5291 Final Project Report

Yinghao Liu yl5645
Yuchen Chen yc4502
Chenyang Ren cr3417
Henry Su hs3539

I. Introduction

Investor sentiment plays a pivotal role in shaping short-term market behavior, particularly in China's A-share and Hong Kong markets, where retail investors dominate trading activity. Unlike more institutionally driven markets, these sentiment-driven environments exhibit frequent deviations from fundamentals, elevated volatility, and momentum-driven reversals. Recognizing this, our project investigates whether sentiment indices can be systematically leveraged to improve the prediction accuracy of short-term stock prices.

To this end, we construct a quantitative sentiment index incorporating liquidity- and expectation-based indicators, such as trading volume, turnover rates, RSI, margin financing balances, and cross-border capital flows. Drawing on methods from prior research—including principal component analysis (PCA), partial least squares (PLS), and correlation-based filtering—we develop a composite index designed to capture both market mood and trading behavior. Exploratory data analysis reveals that peaks and troughs in sentiment often coincide with inflection points in stock prices, indicating potential forecasting value.

Building on this observation, we apply machine learning models—particularly XGBoost classifiers and LSTM networks—to evaluate the extent to which lagged sentiment indicators and market features can enhance the predictive accuracy of stock prices. Model performance is assessed using standard regression metrics such as R^2, RMSE, and MAPE. Our findings contribute to the broader literature on behavioral finance and financial forecasting, offering insights into how sentiment can be quantitatively integrated into price prediction frameworks for sentiment-sensitive markets.

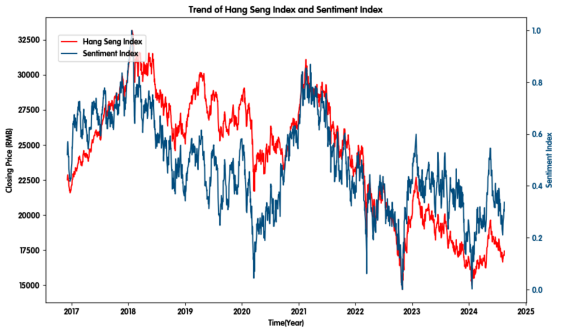

Figure 1: Trend of Hang Seng Index and Sentiment Index

The sentiment index frequently anticipates movements in the Hang Seng Index, highlighting its potential for enhancing market timing strategies. Both indices exhibit considerable volatility, with the Hang Seng Index fluctuating notably between highs around 33,000 HKD in early 2018 and lows near 21,000 HKD during early 2020. The sentiment index similarly ranged significantly, peaking around 0.8 during optimistic periods and dropping below 0.2 during downturns. Notably, the sharp downturn observed around early 2020 coincides closely with the onset of the COVID-19 pandemic, underscoring significant market disruptions. Post-2023, the indices reveal divergent paths, with the Hang Seng Index declining below 20,000 RMB while the sentiment index demonstrates increased volatility but remains around the 0.4 level, suggesting a possible reduction in its predictive

power due to evolving market conditions or investor adaptation.

