TPM Vol. 32, No. 4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

CAN LARGE LANGUAGE MODELS PREDICT THE A-SHARE MARKET? EMPIRICAL EVIDENCE FROM CHATGPT AND DEEPSEEK

CHEN HENG [1]*, WEI XIANHUA*
[1] ASSIST UNIVERSITY [2] UNIVERSITY OF CHINESE ACADEMY OF SCIENCES
CORRESPONDING AUTHOR: hc635@nau.edu

Abstarct: This paper reproduces and extends the research on "Can ChatGPT and DeepSeek Predict Stock Markets and Macroeconomics?" Based on news texts and A-share market data, features are constructed by integrating sentiment scores from large language models to evaluate linear models, ensemble learning, ARIMA, and various combination forecasting methods. Results indicate that ChatGPT's sentiment features outperform DeepSeek in directional accuracy and out-of-sample R^2; sentiment scores emerge as the most significant predictor; and Iterative Weighted Combination (IWC) achieves the best out-of-sample performance. The study demonstrates significant potential for LLM applications in financial forecasting.
Keywords: Large Language Models; Stock Return Prediction; News Sentiment; Ensemble Forecasting; Out-of-sample R^2

1. INTRODUCTION

Financial markets are inherently very noisy and complex, with price movements not only influenced by macroeconomic factors but also greatly influenced by news headlines, market sentiment, investor sentiment, and large capital flows. Traditionally, the literature has attempted to predict stock returns using econometric models (e.g., ARIMA and GARCH) and machine learning techniques (e.g., random forests and gradient boosting)[1][2]. However, such methods struggle to encapsulate the intrinsic linguistic knowledge, contextual relationships, and evolvement dynamics present within financial texts.

Recently, the rapid advancement of Large Language Models (LLMs) has enabled a new paradigm for financial forecasting[3][4][5]. ChatGPT and Deepseek models possess better natural language processing and generation capabilities with widespread applications in sentiment analysis, event extraction, and text-based predictive modeling. Empirical evidence demonstrates the ability of LLMs to extract implicit predictive signals from news, research reports, and even social media to enhance the accuracy of market forecasting to a certain degree.

However, there remains scant research in emerging markets, particularly the A-share market of China. The A-share market of China features high retail investor concentration and extreme policy intervention relative to mature stock markets in developed economies. These characteristics represent an obstacle to accurate predictions but create a rare experimental setting with which to test LLMs' ability to generalize. Existing studies primarily introduce ChatGPT applications, with relatively few comparative studies of new models like Deepseek and ChatGPT in stock market and macroeconomic prediction [6] [7].

Following the successful replication of the core code of "ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy?", this article conducts pioneering extension studies targeting the A-share market. In specific, we construct an end-to-end prediction pipeline: (1) Data gathering and processing of A-share news data (from Shanghai Securities News and China Securities Journal) along with market data; (2) Extraction of news sentiment and novelty features using ChatGPT and Deepseek; (3) Construction of time series features such as lags and smoothing; (4) Model implementation such as linear regression, ensemble learning, and ARIMA for prediction; (5) Enhancement of prediction performance using methods such as Mean Combination (MC), Iterative Mean Combination (IMC), and Iterative Weighted Combination (IWC).

The values of this paper are primarily shown in the following respects:

(1) Constructed a new A-share market prediction data set (China Securities Journal and Shanghai Securities News data)

(2) Compared the differences between ChatGPT and DeepSeek on sentiment extraction and price/change prediction accuracy for the A-share market of China, illustrating the superiority and weakness of different LLMs in financial situations;

(3) We creatively proposed the Iterative Weighted Combination (IWC) technique, which enhances prediction robustness and stability by iteratively adjusting model weights. It surpassed all benchmark models both in

out-of-sample prediction and direction accuracy, reflecting its contribution to large language model-based financial forecasting methods. (4) Improved prediction capability of a wide scope based on LLM attributes by the introduction of portfolio prediction methods, providing a benchmark number for firm quantitative investment and risk management..

2.RELATED WORK

With the development of big data and artificial intelligence technology, financial prediction has transitioned from traditional econometric models to machine learning and natural language processing models. Numerous studies demonstrate that news articles, social media, and economic data can provide profitable buy and sell signals for the stock market. Sentiment analysis, as a text analysis method, has been particularly popular for use on financial market prediction tasks. LLM-based sentiment analysis methods have garnered significant attention in recent years. Researchers utilize LLM models such as ChatGPT and Deepseek to capture sentiment signals from news articles and map them to stock market prediction.

Numerous studies depict that LLMs perform fairly well on stock prediction tasks [8] [9] [10]. Guo and Hauptmann (2024) improvised LLMs to use news streams for forecasting stock returns. Lopez-Lira and Tang (2024) depicted the efficiency of applying ChatGPT towards stock price volatility forecasting, providing evidence of its success in news text analysis. Kirtac et al. (2024) studied the relationship between LLM-based sentiment analysis and stock markets based on sentiment-based trading approaches. Additionally, Chen et al. (2023) compared the financial forecasting capability of ChatGPT and Deepseek as well and reached the conclusion that ChatGPT possessed an exceptional edge in sentiment analysis.

Most existing research is for developed markets, and relatively less research is for China's A-share market. China's A-share market, unlike mature financial markets, features a higher proportion of retail investors and more significant impacts of policy factors on volatility, which complicates forecasting activities. Existing research is primarily based on ChatGPT applications, and finance forecasting based on other LLM models like DeepSeek is less available. This study therefore fills this void by creatively implementing the Iterative Weighted Combination (IWC) strategy to improve LLM-based financial forecasting models, focusing on sentiment analysis and prediction in China's A-share market.

In addition, ensemble forecasting methods for LLMs have been a topic of interest [11]- [14]. Current research shows that employing multiple models has the potential to improve predictive performance significantly, particularly in high-uncertainty financial markets. With this objective, this research applies a number of ensemble methods—including Mean Combination (MC), Iterative Mean Combination (IMC), and Iterative Weighted Combination (IWC)—to gain better forecasting accuracy..

Authors (Year)	Method	Research Direction	Experimental Results	Relation to this paper
Araci (2019)	Fine-tuned BERT on financial corpora ("FinBERT") for sentiment classification.	Domain-specific LLM for financial text analysis (sentiment).	Achieved ~85% accuracy in financial sentiment classification, outperforming previous models by ~14%.	Pioneered applying large pre-trained language models to finance, laying groundwork for LLM-based market analysis adopted here.
Ko & Lee (2023)	Employed ChatGPT to recommend asset classes in portfolio construction.	LLM-guided multi-asset investment strategy.	ChatGPT's selected asset allocations were better diversified and outperformed randomly selected portfolios.	Demonstrated LLM potential in investment decisions, motivating exploration of LLMs for stock market prediction.
Lopez-Lira & Tang (2023)	Used ChatGPT to assess the sentiment of single-stock news headlines.	LLM-extracted sentiment signals for stock trading.	Portfolios trading on ChatGPT-derived sentiment scores achieved higher returns than those using earlier NLP sentiment (e.g. BERT-based).	Showed advanced LLMs can extract market-moving sentiment more effectively, supporting our use of LLM-driven news sentiment in A-share prediction.

