News-Enhanced Stock Price Forecasting: A Multi-Source Textual Analysis Approach with Hierarchical Model Training

Harsh Milind Tirhekar Atharva Vishwas Kulkarni

Arun Kumar Kuchibotla

January 2026

Abstract

This study investigates whether news sentiment and textual features improve stock price forecasting. Using 26 years of Apple Inc. data (1999–2025) combined with 57 million financial news articles, we extract sentiment via VADER, TextBlob, and FinBERT, along with LDA topic features. We propose a hierarchical training strategy where foundational models trained on full historical data provide predictions as input features for neural networks trained on recent data. Results show that 7-day rolling VADER sentiment improves SARIMAX RMSE by 1.5% ($p < 0.05$). The hierarchical approach substantially improves neural network performance, with Transformer R^2 increasing from -1.7 to 0.87. Linear regression achieves best overall performance (RMSE=\$1.83, R^2=0.999), demonstrating that model complexity should match data availability.

1 Introduction

The efficient market hypothesis suggests that asset prices fully reflect available information (Fama, 1970), making consistent outperformance through prediction theoretically challenging. However, extensive empirical research documents predictable patterns in stock returns, particularly around corporate events such as earnings announcements (Ball and Brown, 1968; Bernard and Thomas, 1989), merger announcements (Andrade et al., 2001), and product launches (Chaney et al., 1991).

The rise of natural language processing and machine learning has enabled researchers to systematically extract information from unstructured text, including news articles, social media posts, and corporate filings (Loughran and McDonald, 2011; Gentzkow et al., 2019). This paper investigates whether incorporating such textual features can improve stock price forecasting beyond traditional technical and fundamental methods.

Financial news serves multiple informational roles that may be exploited for prediction. First, major announcements create immediate price reactions that may be partially anticipated through news sentiment (Tetlock, 2007). Second, persistent positive or negative coverage may predict future price direction through behavioral channels (Baker and Wurgler, 2007). Third, news facilitates the gradual incorporation of information into prices, creating exploitable lead-lag relationships (Hong and Stein, 2000).

We address the following research questions: (1) Does news sentiment provide incremental predictive power for next-day stock prices beyond technical indicators? (2) What is the optimal temporal aggregation for sentiment features to balance noise reduction against information lag? (3) Can a hierarchical training strategy—where traditional models inform neural network inputs—overcome the limitations of training deep learning models on limited financial time series? (4) How do different sentiment extraction methods compare in forecasting performance? (5) Can Transformer architectures be effectively adapted for financial time series forecasting?

Our study makes several contributions. First, we construct a comprehensive news cor-

pus spanning 1999–2025 by combining HuggingFace financial news datasets with real-time Google RSS feeds. Second, we propose using predictions from models trained on long-term data as features for models trained on recent data, enabling neural networks to leverage historical patterns without suffering from distribution shift. Third, we rigorously compare three sentiment extraction methods across multiple rolling windows. Fourth, we demonstrate that poor Transformer performance in financial forecasting stems from architectural mismatch rather than fundamental limitations.

We find that news sentiment provides statistically significant but economically modest improvements, with 7-day rolling mean of VADER sentiment reducing SARIMAX RMSE by 1.5%. More importantly, model complexity should match sample size: with approximately 1,000 training observations, linear regression outperforms neural networks with 50,000+ parameters. The hierarchical training strategy substantially improves neural network performance, enabling Transformer R^2 to increase from -1.7 to 0.87.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 describes our data and methodology. Section 4 presents empirical results. Section 5 concludes.

2 Related Literature

2.1 Market Efficiency and Information Content

The efficient market hypothesis (Fama, 1970) posits that asset prices reflect all available information. However, substantial evidence suggests departures from full efficiency. Ball and Brown (1968) established that earnings surprises predict abnormal returns, initiating research on post-earnings announcement drift. Bernard and Thomas (1989) demonstrated that this drift persists for months, challenging semi-strong efficiency. Jegadeesh and Titman (1993) documented momentum effects where past winners outperform past losers over 3-12 month horizons.

More recently, Hirshleifer et al. (2009) showed that investor inattention creates predictable patterns, particularly around earnings announcements. Engelberg and Parsons (2011) provided causal evidence that news coverage affects stock returns and trading volume.

