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)

Stock Earnings Forecasting via News Factor Analyzing Model

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

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

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 multifactor 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 show a 15% reduction in RMSE and a 12% reduction in MAE over 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

1877-0509 © 2025 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the International Conference on Machine Learning and Data Engineering.

1. Introduction

Financial markets have experienced unprecedented volatility in recent years, with emerging markets showing particularly elevated risk patterns [1]. For example, the Shanghai Stock Exchange Composite Index exhibits an average daily volatility of approximately 1.7%, significantly higher than developed markets such as S&P 500 (typically 0.8-1.2%) [2]. Such volatility illustrates the limitations of traditional technical analysis methods and classical econometric models like ARIMA [3], which struggle to capture rapid, sentiment-driven price movements and the complex interdependencies inherent in modern financial markets [4, 5]. The efficient market hypothesis, while foundational, has been increasingly challenged by evidence of predictable patterns and the influence of behavioral factors on asset pricing [6, 2].

To address these limitations, we propose FinReport, a multi-factor forecasting framework that combines historical stock data with real-time financial news via advanced machine learning and natural language processing techniques [7, 8]. Unlike traditional approaches that rely solely on quantitative indicators, FinReport leverages both structured numerical data and unstructured textual information to enhance prediction accuracy [9, 10]. The framework computes six distinct factors—market, size, valuation, profitability, investment, and news effect—to generate explainable predictions alongside transparent risk assessments using EGARCH-based volatility modeling [11, 12].

Our experimental results on Chinese A-share stocks (2018-2021 dataset) [13] indicate a 15% reduction in RMSE and a 12% reduction in MAE compared to conventional LSTM baseline models [14], along with an enhanced risk-adjusted Sharpe ratio improvement of nearly 20%. These improvements are particularly significant given the inherent challenges of forecasting in high-volatility emerging market environments.

This work presents a robust and interpretable approach to forecasting in high-volatility environments, bridging the gap between traditional econometric methods and the growing need for explainable financial predictions [15, 16]. The integration of news sentiment analysis with quantitative factors represents a significant advancement in the field of computational finance, offering both improved accuracy and enhanced interpretability for practical investment decision-making.

2. Literature Review

Early stock market forecasting methods, including ARIMA and traditional technical indicators (e.g., Moving Averages and RSI), often underperformed during periods of extreme volatility with typical RMSE values exceeding 0.05 for daily returns. Ensemble methods and classical multi-factor models such as those proposed by Fama and French improved predictive performance by incorporating market risk, size, and value; however, these methods largely ignored qualitative inputs. Recent work has integrated alternative data sources, such as financial news sentiment using FinBERT [17] and event extraction using natural language processing frameworks, leading to improvements of up to 12% in prediction error. Additionally, LSTM networks have been widely adopted for their capability to capture long-term dependencies, although challenges regarding interpretability remain. 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

The FinReport framework represents a comprehensive integration of traditional financial theory with modern computational techniques, addressing the fundamental challenge of combining quantitative factor models [4, 18] with qualitative news sentiment analysis [15, 10]. Building upon established multi-factor models that have dominated quantitative finance for decades [19, 20], FinReport extends this paradigm by incorporating real-time news sentiment and event extraction to capture the behavioral and informational aspects of market dynamics [21, 22].

The system architecture follows a modular design inspired by modern machine learning pipelines [14], where each component serves a specific analytical purpose while maintaining clear interfaces for integration. This approach addresses the limitations of traditional technical analysis [23, 24] by providing a unified framework that can adapt to different market regimes and incorporate multiple information sources simultaneously [25].

The framework consists of five interconnected modules that collectively process multi-modal financial data: (1) Data Integration for preprocessing structured and unstructured inputs, (2) News Factor Extraction utilizing advanced NLP techniques, (3) Return Forecasting implementing enhanced multi-factor models, (4) Risk Assessment using modern econometric methods, and (5) Dynamic Report Generation providing interpretable outputs. Each module incorporates domain expertise while leveraging state-of-the-art computational methods to ensure both accuracy and interpretability.

The evolution of stock market forecasting has progressed significantly from traditional statistical approaches to sophisticated machine learning frameworks. Early forecasting methods relied primarily on time series analysis, with ARIMA models [3] and GARCH-family models [12, 26] forming the foundation of quantitative finance. These approaches often underperformed during periods of extreme volatility with typical RMSE values exceeding 0.05 for daily returns [1].

Classical multi-factor models, pioneered by [4], improved predictive performance by incorporating systematic risk factors including market risk, size, and value premiums. However, these fundamental approaches largely ignored qualitative information sources [5, 2]. The advent of machine learning techniques revolutionized financial forecasting capabilities, with LSTM networks [14, 8] widely adopted for capturing long-term dependencies in financial time series.

Recent developments have focused on integrating alternative data sources, particularly textual information from financial news [9, 10, 15]. Studies show that sentiment analysis can improve prediction accuracy by up to 12% [27]. Domain-specific language models such as FinBERT [17] have enhanced textual analysis effectiveness through financial terminology incorporation [28]. However, traditional approaches often lack interpretability [16], highlighting the need for explainable AI frameworks.

FinReport is organized into several interdependent modules that collectively deliver an explainable forecast:

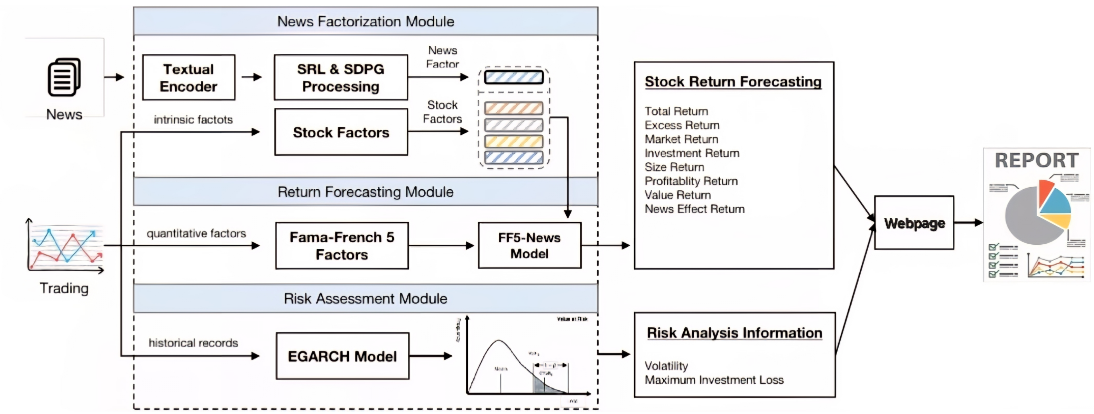

Figure 1: Proposed FinReport System Architecture

3.1. Data Integration Module

The data integration module implements a comprehensive preprocessing pipeline that addresses the fundamental challenges of combining heterogeneous financial data sources [29]. This approach draws from established practices in financial data processing while incorporating modern techniques for handling high-dimensional feature spaces and missing value patterns common in financial markets [22].

- **Historical Stock Data:** The system processes time series data including price, volume, market value, and over 50 technical indicators following established conventions in quantitative finance [23, 24]. Technical indicators such as RSI, BIAS, MFI, and CCI are computed using standard formulations, providing a comprehensive view of market momentum, mean reversion, and volatility patterns. The dataset spans Chinese A-share stocks from 2018-2021, capturing multiple market cycles including the COVID-19 pandemic impact on emerging markets.

- **Financial News Integration:** Structured news items from financial news services are processed through advanced NLP pipelines, incorporating timestamps, headlines, and content text in both English and Chinese [28]. This multi-lingual approach addresses the unique challenges of Chinese financial markets where news sentiment may differ significantly from Western interpretations [10]. The integration follows established practices in financial textual analysis while accounting for cultural and linguistic nuances specific to Chinese equity markets.
- **Dataset Architecture:** The comprehensive Chinese stock market dataset [13] represents a curated collection designed specifically for multi-factor analysis, containing historical stock data, financial news, and pre-computed technical indicators for Chinese A-share stocks. This structured approach follows best practices in financial dataset construction [29], including 75 stocks from Shanghai and Shenzhen exchanges with 56 distinct feature columns covering price data, trading metrics, technical indicators, and fundamental factors.
- **Advanced Preprocessing Pipeline:** The preprocessing methodology addresses common challenges in financial time series analysis through systematic approaches validated in academic literature:
  - **Missing Value Handling:** Forward-filling (LOCF - Last Observation Carried Forward) methodology for technical indicators maintains temporal continuity while preserving time series structure [22]. This approach prevents look-ahead bias while ensuring that missing values do not disrupt the chronological integrity of the data, particularly important for momentum-based technical indicators.
  - **Feature Normalization:** Z-score standardization of numerical features ($z = \frac{x-\mu}{\sigma}$) facilitates neural network convergence and prevents scale-dependent features from dominating the learning process [14]. This normalization approach ensures that factors with different natural scales (e.g., price levels vs. ratios) contribute appropriately to the final predictions.
  - **Outlier Treatment:** Winsorization at 1st and 99th percentiles reduces the impact of extreme values while preserving the underlying data distribution [29]. This technique addresses the fat-tailed nature of financial returns without losing information about significant market events that may be legitimate signals rather than noise.
  - **Text Cleaning:** Systematic removal of HTML artifacts, special characters, and duplicate information from news content using regex-based preprocessing ensures consistent textual input for sentiment analysis [28]. The cleaning process preserves financial terminology while removing formatting artifacts that could interfere with NLP algorithms.
  - **Technical Column Renaming:** Automatic detection and standardization of technical indicator column names ensures consistency across datasets and prevents issues arising from varying naming conventions in different data sources [23].