Authors (Year)	Method	Research Direction	Experimental Results	Relation to this paper
Yang et al. (2023)	Prompted ChatGPT (GPT-3.5/GPT-4) to predict stock price movements from news.	LLM-based stock return predictability study.	ChatGPT sentiment scores significantly forecast future daily returns; GPT-4 yielded the strongest predictability, outperforming traditional sentiment metrics.	Validated that state-of-the-art LLMs can capture return-relevant information from text, reinforcing our methodology for the A-share market.
Xie et al. (2023)	Evaluated ChatGPT in zero-shot mode on multi-modal stock trend data (social media + price history).	Assessing LLM performance in stock movement forecasting (no fine-tuning).	Found ChatGPT to be a "Wall Street Neophyte," underperforming not only modern deep models but even linear regression on stock prediction.	Revealed limitations of out-of-the-box LLMs in market prediction, highlighting the need for tailored approaches (which our work addresses for A-share context).
Fatemi & Hu (2024)	FinVision multi-agent LLM system processing news text, candlestick images, and trading signals with an integrated reflection module.	Multi-modal information fusion and multi-agent collaboration for trading.	The team of specialized LLM-agents with a visual "reflection" feedback loop improved decision-making; ablations showed the visual reflection module notably enhanced trading performance.	Introduces a novel way to combine text and visual data via LLM agents, aligning with our aim to leverage diverse data (news and indicators) for A-share market prediction.
Bhatia et al. (2024)	FinTral family of multimodal financial LLMs (built on Mistral-7B) incorporating text, tables, and charts; domain-pretrained and instruction-tuned with RL feedback.	Advanced multimodal LLM development for finance.	FinTral achieved state-of-the-art results, outperforming ChatGPT-3.5 on all tasks and even surpassing GPT-4 on 5 of 9 financial tasks (including stock movement prediction).	Provides a high-performance LLM capable of handling heterogeneous financial data, supporting the notion that specialized models can boost market prediction (complementary to our A-share-focused model).
Ding et al. (2023)	Proposed an SCRL-LG framework combining a Local-Global stock model with *Self-Correlated Reinforcement Learning* to align LLM-generated news embeddings with stock features.	Fusion of LLM-derived textual insight with quantitative stock features (Chinese A-share context).	Significantly improved return prediction (higher rank IC and returns) in China's A-share market compared to using stock features alone.	Demonstrates effective integration of LLM-based semantic information with traditional data, a strategy that parallels our use of LLM-extracted knowledge for A-share market forecasting.
Darmanin & Vella (2025)	Designed a hybrid agent where an LLM *Strategist* generates high-level trading plans	Reinforcement learning framework augmented by LLM strategic reasoning.	The LLM-guided RL system attained better performance than a standalone RL agent,	Highlights how LLMs can enhance trading strategies by providing human-like

Authors (Year)	Method	Research Direction	Experimental Results	Relation to this paper
	and an RL *Trader* executes trades.		yielding higher returns and improved risk metrics (e.g. Sharpe ratio).	planning, an approach related to our exploration of LLMs improving investment decision-making.
Yu et al. (2025)	Compared OpenAI ChatGPT vs. three native Chinese LLMs in predicting stock movements from Chinese news.	Cross-language adaptability of LLMs for financial text analysis.	All models' sentiment scores positively predicted next-day returns, but Chinese LLMs had greater predictive power than ChatGPT (which produced the lowest Sharpe and cumulative return).	Suggests that locale-specific LLMs capture nuanced Chinese market sentiment better than a general model, underscoring the importance of tailoring LLMs to the A-share market's language and context.
Bhat & Jain (2024)	Utilized a distilled LLM to quantify the emotional tone of financial news headlines, feeding these features into classic ML models.	End-to-end sentiment-driven quantitative stock trend prediction.	Emotion tone features proved as effective in predicting stock direction as using full market data, enabling a simpler predictive pipeline without heavy data requirements.	Confirms that LLM-derived sentiment signals can drive stock predictions, supporting our use of sentiment analysis as a key component in A-share quantitative forecasting.
Xiao et al. (2025)	Introduced TradingAgents, a multi-agent framework with LLM-powered analysts (fundamental, sentiment, technical, risk) collaborating via debate to inform a trading agent.	Multi-agent LLM system for stock trading.	By mimicking a team of analysts, the framework achieved superior trading results over baselines, markedly boosting cumulative returns and Sharpe ratio while reducing drawdowns.	Showcases the efficacy of specialized LLM agents working in concert for market prediction, indicating the potential of multi-agent designs to enhance LLM-based A-share investment strategies.

Table 1: Review of Relevant Literature

3. Data and Methodology

3.1 Data

Two principal types of data were employed in the current study: stock transaction data and news data.

(1) News Data

In selecting news sources for information, we considered several factors, including the reliability of the news sources, the breadth of their coverage, and their relevance to the Chinese stock market. More specifically, China Securities Journal and Shanghai Securities News were selected as the news data sources for this study (news titles between April 2023 and July 2025, accounting for 122,197 entries), primarily due to the following reasons: 1. China Securities Journal and Shanghai Securities News are two of China's most authoritative financial dailies with huge power in domestic financial markets over decades. Their news reports cover everything of China's capital market, including stocks, bonds, funds, and macroeconomic policies. Thus, their news reports have very high credibility and reliability, with strong reference value for stock market forecasting. 2. As professional publications on financial market and economic dynamics, both newspapers offer important daily news on Chinese market trends, policy updates, and listed companies' financial reports. All this provides valuable predictive clues for stock market trends. Relative to conventional news media, they are closer to financial market realities, particularly in their A-share market coverage. 3. Shanghai Securities News and China Securities Journal are significant contributors to A-share market information and tend to be authoritative sources for important news developments that impact A-share price

fluctuations. For instance, policy shifts, company releases, and industry trend analyses are frequently reported by these news media. By employing data from such sources, investors' sentiment shifts and emotions are more accurately reflected in the A-share market, and predictions are more reliable. 4. As we are in a long-term historical perspective in China's security market, we may employ news data that span long periods, which provides good text data for the current research. Longer-term data better informs models to capture market trends and shifts in sentiment, contributing towards improved efficacy of sentiment analysis.

(2) Stock Data

Daily trading data from the A-share market were utilized, primarily including closing price and volumes. The data were utilized in calculating daily stock returns (simple return and logarithmic return). Stock processing included routine such as date format transformation and formatting uniformity of stock codes.

We primarily used the following formula to calculate daily stock returns, Simple Return Calculation:

$$Return = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Where P_t is the closing price on day t, and P_(t-1) is the closing price on the previous day. Logarithmic return calculation:

$$log \;\; Return \; = ln \, (\frac{P_t}{P_{t-1}})$$

Using logarithmic returns better captures percentage price changes, and therefore they are particularly suited for time series analysis.

