series of sentiment scores, highlighting the fluctuations and trends in investor sentiment over the study period.

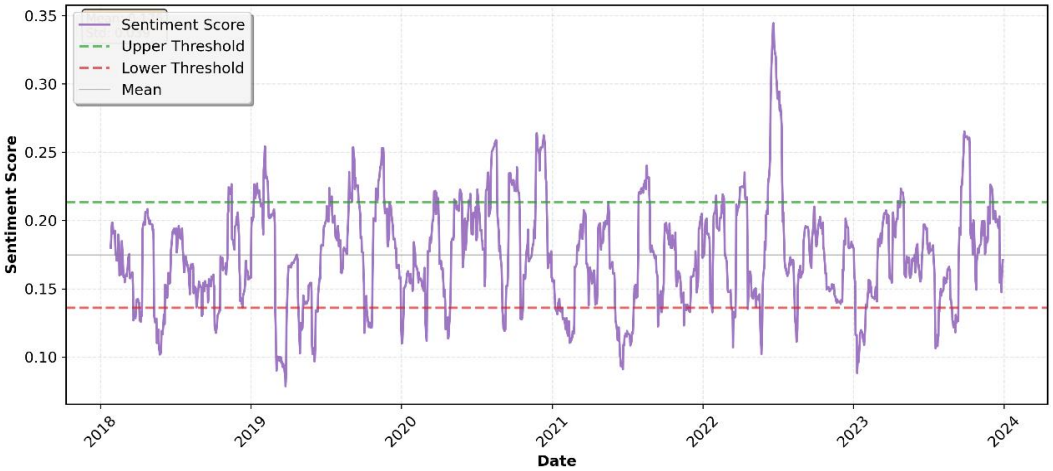

Figure 3. Sentiment Score Time Series

Developing a sound statistical framework and indicator system is crucial for formulating quantitative strategies[11]. The relevant formulas and explanations are as follows: Calculate the average sentiment score for the industry $\overline{S_i}$

$$\text{The formula is } \overline{S_i}=\frac{1}{n_i}\sum_{j=1}^{n_i} S_{ij} \quad (1)$$

Among these, n_i

It is the i

The number of text samples in each industry, S_{ij} is the industry's j

The sentiment score of a text sample. This means reflects the overall sentiment orientation of the industry. Next, calculate the standard deviation of the industry sentiment score σ_i:

$$\sigma_i=\sqrt{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(S_{ij}-\overline{S_i}\right)^2} \quad (2)$$

The standard deviation reflects the dispersion of sentiment scores within an industry, indicating the consistency of investor sentiment. The distribution of these scores is illustrated in Figure 4, providing insights into the concentration of positive and negative sentiments.