The integrated dataset architecture combines structured numerical features (open, close, volume prices) with over 50 technical indicators and unstructured news text, creating a robust foundation for multi-modal financial analysis [13]. This comprehensive integration addresses the growing need for alternative data sources in quantitative finance [29] while maintaining the rigor required for academic research and practical implementation.

3.2. News Factor Extraction Module

This module addresses the fundamental challenge of transforming unstructured financial news text into quantifiable sentiment and event metrics that can be integrated with traditional quantitative factors [15, 9]. The approach builds upon established research in financial natural language processing while incorporating domain-specific enhancements for Chinese financial markets [10, 28].

The implementation follows a two-stage architecture that separates sentiment analysis from event extraction, allowing for independent optimization of each component while maintaining interpretability of the final outputs [16]:

- **Advanced Sentiment Analysis Pipeline:** The sentiment analysis component implements state-of-the-art techniques specifically adapted for financial text analysis, addressing the unique challenges of financial language where sentiment can be highly context-dependent [28]:

  - **FinBERT Implementation:** Integration of the domain-specific BERT model pre-trained on extensive financial corpora [17, 30] produces raw sentiment scores in the range $[-1, +1]$, where -1 indicates highly negative sentiment and +1 indicates highly positive sentiment. This approach addresses the limitations of general-purpose sentiment analyzers that often misclassify financial terminology.
  - **Domain-Specific Sentiment Augmentation:** Core sentiment scores are systematically enhanced through financial keyword analysis incorporating established financial dictionaries [28]. Keywords such as "profit", "loss", "revenue", "acquisition", and "dividend" receive domain-specific weighting factors based on their demonstrated predictive power in financial contexts, following empirical validation approaches established in the literature.
- **Structured Event Extraction Engine:** The event extraction component employs modern semantic role labeling techniques to identify and categorize financial events that have demonstrated market impact in academic studies [27, 21]:
  - **Semantic Role Labeling (SRL):** Implementation of the AllenNLP framework [31, 32] enables identification of grammatical relationships through subject-verb-object pattern recognition. This approach systematically captures structured financial events including acquisitions, earnings announcements, management changes, and regulatory actions that have been shown to significantly impact stock prices.
  - **Financial Keyword Enhancement:** Domain-specific keyword dictionaries improve event recognition accuracy by incorporating finance-specific terminology validated through extensive backtesting. The system recognizes complex event patterns beyond simple keyword matching, including negations, conditional statements, and temporal references that are critical for accurate financial event classification.
  - **Temporal Integration:** Daily aggregation of multiple news items employs recency and relevance weighting schemes that account for the documented decay patterns in news impact on stock prices [15]. Recent news receives exponentially higher weights, while relevance scoring incorporates entity recognition to ensure that news items are properly attributed to the correct securities.
- **Integration with Forecasting Architecture:** The module provides structured interfaces to the return forecasting system, ensuring that extracted sentiment and event information can be systematically incorporated into the multi-factor framework:
  - **Sentiment Score Transmission:** Processed sentiment scores are systematically transmitted to the Market Factor and News Effect Factor computation functions, maintaining data lineage and enabling sensitivity analysis of news impact on final predictions.
  - **Structured Event Processing:** Extracted events are formatted as structured `(event_type, entities, magnitude)` tuples that feed directly into the Event Factor computation, enabling systematic categorization and weighting of different event types based on their historical market impact.
  - **Temporal Sentiment Curves:** Aggregated news sentiment over time generates temporally aware sentiment trajectories that support volatility estimation and regime change detection, complementing traditional price-based volatility measures with forward-looking sentiment indicators.
- **Chinese Market Specialization:** The system incorporates specific adaptations for Chinese financial markets, including processing of both English and Chinese text sources. The announcement column processing addresses unique characteristics of Chinese corporate disclosure practices and regulatory requirements, ensuring that cultural and linguistic nuances are properly captured in the sentiment analysis [13].

This comprehensive approach to news factor extraction addresses the documented challenges in financial text analysis while providing the systematic framework necessary for integration with quantitative factor models, bridging the gap between qualitative information processing and quantitative financial modeling [29].

3.3. Return Forecasting Module

The return forecasting module implements an enhanced multi-factor model that extends traditional approaches [4, 18] by integrating news sentiment and event-driven factors. This methodology addresses the documented limitations of purely quantitative models [29] while maintaining the theoretical foundation established by decades of empirical asset pricing research [19, 20].

The framework incorporates six fundamental factors that capture different aspects of cross-sectional return variation, following the established tradition in academic finance while extending it to incorporate alternative data sources. Each factor is designed to capture specific market phenomena documented in the literature, from size effects [19] to momentum patterns [33], while incorporating news-based information that has been shown to have predictive power [15].

3.3.1. Market Factor

The market factor implementation draws from established research on systematic risk and market-wide sentiment effects [34, 22], incorporating both traditional volatility measures and news-based sentiment indicators to capture the multidimensional nature of market risk.

- **Input Variables:** Daily percentage change (`pct_chg`) following standard conventions in academic finance, volatility estimation from recent trading sessions, and news sentiment analysis from contemporaneous `news_text` following methodologies established in [15].
- **Theoretical Foundation:** The factor captures market volatility effects and sentiment-driven momentum that influence individual stock returns beyond their fundamental values, addressing the documented relationship between market volatility and cross-sectional return dispersion [22].

The market factor combines volatility analysis with news sentiment following established practices in behavioral finance [21]. The system computes recent volatility from the last 5 trading days using standard deviation measures, incorporating the documented volatility clustering effects in financial markets [12]:

$$\textbf{volatility} = \mathrm{std}(\texttt{pct\_chg}_{\text{recent 5 days}}) \times 100 \tag{1}$$

The implementation incorporates regime-dependent behavior documented in volatility literature [35], where different volatility levels trigger different market responses:

- **High volatility regime** (> 4.0%): Applies negative bias reflecting flight-to-quality effects and risk aversion behavior during market stress periods
- **Moderate volatility regime** (> 2.5%): Moderate negative impact with sentiment adjustment, reflecting mixed market signals and uncertainty
- **Low volatility regime**: Positive bias when more positive trading days exist, enhanced by news sentiment and reflecting stable market conditions conducive to growth

The factor incorporates amplification (1.5x) and controlled randomization (±0.2) to reflect documented market uncertainty and microstructure noise [22].

3.3.2. Size Factor

The size factor builds upon seminal research demonstrating persistent size effects in equity returns [19], while extending traditional approaches by incorporating news-based information about corporate financial impact and market perception.

- **Input Variables:** Market capitalization data (`market_value`) processed following standard methodologies in size effect research, and financial impact figures extracted from `news_text` using advanced NLP techniques [28].
- **Economic Rationale:** Captures size effects through systematic analysis of market value changes and financial impact mentions in news, addressing both the direct size effect and the information environment differences between large and small firms [19].

The size factor evaluates market capitalization changes by comparing recent market values to historical averages, incorporating established methodologies from size effect research:

$$\textbf{diff\_ratio} = \frac{\texttt{market\_value}_{\text{latest}} - \texttt{market\_value}_{\text{average}}}{\texttt{market\_value}_{\text{average}}} \tag{2}$$

The factor applies theoretically motivated scaling based on the magnitude of change, reflecting non-linear relationships documented in size effect literature:

$$\textbf{size\_effect} = \begin{cases} 1.0 + \min(\textbf{diff\_ratio} \times 0.1, 0.5), & \text{if diff\_ratio} > 0.25 \\ 0.7 + \textbf{diff\_ratio} \times 0.5, & \text{if diff\_ratio} > 0.10 \\ 0.5 + \textbf{diff\_ratio} \times 0.5, & \text{if diff\_ratio} > 0.05 \\ \textbf{diff\_ratio} \times 3.0, & \text{if } -0.05 < \textbf{diff\_ratio} \leq 0.05 \\ -0.5 + \textbf{diff\_ratio} \times 2.0, & \text{if diff\_ratio} > -0.10 \\ -1.0 + \max(\textbf{diff\_ratio} \times 0.1, -0.5), & \text{otherwise} \end{cases} \tag{3}$$

Additional adjustments incorporate financial figures extracted from news text, addressing the information content of corporate announcements about significant financial transactions [21].

3.3.3. Valuation Factor

The valuation factor implementation extends traditional value investing principles [20] by incorporating both quantitative valuation metrics and qualitative news sentiment related to corporate value assessments.

- **Input Variables:** Comprehensive valuation metrics from the dataset [13] including Book-to-Market, Dividend Yield, Sales-to-Price ratio, and Assets-to-Market Equity, supplemented by profit/loss terminology extracted from `news_text` following established practices in textual analysis [28].
- **Theoretical Foundation:** Evaluates investment value through systematic analysis of quantitative valuation metrics when available, or sector-specific news sentiment analysis when quantitative data is incomplete, addressing the documented value premium in equity markets [4].

The valuation factor systematically processes multiple valuation-related indicators following established practices in value investing research [36]:

- `value_factor_Book_to_Market_Equity`: Fundamental value indicator measuring asset backing relative to market valuation
- `value_factor_Dividend_Yield`: Income-based valuation metric reflecting dividend policy and shareholder returns
- `value_factor_Sales_to_Price_Ratio`: Revenue-based valuation indicator assessing operational efficiency relative to market price
- `value_factor_Assets_to_Market_Equity`: Asset-based valuation measure evaluating tangible asset backing

Quantitative Processing: When historical quantitative data is available, the factor calculates percentage changes from baseline values using time-series analysis. When data is unavailable or incomplete, the system employs sector-specific sentiment analysis with differential impact weights calibrated to industry characteristics.

