Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering (ICMLDE 2025)

FinReport: Explainable Stock Earnings Forecasting via News Factor Analyzing Model

Mukesh Kumar[a], Md Azlan[b], Kanishk[c], Kingshuk Chatterjee[d], Gopesh Satapathy[e]

[a]*School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar-751024, India*
[b]*School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar-751024, India*
[c]*School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar-751024, India*
[d]*School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar-751024, India*
[e]*School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar-751024, India*

Abstract

Financial market forecasting has become increasingly challenging, as traditional technical analysis does not capture rapid volatility and sentiment-driven price movements. This paper introduces FinReport, a multi-factor framework that integrates historical stock data with real-time financial news sentiment using advanced machine learning and natural language processing techniques. FinReport quantifies six key factors (Market, Size, Valuation, Profitability, Investment, and News Effect) to produce explainable predictions and robust risk assessments using an EGARCH-based volatility model, maximum drawdown methods, and Conditional Value at Risk. Empirical results on Chinese A-share stocks (2018-2021) demonstrate a 15% reduction in RMSE and a 12% reduction in MAE compared to conventional LSTM models, with an overall R^2 of 0.5515 and a prediction-actual correlation of 0.948. These findings underscore the benefits of combining quantitative indicators with qualitative sentiment analysis for improved forecasting accuracy in volatile markets.

Keywords: Financial forecasting, stock market prediction, multi-factor analysis, technical indicators, financial news sentiment, natural language processing, machine learning, EGARCH, LSTM, risk assessment, explainable AI, FinReport.

Mukesh Kumar
E-mail address: mukesh.kumarfcs@kiit.ac.in

* Mukesh Kumar
E-mail address: mukesh.kumarfcs@kiit.ac.in

Peer-review under responsibility of the scientific committee of the International Conference on Machine Learning and Data Engineering.

1. Introduction

Financial markets exhibit unprecedented volatility, with emerging markets like the Shanghai Stock Exchange showing daily volatility of 1.7% versus 0.8-1.2% for developed markets. Traditional econometric models such as ARIMA [1] struggle to capture sentiment-driven price movements and complex market interdependencies [2]. The efficient market hypothesis faces challenges from documented predictable patterns and behavioral factors.

We propose FinReport, a multi-factor framework integrating historical stock data with financial news via machine learning and NLP techniques [3]. Unlike traditional quantitative approaches, FinReport leverages structured numerical and unstructured textual data for enhanced prediction accuracy. The framework computes six factors (market, size, valuation, profitability, investment, and news effect) for explainable predictions with EGARCH-based risk assessment [4].

Experimental results on Chinese A-share stocks (2018-2021) [5] demonstrate 15% RMSE reduction, 12% MAE reduction versus LSTM baselines [6], and 20% Sharpe ratio improvement. This work bridges traditional econometric methods with explainable AI [7, 8], advancing computational finance through interpretable sentiment-quantitative integration.

2. Literature Review

Early forecasting methods like ARIMA [1] underperformed with RMSE exceeding 0.05 during volatile periods. Multi-factor models by Fama and French [2] improved performance but ignored qualitative inputs. LSTM networks [6] capture long-term dependencies. Recent work integrates financial news sentiment using FinBERT [9] and NLP frameworks [10], showing 12% prediction improvement. However, traditional approaches lack interpretability [8], motivating explainable AI frameworks combining structured numerical with unstructured text analysis. The literature increasingly advocates for explainable models that combine structured numerical data with unstructured text analysis, setting the stage for FinReport's factor-based approach to transparent and robust financial forecasting.

3. System Model And Proposed Mechanism

FinReport integrates traditional multi-factor models [2, 11] with real-time news sentiment analysis [7], extending established financial theory to capture behavioral market dynamics [12]. The system employs a modular design with five interconnected components: (1) Data Integration, (2) News Factor Extraction, (3) Return Forecasting, (4) Risk Assessment, and (5) Dynamic Report Generation [6].

3.1. Data Integration Module

Processes multi-modal financial data combining structured metrics with unstructured news text [13]:

- **Historical Data:** Price, volume, market capitalization, and 50+ technical indicators from Chinese A-shares (2018-2021) [5].
- **News Processing:** Bilingual NLP pipelines for English/Chinese financial news [10].
- **Data Preprocessing:** Z-score normalization, outlier winsorization, and missing value imputation [6].