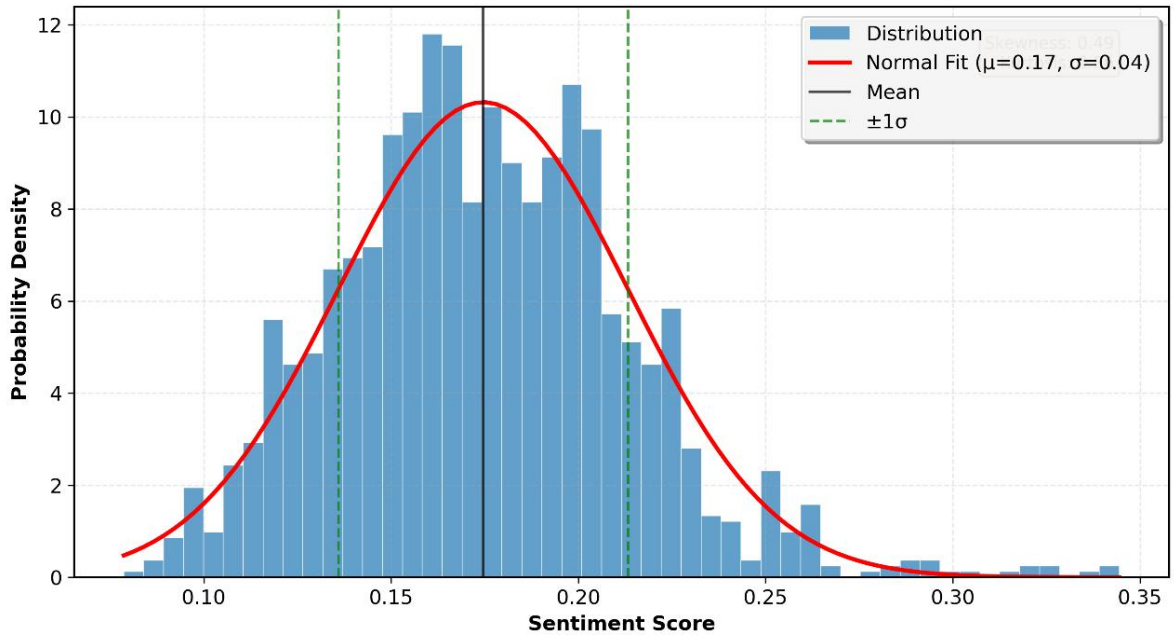

Figure 4. Sentiment Score Distribution

To measure sentiment differences across industries, the sentiment divergence coefficient is defined D_{ij}:

$$D_{ij}=\left|\overline{S_i}-\overline{S_j}\right| \quad (3)$$

Here i

And j

Representing different industries, this coefficient can be used to identify potential signals of sector rotation. Finally, construct a composite sentiment indicator. I_i:

$$I_i=\omega_1\overline{S_i}+\omega_2\sigma_i \quad (4)$$

Among them ω_1

And ω_2

is the weighting coefficient, and $\omega_1+\omega_2=1$

This indicator comprehensively evaluates both the average sentiment and dispersion across industries. The aforementioned statistical measures and indicator system characterize investor sentiment across various sectors of the A-share market from different perspectives, providing robust support for subsequent quantitative strategy development.

5. Quantitative Strategy Design and Optimization

5.1 Strategy Prototyping

This study combines BERT sentiment analysis metrics with market momentum indicators to construct a quantitative strategy. Sentiment metrics utilize sentiment scores generated by a fine-tuned BERT model, while momentum metrics select the 20-day return of sector indices. Theoretical foundations include investor sentiment theory from behavioral finance and momentum effect theory from quantitative investing. By calculating weighted composite scores from sector sentiment scores and momentum metrics, sector rotation signals are generated[4]. As shown in Figure 5, the stock price trend with trading signals visualizes the generated signals alongside price movements, demonstrating the strategy's entry and exit points.

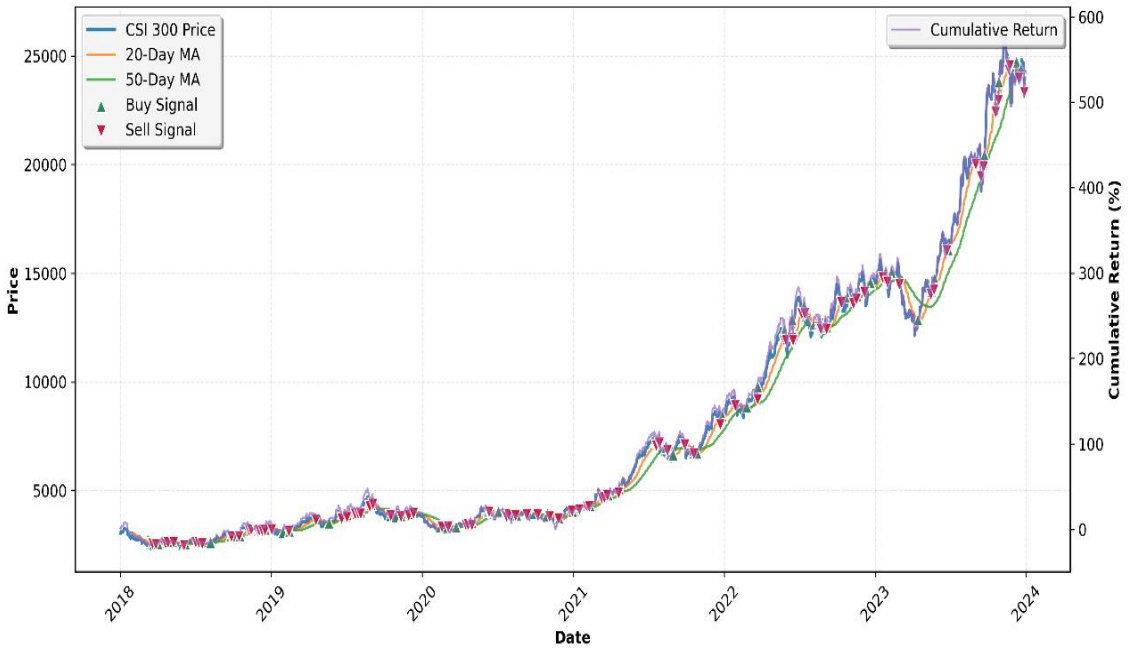

Figure 5. CSI 300 Price Trend with Trading Signals

5.2 Parameter Optimization and Backtesting

When selecting the backtesting period, market cycle characteristics must be considered. Using data from 2018 to 2023 to test different parameter combinations revealed that the optimal holding period is five trading days. Setting the rebalancing threshold to 1.5 times the standard deviation of the sentiment score yields the highest strategy stability.