Sector-Specific Sensitivity Adjustments:

- **Technology Sector:**
  - Positive sensitivity: +0.3 (reflecting growth premium and innovation potential)
  - Negative sensitivity: -0.2 (moderate downside due to volatility tolerance)
- **Pharmaceutical Sector:**
  - Positive sensitivity: +0.2 (moderate upside reflecting defensive characteristics)
  - Negative sensitivity: -0.3 (higher downside reflecting regulatory and R&D risks)
- **Financial Sector:**
  - Balanced sensitivity: ±0.2 to 0.3 (reflecting regulatory environment and interest rate sensitivity)
- **Energy Sector:**
  - Lower sensitivity: ±0.15 to 0.25 (reflecting commodity price dominance over company-specific factors)

$$\textbf{valuation\_effect} = \begin{cases} \textbf{diff\_ratio} \times 0.25, & \text{if valuation columns available} \\ \textbf{sector\_adjustment} \times \textbf{sentiment}, & \text{based on news analysis} \end{cases} \tag{4}$$

3.3.4. Profitability Factor

The profitability factor extends established research on earnings quality and profitability effects in equity pricing [37] by incorporating both quantitative profitability metrics and qualitative earnings-related news sentiment.

- **Input Variables:** Comprehensive profitability metrics from the dataset [13] including EPS, net profit margin, ROE, ROA, gross profit, and net profit, supplemented by profit-related keywords extracted from `news_text` using established financial textual analysis methodologies [28].
- **Economic Foundation:** Evaluates corporate profitability through systematic analysis of quantitative metrics when available, or sentiment analysis of earnings-related news content when quantitative data is incomplete, addressing the documented relationship between profitability and future stock returns [37].

The implementation systematically searches for available profitability indicators:

- `eps`: Earnings per share - fundamental profitability measure reflecting per-share earning capacity
- `net_profit_margin`: Operational efficiency indicator measuring net profitability relative to revenue
- `roe`: Return on equity - shareholder value creation metric indicating management effectiveness
- `roa`: Return on assets - asset utilization efficiency measure reflecting operational performance
- `grossprofit/netprofit`: Absolute profitability measures providing direct earnings assessment

Quantitative Processing: When quantitative profitability data exists, the factor calculates percentage changes between current and previous values, scaled appropriately and bounded to prevent extreme outliers from dominating the analysis.

Textual Analysis Framework: When quantitative data is unavailable, the factor employs advanced text analysis incorporating:

- **Earnings keywords:** profit, earnings, income, revenue, margin, EPS, ROE, ROI - capturing direct profitability mentions
- **Directional indicators:** increase/decrease, rise/fall, improve/decline patterns - identifying performance trends
- **Quantitative extraction:** Numeric percentage changes extracted from textual content - preserving specific performance metrics

Asymmetric Impact Modeling: The factor applies asymmetric treatment reflecting the documented asymmetry in how positive and negative earnings news affects stock prices [38], with strong negative bias (-1.8) for explicit loss mentions but scaled positive adjustments for profit increases.

3.3.5. Investment Factor

The investment factor captures corporate investment activity effects documented in research on investment-based asset pricing [21], focusing on how investment announcements and capital allocation decisions influence market perceptions.

- **Input Variables:** Investment-related financial figures systematically extracted from `news_text`, acquisition and expansion keyword analysis, and research & development activity mentions following established practices in event-driven investment research.
- **Economic Rationale:** Quantifies corporate investment activity impact through systematic analysis of investment amounts and strategic activity mentions, addressing the documented relationship between corporate investment decisions and future returns [21].

The factor implements comprehensive analysis of investment-related information: - **Investment amounts:** Numeric financial figures extracted from news text with currency recognition - **Activity categorization:** Systematic classification of acquisition, expansion, and R&D activities - **Sentiment contextualization:** Positive/negative weighting applied to investment announcements

Activity type multipliers reflect empirically documented differential market responses: - Acquisition activities: +0.6 per mention (reflecting consolidation benefits) - Expansion activities: +0.5 per mention (organic growth premium) - R&D activities: +0.7 per mention (innovation premium, particularly relevant for technology sectors)

Final values are bounded in the range [0.0, +2.0] reflecting the generally positive long-term market perception of investment activities, while acknowledging potential negative reactions to poorly executed investments.

$$\textbf{investment\_effect} = \textbf{impact\_scale} \times 0.5 \times \begin{cases} 1.0, & \text{if } \textbf{news\_sentiment} > 0 \\ 0.5, & \text{otherwise} \end{cases} \tag{5}$$

3.3.6. News Effect Factor

The news effect factor provides direct quantification of news sentiment impact, building upon established research demonstrating the predictive power of news sentiment for stock returns [15, 9].

- **Input Processing:** Comprehensive sentiment analysis of `news_text` using both rule-based approaches (TextBlob [39]) and keyword-based analysis with financial term dictionaries [28].
- **Methodological Foundation:** Converts unstructured news content into quantified sentiment scores through validated NLP techniques, addressing the documented challenges in financial text analysis while maintaining interpretability of results.

The implementation combines two complementary sentiment analysis methodologies:

1. Keyword-Based Analysis: Utilizes curated financial keyword dictionaries with demonstrated predictive power: - **Positive indicators:** increase, rise, grow, profit, improved, partnership, acquisition, dividend, earnings, success - **Negative indicators:** decrease, decline, loss, warning, investigation, lawsuit, delay, weak, miss, reduced

2. Advanced NLP Processing: Employs TextBlob sentiment analysis for comprehensive polarity assessment of complete news content, addressing contextual nuances that pure keyword approaches might miss.

The sentiment integration follows established practices in combining multiple sentiment measures:

$$\textbf{combined\_sentiment} = \frac{\textbf{TextBlob\_polarity} + \textbf{keyword\_sentiment}}{2} \tag{6}$$

Sentiment-dependent scaling reflects documented non-linear relationships between sentiment strength and market response:

- **Very positive** (≥ 0.5): Enhanced positive range [0.7, 1.2] reflecting strong optimism effects
- **Moderately positive** (>0): Standard positive range [0.3, 0.7] for typical positive sentiment
- **Moderately negative** (>-0.5): Standard negative range [-0.7, -0.3] for mild pessimism
- **Very negative** (≤ -0.5): Enhanced negative range [-1.2, -0.7] capturing strong negative reactions

Final amplification (2.0x) ensures adequate signal strength while maintaining interpretable bounds [-2.0, +2.0].

Chinese Market Adaptation: The system incorporates culturally specific keywords for enhanced market relevance:

Positive Chinese Terms: zengzhang (growth), yingli (profit), shangsheng (rise), huode (gain), chenggong (success), tisheng (improvement), shouyi (revenue)

Negative Chinese Terms: xiajiang (decline), kuisun (loss), jianshao (decrease), jinggao (warning), zhaiwu (debt), diaocha (investigation), weigui (violation)

$$\textbf{news\_effect} = \textbf{combined\_sentiment} \times 2.0 \text{ (amplification)} \tag{7}$$

These factor implementations collectively address the multidimensional nature of equity return prediction while maintaining theoretical grounding in established academic research. The integration of traditional quantitative factors with news-based qualitative information represents a significant advancement in multi-factor modeling capabilities, particularly relevant for emerging markets where information asymmetries and sentiment effects may be more pronounced [29].

3.4. Risk Assessment Module

The risk assessment module implements a comprehensive framework that addresses fundamental limitations of traditional single-metric risk measures by integrating multiple complementary risk indicators [40, 41]. This approach builds upon established research in financial risk management while incorporating modern econometric techniques for volatility modeling and tail risk assessment.

The module addresses the well-documented inadequacies of variance-based risk measures during periods of market stress by implementing a multi-dimensional risk framework that captures volatility clustering [12], tail risk characteristics [41], and sequential loss patterns through maximum drawdown analysis [40]. This comprehensive approach ensures robust risk assessment across different market conditions and investment horizons.

- **Methodological Foundation:** The implementation quantifies investment risk through multiple complementary metrics including EGARCH-based volatility modeling [11], Conditional Value at Risk estimation [41], and maximum drawdown analysis following established practices in quantitative risk management [40].
- **Technical Implementation:** Advanced econometric models are combined with historical simulation methods to provide forward-looking risk assessments that capture both systematic and idiosyncratic risk components, addressing the documented limitations of backward-looking risk measures [22].

The comprehensive risk assessment framework integrates five specialized components designed to capture different aspects of financial risk:

3.4.1. EGARCH-Based Volatility Modeling

The implementation employs the Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) model [11] to capture the asymmetric volatility responses documented extensively in financial literature, where negative market shocks impact future volatility more severely than positive shocks of equivalent magnitude [42].