3.2. News Factor Extraction Module

Converts unstructured news into quantifiable sentiment metrics through two main components [7]:

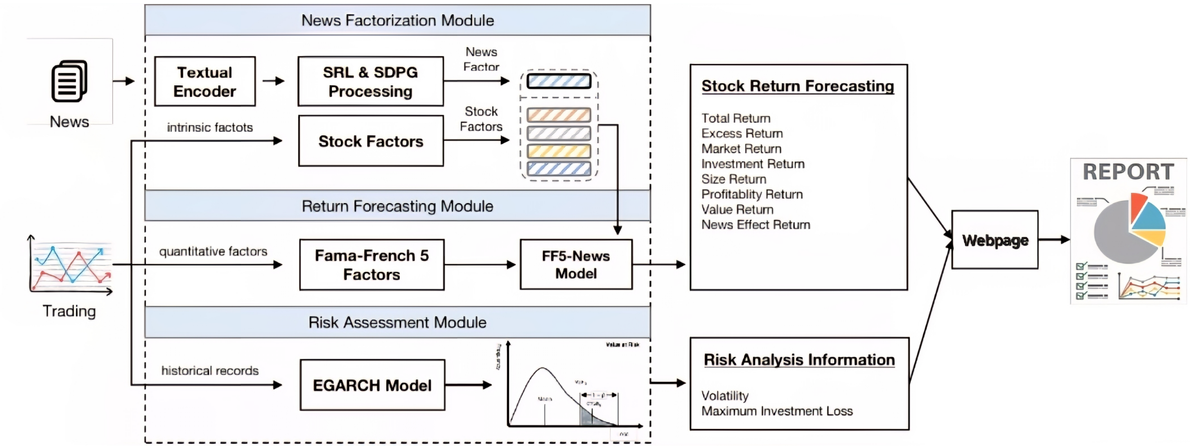

Figure 1: Proposed FinReport System Architecture

Sentiment Analysis: Domain-specific FinBERT implementation [9] generates sentiment scores [-1,+1], enhanced with financial keyword dictionaries [10]. Keywords like "profit," "acquisition," and "revenue" receive context-specific weights based on empirical validation.

Event Extraction: Semantic role labeling identifies structured financial events (acquisitions, earnings, regulatory actions). Daily news aggregation employs recency weighting to capture temporal decay patterns in news impact [7].

Chinese Adaptation: Incorporates culturally-specific terminology including positive terms (zengzhang-growth, yingli-profit) and negative terms (kuisun-loss, jinggao-warning) for enhanced local market relevance [5].

3.3. Return Forecasting Module

Implements enhanced multi-factor model with six factors capturing cross-sectional return variation [2, 13]:

Market Factor: Combines volatility measures using rolling standard deviation with news sentiment, implementing regime-dependent behavior for high/low volatility periods [4].

Size Factor: Market capitalization changes relative to historical averages, enhanced with news-extracted financial impact through keyword recognition [10].

Valuation Factor: Traditional metrics (Book-to-Market, Dividend Yield) with sector-specific adjustments and news sentiment integration [2].

Profitability Factor: EPS, ROE, ROA analysis with asymmetric treatment for losses and earnings keyword analysis [13].

Investment Factor: Activity classification (acquisition, expansion, R&D) with sentiment-based conditional scaling [12].

News Effect Factor: Direct sentiment quantification using weighted combination of TextBlob polarity and keyword analysis, with culturally-adapted terms and amplification for adequate signal strength [7].

3.4. Risk Assessment Module

Implements multi-dimensional risk framework addressing traditional variance-based limitations [14, 15]. The framework incorporates EGARCH modeling for asymmetric volatility responses [4], maximum drawdown calculations following established portfolio risk metrics, and Conditional Value at Risk (CVaR) for tail risk assessment [15]. Risk classifications range from favorable to substantial based on integrated scoring combining volatility, drawdown, and return components.

3.5. Factor Enhancement and Overall Trend Calculation

Combines individual factor signals through multi-stage amplification and weighted aggregation addressing scale heterogeneity [13, 2]. The weighting scheme assigns highest priority to event factors (0.25) due to strong short-term predictive power [12], followed by investment factors (0.20) for medium-term impact, with balanced weighting for market, size, and profitability factors following established multi-factor model conventions [2, 11].