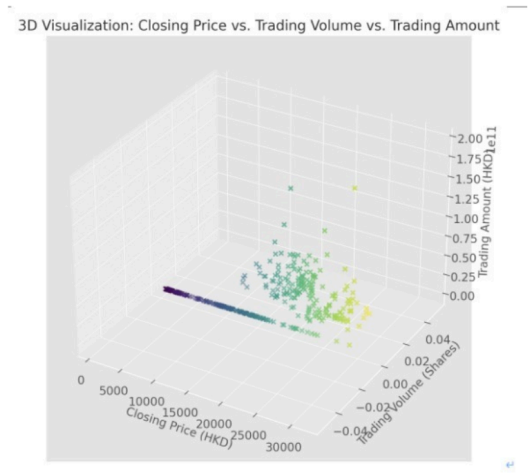

Figure 2: 3D Visualization: Closing Price vs. Trading Volume vs. Trading Amount

The visualization reveals a distinct relationship among closing price, trading volume, and trading amount. Generally, as the closing price increases from approximately 5,000 HKD to 30,000 HKD, there is a noticeable decline in the trading volume of shares, illustrating an inverse linear relationship between these two variables. Cluster analysis indicates that higher trading amounts—ranging up to around 2.0e11 HKD—tend to concentrate in regions of moderate closing prices (approximately 15,000 to 20,000 HKD) and moderate trading volumes (around 0.01M to 0.02M shares). This clustering suggests that mid-priced stocks have a disproportionately large impact on overall trading values. Additionally, the visualization highlights several outliers at higher price levels (above 25,000 HKD) with relatively lower trading volumes, possibly representing high-value but less frequently executed transactions.

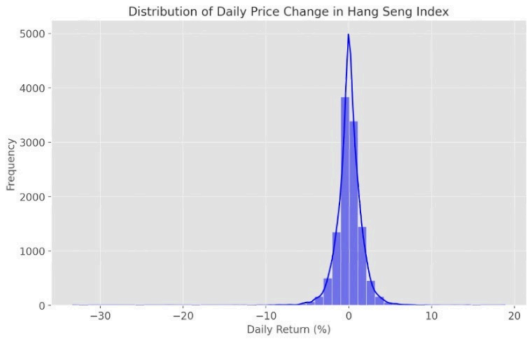

Figure 3: Distribution of Daily Price Change in Hang Seng Index

The daily returns of the Hang Seng Index display a sharply peaked distribution centered closely around zero, with the majority of daily returns falling within ±2%, suggesting minimal daily market movement as typically seen in major stock indices. The distribution demonstrates symmetry, indicating an equal likelihood of experiencing gains or losses each day. However, the pronounced "fat tails" are apparent, extending beyond ±5% in daily returns, indicating frequent occurrences of extreme market movements compared to what would be expected under a normal distribution. This reflects a high kurtosis level, highlighting an elevated risk of significant market fluctuations and volatility.

Figure 4: Highest vs. Lowest Price in Hang Seng Index Futures

The graph reveals significant volatility patterns, with distinct spikes in both highest and lowest prices clearly aligning with major financial crises, notably during the Asian financial crisis in 1997, the global financial

crisis in 2008, and the COVID-19 pandemic in early 2020. Despite these dramatic market events, a clear long-term upward trend emerges, with prices rising from below 10,000 HKD in the early 1990s to consistently exceeding 20,000 HKD post-2010, suggesting overall market growth. Sharp downturns followed by rapid recoveries, particularly pronounced around 2008 and 2020, highlight the sensitivity of market prices to external economic shocks and underline their temporary but severe impacts on investor sentiment and market stability.

II. Data Collection and Description

This project investigates the predictive power of investor sentiment in China's A-share and Hong Kong stock markets. To that end, we compiled a multi-dimensional dataset consisting of daily and monthly market indicators that reflect liquidity conditions, risk pricing, trading activity, and cross-border capital flows. Each market's sentiment index is constructed from five key variables, chosen based on their theoretical relevance and empirical importance in prior literature.

For the A-share market, the sentiment index incorporates the turnover rate, implied risk premium, equity-bond yield spread, stock index futures basis, and total initial public offering (IPO) fundraising. These variables jointly capture investor participation levels, risk expectations, and market valuation discrepancies. Meanwhile, the sentiment index for the Hong Kong market is constructed using trading amount, implied risk premium, equity-bond yield spread, stock index futures basis, and net capital inflows through the Stock Connect program, reflecting both domestic and cross-border investment sentiment.

To reduce dimensionality and extract the dominant sentiment signal, we apply Principal Component Analysis (PCA) to each group of indicators. This technique enables us to condense multiple correlated measures into a single, interpretable sentiment index that captures the most significant variation in investor mood over time.