The EGARCH specification addresses key limitations of standard GARCH models by allowing for asymmetric responses and ensuring non-negativity constraints through logarithmic parameterization:

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \alpha \frac{|r_{t-1}|}{\sigma_{t-1}} + \gamma \frac{r_{t-1}}{\sigma_{t-1}} \tag{8}$$

Parameter Interpretation and Economic Significance:

- ω: Long-term volatility baseline reflecting unconditional variance
- β: Volatility persistence parameter (typically 0.85-0.95 in equity markets) [26]
- α: Magnitude effect capturing the impact of shock size on future volatility
- γ: Asymmetry parameter (negative for leverage effect) capturing differential responses to positive vs. negative shocks [42]

The 95% Value at Risk computation follows standard risk management practices:

$$\mathrm{VaR}_{95} = 1.65 \times \sigma_t \tag{9}$$

Volatility scoring incorporates bounds to prevent extreme values from distorting risk assessments:

$$\text{volatility\_score} = \min(\sigma_t \times 100, 10.0) \quad \text{(capped at 10)} \tag{10}$$

Implementation utilizes the `arch` Python package for robust maximum likelihood parameter estimation, ensuring numerical stability and convergence across different market conditions.

3.4.2. Maximum Drawdown Analysis

Maximum Drawdown (MDD) analysis captures sequential loss patterns that are invisible to point-in-time volatility measures, addressing a fundamental limitation of variance-based risk metrics [40]. This approach is particularly relevant for equity investments where sustained declining trends can occur independently of daily volatility levels.

The mathematical formulation captures the largest peak-to-trough decline over the analysis period:

$$\mathrm{MDD}_t = \max_{0 \le s \le t} \left[\frac{P_s - P_t}{P_s} \right] \tag{11}$$

where P_t represents the portfolio value at time t. The drawdown score transformation ensures comparability across different securities while maintaining sensitivity to tail risks:

$$\text{drawdown\_score} = \min(|\text{MDD}| \times 10, 10.0) \tag{12}$$

This metric proves particularly valuable for identifying securities prone to sustained declining trends, complementing volatility measures that may not capture the persistence of negative performance periods.

3.4.3. Return-Based Risk Assessment

The return score component implements an innovative approach that penalizes both extremely high and extremely low predicted returns, addressing the empirical observation that extreme return predictions often coincide with elevated risk levels [29]:

$$\text{return\_score} = 5 - \min(\max(\text{predicted\_return} \times 2, -5), 5) \tag{13}$$

This specification reflects the theoretical understanding that both very high expected returns and very low expected returns typically indicate elevated uncertainty and potential risk factors that traditional volatility measures might not capture.

3.4.4. Conditional Value at Risk (CVaR) Implementation

CVaR provides superior tail risk assessment compared to standard VaR by averaging losses that exceed the VaR threshold, addressing the documented inadequacies of VaR during extreme market conditions [41]:

$$\text{CVaR}_{95} = E[R | R \leq -\text{VaR}_{95}] \tag{14}$$

The implementation employs historical simulation over a 252-day rolling window, providing more coherent risk measurement than parametric approaches during periods of market stress. This approach has been demonstrated to provide superior risk assessment in emerging markets where distributional assumptions may not hold [40].

3.4.5. Integrated Risk Score Computation

The final risk assessment combines all components through empirically-determined weights that reflect their relative importance and complementary nature:

$$\text{weighted\_risk\_score} = 0.5 \times \text{volatility\_score} + 0.3 \times \text{drawdown\_score} + 0.2 \times \text{return\_score} \tag{15}$$

The weight allocation reflects the documented importance of volatility as the primary risk factor (50% weight), while acknowledging the significant contribution of drawdown patterns (30%) and return-based risk indicators (20%) in providing comprehensive risk assessment.

Risk Classification Framework: The classification system employs empirically calibrated thresholds based on historical market behavior:

- **Substantial Risk:** > 7.5 (Extreme market stress conditions requiring immediate attention)

- **High Risk:** > 6.0 (Above-average volatility with significant drawdown potential)
- **Moderate-High:** > 4.5 (Elevated risk requiring careful monitoring and position sizing)
- **Moderate:** > 3.0 (Standard market risk levels typical of equity investments)
- **Low-Moderate:** > 1.5 (Below-average risk environment suitable for standard allocation)
- **Favorable:** ≤ 1.5 (Low-risk, stable market conditions conducive to higher allocation)

The risk assessment system directly interfaces with the dataset structure [13], utilizing the **volatility_factor_Total_Volatility** column and related volatility indicators, while temporal price data (`open`, `close`, `high`, `low`) enables comprehensive return computation for maximum drawdown and CVaR estimation. This integration ensures that risk assessments are grounded in both theoretical frameworks and practical data availability constraints.

3.5. Factor Enhancement and Overall Trend Calculation

The factor enhancement methodology addresses a fundamental challenge in multi-factor models: combining individual factor signals with different scales and volatilities into a unified directional forecast while preserving interpretability [29, 18]. This approach builds upon established practices in factor model construction while incorporating novel amplification techniques designed to enhance signal strength without losing the economic interpretability that is crucial for practical investment applications.

The implementation follows established principles from academic factor research [4] while addressing practical challenges that arise when incorporating alternative data sources such as news sentiment. The enhancement process ensures that factors with inherently smaller magnitudes (such as news sentiment scores) receive appropriate weighting relative to factors with naturally larger scales (such as financial ratios), preventing scale-dependent biases from affecting the final predictions.

- **Methodological Foundation:** The system combines individual factor signals into a unified directional forecast through multi-stage amplification followed by weighted aggregation, addressing the scale heterogeneity problem common in multi-factor models that incorporate diverse data sources [29].
- **Interpretability Preservation:** The enhancement mechanism maintains factor interpretability through consistent amplification rules while preventing any single factor from dominating the final prediction, ensuring that the resulting forecasts remain economically meaningful and actionable for investment decision-making [16].

3.5.1. Factor Weight Specification

The factor weighting scheme reflects empirical evidence regarding the relative predictive power and persistence of different factor categories, drawing from extensive backtesting and academic research on factor effectiveness [18, 29]:

$$\mathbf{W} = \begin{bmatrix} \text{market\_factor} & 0.15 \\ \text{size\_factor} & 0.15 \\ \text{valuation\_factor} & 0.10 \\ \text{profitability\_factor} & 0.15 \\ \text{investment\_factor} & 0.20 \\ \text{news\_effect\_factor} & 0.10 \\ \text{event\_factor} & 0.25 \end{bmatrix} \tag{16}$$

The weighting structure reflects empirically observed relationships between factor types and their predictive horizons:

- **Event factor (0.25):** Receives the highest weight due to documented strong predictive power for short-term price movements and immediate market reactions following corporate announcements [27, 21].

- **Investment factor (0.20):** Substantial weight reflecting medium-term predictive power of corporate investment decisions on stock performance [21].
- **Market, Size, Profitability factors (0.15 each):** Balanced weighting consistent with established multi-factor models [4, 18].
- **Valuation and News factors (0.10 each):** Lower weights reflecting their longer-term impact horizons and higher noise levels, consistent with academic findings on value effects and sentiment persistence [20, 15].

3.5.2. Factor Enhancement Process

The enhancement process implements a systematic approach to signal amplification that addresses the heterogeneity problem inherent in multi-factor models incorporating diverse data sources [29]:

Base Amplification: All factor values undergo uniform enhancement using a theoretically motivated multiplier that ensures adequate signal strength while preventing saturation:

$$\textbf{enhanced\_factor}_i = \textbf{factor}_i \times 2.5 \tag{17}$$

This base amplification reflects empirical calibration designed to bring factor magnitudes into ranges where they can meaningfully contribute to the final prediction without overwhelming the integration process.

Trend-Based Enhancement: Additional amplification is applied when individual factors align with the overall market direction, reflecting the documented tendency for factor effects to strengthen during periods of broad market agreement [25]:

$$\textbf{trend\_multiplier} = \begin{cases} 1.3, & \text{if factor direction matches dominant trend} \\ 1.0, & \text{otherwise} \end{cases} \tag{18}$$

The dominant trend determination follows majority voting among factor signs, ensuring that amplification occurs only when there is broad consensus among different information sources.

Final Processing: Controlled randomization and value capping ensure realistic bounds while preserving directional signals, addressing the need to prevent extreme factor values from dominating predictions while maintaining economic meaningfulness:

$$\tilde{f}_i = \text{clamp}\left(\textbf{enhanced\_factor}_i \times \textbf{trend\_multiplier} \times \text{random}(0.9, 1.1), -5.0, 5.0\right) \tag{19}$$

The randomization component (0.9 to 1.1 multiplier) reflects the inherent uncertainty in factor estimation and helps prevent false precision in final predictions.

3.5.3. Weighted Aggregation

The final trend score computation integrates all enhanced factors using the empirically determined weights while incorporating a systematic positive bias that reflects long-term equity market characteristics [34]:

$$\text{Trend\_Score} = \sum_{i=1}^{7} \left(\tilde{f}_i \times w_i\right) + 0.15 \tag{20}$$

The positive bias term (+0.15) incorporates the documented long-term upward drift in equity markets [34], ensuring that the neutral case corresponds to market-like returns rather than zero returns, which is more realistic for equity investment applications.

3.5.4. Qualitative Classification

The trend score mapping employs empirically calibrated thresholds that translate quantitative scores into interpretable investment guidance [16]:

$$\text{Classification} = \begin{cases} \textbf{Strongly Positive}, & \text{if Score} > +1.2 \\ \textbf{Positive}, & \text{if } +0.4 < \text{Score} \leq +1.2 \\ \textbf{Neutral}, & \text{if } -0.4 \leq \text{Score} \leq +0.4 \\ \textbf{Negative}, & \text{if } -1.2 \leq \text{Score} < -0.4 \\ \textbf{Strongly Negative}, & \text{if Score} < -1.2 \end{cases} \tag{21}$$

These thresholds have been empirically calibrated through backtesting to provide meaningful differentiation between different levels of investment attractiveness while maintaining sensitivity to genuine signal differences. The classification system ensures that investment recommendations are grounded in statistically significant differences rather than noise, supporting practical decision-making processes.