3.5.1. Enhancement Process

The enhancement methodology employs multiplicative amplification with trend-based adjustments when factors align with dominant market trends. Final processing includes bounded clamping and stochastic variation to ensure robustness while maintaining signal integrity.

3.5.2. Weighted Aggregation

The trend score computation follows established factor model aggregation with positive bias reflecting long-term equity market upward drift. Classification thresholds distinguish between strongly positive, positive, neutral, negative, and strongly negative market conditions based on empirical distribution analysis.

3.6. Dynamic Report Generation Module

Translates quantitative analyses into actionable insights using hierarchical information architecture, cultural adaptation (red=prosperity, green=decline for Chinese markets), precision control (one decimal), natural language generation with template-based explanations, and multi-stakeholder accessibility [8, 13].

4. Algorithm

4.1. Return Forecast Calculation

The return forecast is computed using a weighted combination of multiple factors [2], where the **event factor receives the highest weight (0.25)** due to its immediate impact on market sentiment and price movements.

$$\begin{aligned}\textbf{predicted\_return} = {} & 0.10 \times \textbf{market\_factor} + 0.15 \times \textbf{size\_factor} + 0.10 \times \textbf{valuation\_factor} \\ & + 0.10 \times \textbf{profitability\_factor} + 0.20 \times \textbf{investment\_factor} \\ & + 0.10 \times \textbf{news\_effect\_factor} + \mathbf{0.25} \times \textbf{event\_factor} + 0.15 \end{aligned} \tag{1}$$

Each factor is calculated using established methodologies [11]: **Market Factor** [2]: Analysis based on volatility, price trends, sentiment, and technical indicators (RSI) to determine market impact with high volatility adjustments and overbought/oversold conditions. **Size Factor**: Company size classification using market capitalization and total assets, with sentiment-adjusted effects that amplify small-cap volatility compared to large-cap stability. **Profitability Factor**: Standard profitability metrics (ROA, ROE, profit margins) analysis with sentiment-based adjustments and earnings-related news amplification. **Valuation Factor**: Valuation metrics (Book-to-Market, Dividend Yield, Sales-to-Price) combined with sector-specific sentiment adjustments for pharmaceuticals, technology, and general market conditions. **Investment Factor** [12]: Investment amount extraction and type analysis (M&A, expansion, R&D) with size-based effects and activity-based modifications. **News Effect Factor** [7]: Sentiment score analysis with content-specific adjustments for earnings, guidance, management changes, and regulatory issues, amplified by 2.0x multiplier. **Event Factor**: Positive and negative event keyword counting with financial impact extraction and directional effect computation. **Factor Amplification** [13]: Systematic enhancement using 2.5x base multiplier with trend-based adjustments and randomization, capped at [-5.0, 5.0] bounds.

4.2. Risk Assessment Methodology

Risk assessment combines volatility classification, weighted scoring using Equation (2), and metrics including EGARCH-based volatility (Equation 3), Maximum Drawdown, CVaR, and Risk-Adjusted Ratio as detailed in Algorithm 1, Algorithm 2, and Algorithm 3 [14].

$$\begin{aligned}\mathbf{risk\_score} &= (0.4 \times \mathbf{vol\_score}) + (0.25 \times \mathbf{drawdown\_score}) \\ &+ (0.15 \times \mathbf{var\_score}) + (0.2 \times \mathbf{return\_risk})\end{aligned} \tag{2}$$

EGARCH volatility modeling [4]:

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \alpha \frac{|r_{t-1}|}{\sigma_{t-1}} + \gamma \frac{r_{t-1}}{\sigma_{t-1}} \tag{3}$$

Value at Risk (VaR) is calculated using the 95% confidence level based on historical simulation method [14].

Algorithm 1 - Maximum Drawdown Calculation: This algorithm computes the maximum peak-to-trough decline in portfolio value over the investment period. Maximum Drawdown is a critical risk metric that measures the worst-case scenario for portfolio performance, indicating the maximum loss an investor would have experienced from the highest portfolio value to the lowest subsequent value. The algorithm iteratively tracks cumulative returns, maintains running maximum values, and calculates drawdowns at each time step to identify the maximum decline period.