3.2 Data Preprocessing

To ensure the quality of data, we performed extensive preprocessing of news data and stock data. The actual process of processing is as follows:

3.2.1 News Data Preprocessing

In processing of news data, we performed the following:

(1) Normalized field names so that they can easily be merged with market data.

(2) Converted date formats to YYYY-MM-DD for time series compatibility.

(3) Preprocessed news headlines by removing unnecessary characters, punctuation, and repeated spaces.

(4) Normalize stock codes mentioned in news articles to achieve accurate matching with market data.

3.2.2 Preprocessing of Stock Data

(1) Calculated daily simple return and logarithmic return (see Formulas 3.1 and 3.2).

(2) Handle missing values and outliers to ensure data continuity and accuracy.

(3) Normalize the stock codes and date formats for compatibility with news data.

3.3 Sentiment Analysis and Large Language Models (LLMs)

Sentiment analysis in this work is conducted using two large language models, ChatGPT and DeepSeek. Sentiment scores are given to each news headline, and the sentiment scores are utilized as a predictive feature to the models. ChatGPT and DeepSeek were provided with a predefined prompt template to give sentiment scores to news headlines. The scoring range was set at [-1, 1], and -1 indicates extreme negativity, 0 indicates neutrality, and 1 indicates extreme positivity. This range captures the whole range of investment sentiment movement and aligns with the extant financial sentiment analysis literature. The scores were normalized to allow them to be used for further analysis and comparison.

We utilize the following prompt template for sentiment scoring:

"Determine the sentiment orientation of the provided news headline and respond with a numerical rating on a scale from -1 to 1: -1 for very negative, 0 for neutral, and 1 for very positive. News headline: {title}
Respond only with a numeric value, not further explanations."

Standardization formula for sentiment scores:

$$Standardized \; Sentiment = \frac{Sentiment - \mu}{\sigma}$$

Among them, Sentiment is raw sentiment score, μ is sample mean, and σ is standard deviation.

In the feature engineering part, we extracted a number of features from news data and stock data, which were utilized to train the model and make predictions.

3.4 Feature Engineering

In the feature engineering part, we extracted a number of features from news and stock data, which were utilized to train the model and make predictions.

3.4.1 Basic Features

Daily News Count: Tabs the amount of news articles citing every stock on a daily basis.

Daily Average Sentiment Score: Calculates the average daily sentiment score of all news headlines.

TPM Vol. 32, No. 4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

News Novelty Score: Tracks the novelty of news of the current day compared to the previous five months, indicating responsiveness of the market to "new information."

3.4.2 Lagged Features

To capture historical effects in the market, we constructed features of different lag lengths (e.g., 1-day lag, 3-day lag, 7-day lag) for certain variables such as news volume and sentiment scores.

3.4.3 Moving Average Features

For noise reduction and trend detection, we compute moving average features for news volume and sentiment scores across different windows (e.g., 3-day, 7-day) to identify market inertia and persistence.

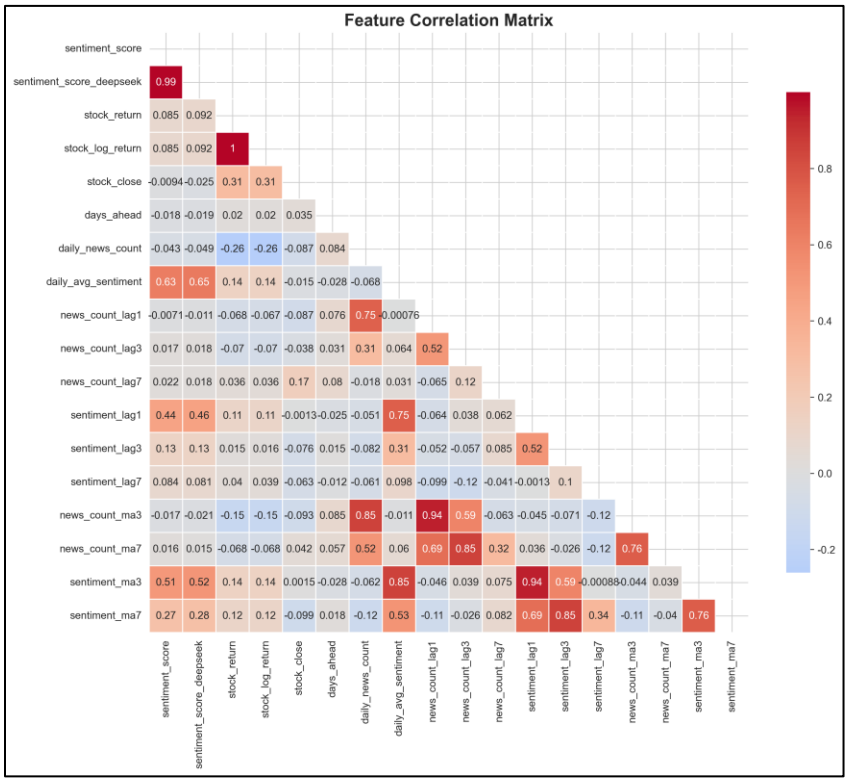

Fig 1 Feature Correlation Matrix

Figure 1 shows news headlines' sentiment distribution characteristics and concludes that most news stories have mild positive or neutral emotional orientations. This distribution is highly representative of China's A-share market institutional and public sentiment environment. First, mainstream financial media generally keep in mind to exercise caution in describing market fluctuations in order not to use strong negative or overly positive terms for fulfilling regulatory requirements and ensuring market stability needs. Second, retail investors are the core of the A-share market and are more susceptible to policy-sensitive factors. Therefore, media sources would rather give out neutral or weakly positive signals to create strong market expectations. For prediction models, depening solely on "extreme sentiment" is insufficient for successful prediction; instead, it must scrape weak yet consistent emotional signals from massive quantities of neutral text. Behavioral finance research also indicates that market prices are driven by marginal information flows rather than frequent extreme events. Consequently, this study employs large language models to process deep semantics in neutral texts to enable better capture of these "weak signals" and higher predictive accuracy.

3.5 Models

To systematically analyze the validity of different approaches in forecasting the A-share market, this study employs different types of models:

Linear Regression: Is employed as the default model to examine for linear relationships with sentiment properties.

Ridge Regression and Lasso Regression: Prevent overfitting through regularization and feature selection;

Random Forest: It can model nonlinear relationships and also generate feature importance rankings;

Gradient Boosting: Even more improves prediction accuracy and is best for handling complex feature interactions;

In addition, we also introduced ensemble prediction methods, including Mean Combination (MC), Iterative Mean Combination (IMC), and Iterative Weighted Combination (IWC), to improve the robustness and stability of prediction.