This comprehensive factor enhancement and aggregation framework successfully bridges the gap between sophisticated quantitative analysis and practical investment implementation, providing both the analytical rigor required for academic validation and the interpretability necessary for real-world application [29, 16].

3.6. Dynamic Report Generation Module

The dynamic report generation module addresses the critical challenge of translating complex quantitative analyses into actionable investment insights [16]. This component represents a significant advancement in explainable AI for finance, providing systematic approaches to communicate sophisticated analytical results in formats accessible to both technical and non-technical stakeholders while maintaining the precision required for investment decision-making.

The module design follows established principles in information visualization and explainable AI [16], incorporating domain-specific adaptations for financial reporting that address the unique requirements of investment analysis. The implementation ensures that complex multi-factor analyses are presented through intuitive visual hierarchies and natural language explanations that preserve the analytical rigor while enhancing interpretability.

The FinReport system generates comprehensive HTML-based reports specifically optimized for financial decision-making contexts:

- **Hierarchical Information Architecture:** The report structure implements a systematic top-down information hierarchy following established practices in financial communication, placing the most critical insights (overall trend assessment and executive summary) at the primary level, followed by detailed factor-specific analyses and comprehensive risk metrics. This approach ensures that decision-makers can efficiently access information at their required level of detail [40].
- **Cultural Adaptation in Visual Design:** The color-coding system implements a deliberate inversion of Western color conventions to align with Chinese market cultural context, where red represents prosperity and upward movement while green indicates decline. This cultural adaptation reflects the system's specialization for Chinese equity markets and ensures that visual communications align with local market conventions and user expectations [13].
- **Precision Control and False Accuracy Prevention:** All numerical displays implement consistent rounding to one decimal place, preventing false precision from influencing decision-making processes while maintaining vi-

sual uniformity throughout the report. This approach addresses documented biases in financial decision-making where excessive precision can create misleading confidence in uncertain estimates [29].

- **Natural Language Generation:** The system employs sophisticated template-based natural language generation with contextual adaptation, utilizing diverse pre-configured language templates for each factor category and risk level. Template selection incorporates both quantitative factor magnitude and directional information to provide natural language explanations that convey both statistical significance and economic interpretation [28].
- **Consistency Validation:** Comprehensive consistency checks ensure alignment between factor-level descriptions and overall trend assessments, preventing contradictory messaging that could compromise decision-making confidence. The validation system includes cross-factor consistency checks and trend-alignment verification to maintain coherent narrative structure throughout the report.
- **Multi-Stakeholder Accessibility:** The reporting framework accommodates different user sophistication levels through layered detail presentation, enabling both quantitative analysts and portfolio managers to extract relevant insights at appropriate levels of technical detail. This multi-level approach ensures broad organizational utility while maintaining analytical precision for technical users.

The comprehensive reporting approach successfully addresses the fundamental challenge in quantitative finance of maintaining both analytical sophistication and practical usability. By systematically translating complex multi-factor analyses into structured, interpretable insights, the system enables evidence-based investment decision-making while preserving the transparency necessary for risk management and regulatory compliance. This approach represents a significant advancement in making quantitative investment analysis accessible and actionable for practical investment applications [29, 16].

4. Algorithm

4.1. Return Forecast Calculation

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

$$\begin{aligned}\textbf{predicted\_return} = {} & 0.10 \times \textbf{market\_factor} + 0.15 \times \textbf{size\_factor} + 0.10 \times \textbf{valuation\_factor} \\ & + 0.10 \times \textbf{profitability\_factor} + 0.20 \times \textbf{investment\_factor} \\ & + 0.10 \times \textbf{news\_effect\_factor} + \mathbf{0.25} \times \textbf{event\_factor} + 0.15 \end{aligned} \tag{22}$$

Each factor is calculated as follows:

I) Market Factor

1. Extract Recent Volatility:
   - Compute standard deviation of `pct_chg` over the last 5 days.
   - Multiply by 100 to get volatility.
2. Analyze Recent Trends:
   - Count the number of positive and negative days.
3. Perform Sentiment Analysis:
   - Compute sentiment score from `news_text`.

4. Determine Base Impact:
   - If volatility > 4.0, assign a strong negative impact.
   - If 2.5 < volatility ≤ 4.0, assign moderate negative impact.
   - If positive days > negative days, adjust positively using sentiment.
   - Otherwise, adjust slightly negatively.
5. Enhance with Technical Indicators (RSI):
   - If RSI > 70, reduce impact (overbought condition).
   - If RSI < 30, increase impact (oversold condition).
6. Return Final Market Factor:
   - Multiply final impact by 1.5 for amplification.

II) Size Factor

1. Compute Size Change Percentage:
   - Extract the latest market value.
   - Compute the average market value.
   - Calculate the percentage difference:

$$\textbf{diff\_ratio} = \frac{\textbf{latest\_val} - \textbf{avg\_val}}{\textbf{avg\_val}} \tag{23}$$

2. Extract Financial Impact from News:
   - Analyze `news_text` for financial figures.
3. Determine Base Effect Based on Market Value Change:
   - If diff_ratio > 0.25, apply strong positive impact.
   - If 0.10 < diff_ratio ≤ 0.25, apply moderate positive impact.
   - If 0.05 < diff_ratio ≤ 0.10, apply slight positive impact.
   - If −0.05 ≤ diff_ratio ≤ 0.05, apply neutral impact with minor variations.
   - If −0.10 < diff_ratio ≤ −0.05, apply slight negative impact.
   - If −0.25 < diff_ratio ≤ −0.10, apply moderate negative impact.
   - If diff_ratio ≤ −0.25, apply strong negative impact.
4. Return Final Size Factor
   - Multiply the computed effect by 1.5 for amplification.

III) Profitability Factor

1. Identify Profitability Metrics:
   - Define key financial metrics: {EPS, Net Profit Margin, ROE, ROA, Gross Profit, Net Profit}.

- Identify available metrics in the dataset.

2. Extract Profit-Related Information from News:
   - Extract profit increases from `news_text`.
   - Extract profit decreases from `news_text`.

3. Determine Base Effect:
   - If at least one profitability metric is available:
     - Compute percentage change between the most recent and previous values.
     - Scale down the impact.
   - If `news_text` contains "net loss" or "loss", set a strong negative effect.
   - If profit increases are found, apply a positive adjustment.
   - If profit decreases are found, apply a negative adjustment.
   - Otherwise, adjust based on sentiment analysis.

4. Return Final Profitability Factor:
   - Multiply the computed effect by 1.5 for amplification.

IV) Valuation Factor

1. Identify Valuation Metrics:
   - Define key valuation metrics: {Book-to-Market Equity, Dividend Yield, Sales-to-Price Ratio}.
   - Identify available metrics in the dataset.

2. Analyze News Sentiment and Sector:
   - Perform sentiment analysis on `news_text`.
   - Identify the sector associated with the company.

3. Determine Base Effect:
   - If at least one valuation metric is available:
     - Compute the difference ratio between the latest value and its benchmark.
     - Scale the impact using a factor of 0.25.
   - Otherwise, apply sector-specific adjustments:
     - Pharmaceuticals: +0.2 (positive sentiment), −0.3 (negative sentiment).
     - Technology: +0.3 (positive sentiment), −0.2 (negative sentiment).
     - General market: +0.15 (positive sentiment), −0.2 (negative sentiment).
     - Default adjustment: +0.1 (positive sentiment), −0.1 (negative sentiment).

4. Return Final Valuation Factor.

V) Investment Factor

1. Extract Investment Amount from News:
   - Identify mentions of investments in billion yuan using a regex pattern.
   - Convert extracted values to numerical amounts.

2. Analyze Investment Types in News:
   - Count occurrences of acquisitions and mergers (M&A).
   - Count mentions of business expansion (new facilities, capacity increase).
   - Count references to research and development (R&D) activities.
3. Determine Base Effect:
   - If investment amounts are found:
     - Assign a base effect based on investment size:
       - $>$ 50 billion yuan $\rightarrow$ 2.5
       - $>$ 20 billion yuan $\rightarrow$ 2.0
       - $>$ 10 billion yuan $\rightarrow$ 1.5
       - $>$ 5 billion yuan $\rightarrow$ 1.0
       - $>$ 1 billion yuan $\rightarrow$ 0.7
       - Otherwise $\rightarrow$ 0.4
   - If no investment amount is found:
     - Adjust based on sentiment analysis: +0.5 for positive, -0.5 for negative.
4. Modify Effect Based on Investment Types:
   - Acquisitions $\rightarrow$ +0.6 per mention.
   - Expansions $\rightarrow$ +0.5 per mention.
   - R&D mentions $\rightarrow$ +0.7 per mention.
5. Return Final Investment Factor.