Descriptive statistics further illuminate the characteristics of the underlying variables. In the A-share market, the average daily turnover rate is approximately 3.5%, with a mean implied risk premium of 2.1% and an average equity-bond yield spread of 1.8%. The annual average IPO fundraising volume stands at around ¥12 billion. For the Hong Kong market, the average daily trading amount is HK$85 billion, the implied risk premium averages 1.9%, and the equity-bond yield spread is approximately 1.5%. Additionally, the Stock Connect program channels an average of HK$30 billion in annual net capital inflows, providing a crucial gauge of international investor sentiment.

Overall, the dataset offers a rich foundation for quantifying sentiment across different market environments and evaluating its relevance for short-term stock price movements.

III. Statistical Model(s)

Given our focus on accurately forecasting short-term stock prices, we centered our analysis on two modeling approaches that are particularly well-suited for time-dependent financial data: Long Short-Term Memory (LSTM) neural networks and traditional time series models, specifically VAR (Vector Autoregression) and ARIMA (AutoRegressive Integrated Moving Average).

LSTM networks are designed to capture long-range dependencies and nonlinear patterns in sequential data, making them well-suited for modeling the complex temporal dynamics observed in financial markets. Their architecture allows for flexible incorporation of multiple features—such as sentiment indicators, turnover rates, and past prices—and can adapt to the nonstationary, high-noise nature of stock price movements. This makes LSTM an attractive choice for learning intricate behavioral and technical relationships that influence future price trajectories.

In parallel, we also implemented VAR and ARIMA models to provide a more interpretable and statistically grounded framework for time series forecasting. VAR allows us to jointly model multiple interrelated time series—such as price and sentiment—while capturing their lagged interactions and feedback effects. ARIMA, on the other hand, offers a robust approach for univariate time series forecasting, incorporating autoregressive and moving average components as well as differencing to address nonstationarity.

By leveraging both deep learning and classical econometric models, our study aims to compare and evaluate their effectiveness in stock price prediction across different market conditions. This dual-model approach enables a more comprehensive understanding of the predictive value of sentiment and market features, offering both flexibility in capturing complex dynamics and clarity in interpreting temporal structures.

IV. LSTM Model

To assess the effectiveness of deep learning in capturing the nonlinear dynamics of stock prices, we implemented a Long Short-Term Memory (LSTM) network trained on lagged price, sentiment, and turnover-based features. The model's performance is evaluated both quantitatively and visually.

Figure 5 plots the predicted stock prices alongside the actual prices across the testing period. The LSTM model demonstrates strong alignment with the true price series, successfully capturing both trend directions and the timing of turning points. Despite some lag in high-volatility regions, the predicted trajectory remains close to the observed data throughout the series, indicating robust learning of temporal dependencies.

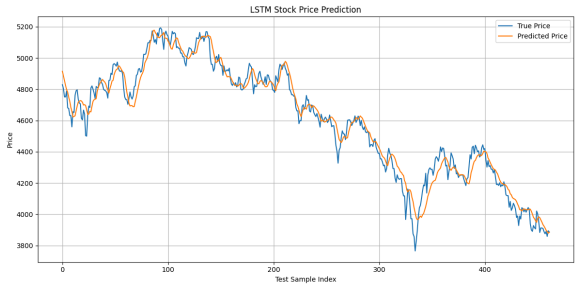

Figure 5: Predicted vs. Actual Stock Price Trajectory

To further investigate the distribution and behavior of prediction errors, Figure 6 presents the residual plot, where each point represents the difference between true and predicted prices at a given time. The residuals are largely centered around zero, suggesting that the model is unbiased on average. However, we observe volatility clusters and outliers in certain periods, particularly during rapid market movements, where the LSTM struggles to adjust in real time.