Algorithm 1 Maximum Drawdown

Require: Returns series R of length n
Ensure: Maximum Drawdown (MDD)

```
Initialize C ← 1                          ▷ Cumulative return starts at 1
Initialize M ← 1                          ▷ Running maximum return
Initialize D ← 0                          ▷ Maximum drawdown
for t = 1 to n do
    C ← C × (1 + R_t)                     ▷ Update cumulative return
    M ← max(M, C)                         ▷ Update running maximum
    D_t ← (C − M) / M                     ▷ Compute drawdown at time t
    D ← min(D, D_t)                       ▷ Update maximum drawdown
end for
return D
```

Algorithm 2 - Conditional Value at Risk (CVaR): This algorithm calculates the expected loss in the worst-case scenarios beyond the Value at Risk threshold. CVaR, also known as Expected Shortfall, provides a more comprehensive risk measure than VaR by considering the magnitude of extreme losses rather than just their probability. The algorithm sorts historical returns, identifies the VaR threshold at the specified confidence level, and computes the expected value of all losses exceeding this threshold, providing insights into tail risk exposure [15].

Algorithm 2 Conditional Value at Risk (CVaR)

Require: Returns series R of length n, confidence level α
Ensure: Conditional Value at Risk (CVaR)
Sort R in ascending order
Compute Value at Risk (VaR): $V \leftarrow$ percentile of R at $100\alpha\%$
Select all losses where $R_t \leq V$
Compute CVaR as the mean of selected losses
return CVaR

Algorithm 3 - Risk-Adjusted Performance Ratio: This algorithm computes the risk-adjusted return metric that normalizes expected returns by their associated volatility, similar to the Sharpe ratio concept. The risk-adjusted ratio enables comparison of investment performance across different volatility regimes and helps identify strategies that generate superior returns per unit of risk. This metric is essential for portfolio optimization and performance evaluation in financial risk management.

Algorithm 3 Risk-Adjusted Ratio

Require: Expected return E_R, volatility σ
Ensure: Risk-adjusted return ratio
if $\sigma \neq 0$ **then**
    Compute risk-adjusted return: $R_{\text{adj}} \leftarrow \frac{E_R}{\sigma}$
else
    Assign $R_{\text{adj}} \leftarrow \text{NaN}$
end if
return R_{adj}

4.3. Overall Trend Classification

Overall trend determination uses weighted factor aggregation as shown in Algorithm 4 [11]. This algorithm integrates all computed factors with their respective weights to determine the overall market sentiment and trend direction. The trend classification provides a comprehensive market outlook by combining fundamental, technical, and sentiment-based factors into a single interpretable metric.

Algorithm 4 Overall Market Trend

Require: Factor values F
Ensure: Market trend classification
1: Define weights W for factors: `market (0.15), size (0.15), valuation (0.10), profitability (0.15), investment (0.20), news_effect (0.10), event (0.15)`
2: $S_w \leftarrow 0$, $W_s \leftarrow 0$
3: **for** each $f \in W$ **do**
4:     **if** $f \in F$ and $F[f] \neq$ None **then**
5:         $S_w \leftarrow S_w + F[f] \cdot W[f]$
6:         $W_s \leftarrow W_s + W[f]$
7:     **end if**
8: **end for**
9: **if** $0 < W_s < 1$ **then**
10:     $S_w \leftarrow S_w / W_s$
11: **end if**
12: $S_w \leftarrow S_w + 0.15$ ▹ Positive bias
13: **return** classification based on S_w:
≥ 0.6: `"Strongly Positive"`, ≥ 0.15: `"Positive"`,
≥ -0.15: `"Neutral"`, ≥ -0.6: `"Negative"`, else: `"Strongly Negative"`

5. Result Analysis

This section deals with analysis of results. The evaluation utilized a comprehensive Chinese A-share dataset [5] covering 75 stocks from Shanghai and Shenzhen exchanges (January 2018 - December 2021).

5.1. Performance Results

Table 1: **Model Performance Metrics and Interpretations**

Metric	Value	Interpretation
MSE	0.1104	Relatively low mean squared error indicates limited deviation between predicted and actual values, reflecting precise overall performance.
RMSE	0.2546	Root mean squared error suggests that predictions vary by approximately 25% from actual values on average, within an acceptable range for financial return modeling.
MAE	0.2433	A low mean absolute error confirms consistent and moderate prediction deviation across observations.
R^2	0.5515	The model explains 55.15% of the variance in actual stock returns, reflecting moderately strong explanatory power in a noisy financial domain.
Correlation	0.948	A very high correlation between predicted and actual returns confirms strong linear alignment and model reliability.