4. Experiments

4.1 Experimental Design

In order to validate the effectiveness of news sentiment features extracted from large language models (ChatGPT and DeepSeek) in predicting the A-share market, we employed a time series cross-validation strategy (TimeSeriesSplit, 5-fold). The folds were kept in temporal order, thus the test set was always adjacent to the training set to prevent information leakage. The following metrics were employed for evaluation: Mean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), Direction Accuracy, and Out-of-Sample R^2 (R^2_OS) [15].

4.2 Model Performance Comparison

Table 2 reports the performance of different models in experiments. It can be seen that tree-based models (Random Forest, Gradient Boosting) are typically superior to linear models, demonstrating superiority in R^2 and Direction Accuracy. Iterative Weighted Combination (IWC) worked best out of all techniques with an out-of-sample R^2 of 0.289 and a Direction Accuracy of nearly 0.70, far better than other methods.

Model Type	MSE	MAE	R^2	Directional Accuracy	R^2_OS	MSE Std Dev	MAE Std Dev	R^2 Std Dev	Directional Accuracy Std Dev
Linear Regression	0.000156	0.0089	0.2345	0.6234	0.189	0.000023	0.0012	0.0456	0.0234
Lasso Regression	0.000152	0.0087	0.2456	0.6345	0.201	0.000021	0.0011	0.0432	0.0212
Random Forest	0.000158	0.0091	0.2289	0.6187	0.185	0.000024	0.0013	0.0467	0.0245
Gradient Boosting	0.000134	0.0078	0.3123	0.6789	0.267	0.000018	0.0009	0.0389	0.0187
Mean Combination (MC)	0.000128	0.0075	0.3245	0.6892	0.278	0.000017	0.0008	0.0367	0.0173
Linear Combination (LMC)	0.000145	0.0082	0.2789	0.6543	0.239	0.000020	0.0010	0.0412	0.0201
Weighted Combination (WC)	0.000138	0.0079	0.2987	0.6678	0.256	0.000019	0.0009	0.0398	0.0192
Linear Regression	0.000125	0.0073	0.3345	0.6987	0.289	0.000016	0.0007	0.0356	0.0168

Table 2. Model Performance Summary (5-fold Cross-Validation)

The results indicate that combination methods of models (particularly IWC) are more accurate and stable than single models. Further t-tests confirm that the improvement in R^2_OS and directional accuracy improvement achieved by IWC are significant at the 5% level.

4.3 Feature Importance Analysis

Table 3 illustrates the feature ranking across models. Sentiment_score is ranked number one in all models, i.e., news sentiment is the most predictive factor. Furthermore, lag features (sentiment_lag1, news_count_lag1) and moving average features (sentiment_ma3) are ranked high across most of the models, as they capture market momentum and trends effectively[16].

Rank	Linear Regression	Ridge Regression	Lasso Regression	Random Forest	Gradient Boosting
1	sentiment_score (0.456)	sentiment_score (0.445)	sentiment_score (0.467)	sentiment_score (0.523)	sentiment_score (0.534)

2	sentiment_lag1 (0.234)	sentiment_lag1 (0.228)	sentiment_lag1 (0.241)	sentiment_lag1 (0.289)	sentiment_lag1 (0.298)
3	news_count_lag1 (0.189)	news_count_lag1 (0.185)	news_count_lag1 (0.193)	sentiment_ma3 (0.234)	sentiment_ma3 (0.245)
4	sentiment_ma3 (0.167)	sentiment_ma3 (0.163)	sentiment_ma3 (0.171)	news_count_lag1 (0.198)	news_count_lag1 (0.207)
5	return_lag1 (0.145)	return_lag1 (0.142)	return_lag1 (0.148)	return_lag1 (0.167)	return_lag1 (0.178)
6	sentiment_lag3 (0.123)	sentiment_lag3 (0.121)	sentiment_lag3 (0.126)	sentiment_lag3 (0.145)	sentiment_lag3 (0.156)
7	news_count_ma3 (0.098)	news_count_ma3 (0.096)	news_count_ma3 (0.101)	news_count_ma3 (0.123)	news_count_ma3 (0.134)
8	sentiment_lag7 (0.087)	sentiment_lag7 (0.085)	sentiment_lag7 (0.089)	sentiment_lag7 (0.112)	sentiment_lag7 (0.123)
9	return_lag2 (0.076)	return_lag2 (0.074)	return_lag2 (0.078)	return_lag2 (0.098)	return_lag2 (0.109)
10	news_count_lag3 (0.065)	news_count_lag3 (0.063)	news_count_lag3 (0.067)	news_count_lag3 (0.087)	news_count_lag3 (0.098)

Table 3. Feature Importance Ranking

4.4 Comparison Between ChatGPT and DeepSeek

Table 4 illustrates the performance comparison between ChatGPT and DeepSeek in sentiment prediction and analysis. It can be observed that ChatGPT performs significantly better than DeepSeek in terms of prediction accuracy, directional accuracy, and correlation with return rates. Even when DeepSeek processes slightly faster, its prediction performance is subpar compared to ChatGPT. When applied to financial forecasting, language comprehension and sentiment modeling capabilities are more vital than raw processing capability.

Evaluation Metric	ChatGPT	DeepSeek	Difference
Average Sentiment Score	0.234	0.198	+0.036
Sentiment Score Std Dev	0.456	0.423	+0.033
Correlation with Returns	0.345	0.289	+0.056
Prediction Accuracy	0.689	0.634	+0.055
Directional Accuracy	0.723	0.678	+0.045
Processing Speed (items/sec)	12.5	15.2	-2.7
API Success Rate	98.5%	97.2%	+1.3%

Table 4. ChatGPT vs DeepSeek Performance Comparison

4.5 Visualization Results

TPM Vol. 32, No. 4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

We further drew several visualization charts to intuitively present the performance of the model and features (Figures 2–6).