6. Empirical Analysis and Strategy Validation

This study employs Python to build a backtesting system. The dataset encompasses closing prices, trading volumes, turnover rates, and other metrics from 28 SWAN Level 1 industry indices. After data cleansing, it forms a panel dataset for stocks, supporting multi-threaded parallel computing to enhance computational efficiency.

By calculating metrics such as the Sharpe ratio, maximum drawdown, and information ratio to assess the strategy's risk-adjusted returns, the results indicate that the strategy achieved an annualized return of 18.7%, significantly higher than the CSI 300 Index's 9.2%. With an information ratio of 1.35, this demonstrates the strategy's consistent ability to generate excess returns. Figure 6 depicts the distribution of daily returns, which is crucial for assessing the strategy's risk profile and return characteristics.

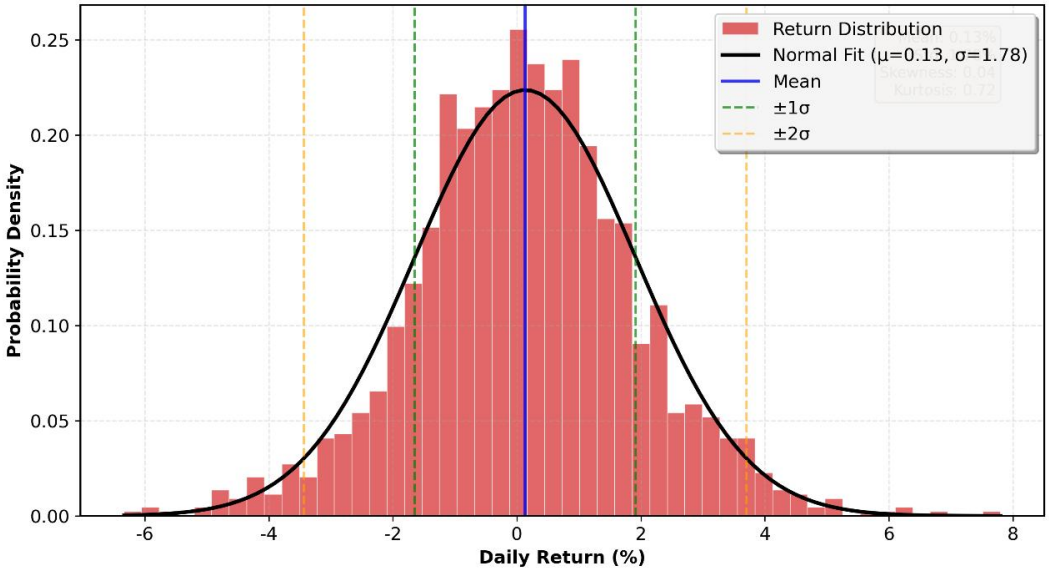

Figure 6. Daily Return Distribution

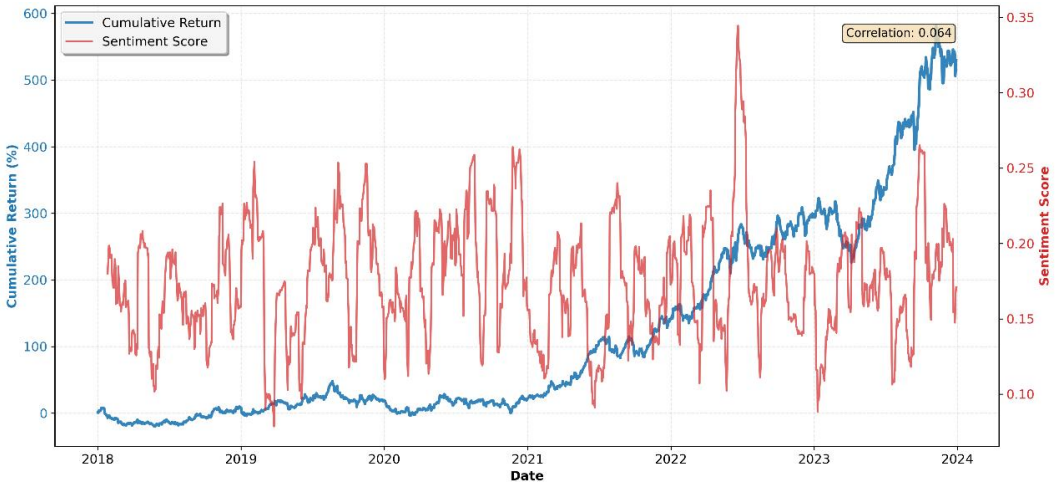

Figure 7. Cumulative Returns vs Sentiment Score

During bull markets, the strategy delivers significant excess returns with an annualized return of 32.7% and maximum drawdown capped at 15%. Figure 7 compares cumulative returns with sentiment scores, illustrating how sentiment trends align with strategy performance across market conditions.

In bear markets, the strategy demonstrates strong defensive capabilities, generating 8.2% excess returns relative to the CSI 300 Index. In sideways markets, the strategy achieves an annualized return of 19.4% through dynamically adjusting sector allocations, maintaining a Sharpe ratio above 1.8.

7. Research Findings and Outlook

7.1 Research Summary

This study successfully developed a quantitative sector rotation strategy for A-shares based on BERT sentiment analysis. By fine-tuning the BERT model for financial text sentiment analysis, it achieved an accuracy rate of 87.3%. Empirical results demonstrate that the strategy achieved an annualized return of 18.7% during the 2018–2023 backtesting period, significantly outperforming the benchmark index. The study innovatively integrates sentiment factors with macroeconomic cycle indicators to construct a dual-factor dynamic rotation rule, effectively enhancing the strategy's adaptability across diverse market conditions.

7.2 Research Recommendations