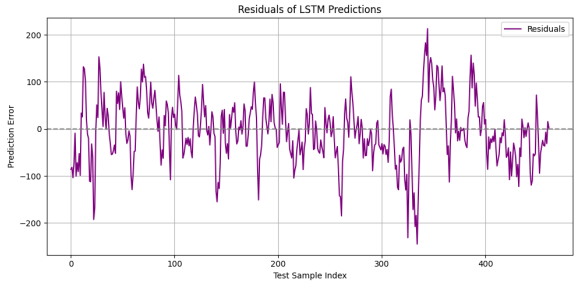

Figure 6: Time Series of Prediction Residuals

Figure 7 provides a scatter plot comparing predicted and true prices. The points tightly cluster around the ideal y=xy = xy=x reference line, indicating a high degree of linear concordance between the model's output and actual price levels. The symmetry and compactness of the scatter cloud affirm the model’s ability to generalize well within the observed price range.

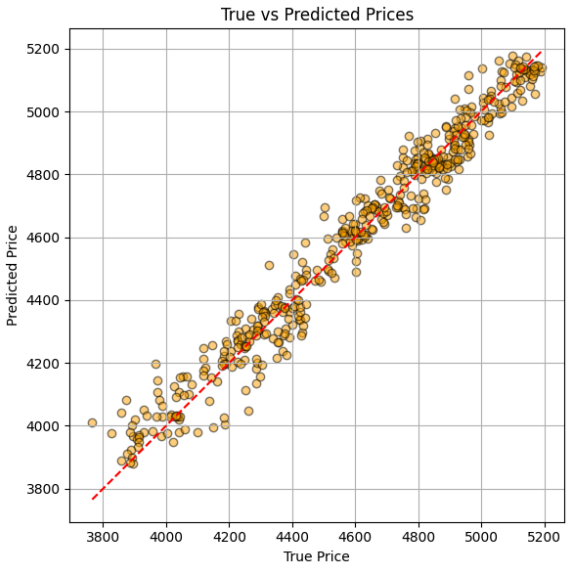

Figure 7：Scatter Plot of True vs. Predicted Prices

The predictive accuracy of the model is further evidenced by several key regression metrics: the R^2 score reaches 0.9640, indicating that the model explains over 96% of the variance in the target variable; the mean absolute error (MAE) is 53.12, while the mean squared error (MSE) stands at 4648.30; the root mean squared error (RMSE) is 68.18; and the mean absolute percentage error (MAPE) is as low as 1.18%, highlighting the model's strong precision across varying price levels.

These results suggest that the LSTM model captures the structure of stock price movements with high precision, achieving a very strong fit between predicted and actual values. Notably, the low MAPE indicates excellent relative accuracy, even across varying price levels.

Taken together, the visualizations and statistical indicators confirm that the LSTM framework is highly effective for short-term price forecasting, although its ability to react to sudden market shocks may be further improved by incorporating volatility-sensitive features or attention-based mechanisms.

V. Time Series Models

Since various studies suggest that time series models are better suited for analyzing financial data, we introduce two classical models to examine the influence of investor sentiment on the stock market: the Autoregressive Integrated Moving Average (ARIMA) model and the Vector Autoregression (VAR) model. The ARIMA model focuses on a single time series, combining autoregression, differencing, and moving average components to effectively model and forecast non-stationary univariate data. In contrast, the VAR model captures the interdependencies among multiple variables by modeling each as a function of its own lagged values and those of other variables in the system, making it especially useful in multivariate macroeconomic and financial analysis.

(i) Data Preprocessing and Stationarity Assessment

We begin by detecting and correcting outliers in the stock closing price series(close(CNY)). Outliers were identified using a 30-day rolling window with ±2 standard deviations as the threshold. Identified outliers were replaced with the previous day's closing price, resulting in a smoother series denoted as close_cleaned. This preprocessing step helps eliminate extreme fluctuations and ensures stability for subsequent trend and model analysis.

To test whether the cleaned series is stationary, we applied the Augmented Dickey-Fuller (ADF) test. The resulting p-value was approximately 0.07, which exceeds the 0.05 threshold, suggesting non-stationarity. Therefore, differencing is required to transform the series into a stationary process.