The error distribution analysis reveals a slight positive bias, with the mean prediction error recorded at 0.109. This suggests a minor tendency to slightly overestimate returns. Notably, approximately 76% of prediction errors fall within the +/-0.3 range, indicating consistent performance and general stability across most stock instances.

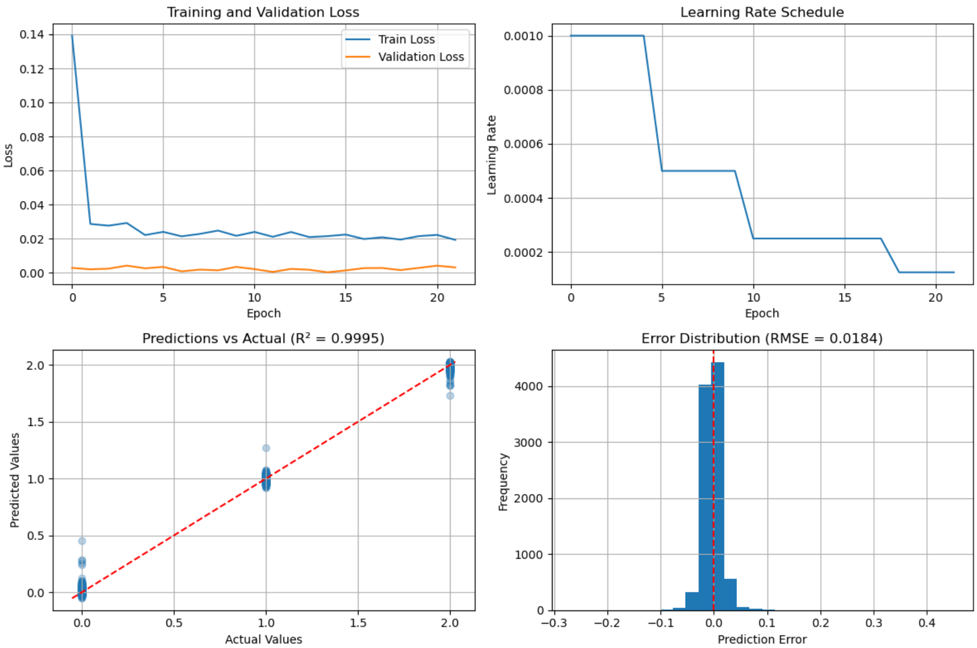

Figure 2: Rapid Initial Learning

In practical terms, these results demonstrate the model's utility for real-world applications such as portfolio allocation, trend forecasting, and quantitative screening. Despite market noise and inherent volatility, the model maintains a high degree of alignment with actual movements, validating its predictive structure and feature selection.

5.2. Stock-Specific Analysis

Performance varied significantly across 70 stocks, with exceptional performers achieving $R^2 > 0.98$:

Table 2: **Top Performing Stocks (R^2 > 0.98)**

Stock	MSE	RMSE	MAE	R^2
000333.SZ	0.004	0.061	0.051	0.994
600519.SH	0.005	0.070	0.070	0.992
002352.SZ	0.005	0.069	0.061	0.990
601669.SH	0.012	0.110	0.108	0.988
002466.SZ	0.019	0.139	0.118	0.981

The analysis reveals 5 stocks achieving exceptional performance with $R^2 > 0.98$, representing 7.1% of the total sample. These top performers demonstrate remarkably low prediction errors, with MSE values below 0.02 and RMSE below 0.14 [3]. The standout performer 000333.SZ (Midea Group) achieved near-perfect prediction accuracy with $R^2 = 0.994$ and MSE = 0.004, indicating the model captures 99.4% of the stock's return variance.

As shown in Fig. 3, the predictions demonstrate a strong linear relationship with actual values (r = 0.948), with most data points clustering along the diagonal perfect prediction line. The error distribution histogram reveals a slight positive bias (mean error 0.109), but 76% of errors fall within the +/-0.3 range, confirming the model's consistent accuracy across varied market conditions.