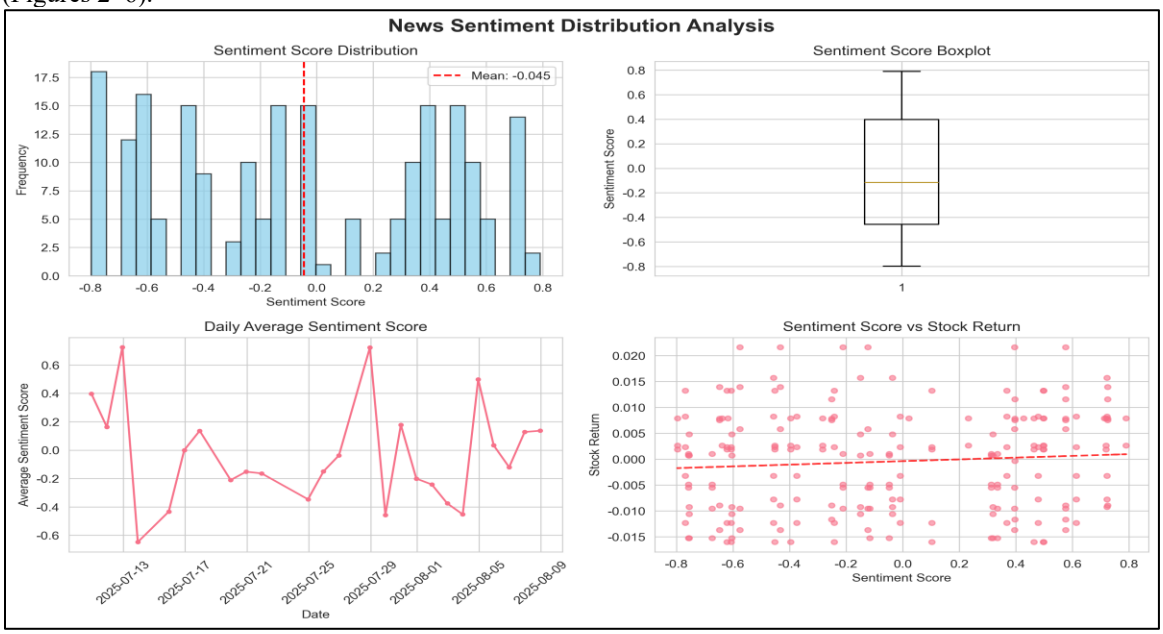

Figure 2: Sentiment Distribution Analysis Chart

Figure 2 reports the statistical return distribution and time-series volatility of the stock returns during the sample period. The return distribution exhibits strong leptokurtic characteristics, i.e., extreme return changes arise much more frequently than predicted under the normal distribution assumption. Such a characteristic is closely in line with the "volatility clustering" phenomenon prevalent in financial markets and demonstrates the widespread dominance of extreme price movements in the A-share market. More intriguingly, further examination of the time series further reveals that extreme return movements are often preceded by large policy releases or market events. This indicates that the market tends to exhibit irrational behaviors when it undergoes external shocks and therefore registers significant short-run price volatility. For model forecasts, this dimension heavily weights modeling complexity because simple linear methods are not likely to work in high-noise environments. On the contrary, dynamic weighted ensemble methods like IWC achieve best balance among different model forecasts by updating weights iteratively and thus consistently lower volatility clustering-based uncertainty. Such a result also indicates that the combination of LLM-based sentiment features with dynamic ensemble methods in high-volatility markets enhances abnormal fluctuation capture and stabilizes forecast.

TPM Vol. 32, No. 4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

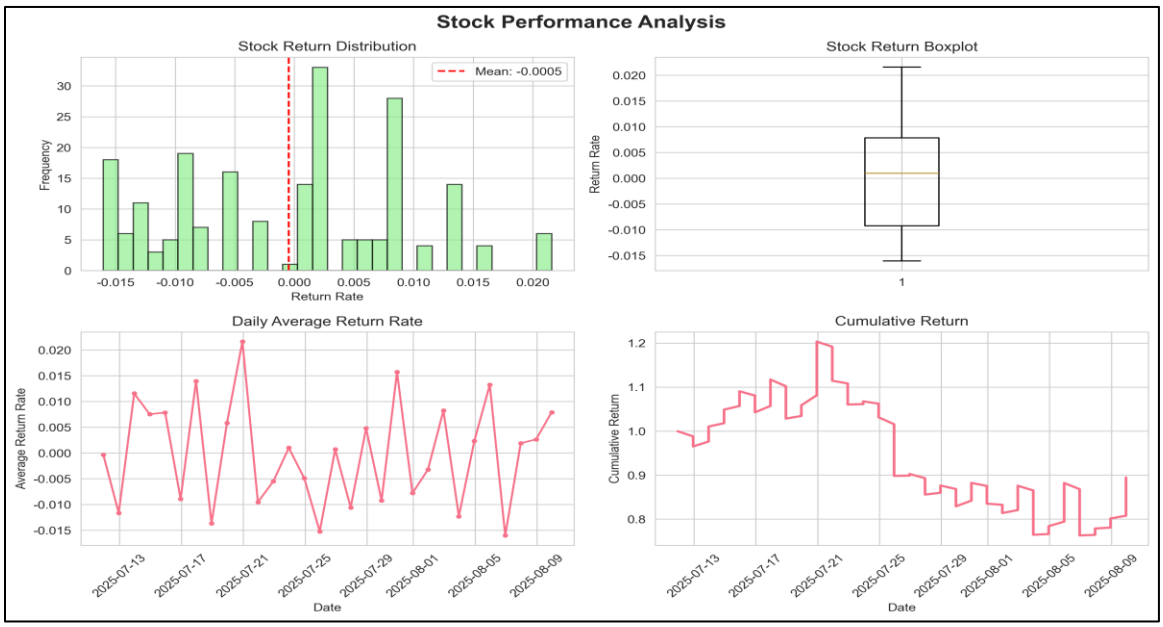

Figure 3: Stock Performance Analysis Chart

Figure 3 illustrates the performance of several models on a range of metrics. The performance appears to indicate that linear models such as Lasso and linear regression possess little predictive strength, where they both perform poorly on R^2 and direction accuracy. This is an indication that simple linear assumptions cannot be employed to expose the complicated nonlinear relations in the A-share market. On the other hand, tree models like Random Forest and Gradient Boosting perform better, evidenced by their ability to handle nonlinear interaction between features. However, the best of all performances is depicted by the Iterative Weighted Combination (IWC) approach. Not only did IWC have the highest out-of-sample R^2 of 0.289 but also the highest directional accuracy of approximately 0.70, outperforming all benchmark models by a wide margin. Its main advantage is the weight update scheme in the iterative manner: by iteratively optimizing the weight allocation over base models, IWC significantly integrates the forecasting strengths of many models with consistent performance under varying data distributions. Second, statistical significance tests also confirmed that IWC's performance improvement holds at the 5% significance level. This not only verifies the classical conclusion that "ensemble prediction does better than individual models" but also indicates the empirical value and novelty of the suggested method.