To further understand the underlying structure of the time series, we performed an additive seasonal decomposition. As shown in Figure 6, the trend component exhibits long-term fluctuations, confirming the presence of a non-stationary trend, while the seasonal component reveals strong annual cycles, indicating seasonal non-stationarity as well. Based on these findings, we applied both a first-order difference to eliminate the trend and a seasonal difference with a

365-day period to remove annual cycles. These transformations stabilize the mean and variance, making the series suitable for ARIMA modeling.

(i) ARIMA Model

Before fitting the ARIMA model, we examined the autocorrelation function (ACF) and partial autocorrelation function (PACF) to gain insights into the temporal structure of the series. As shown in Figures 7 and 8, the ACF decays slowly, reinforcing the need for differencing, while the PACF displays significant spikes at early lags, guiding the selection of model parameters.

Based on this analysis, we fitted an ARIMA(2,1,2) model to the differenced series. Residual diagnostics (Figure 9) show that the residuals fluctuate around zero with no discernible structure, and autocorrelations remain within the 95% confidence bounds, supporting the white noise assumption. The Q-Q plot and residual histogram indicate approximate normality, with only minor deviations in the tails. These results suggest that the ARIMA(2,1,2) model provides a satisfactory fit.

To rigorously evaluate out-of-sample performance, we implemented an expanding window forecast, retraining the model at each step and predicting the next value. As shown in Figure 11, the predicted values closely follow the actual trajectory. Evaluation metrics confirm the model's accuracy, with a Mean Squared Error (MSE) of 1882.12, Root Mean Squared Error (RMSE) of 43.38, Mean Absolute Error (MAE) of 32.48, and R^2 of 0.9854.

(ii) Var Model

To capture multivariate dynamics, we also fitted a Vector Autoregression (VAR) model using close_cleaned, Senti Index, and turnover rate. In-sample performance was evaluated using R^2, yielding values of 0.9886 for close_cleaned, 0.9790 for Senti Index, and 0.8838 for turnover rate, indicating excellent fit across all variables.

To further interpret the model, we conducted impulse response analysis (Figure 12). A positive shock to the sentiment index results in a sharp and immediate increase in the closing price, highlighting the significant influence of sentiment on market valuation. In contrast, a shock to turnover rate causes a temporary decline in price, potentially reflecting panic or corrective behavior. Meanwhile, sentiment itself appears largely exogenous, showing minimal response to other variables. Turnover rate, however, responds strongly to sentiment shocks, indicating that rising optimism fuels market activity.

These findings support the hypothesis that investor sentiment plays a leading role in driving both price dynamics and trading intensity in the short term.

(iii) Model Comparison

To compare forecasting and interpretability, we evaluated both the ARIMA and VAR models. The ARIMA(2,1,2) model, applied to the univariate price series, demonstrated high out-of-sample accuracy ($R^2 = 0.9854$), capturing both long-term trends and short-term fluctuations. The VAR model, by contrast, provided a comprehensive view of the interdependence among variables, with strong in-sample performance and valuable insights from impulse response analysis.

Overall, ARIMA is well-suited for univariate forecasting and trend modeling, while VAR excels in understanding the dynamic relationships across multiple financial indicators. These models serve complementary purposes—ARIMA for accurate prediction, and VAR for structural insight and policy analysis.