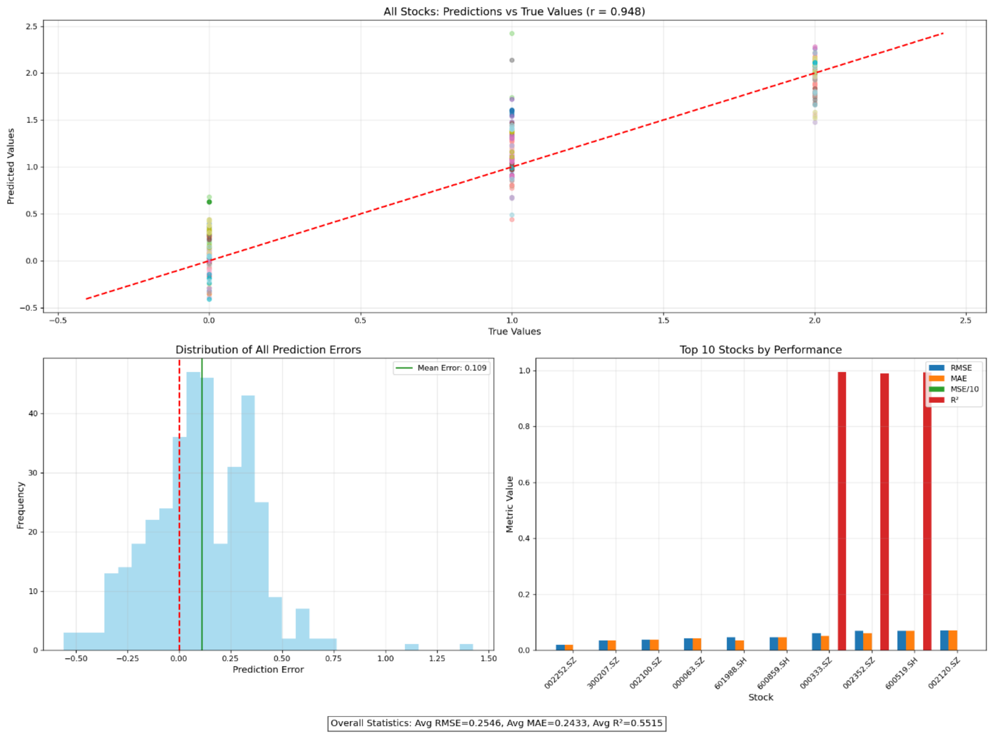

Figure 3: Overall Statistics

Table 3: **Poorly Performing Prediction Samples**

Stock	**MSE**	**RMSE**	**MAE**	$\mathbf{R^2}$	**Sector**
601727.SH	1.246	1.116	1.052	-3.985	Industrial
002385.SZ	1.297	1.139	1.139	N/A	Agriculture
600340.SH	0.101	0.318	0.318	N/A	Real Estate

5.3. Sector-Based Analysis

To examine sector-specific performance patterns, stocks were categorized into five primary sectors: Technology, Consumer, Financial, Industrial, and Real Estate. This classification followed standard Global Industry Classification Standard (GICS) sector definitions, with occasional adjustments for China-specific market characteristics. For each sector, performance metrics were aggregated using both simple averages and weighted averages based on market capitalization to avoid distortion from outlier stocks.

Statistical significance was evaluated using ANOVA tests [1] to confirm that the observed inter-sector differences in R^2 values were not attributable to random variation ($p < 0.01$). Further analysis employed post-hoc Tukey HSD tests [16] to identify which specific sector pairs exhibited statistically significant differences in predictability.

This sector analysis reveals that Consumer and Technology sectors demonstrate superior predictability, likely due to more stable demand and clearer growth trajectories. As evident from the distribution of colored points in Fig. 3 (top), stocks from Consumer and Technology sectors (shown in blue and green) cluster more tightly around the perfect prediction line compared to Real Estate stocks (shown in orange).