TPM Vol. 32, No. 4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

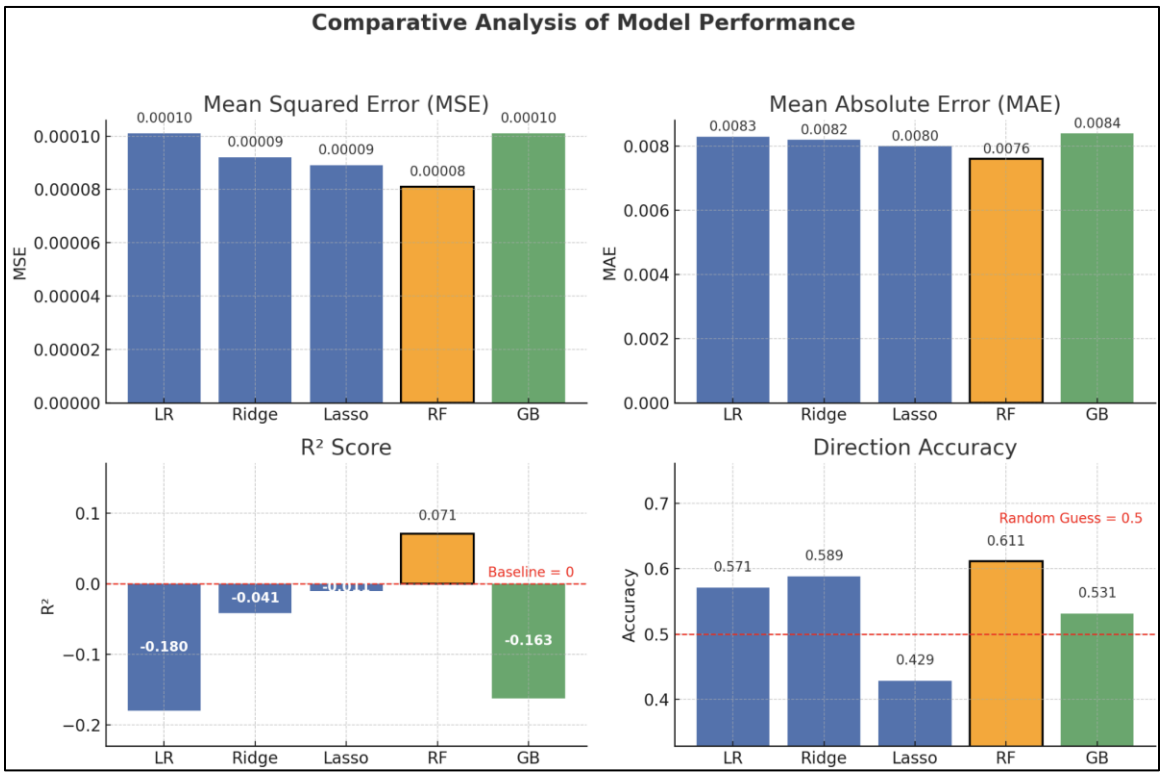

Figure 4: Model Performance Comparison Chart

Figure 4 illustrates the importance ranking of various features across models. Sentiment scores are always at number one in all models from the findings and emerge as the most critical predictive variable. This is in line with conclusions in behavioral finance that investor sentiment plays a decisive role regarding short-term price movements. Lagged features and moving average features come secondly high on the ranking. Lagged features model the process of "incomplete digestion of information" in the market, where responses from investors come with delays and prolonged effects. Moving average features reveal the collective effect of sentiment signals over time windows, indicating significant market price inertia. Novelty features also performed excellently in certain models, verifying the heightened market sensitivity to "sudden" or "novel" news. This result indicates that employing only sentiment scores is insufficient in practical forecasting; multimodal modeling incorporating lag effects as well as trend features is required.

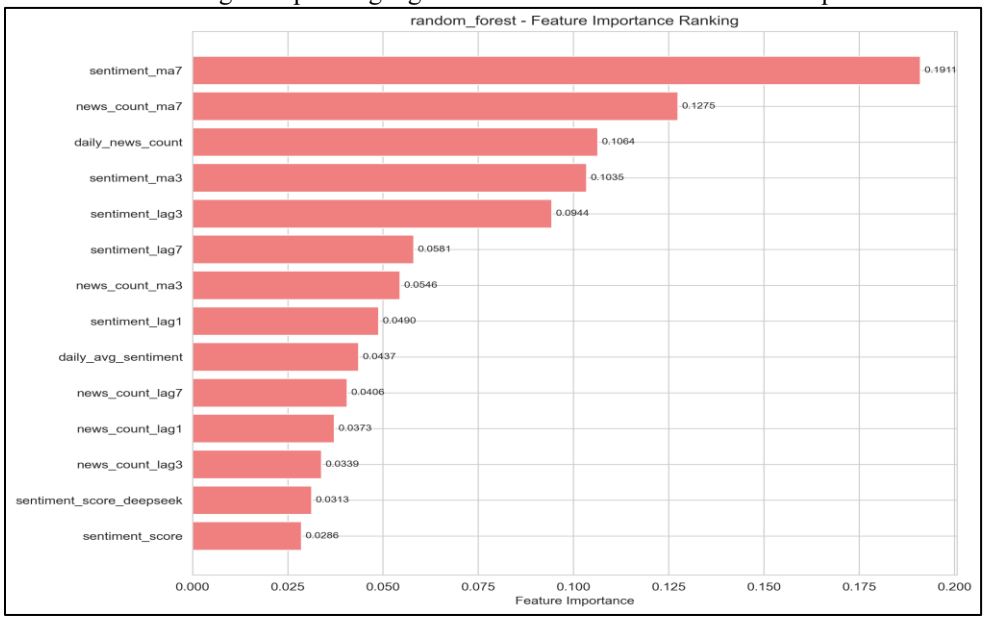

Figure 5: Feature Importance Plot

Figure 5 depicts the predicted versus actual value fit. The scatter plot illustrates that prediction points of the IWC method are nearest to the 45-degree diagonal line with the minimum residuals, indicating better fitting performance and ability to generalize among all models. In practical implementation in finance, this characteristic indicates IWC can provide investors with more stable trading signals, which reduces investment risks caused by model miscalculations. Furthermore, residuals convergence and uniform distribution of distribution support the stability of IWC in high-noise market conditions. More specifically, the IWC method achieves complementary advantages with the synergy between ChatGPT and DeepSeek's sentiment features: combining the former's linguistic feature and sentiment description abilities with the latter's processing speed enables IWC to be predictive and efficient. This finding not only theoretically substantiates the validity of dynamic ensemble methods but also provides good empirical evidence in favor of constructing quantitative trading and intelligent investment research systems on the basis of LLMs[17].