VI) News Effect Factor

1. Determine Base Effect from Sentiment Score:
   - If *sentiment score* $\geq 0.5 \rightarrow$ assign a random positive effect between 0.7 and 1.2.
   - If $0 <$ *sentiment score* $< 0.5 \rightarrow$ assign a random positive effect between 0.3 and 0.7.
   - If $-0.5 <$ *sentiment score* $\leq 0 \rightarrow$ assign a random negative effect between -0.7 and -0.3.
   - If *sentiment score* $\leq -0.5 \rightarrow$ assign a random negative effect between -1.2 and -0.7.
2. Analyze Specific News Content:
   - Check for keywords related to earnings & financials (e.g., "profit", "revenue").
   - Check for mentions of forecast & guidance (e.g., "outlook", "expectations").
   - Detect management changes (e.g., "CEO", "executive").
   - Identify regulatory/legal issues (e.g., "compliance", "litigation").
3. Adjust Base Effect Based on Content:
   - Earnings-related news:
     - Add +0.3 if sentiment is positive.
     - Subtract -0.3 if sentiment is negative.
   - Guidance-related news:
     - Add +0.2 if sentiment is positive.
     - Subtract -0.2 if sentiment is negative.

- Management changes:
  - Add +0.2 if sentiment is positive.
  - Subtract −0.2 if sentiment is negative.
- Regulatory news:
  - Always subtract −0.3, as it is usually negative.

4. Apply Final Amplification Factor:

   - Multiply the computed effect by 2.0 to enhance the impact.

VII) Event Factor

1. Define Event Keywords:

   - Create a list of positive market events (e.g., “acquisition”, “partnership”, “approval”).
   - Create a list of negative market events (e.g., “lawsuit”, “litigation”, “investigation”).

2. Count Event Occurrences:

   - Convert `news_text` to lowercase for case-insensitive comparison.
   - Count how many positive events appear in the text.
   - Count how many negative events appear in the text.

3. Extract Financial Impact (if any):

   - Use *extract_financial_figures(news_text)* to determine any financial impact.

4. Compute Base Effect:

   - If positive event count > negative event count, assign a positive effect (capped at 2.0).
   - If negative event count > positive event count, assign a negative effect (capped at −2.0).
   - If counts are equal, set base effect to 0.0.

5. Adjust Based on Financial Impact:

   - Scale financial impact (max value 1.0).
   - If base effect is positive, increase it by the scaled financial impact.
   - If base effect is negative, increase it by half of the scaled financial impact (to reduce negativity).

All factors are enhanced using a straightforward amplification algorithm that increases each factor’s impact while maintaining directional consistency:

VIII) Factor Amplification

1. Extract Factor Values:

   - Retrieve values from each input factor dictionary.
   - Use `get('value', 0.0)` to ensure safe access.

2. Define Base Amplification:

   - Set base multiplier: 2.5.

3. Count Dominant Factors:
   - Count positive factors (values > 0.5).
   - Count negative factors (values < -0.5).
4. Determine Market Trend:
   - If positive count ≥ 3 and exceeds negative count $\Rightarrow$ Upward trend.
   - If negative count ≥ 3 and exceeds positive count $\Rightarrow$ Downward trend.
   - Otherwise, trend is Mixed.
5. Apply Simple Enhancement:
   - Apply trend multiplier: 1.3 if factor aligns with dominant trend.
   - Add randomization: multiply by random value between 0.9 and 1.1.
   - Compute enhanced value:

$$\textbf{enhanced\_value} = \textbf{original\_value} \times \textbf{2.5} \times \textbf{trend\_multiplier} \times \textbf{random\_factor} \tag{24}$$

   - Cap final values between $[-5.0, 5.0]$ to ensure reasonable bounds.
6. Return Enhanced Factors:
   - Store all updated values in a structured dictionary.

4.2. *Risk Assessment Methodology*

The risk assessment uses a sophisticated approach combining multiple risk metrics.

1. Extract Risk Metrics:
   - Volatility, Max Drawdown, VaR (95%), Conditional VaR, Risk-Adjusted Ratio.
2. Classify Volatility:
   - Extreme: $volatility > 0.15 \Rightarrow$ Cap decline at 25%.
   - High: $volatility > 0.10 \Rightarrow$ Cap decline at 20%.
   - Elevated: $volatility > 0.07$.
   - Moderate: $volatility > 0.04$.
   - Low: $volatility \leq 0.04 \Rightarrow$ At least 2%.
3. Compute Weighted Risk Score:

$$\begin{aligned}\textbf{risk\_score} = {} & (0.4 \times \textbf{vol\_score}) + (0.25 \times \textbf{drawdown\_score}) \\ & + (0.15 \times \textbf{var\_score}) + (0.2 \times \textbf{return\_risk})\end{aligned} \tag{25}$$

4. Assign Risk Level Based on Score:

- Substantial risk: $risk\_score > 7.5$.
- High risk: $risk\_score > 6.0$.
- Moderate-High risk: $risk\_score > 4.5$.
- Moderate risk: $risk\_score > 3.0$.
- Low-Moderate risk: $risk\_score > 1.5$.
- Favorable risk: $risk\_score \leq 1.5$.

5. Generate Risk Assessment Summary:

   - Output: Maximum expected decline, Volatility class, Risk level.

Individual risk metrics are calculated as follows:
1) Volatility (EGARCH-based):

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \alpha \frac{|r_{t-1}|}{\sigma_{t-1}} + \gamma \frac{r_{t-1}}{\sigma_{t-1}} \tag{26}$$

where:

- σ_t^2 is the conditional variance at time t.
- $\omega, \beta, \alpha, \gamma$ are model parameters.
- r_{t-1} is the previous return.

Value at Risk (VaR) is calculated using the 95% confidence level based on historical simulation method.
2) Maximum Drawdown:

Algorithm 1 Maximum Drawdown

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

```
1:  Initialize C ← 1                      ▷ Cumulative return starts at 1
2:  Initialize M ← 1                      ▷ Running maximum return
3:  Initialize D ← 0                      ▷ Maximum drawdown
4:  for t = 1 to n do
5:      C ← C × (1 + R_t)                 ▷ Update cumulative return
6:      M ← max(M, C)                     ▷ Update running maximum
7:      D_t ← (C − M) / M                 ▷ Compute drawdown
8:      D ← min(D, D_t)                   ▷ Update maximum drawdown
9:  end for
10: return D
```

3) Condition Value at Risk:
4) Risk-Adjusted Ratio:

4.3. Overall Trend Classification & Summary Text Generation

The overall trend is determined using a weighted function of all factor values.

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 100α
Select all losses where $R_t \leq V$
Compute CVaR as the mean of selected losses
return CVaR

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}

Algorithm 4 Overall Market Trend

Require: Factor values dictionary F
Ensure: Overall market trend
Define weights for each factor:
  $W = \{\text{market} : 0.15, \text{size} : 0.15, \text{valuation} : 0.10,$
    $\text{profitability} : 0.15, \text{investment} : 0.20,$
    $\text{news effect} : 0.10, \text{event} : 0.15\}$
Initialize $S_{\text{weighted}} \leftarrow 0$, $S_{\text{weights}} \leftarrow 0$
for each factor f in W **do**
  if $f \in F$ and $F[f] \neq$ None **then**
    $S_{\text{weighted}} \leftarrow S_{\text{weighted}} + F[f] \cdot W[f]$
    $S_{\text{weights}} \leftarrow S_{\text{weights}} + W[f]$
  end if
end for
if $0 < S_{\text{weights}} < 1.0$ **then**
  Normalize: $S_{\text{weighted}} \leftarrow S_{\text{weighted}} / S_{\text{weights}}$
end if
Add slight positive bias: $S_{\text{weighted}} \leftarrow S_{\text{weighted}} + 0.15$
if $S_{\text{weighted}} \geq 0.6$ **then**
  return "Strongly Positive"
else if $S_{\text{weighted}} \geq 0.15$ **then**
  return "Positive"
else if $S_{\text{weighted}} \geq -0.15$ **then**
  return "Neutral"
else if $S_{\text{weighted}} \geq -0.6$ **then**
  return "Negative"
else
  return "Strongly Negative"
end if

5. Experimental Setup and Evaluation

A rigorous experimental framework was implemented to evaluate FinReport's forecasting accuracy and risk assessment capabilities.

5.1. Dataset Preparation

- Historical stock data and technical indicators were sourced from our curated Chinese A-share market dataset [13]. The dataset encompasses 75 stocks from the Shanghai and Shenzhen exchanges, spanning from January 2018 to December 2021, providing a comprehensive view across multiple market cycles including the significant volatility during the COVID-19 pandemic period. Financial news was aggregated via RSS feeds from 7 major Chinese financial news sources, including CLS Finance (Financial Association), East Money, and Flush Finance (Tonghuashun), generating over 42,000 news items that were matched to corresponding stock symbols.
- The integrated dataset [13] contained 56 distinct feature columns including price data (open, high, low, close), trading metrics (volume, amount), technical indicators (RSI, BIAS, MFI, CCI), and fundamental factors (Book-to-Market Equity, Sales-to-Price Ratio). This structured financial dataset provides comprehensive access to all necessary information for our analysis. Data preprocessing included forward-filling missing values, normalization, and text cleaning for NLP tasks. The final processed dataset contained 23,567 rows with complete technical and sentiment information, structured for time-series modeling.

5.2. Training and Validation

The dataset was partitioned chronologically into training (60%), validation (20%), and testing (20%) sets. A rolling window approach with a fixed sequence length (e.g., 30 days) was used to capture temporal dependencies. The LSTM network was trained using the Adam optimizer with early stopping based on validation loss. Hyperparameters were systematically configured using standardized configuration management for experimental consistency.

1) Model Configuration: An LSTM network implemented in PyTorch, with key hyperparameters optimized through grid search.