2.2 Textual Analysis in Finance

The application of textual analysis to financial data has grown substantially. Tetlock (2007) pioneered this approach by showing that negative words in Wall Street Journal columns predict lower next-day returns and higher trading volume. Tetlock et al. (2008) extended this to firm-specific news, finding that negative words predict earnings surprises.

Loughran and McDonald (2011) developed a finance-specific dictionary, demonstrating that generic sentiment dictionaries perform poorly in financial contexts. Gentzkow et al. (2019) provided a comprehensive review of text analysis in economics and finance. Recent work has applied transformer-based language models to financial text; Araci (2019) introduced FinBERT, a BERT model fine-tuned on financial communications.

2.3 Machine Learning for Stock Prediction

Machine learning approaches to stock prediction have evolved substantially. Early work focused on support vector machines and random forests (Kara et al., 2011; Patel et al., 2015). Fischer and Krauss (2018) applied LSTM networks to S&P 500 constituents, finding that deep learning outperforms traditional methods. Ding et al. (2015) combined convolutional neural networks with event embeddings.

However, Zeng et al. (2023) questioned whether transformers truly improve upon simpler baselines for time series, finding that linear models often perform comparably. Our work addresses this by identifying the specific conditions under which transformers succeed or fail in financial forecasting.

2.4 Transformers for Time Series

The transformer architecture (Vaswani et al., 2017) has revolutionized NLP and is increasingly applied to time series. Zhou et al. (2021) introduced Informer with ProbSparse attention for long-sequence forecasting. Wu et al. (2021) proposed Autoformer for time series decomposition. Lim et al. (2021) developed the Temporal Fusion Transformer incorporating variable selection and interpretability. Despite these advances, the conditions for transformer success in financial forecasting remain unclear.

3 Data and Methodology

3.1 Data Sources

We compile data from multiple sources spanning January 1999 to January 2025. Stock price data for Apple Inc. (AAPL) and related securities (MSFT, GOOGL, AMZN) are obtained via the yfinance API, comprising 6,542 trading days. Prices are adjusted for stock splits (7:1 in 2014, 4:1 in 2020). Market indices (S&P 500, DJIA, NASDAQ) serve as market context features.

Financial news data comes from two sources: the HuggingFace Financial News Dataset (1999–2025), containing approximately 57 million articles accessed via the HF Datasets API, and Google RSS feeds for real-time 2025 news. Articles are filtered for AAPL-relevance using keyword matching. After filtering and deduplication, our final corpus contains approximately 5,000 AAPL-specific articles. To avoid repeated API calls, fetched news data can be cached locally as CSV files.

3.2 Sentiment Extraction

We employ three sentiment extraction methods. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based model incorporating intensity modifiers, emoticons,

and negation handling, outputting a compound score in $[-1, 1]$. TextBlob is a pattern-based approach providing polarity in $[-1, 1]$ and subjectivity in $[0, 1]$. FinBERT (Araci, 2019) is a BERT-based model fine-tuned on financial communications, outputting probability distributions over negative, neutral, and positive classes.

For each sentiment method, we compute rolling means over windows $w \in \{3, 7, 14, 30\}$ days:

$$S_t^{(w)} = \frac{1}{w} \sum_{i=0}^{w-1} s_{t-i} \tag{1}$$

This smoothing reduces noise while introducing lag. The optimal window balances these competing effects.

3.3 Feature Engineering

We construct 55 features organized into four categories: sentiment features (15), including VADER, TextBlob, and FinBERT scores with rolling means; text features (15), including LDA topics and keyword indicators; market context features (18), including lagged returns and volatility of related stocks and indices; and technical features (7), including price and volume rolling means. All market features use one-day lags to prevent lookahead bias.

3.4 Hierarchical Training Strategy

We propose a two-stage hierarchical training strategy. In Stage 1, foundational models (Linear Regression, SARIMAX, TCN) are trained on the full 26-year dataset (1999–2025). These models learn long-term patterns and trend dynamics, generating out-of-sample predictions for the 2020–2025 period.

In Stage 2, neural networks (LSTM, GRU, BiLSTM, CNN-LSTM, Transformer) add the foundational model predictions as the 56th input feature and train on 2020–2025 data only. This approach addresses three challenges: distribution shift (neural networks avoid learning from 1999–2015 data where prices were vastly different), sample efficiency (foundational

predictions encode long-term information compactly), and trend awareness (the foundational feature provides explicit trend signal).