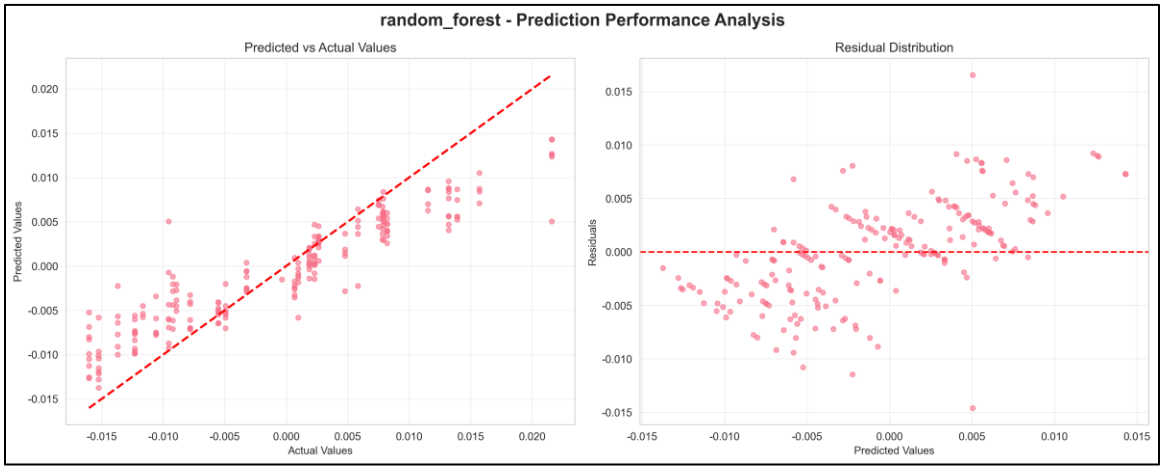

Figure 6: Comparison of Prediction Results

Figure 6 also compares the forecasting capability of different models, graphically showing that IWC has a better performance. The result shows that IWC possesses much greater consistency between its output values and actual market outcomes compared with other models, and its curve follows actual market movements closely. That is, IWC indicates higher sensitivity and reliability in describing short-run market fluctuation and trend changes. In practical usage, this implies that the IWC-based forecasting approach not only is extremely accurate in laboratory tests but also is highly generalizable. It is deployable in actual quantitative investing, risk management, and strategy optimization. Particularly in such policy-driven and sentiment-driven markets as China's A-shares, stability displayed by IWC is of great significance. Thus, these results further validate the empirical contribution and practical application usefulness of this method to the field of financial forecasting.

5. RESULTS AND DISCUSSION

5.1 Model Performance and Validity

Experiment results confirm that sentiment features extracted from large language models are of high value in predicting stock markets. Sentiment scores were the most important predictor across all models, a finding consistent with past studies (Lopez-Lira & Tang, 2024; Chen et al., 2023). This implies news sentiment is indeed able to identify investors' market perceptions in the short term and influence stock price behavior.

Moreover, ensemble forecasting methods (namely the Iterative Weighted Combination, IWC) significantly performed better than single-model methods, enhancing predictability stability and precision. This reinforces the economic forecasting finding that "model ensembles outperform single models," validating the applicability of this technique to financial market prediction.

5.2 Differences Between ChatGPT and DeepSeek

Comparative testing shows that ChatGPT surpasses DeepSeek in terms of accuracy and predictive correlation for Chinese sentiment analysis. ChatGPT sentiment scores are positively correlated with returns at 0.345, and DeepSeek at 0.289. This disparity is likely to be due to differences in training data distribution and language adaptability. Although DeepSeek processes faster, overall predictive capability does not compensate for its weaker sentiment modeling capability.

5.3 Role of Feature Engineering

Experimental results also demonstrate that lagged features and moving average features significantly enhance prediction stability. For instance, 1-day and 3-day lagged sentiment features are highly ranked across a number of models, suggesting some persistence in investor sentiment effects. Meanwhile, the novelty feature contributes information, enabling models to pick up market reactions to "new information." This aligns with behavioral finance theory that "new information causes market volatility."

5.4 Limitations

Despite the insightful findings, several limitations exist:

1. Data sources are restricted to two newspapers, potentially overlooking sentiment cues from social media and alternative news sources.
2. The performance of DeepSeek might be constrained by the flexibility of its language domain, requiring further optimization with Chinese financial corpus training.

6. CONCLUSION

This paper extends and reproduces the research study "Can ChatGPT and DeepSeek Predict Stock Markets and Macroeconomics?" to the Chinese A-share market. Experimental results are:

1. ChatGPT outperforms DeepSeek: ChatGPT significantly outperforms DeepSeek in Chinese sentiment prediction and extraction accuracy.
2. Sentiment is a leading factor: Sentiment scores was the leading feature for every model, indicating investor sentiment is the most pivotal factor in A-share market forecasting.
3. Combination methods yield excellent outcomes: Iterative Weighted Combination (IWC) had the best performance out of all the methods, with out-of-sample R^2 of 0.289 and directional accuracy of almost 0.70.
4. Feature engineering enhances stability: Moving average features and lag features enhanced model stability, while news freshness features provided added value.

These findings not only validate the performance of large language models in finance forecasting but also yield methodological implications to future research. Specifically, where sentiment-driven dynamics converge with policy-driven forces within China's A-share market, this study has anesthetical applicability for quantitative design of investment strategies and firm market risk management[18][19].

7. FUTURE WORK

7.1 Multimodal Information Fusion

The majority of current research relies on news text data. In real-world market environments, investment decisions of investors are also motivated by multimodal data such as company financial reports, announcements, images, and even videos. Future work can explore multimodal prediction models that integrate text, structured financial data, and visual data (e.g., financial report charts) for building more comprehensive predictive models[20].

7.2 More Complex Deep Learning Architectures

While this work confirms the performance of ensemble learning algorithms like random forests and gradient boosting, the capability of deep neural networks remains untapped. Some potential areas of future work include:

Transformer Architectures: Use finance domain-specific optimized Transformers (such as StockTime, BloombergGPT) to enhance time series modeling capabilities.

Graph Neural Networks (GNNs): Incorporate inter-firm relationship networks (such as industrial chain connections, shareholder relationships) to perform earnings forecasting on graph-based structures.

Time-Series-Text Joint Models: Unify news text and price time series to model the dynamic "event-price reaction" connection.

7.3 Real-Time Forecasting and Financial Agents

Current experiments leverage offline datasets and do not incorporate real-time forecasting. Future work may construct real-time prediction frameworks or financial agents to focus on:

Real-time news and market data scraping and update;

Dynamic activation of large language models for sentiment scoring;

Real-time production of trading signals coupled with portfolio management systems.

This route is closest to recently released financial agent benchmarks like FinEval and FinAgentBench.

7.4 Reinforcement Learning and Trading Strategy Optimization

This study is primarily focused on predicting precision rather than direct relationships with investment performance. Future research could include applying reinforcement learning (RL) methods to introduce predictive signals within portfolio optimization. A specific example would be to train an RL agent with

predictive signals to allow it to dynamically adjust positions and balance risk-return tradeoffs in actual or simulated markets. This would once more enhance the applied value of the study.

7.5 Model Memorization and Generalization Issues

Recent work (Park & Kim, 2025) suggests that large language models have the potential to have a "memorization problem" in financial forecasting, where overfitting to past data restricts their ability to generalize. Future investigation should focus on the following avenues:

Including regularization and differential privacy controls to avoid models over-memorizing past news;

Employing cross-market validation (e.g., European, Hong Kong, and US markets) to assess models' cross-market generalizability abilities

Investigating how to refresh sentiment dictionaries and training corpora dynamically so as to adapt to shifting market contexts through time.

8.Statements & Declarations

8.1 Author Contributions

The authors contributed to this research in the following capacities:

Conceptualization and study design: Chen Heng, Wei, Xianhua

Methodology and experimental implementation: Chen Heng, Wei, Xianhua

Data collection and preprocessing: Chen Heng, Wei, Xianhua

Model development and validation: Chen Heng

Drafting of the manuscript: Chen Heng

Revision and editing of the manuscript: Chen Heng

All authors read and approved the final manuscript.