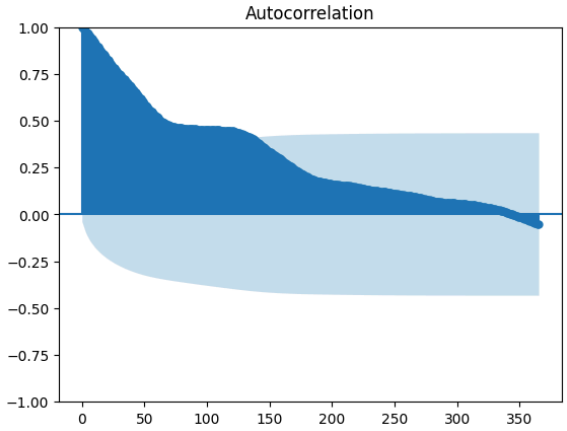

Figure 7: Decomposition trend, seasonality, and residuals

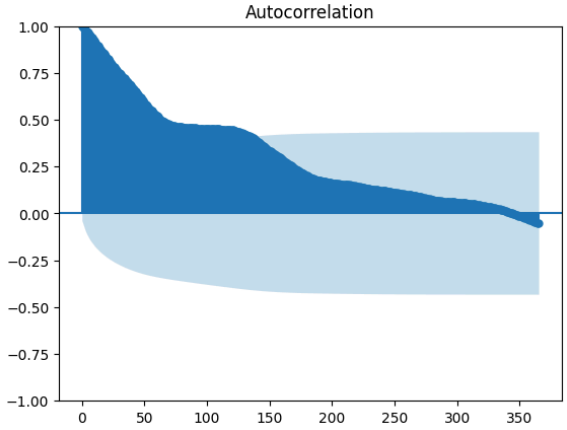

Figure 8: ACF

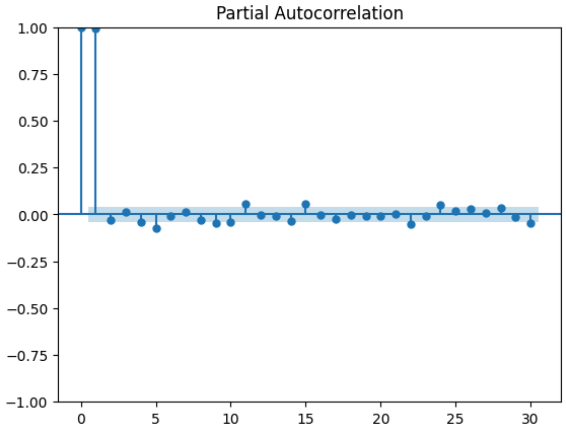

Figure 9 : PACF

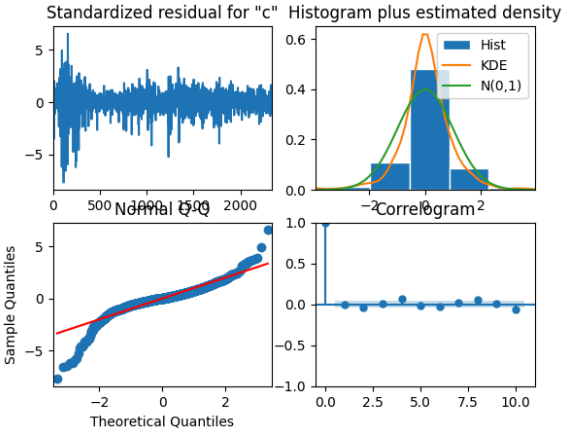

Figure 10: Standard diagnostic plots of ARIMA model

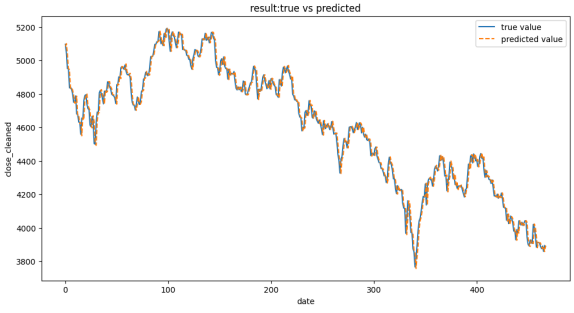

Figure 11: Prediction results of the ARIMA model

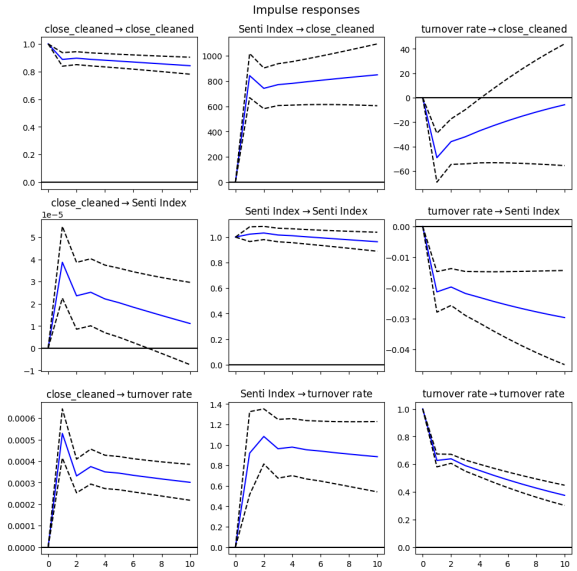

Figure 12: IRF Of VAR

VI. Conclusion

This study evaluates the effectiveness of sentiment-based forecasting using LSTM, ARIMA, and VAR models. The LSTM model achieved strong predictive performance (R^2 = 0.9640), effectively capturing nonlinear temporal patterns. The ARIMA(2,1,2) model outperformed in univariate forecasting (R^2 = 0.9854), providing accurate and robust price predictions. In contrast, the VAR model offered valuable interpretability, revealing that sentiment leads short-term price and turnover dynamics. Overall, LSTM and ARIMA are well-suited for prediction, while VAR excels in understanding structural relationships, highlighting the complementary strengths of these approaches.

VII. References

Chen, G., & Xia, F. (2022). Construction and Application of Investor Sentiment Index: How Do Investors Currently View the Market? China Securities.

Huang, J. (2015). Construction of the Investor Sentiment Index: A Study on Measuring Investor Sentiment in China's Stock Market. Century Securities.

Lai, Y. (2024). Construction and Application of the A-Share Market Sentiment Index. Pu Yin International.

Wang, K. (2022). Construction and Application of the A-Share Sentiment Index. Guosen Securities.

VIII. Appendix

Literature reviews:

1. The research paper *"Construction and Application of Investor Sentiment Index: How Do Investors Currently View the Market?"* by China Securities delves into the impact of investor sentiment on stock market behavior and its ability to predict market trends. Based on behavioral finance theory, the study posits that stock prices do not always align with intrinsic values, primarily due to the influence of investor sentiment. This influence persists even in highly liquid and transparent markets, where emotions often dictate short-term price movements, highlighting the importance of sentiment in market analysis. The