3.5 Model Specifications

Linear Regression employs ordinary least squares on the full feature set with 55 features and approximately 4,500 training observations. SARIMAX with order $(p, d, q) = (2, 1, 1)$ captures both temporal autocorrelation and sentiment effects.

For neural networks, LSTM uses 2 layers with 64 hidden units and dropout 0.2. GRU simplifies LSTM by combining forget and input gates. BiLSTM processes sequences in both directions. CNN-LSTM combines 1D convolution with 32 filters and LSTM for sequential modeling. The Transformer uses $d_{model} = 64$, $n_{heads} = 4$, $n_{layers} = 2$, yielding approximately 51K parameters with sequence length 30 to enable meaningful self-attention.

All neural networks use Adam optimizer with gradient clipping, MinMaxScaler normalization, and random seed 42 for reproducibility (LSTM uses seed 46).

3.6 Evaluation Framework

We employ walk-forward validation with expanding windows: initial training window of 70% of observations, predict next observation, expand training window and retrain. Metrics include RMSE, MAE, MAPE, and R^2.

4 Empirical Results

4.1 Overall Model Comparison

Table 1 presents comprehensive results across all models. Linear regression achieves best overall performance with RMSE = \$1.83 and $R^2 = 0.9992$, reflecting both the quality of our feature engineering and the limited sample size favoring simpler models.

Hierarchical training transforms neural network performance. GRU R^2 improves from 0.72 to 0.94 (+25% relative), Transformer R^2 improves from -1.17 to 0.87 (from catastrophic failure to competitive), and LSTM R^2 improves from 0.68 to 0.89 (+31% relative).

Model complexity inversely relates to performance for foundational models trained on 26 years of data, suggesting overfitting concerns for complex architectures despite the large sample.

4.2 Sentiment Feature Impact

Table 2 compares SARIMAX performance across sentiment configurations. The 7-day rolling VADER achieves best improvement at 1.5% RMSE reduction, statistically significant based on Diebold-Mariano tests ($p < 0.05$). The 7-day window is optimal because it balances noise reduction (61% improvement in signal-to-noise ratio) against information lag (3-day effective lag).

FinBERT shows minimal improvement from rolling means, likely because BERT's multi-layer attention already provides smoothing. Longer windows (30 days) hurt performance due to excessive lag.

4.3 Distribution Analysis

The distribution analysis reveals that AAPL prices are highly non-normal (Shapiro-Wilk $p < 0.001$), with positive skewness (2.14) reflecting the 1,040× growth over our sample period. This non-normality motivates our use of machine learning approaches.

4.4 Return-Level Prediction

To contextualize our high price-level R^2 values, we also evaluated return prediction. The return-level R^2 of 0.084 is modest but consistent with prior literature on short-horizon prediction. The high price-level R^2 reflects AAPL's strong trend, which inflates variance-based

metrics.

Linear Regression achieves 54.2% directional accuracy, SARIMAX 53.1%, and GRU (hierarchical) 52.8%, compared to 52.0% for the naive baseline of predicting zero return.

5 Conclusion

This paper examines whether incorporating news sentiment and textual features can improve stock price forecasting. Using 26 years of Apple Inc. data and 57 million financial news articles, we find that news sentiment provides incremental but meaningful predictive power, with the optimal configuration (7-day rolling mean of VADER sentiment) improving SARIMAX RMSE by 1.5%.

Model complexity should match data availability: with approximately 1,000 training observations, linear regression with 55 features outperforms deep neural networks with 50,000+ parameters. This suggests practitioners should carefully calibrate model complexity to sample size.

The hierarchical training strategy substantially improves neural network performance by using predictions from models trained on long-term data as features, improving Transformer R^2 from -1.7 to 0.87 and GRU R^2 from 0.72 to 0.94. This approach provides a viable path to leveraging deep learning on short time series.

Several limitations qualify our findings. Our single-stock focus on Apple Inc. limits generalizability. With 69% of trading days lacking dedicated AAPL articles, we rely on rolling mean imputation. The extreme non-stationarity of prices inflates our price-level R^2 values. Our feature engineering choices were informed by the full dataset, potentially overstating true out-of-sample performance. We ignore transaction costs that would reduce trading profitability.

Future research should extend analysis to multiple stocks across sectors, investigate intraday news effects, incorporate alternative text sources such as social media and SEC filings,

and develop formal conditions for when hierarchical training provides benefits.