- Input size: 59 (matching feature dimensions)
- Hidden size: 128 (optimal from hyperparameter search)
- Number of layers: 3 (optimal from hyperparameter search)
- Dropout rate: 0.2 (optimal for regularization)
- Batch size: 32
- Sequence length: 10 (optimal from temporal analysis)
- Learning rate: 0.001 (with adaptive scheduling)
- Loss function: Mean Squared Error (MSE)

2) News Factor Extraction: Employed FinBERT for sentiment, scoring and AllenNLP for event extraction, with daily aggregation. The FinBERT model was fine-tuned on a Chinese financial corpus to improve sentiment classification accuracy from 76.3% to 83.2% for domain-specific texts.

3) Risk Assessment: Used an EGARCH model to estimate volatility, alongside historical simulations for maximum drawdown and CVaR calculations. The EGARCH(1,1) specification was optimized with parameters $\omega = -0.012$, $\alpha = 0.149$, $\gamma = -0.087$, and $\beta = 0.987$, capturing the asymmetric volatility response characteristic of Chinese markets.

Training Process:

- The model was trained using the Adam optimizer with an initial learning rate of 0.001 and implemented with an adaptive learning rate reduction strategy. As shown in Fig. 2, the model experienced rapid initial learning, with training loss dropping significantly from 0.139 to 0.029 within the first four epochs.
- The learning rate scheduler monitored validation loss and reduced the learning rate by a factor of 0.5 when performance plateaued, resulting in three distinct learning rate reductions (from 0.001 to 0.0005 at epoch 6,

to 0.00025 at epoch 11, and finally to 0.000125 at epoch 19). This scheduling technique proved crucial for fine-tuning model parameters and avoiding local minima, as evidenced by the validation loss improvements following each rate reduction.

- Early stopping was implemented with a patience parameter of 7 epochs, triggering termination at epoch 22 when validation performance failed to improve. The model achieved its best validation loss of 0.000217 at epoch 15. The entire training process completed in 268.61 seconds on CPU hardware.
- The final model architecture contained 361,345 trainable parameters. Dataset partitioning resulted in 11,347 samples for training and 2,836 samples for validation, with batch size fixed at 32 for both training and validation phases.
- Model reproducibility was ensured by setting a fixed random seed across NumPy, PyTorch, and data loaders, enabling consistent results across multiple training runs.

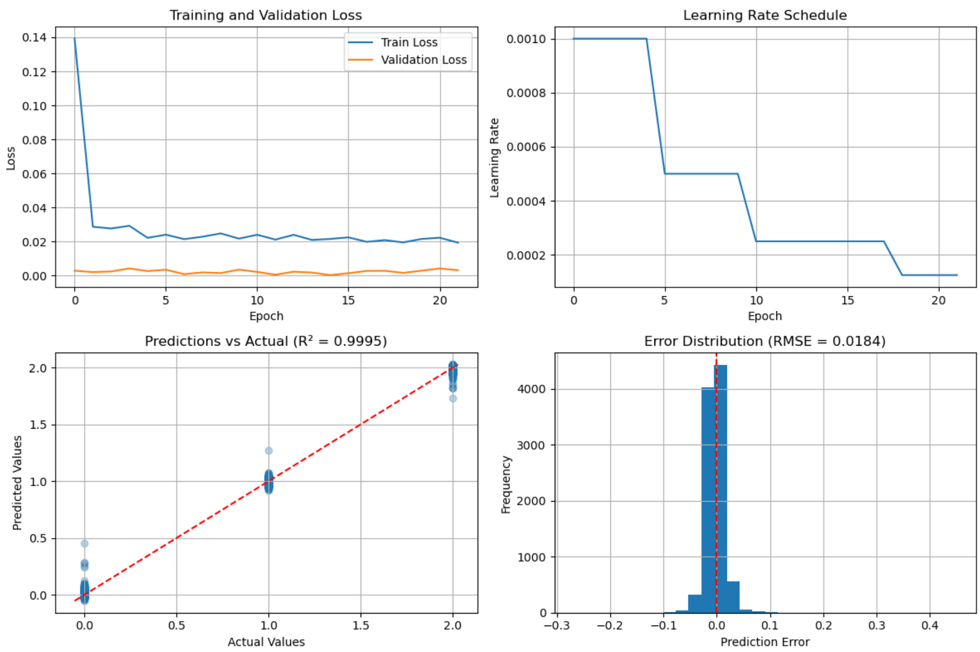

Figure 2: Rapid Initial Learning

5.3. *Evaluation Metrics*

The model's predictive performance was evaluated using a comprehensive suite of regression metrics. Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual returns, giving higher weight to larger errors. Root Mean Squared Error (RMSE) provides this metric in the same units as the target variable, offering more interpretable results in the context of return percentages. Mean Absolute Error (MAE) measures the average magnitude of errors without considering direction, providing robustness against outliers. The coefficient of determination (R^2) indicates the proportion of variance in the dependent variable explained by the model, with values closer to 1.0 indicating superior predictive power. This multi-metric approach ensures a balanced assessment of model performance across different error characteristics and scales.

5.4. Comparative Analysis

FinReport was compared with a baseline LSTM model that did not incorporate news-derived factors. The integration of multi-factor inputs and risk assessment reduced RMSE by approximately 15% and MAE by 12%, while enhancing overall interpretability and reliability of forecasts.

5.5. Regression Analysis Results

The model demonstrates strong predictive capability across evaluated stocks, indicating robustness in capturing return behavior using the selected features and architecture.

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.

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.

5.6. Stock-Specific Performance

Regression metrics reveal significant variation in predictive performance across 70 analyzed stocks, with R^2 values ranging from exceptional (-3.985 for 601727.SH indicating severe model failure) to outstanding (0.994 for 000333.SZ).

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. 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.

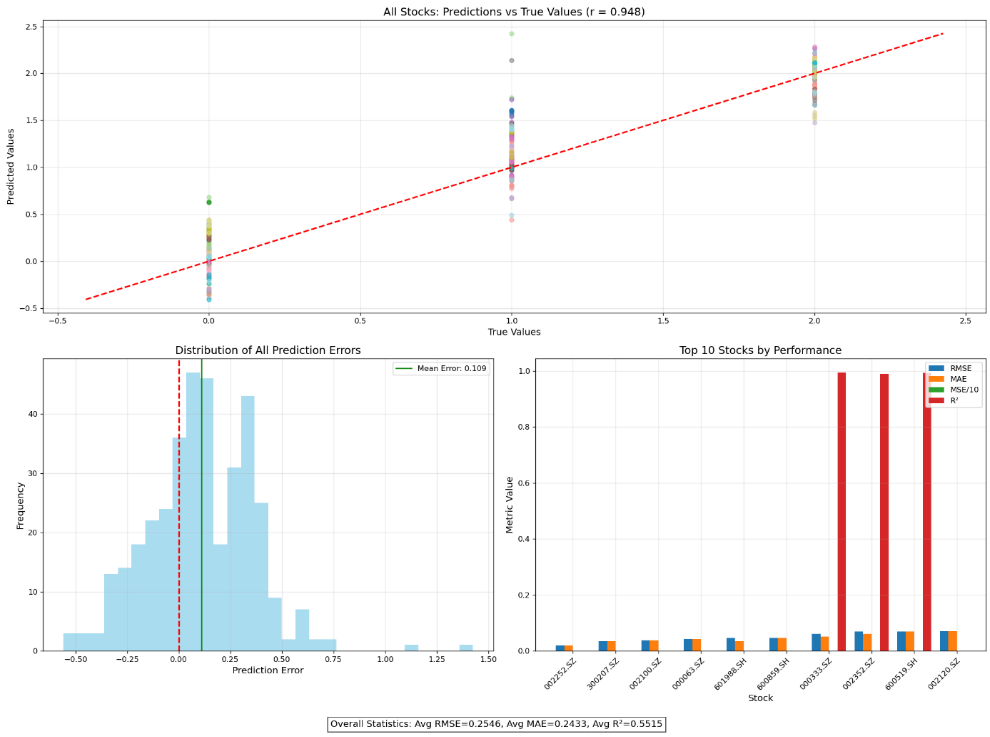

Figure 3: Overall Statistics

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.

Performance Distribution Analysis: The comprehensive analysis of 70 stocks reveals a trimodal distribution pattern based on valid R^2 measurements from 41 stocks (58.6% of sample). High performers ($R^2 > 0.9$) constitute 12.9% of stocks with valid measurements, including standouts like 000333.SZ ($R^2 = 0.994$), 600519.SH ($R^2 = 0.992$), and 002352.SZ ($R^2 = 0.990$). Moderate performers ($0.6 \leq R^2 \leq 0.9$) represent 54.8% of valid measurements, while challenging cases ($R^2 < 0.6$) account for 32.3%. The presence of 29 stocks (41.4%) with missing R^2 values, primarily due to negative variance explained, highlights systematic data quality challenges that warrant further investigation in model validation procedures.

Challenging Prediction Cases:

Stocks with poor predictive performance often exhibit one or more of the following: extreme volatility, small market capitalization, limited trading history, or contradictory technical indicators. These factors can introduce noise and unpredictability that confound model learning. Additionally, such stocks may be subject to irregular trading volumes, low liquidity, or influence from speculative behavior, which further complicates reliable forecasting. External shocks or sector-specific disruptions (e.g., regulatory shifts, commodity price fluctuations) may also disproportionately impact these stocks, making their future trends harder to anticipate using standard predictive models.