paper develops the China Securities Strategy-Investor Sentiment Index using eight objective market indicators split into volume-based and price-based categories. Volume-based indicators, such as turnover rate, new equity fund issuances, and margin trading volume, capture aspects of market liquidity and investor enthusiasm for trading. Price-based indicators, including the implied risk premium, stock-bond yield gap, stock index futures basis, moving average deviations, and overbought-oversold levels, reflect investor preferences and expectations regarding equities relative to other asset classes.

The findings suggest that the investor sentiment index is a strong predictor of future market directions. Typically, the index aligns with market movements. However, extreme sentiment levels—above 90 in euphoric states or below 10 during panic—often precede market

reversals. A sharp decline in sentiment from high levels can indicate imminent market corrections, while a surge from low levels may signal a potential recovery. The research highlights three primary uses for the sentiment index: forecasting market peaks, spotting potential market bottoms, and identifying transitional phases. Historical data validate these uses, showing that high sentiment levels predicted major market downturns in 2015 and 2019, while low levels signaled buying opportunities in early 2019 and 2022.

Currently, the sentiment index by China Securities is at a historically low point, suggesting the absence of imminent buying opportunities. Early in 2022, the index fell below 10, reflecting a bearish outlook due to geopolitical uncertainties and the pandemic. The study stresses that while sentiment analysis is insightful, its predictive capability is limited to historical contexts and should be used alongside other analytical tools. In conclusion, the paper presents a comprehensive method for measuring investor sentiment, proving its effectiveness in predicting market movements. The China Securities sentiment index, with its diverse indicator set, provides a crucial tool for market timing and risk assessment. Nevertheless, the study also recognizes the limitations of relying solely on sentiment, advocating for its integration with fundamental and macroeconomic analyses to achieve a thorough market perspective.