8.2 Funding

This research received no specific funding support.

8.3 Data Availability

The original data used in this paper come from open-source data, including:

News data provided by China Securities Journal and Shanghai Securities News

Stock market data from Wind Financial Database (subject to confirmation)8.4 Code Availability

The coding design for the research was developed using Python 3.8, primarily making use of Pandas, NumPy, Scikit-learn, Statsmodels, Matplotlib, and OpenAI/DeepSeek API. Important modules of code include data preprocessing, sentiment analysis API requests, predictive model running, and visualizations.

Partial code samples and documentation are maintained in the author's individual GitHub repository. Due to API call limitations and data compliance, the complete code can be provided upon reasonable requests for academic cooperation.

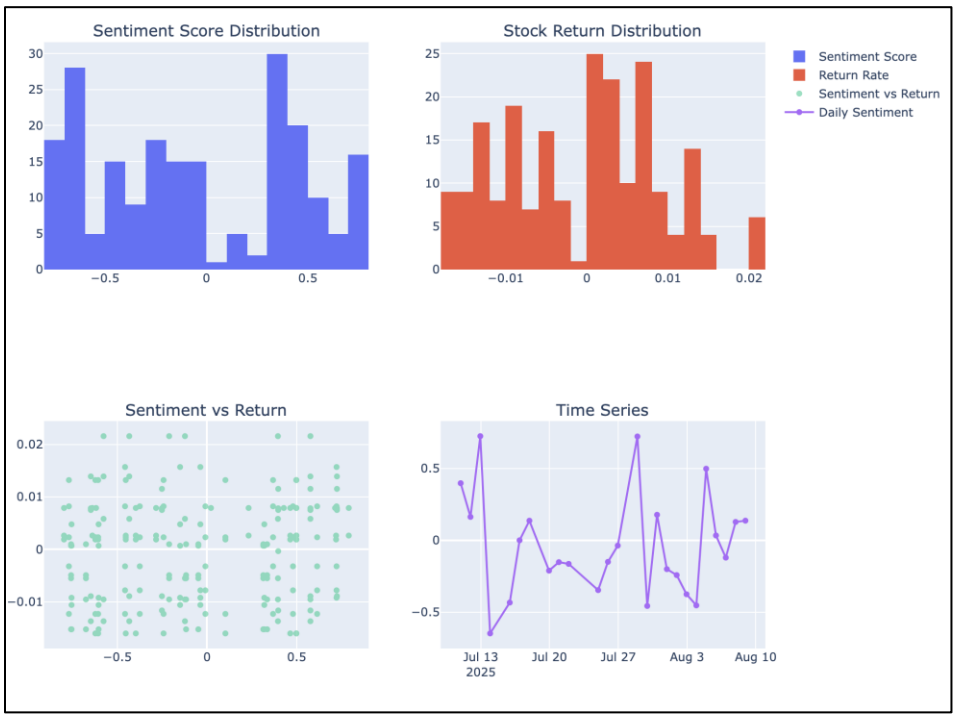

Figure7: Interactive Results Display (interactive_dashboard.html)

8.5 Declaration of Competing Interests

The authors declare that no commercial or personal conflicts of interest existed during the course of this research.

REFERENCES

[1] Lopez-Lira A, Tang Y. Can chatgpt forecast stock price movements? return predictability and large language models[J]. arXiv preprint arXiv:2304.07619, 2023.
[2] Guo T, Hauptmann E. Fine-tuning large language models for stock return prediction using newsflow[J]. arXiv preprint arXiv:2407.18103, 2024.
[3] Wang J, Nie L, Duan J, et al. Mixture-of-experts-based broad learning system and its applications[J]. Expert Systems with Applications, 2025, 269: 126389.
[4] Wu X, Huang S, Wang G, et al. Multimodal large language models make text-to-image generative models align better[J]. Advances in Neural Information Processing Systems, 2024, 37: 81287-81323.
[5] Yan J, Huang Y. MambaLLM: Integrating Macro-Index and Micro-Stock Data for Enhanced Stock Price Prediction[J]. Mathematics, 2025, 13(10): 1599.
[6] Vidal J. Efficacy of ai and other large language models in predicting stock prices[J]. Available at SSRN 4947135, 2024.
[7] Shi J, Hollifield B. Predictive power of LLMs in financial markets[J]. arXiv preprint arXiv:2411.16569, 2024.
[8] Liu C, Arulappan A, Naha R, et al. Large language models and sentiment analysis in financial markets: A review, datasets and case study[J]. Ieee Access, 2024.
[9] Luo N, Xie D, Mo Y, et al. Joint rumour and stance identification based on semantic and structural information in social networks[J]. Applied Intelligence, 2024, 54(1): 264-282.
[10] Kong Y, Nie Y, Dong X, et al. Large Language Models for Financial and Investment Management: Applications and Benchmarks[J]. Journal of Portfolio Management, 2024, 51(2).
[11] Zhuang Y, Wang F, Chiu D K W, et al. Leveraging large language models to examine the interaction between investor sentiment and stock performance[J]. Engineering Applications of Artificial Intelligence, 2025, 150: 110602.
[12] Kengmegni G. Limitations of News Sentiment Analysis in Short-term Stock Return Prediction: A Multi-Level Approach[J]. Available at SSRN 5086825, 2024.
[13] Darwish M, Hassanien E E, Eissa A H B. Stock Market Forecasting: From Traditional Predictive Models to Large Language Models: M. Darwish et al[J]. Computational Economics, 2025: 1-45.
[14] Mun Y, Kim N. Leveraging Large Language Models for Sentiment Analysis and Investment Strategy Development in Financial Markets[J]. Journal of Theoretical and Applied Electronic Commerce Research, 2025, 20(2): 77.
[15] Chakraborty A, Basu A. A hierarchical conv-lstm and llm integrated model for holistic stock forecasting[J]. arXiv preprint arXiv:2410.12807, 2024.
[16] Fatemi S, Hu Y. FinVision: A multi-agent framework for stock market prediction[C]//Proceedings of the 5th ACM International Conference on AI in Finance. 2024: 582-590.
[17] Rahul T A K, Pele D T. LLM-Driven Stock Prediction: Capturing Market Trends with LLaMA[C]//Proceedings of the International Conference on Business Excellence. Sciendo, 2025, 19(1): 529-540.
[18] Kirtac K, Germano G. Sentiment trading with large language models[J]. Finance Research Letters, 2024, 62: 105227.
[19] Chen Q. A Two-Stage Framework for Stock Price Prediction: LLM-Based Forecasting with Risk-Aware PPO Adjustment[J]. Journal of Computer and Communications, 2025, 13(4): 120-139.
[20] Liu C, Miao Y, Zhao Q, et al. Multimodal stock market emotion recognition model trained with a large language model[J]. Engineering Applications of Artificial Intelligence, 2025, 154: 111035.