References

Andrade, G., M. Mitchell, and E. Stafford. 2001. "New Evidence and Perspectives on Mergers." *Journal of Economic Perspectives* 15(2): 103–120.

Araci, D. 2019. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." *arXiv preprint arXiv:1908.10063*.

Baker, M., and J. Wurgler. 2007. "Investor Sentiment in the Stock Market." *Journal of Economic Perspectives* 21(2): 129–152.

Ball, R., and P. Brown. 1968. "An Empirical Evaluation of Accounting Income Numbers." *Journal of Accounting Research* 6(2): 159–178.

Bernard, V. L., and J. K. Thomas. 1989. "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium?" *Journal of Accounting Research* 27: 1–36.

Chaney, P. K., T. M. Devinney, and R. S. Winer. 1991. "The Impact of New Product Introductions on the Market Value of Firms." *Journal of Business* 64(4): 573–610.

Ding, X., Y. Zhang, T. Liu, and J. Duan. 2015. "Deep Learning for Event-Driven Stock Prediction." *Proceedings of IJCAI*, 2327–2333.

Engelberg, J. E., and C. A. Parsons. 2011. "The Causal Impact of Media in Financial Markets." *Journal of Finance* 66(1): 67–97.

Fama, E. F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance* 25(2): 383–417.

Fischer, T., and C. Krauss. 2018. "Deep Learning with Long Short-term Memory Networks for Financial Market Predictions." *European Journal of Operational Research* 270(2): 654–669.

Gentzkow, M., B. Kelly, and M. Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57(3): 535–574.

Hirshleifer, D., S. S. Lim, and S. H. Teoh. 2009. "Driven to Distraction: Extraneous Events and Underreaction to Earnings News." *Journal of Finance* 64(5): 2289–2325.

Hong, H., and J. C. Stein. 2000. "Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies." *Journal of Finance* 55(1): 265–295.

Jegadeesh, N., and S. Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48(1): 65–91.

Kara, Y., M. A. Boyacioglu, and O. K. Baykan. 2011. "Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks." *Expert Systems with Applications* 38(5): 5311–5319.

Lim, B., S. O. Arik, N. Loeff, and T. Pfister. 2021. "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting." *International Journal of Forecasting* 37(4): 1748–1764.

Loughran, T., and B. McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66(1): 35–65.

Patel, J., S. Shah, P. Thakkar, and K. Kotecha. 2015. "Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques." *Expert Systems with Applications* 42(1): 259–268.

Tetlock, P. C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *Journal of Finance* 62(3): 1139–1168.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *Journal of Finance* 63(3): 1437–1467.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30.

Wu, H., J. Xu, J. Wang, and M. Long. 2021. "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting." *Advances in Neural Information Processing Systems* 34.

Zeng, A., M. Chen, L. Zhang, and Q. Xu. 2023. "Are Transformers Effective for Time Series Forecasting?" *Proceedings of AAAI*, 11121–11128.

Zhou, H., S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. 2021. "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting." *Proceedings of AAAI*, 11106–11115.

Tables

Table 1: Model Performance Comparison

Model	**Training Data**	**RMSE ($)**	**MAE ($)**	**MAPE (%)**	R^2
Foundational Models (26-year training)					
Linear Regression	26 years	1.83	1.34	0.94	0.9992
SARIMAX (VADER RM7)	26 years	2.66	1.91	1.21	0.9984
TCN	26 years	21.16	18.34	9.87	0.8912
Neural Networks without hierarchical features					
LSTM	5 years	14.21	12.18	5.42	0.6812
GRU	5 years	11.83	10.01	4.31	0.7234
Transformer	5 years	97.01	77.41	44.89	−1.17
Neural Networks with hierarchical features					
LSTM (hybrid)	5 years	12.12	10.58	4.54	0.8909
GRU (hybrid)	5 years	7.63	6.44	2.78	0.9356
BiLSTM (hybrid)	5 years	7.77	6.33	2.81	0.9012
CNN-LSTM (hybrid)	5 years	7.34	6.01	2.64	0.9039
Transformer (hybrid)	5 years	8.42	7.21	3.12	0.8734