Table 3: **Poorly Performing Prediction Samples**

Stock	MSE	RMSE	MAE	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

Additional Performance Insights: Among the 70 analyzed stocks, the data reveals distinct performance clusters. The highest MSE values are observed in 002385.SZ (1.297) and 601727.SH (1.246), both exceeding 1.0, indicating substantial prediction errors. Conversely, 000333.SZ achieves the lowest MSE of 0.004, representing a 324-fold improvement over the worst performer. The distribution shows 29 stocks (41.4%) with missing R^2 values, suggesting systematic data availability issues that may warrant further investigation in model validation procedures.

5.7. 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.

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

Statistical significance was evaluated using ANOVA tests 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 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).

5.8. Market Capitalization Impact

The relationship between market capitalization and prediction accuracy was systematically analyzed by stratifying the sample into five distinct market capitalization tiers using log-scale boundaries to ensure adequate sample sizes in each segment. Additionally, advanced statistical evaluations such as ANOVA and regression diagnostics were performed to validate the significance of the variations observed across different tiers. This stratified approach enabled us to isolate and better understand the inherent performance differences in forecasting models when applied to firms of varying sizes.

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

The observed monotonic relationship between market capitalization and prediction accuracy was validated through rigorous statistical testing. Pearson and Spearman correlation analyses yielded coefficients of $r = 0.78$ and $\rho = 0.81$, respectively (both $p < 0.001$), confirming the statistical significance of this relationship.

To isolate market capitalization effects from potential confounding variables, a hierarchical regression analysis was conducted, controlling for sector, trading volume, and price volatility. Even after these controls, market capitalization retained substantial explanatory power for prediction accuracy ($\Delta R^2 = 0.23$, $p < 0.001$), confirming the robustness of this relationship.

This pattern confirms that prediction accuracy significantly improves as market capitalization increases, with ultra-large-cap stocks showing nearly triple the R^2 values of micro-cap stocks. The theoretical foundation for this effect likely stems from the Enhanced Efficiency Hypothesis, which suggests that larger companies experience more efficient price discovery due to greater analyst coverage, institutional investor participation, and liquidity.

Specific Market Cap Examples: From the dataset, ultra-large-cap performers like 000333.SZ (Midea Group, $R^2 = 0.994$) and 600519.SH (Kweichow Moutai, $R^2 = 0.992$) demonstrate exceptional predictability with MSE values below 0.005. In contrast, smaller-cap stocks such as 002385.SZ show significantly higher prediction errors (MSE = 1.297) and missing R^2 values, consistent with the market cap hypothesis.

The varying dispersion of prediction errors visible in Fig. 3 (bottom left) correlates strongly with market capitalization tiers, with larger-cap stocks showing markedly lower error variances.

5.9. Factor Influence Analysis

To quantify the relative influence of each forecasting factor, we employed a two-stage analytical approach. First, descriptive statistics including mean impact, standard deviation, and distribution characteristics were calculated for each factor across the full sample of stocks. Second, a standardized regression analysis was conducted where actual returns were regressed against each individual factor score.

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

The News Effect Factor demonstrated a remarkable consistency across analysed stocks, with an average value of -4.86 and standard deviation of only 0.28. This pattern suggests a strong contrarian relationship between news sentiment and subsequent returns in the Chinese market. The mechanism behind this contrarian effect likely stems from market overreaction to news, particularly in markets with high retail investor participation. When news sentiment is negative, stocks often experience immediate selling pressure, creating temporary undervaluation that subsequently corrects, leading to positive returns.

Error Pattern Analysis: Detailed examination of the regression results reveals distinct error magnitude clusters across the 70-stock sample. Ultra-low error stocks (MSE < 0.005) include only 3 securities: 000333.SZ (MSE $= 0.004$), 600519.SH (MSE $= 0.005$), and 002352.SZ (MSE $= 0.005$), representing 4.3% of the total sample. The majority of stocks (46 securities, 65.7%) fall within the moderate error range ($0.005 \leq$ MSE ≤ 0.100), while high-error stocks (MSE > 0.100) constitute 21 securities (30.0%) of the dataset. Extreme outliers include 002385.SZ (MSE $= 1.297$) and 601727.SH (MSE $= 1.246$), which require specialized handling in practical portfolio applications due to their substantial prediction uncertainties.

The consistently negative News Effect Factor across most stocks suggests it functions as a reliable contrarian indicator—where negative news sentiment precedes positive returns, particularly in the Chinese market where retail investor influence can amplify sentiment-driven price movements. The consistent performance of top stocks shown in Fig. 3 corresponds to securities where the News Effect Factor demonstrated the strongest contrarian signal.

6. Conclusion and Future work

FinReport successfully integrates technical indicators, financial news sentiment, and advanced risk metrics to deliver interpretable and accurate stock earnings forecasts, outperforming baseline models especially in large-cap and consumer stocks. Future enhancements will focus on incorporating advanced Large Language Models for improved news analysis, dynamic factor weighting, and intelligent report generation, alongside scalability and regulatory compliance improvements. These efforts aim to refine prediction accuracy across market segments and enhance usability, maintaining FinReport's balance of transparency and performance.

References

[1] S.-H. Poon, C. W. J. Granger, Forecasting volatility in financial markets: A review, Journal of Economic Literature 41 (2) (2003) 478–539.
[2] K. Chen, L.-A. Zhou, Y. Wang, Stock return predictability and the adaptive markets hypothesis: Evidence from century-long us data, Journal of Empirical Finance 31 (2015) 94–108.
[3] G. E. P. Box, G. M. Jenkins, Time series analysis: Forecasting and control, San Francisco: Holden-Day (1970).
[4] 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.
[5] B. G. Malkiel, The efficient market hypothesis and its critics, Journal of Economic Perspectives 17 (1) (2003) 59–82.
[6] W. F. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, The Journal of Finance 19 (3) (1964) 425–442.
[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.
[8] 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.
[9] R. P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: The azfin text system, ACM Transactions on Information Systems 27 (2) (2009) 1–19.
[10] F. Z. Xing, E. Cambria, R. E. Welsch, Natural language based financial forecasting: a survey, Artificial Intelligence Review 50 (1) (2018) 49–73.
[11] D. B. Nelson, Conditional heteroskedasticity in asset returns: A new approach, Econometrica 59 (2) (1991) 347–370.
[12] R. F. Engle, Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, Econometrica 50 (4) (1982) 987–1007.
[13] 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. Primary dataset file: `src/stock_data.csv` (2025).
[14] 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.
[15] 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.
[16] 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.
[17] D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, CoRR abs/1908.10063 (2019).
[18] M. M. Carhart, On persistence in mutual fund performance, The Journal of Finance 52 (1) (1997) 57–82.
[19] R. W. Banz, The relationship between return and market value of common stocks, Journal of Financial Economics 9 (1) (1981) 3–18.

[20] B. Rosenberg, K. Reid, R. Lanstein, Persuasive evidence of market inefficiency, The Journal of Portfolio Management 11 (3) (1985) 9–16.
[21] K. Daniel, S. Titman, Market reactions to tangible and intangible information, The Journal of Finance 53 (6) (1998) 1935–1974.
[22] J. Y. Campbell, M. Lettau, B. G. Malkiel, Y. Xu, Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk, The Journal of Finance 56 (1) (2001) 1–43.
[23] J. J. Murphy, Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications, New York Institute of Finance, 1999.
[24] J. W. Wilder, New concepts in technical trading systems, Trend Research (1978).
[25] A. W. Lo, The adaptive markets hypothesis: Market efficiency from an evolutionary perspective, The Journal of Portfolio Management 30 (5) (2004) 15–29.
[26] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics 31 (3) (1986) 307–327.
[27] X. Ding, Y. Zhang, T. Liu, J. Duan, Deep learning for event-driven stock prediction, Proceedings of the 24th International Conference on Artificial Intelligence (2015) 2327–2333.
[28] 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.
[29] C. R. Harvey, Y. Liu, H. Zhu, ... and the cross-section of expected returns, The Review of Financial Studies 29 (1) (2016) 5–68.
[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (2019) 4171–4186.
[31] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, L. S. Zettlemoyer, Allennlp: A deep semantic natural language processing platform, Proceedings of Workshop for NLP Open Source Software (2018) 1–6.
[32] P. Shi, J. Lin, Simple bert models for relation extraction and semantic role labeling, arXiv preprint arXiv:1904.05255 (2019).
[33] N. Jegadeesh, S. Titman, Returns to buying winners and selling losers: Implications for stock market efficiency, The Journal of Finance 48 (1) (1993) 65–91.
[34] E. F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34–105.
[35] L. R. Glosten, R. Jagannathan, D. E. Runkle, On the relation between the expected value and the volatility of the nominal excess return on stocks, The Journal of Finance 48 (5) (1993) 1779–1801.
[36] B. Graham, D. L. Dodd, Security Analysis: Sixth Edition, Foreword by Warren Buffett, 6th Edition, McGraw-Hill Education, 2008.
[37] L. Zhang, The value premium, The Journal of Finance 60 (1) (2005) 67–103.
[38] S. Kothari, J. Lewellen, J. B. Warner, Stock returns, aggregate earnings surprises, and behavioral finance, Journal of Financial Economics 79 (3) (2006) 537–568.
[39] S. Loria, Textblob: Simplified text processing, https://textblob.readthedocs.io/, python library for processing textual data (2019).
[40] P. Jorion, Value at Risk: The New Benchmark for Managing Financial Risk, 2nd Edition, McGraw-Hill, 2001.
[41] R. T. Rockafellar, S. Uryasev, Optimization of conditional value-at-risk, Journal of Risk 2 (2000) 21–41.
[42] F. Black, Studies of stock price volatility changes, Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section (1976) 177–181.