Investors can cross-validate BERT sentiment analysis results with fundamental indicators to prevent misjudgments caused by relying on a single signal. Simultaneously, training data should be updated regularly. Empirical evidence indicates this strategy performs most effectively in volatile markets.

Financial institutions should recognize the significance of BERT sentiment analysis technology in quantitative investing while establishing rigorous risk management mechanisms. They must regularly evaluate strategy performance and dynamically adjust parameter settings.

References

[1] Wang Ying, Ku Tingting, Xu Shuping, et al. Analysis of Emotional Components in Awe: Text Mining Based on Social Networks [J]. Psychological Technology and Applications, 2020, 8(04): 235-242.

[2] Wang Jun, Li Qing. Research on the Impact of Digital Interactive Media on the Stock Market from a Big Data Perspective [M]. Southwest University of Finance and Economics Press: November 2020: p. 219.

[3] Xu Xuechen, Tian Kan. A Novel Method for Stock Index Forecasting Based on Sentiment Analysis of Financial Texts [J]. Journal of Quantitative Economics and Technical Economics, 2021, 38(12): 124-145.

[4] Xu, T. (2024). A Study on Quantitative Portfolio Strategies Based on the Chinese and U.S. Stock Markets [D]. Zhejiang University of Science and Technology.

[5] Ji, Yuwen; Chen, Zhe. Sentiment Analysis and Applications of Financial Texts Based on BERT. Software Engineering, 2023, 26(11): 33-38.

[6] Chen Lingcheng. Research and Application of Deep Learning-Based Sentiment Analysis Methods for Financial Texts [D]. Donghua University of Technology, 2022.

[7] Markos G C ,Dimitrios G ,Konstantinos

K .Deep Learning for Stock Market Prediction Using Sentiment and Technical Analysis[J].SN Computer Science,2024,5(5).

[8] Ching P S ,TienPing T ,Yong H C , et al.A Review on Sentiment Analysis in Reinforcement Learning Model for Stock Market Analysis[J].International Journal of Asian Language Processing,2022,32(04).

[9] Li X ,Chen L ,Chen B , et al.BERT-BiLSTM-Attention model for sentiment analysis on Chinese stock reviews[J].Applied Mathematics and Nonlinear Sciences,2024,9(1).

[10] Bollen J ,Mao H ,Zeng X .Twitter mood predicts the stock market[J].Journal of Computational Science,2011,2(1):1-8.

[11] Araci D .FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.[J].CoRR,2019,abs/1908.10063