2. The research paper titled *"Construction of the Investor Sentiment Index: A Study on Measuring Investor Sentiment in China's Stock Market"* by Century Securities delves into the significant influence of investor sentiment on short-term market dynamics. Traditional asset pricing models, which typically emphasize macroeconomic indicators, market liquidity, and company fundamentals, often fall short in the short term, particularly for stocks that are hard to value, have high arbitrage costs, or are speculative—traits prevalent in small-cap stocks, new issues, and high-growth firms. This backdrop underscores the growing importance of accurately measuring investor sentiment.

To develop a comprehensive investor sentiment index, the study utilizes factor analysis and partial least squares (PLS) regression. Initially, it identifies nine crucial sentiment proxies, such as the China Securities Investor Confidence Index, the Large-Cap Optimism Index, and the Hedge Fund Manager Confidence Index, along with market activity indicators like the weighted average price-to-earnings ratio of the Shenwan 300 Index and trading volume of the Shanghai Composite Index. These are categorized into subjective measures (survey-based) and objective measures (market-based). Factor analysis helps to reduce complexity by isolating the most impactful variables that reflect investor sentiment.

Unlike the commonly used principal component analysis (PCA), this study employs PLS regression for its ability to discard irrelevant factors and focus on those most pertinent to investor sentiment, enhancing the precision and forecasting utility of the index. The findings reveal that the PLS-based index outperforms traditional indices built solely on PCA, demonstrating superior predictive power for market trends and movements.

The research highlights the strong correlation between investor sentiment and short-term market fluctuations: positive sentiment typically precedes market rallies, while negative sentiment can signal upcoming corrections or consolidation phases. The study finds that subjective sentiment indicators, like confidence surveys, are particularly effective at pinpointing market turning points, whereas objective measures like trading volume and price-to-earnings ratios confirm these trends, thereby boosting the reliability of sentiment analysis.

Overall, the study successfully constructs an investor sentiment index that merges survey-based and market-driven data, offering a nuanced perspective on market dynamics. This index surpasses the predictive capabilities of any single indicator, providing a thorough tool for market analysis. The research underscores the critical role of investor sentiment in driving short-term market fluctuations, advocating for its integration into investment strategies. This approach, which diverges from traditional methods focused solely on economic fundamentals, provides a dynamic framework

for investors, analysts, and policymakers aiming to enhance their market navigation by incorporating sentiment analysis into their decision-making processes.

3. The research paper *"Construction and Application of the A-Share Market Sentiment Index"* by Pu Yin International presents a quantitative sentiment index tailored to China's A-share market. Motivated by the increasing influence of sentiment on short-term market fluctuations, the study constructs an index using 14 liquidity- and expectation-based indicators, including trading volume, margin trading, RSI, fund flows, and earnings revisions.

The index is standardized using correlation and regression analyses, with three weighting methods: equal weight, R^2-based weight, and a hybrid approach. All versions show strong positive correlations with the Shanghai Composite Index (up to 0.76). Granger causality tests and backtesting confirm the index's predictive power, particularly under the combined weighting method.

The paper concludes that the sentiment index effectively signals market turning points and recommends its use alongside macro and fundamental indicators. It also suggests adopting defensive investment strategies during weak sentiment periods, emphasizing the index's practical relevance for market timing and risk management.

4. The paper *"Construction and Application of the A-Share Sentiment Index"* by Guosen Securities proposes a sentiment index tailored to China's A-share market, emphasizing the role of retail investors in driving short-term price fluctuations. Unlike traditional survey-based approaches, the index is built on transactional data to avoid biases and better reflect market psychology.

The methodology combines industry beta coefficients and the Spearman rank correlation between industry returns and risk exposure to quantify investor optimism or pessimism. This composite index, GSISI, effectively identifies market turning points, with historical backtests capturing major reversals in 2016, 2018, and 2022.

The study further extends the index to other market segments (GSISI II and III), confirming its robustness. It concludes that sentiment-based indicators, when used alongside fundamentals, offer valuable foresight for market participants navigating the sentiment-sensitive A-share environment.