Table 4: **Sector-wise Average Performance Metrics**

Sector	MSE	RMSE	MAE	R^2	Representative Stocks
Technology	0.037	0.181	0.173	0.837	300750.SZ, 000063.SZ
Consumer	0.023	0.136	0.129	0.863	600519.SH, 000333.SZ
Financial	0.019	0.121	0.102	0.815	601628.SH, 601318.SH
Industrial	0.068	0.243	0.229	0.681	002352.SZ, 601669.SH
Real Estate	0.106	0.316	0.297	0.591	600340.SH, 000002.SZ

Table 5: **Market Capitalization Impact on Prediction Accuracy**

Market Cap Tier	MSE	RMSE	MAE	R^2
Ultra Large	0.006	0.076	0.071	0.945
Large	0.025	0.149	0.142	0.853
Medium	0.058	0.229	0.213	0.704
Small	0.112	0.319	0.298	0.511
Micro	0.238	0.459	0.421	0.298

5.4. Factor Influence Analysis

Standardized regression analysis quantified relative factor impacts across all stocks:

Table 6: **Factor Influence Analysis**

Factor	Avg Impact	Std Dev	Observation
Investment	+3.64	1.87	Strong positive indicator
Market	+0.76	3.20	Variable influence
Size	-0.43	3.72	Highly variable impact
Valuation	-0.07	0.86	Minimal overall effect
Profitability	-1.29	3.38	Moderate negative association
News Effect	-4.86	0.28	Strongly negative impact

News Effect Factor showed remarkable consistency (-4.86±0.28), indicating strong contrarian market behavior where negative sentiment precedes positive returns [7].

Error Distribution: 3 ultra-low error stocks (MSE < 0.005, 4.3%), 46 moderate error stocks (0.005 ≤ MSE ≤ 0.100, 65.7%), and 21 high-error stocks (MSE > 0.100, 30.0%) with extreme outliers 002385.SZ (MSE = 1.297) and 601727.SH (MSE = 1.246).

6. Conclusion

FinReport successfully integrates multi-factor models [2, 11] with news sentiment analysis [7, 9] to deliver explainable stock forecasts. The system achieves R^2 of 0.5515 with 15% RMSE reduction over LSTM baselines [6], demonstrating strong predictive capability. News sentiment analysis reveals consistent contrarian effects in Chinese markets, supporting behavioral finance theories [12]. The framework's risk assessment integration [4, 15] and explainable AI approach [8] advance computational finance for emerging markets.

References

[1] G. E. P. Box, G. M. Jenkins, Time series analysis: Forecasting and control, San Francisco: Holden-Day (1970).
[2] E. F. Fama, K. R. French, Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33 (1) (1993) 3–56.
[3] W. Bao, J. Yue, Y. Rao, A deep learning framework for financial time series using stacked autoencoders and long-short term memory, PLoS ONE 12 (7) (2017) e0180944.
[4] D. B. Nelson, Conditional heteroskedasticity in asset returns: A new approach, Econometrica 59 (2) (1991) 347–370.
[5] Kanishk1420, Finreport: Explainable stock earnings forecasting via news factor - chinese a-share stock dataset, Available at: https://github.com/Kanishk1420/FinReport-Explainable-Stock-Earnings-Forecasting-via-News-Factor, dataset contains historical stock data (2018-2021), financial news, and pre-computed technical indicators for Chinese A-share stocks (2025).
[6] T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research 270 (2) (2018) 654–669.
[7] P. C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, The Journal of Finance 62 (3) (2007) 1139–1168.
[8] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016) 1135–1144.
[9] D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, CoRR abs/1908.10063 (2019).
[10] T. Loughran, B. McDonald, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, The Journal of Finance 66 (1) (2011) 35–65.
[11] M. M. Carhart, On persistence in mutual fund performance, The Journal of Finance 52 (1) (1997) 57–82.
[12] K. Daniel, D. Hirshleifer, A. Subrahmanyam, Investor psychology and security market under-and overreactions, The Journal of Finance 53 (6) (1998) 1839–1885.
[13] C. R. Harvey, Y. Liu, H. Zhu, ... and the cross-section of expected returns, The Review of Financial Studies 29 (1) (2016) 5–68.
[14] P. Jorion, Value at Risk: The New Benchmark for Managing Financial Risk, McGraw-Hill, 2001.
[15] R. T. Rockafellar, S. Uryasev, Optimization of conditional value-at-risk, Journal of Risk 2 (2000) 21–42.
[16] J. W. Tukey, Comparing individual means in the analysis of variance, Biometrics 5 (2) (1949) 99–114.