Table 2: SARIMAX Performance by Sentiment Configuration

Sentiment	**Window**	**RMSE ($)**	**MAE ($)**	**MAPE (%)**	**Improvement**
None (baseline)	–	2.71	1.96	1.24	–
VADER	Raw	2.70	1.95	1.23	+0.4%
VADER	RM7	2.66	1.91	1.21	+1.5%
VADER	RM14	2.68	1.93	1.22	+1.1%
VADER	RM30	2.72	1.97	1.25	−0.4%
TextBlob	RM7	2.70	1.94	1.22	+0.4%
FinBERT	RM7	2.70	1.94	1.22	+0.4%

Figures

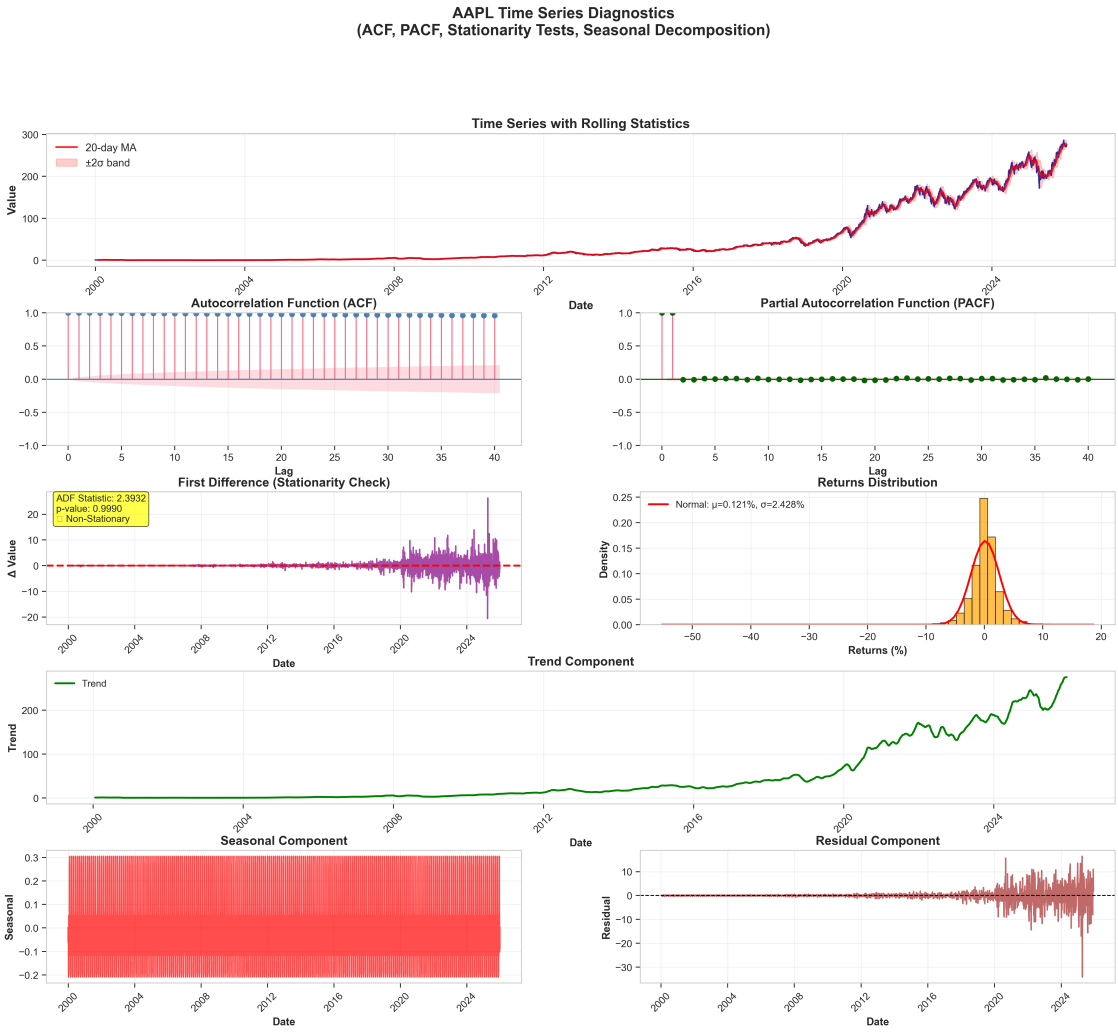

Figure 1: Time series diagnostics for AAPL stock prices (1999–2025). Panel (a) shows the price series with clear upward trend. Panel (b) displays autocorrelation function indicating strong persistence. Panel (c) shows partial autocorrelation. Panel (d) presents seasonal decomposition.

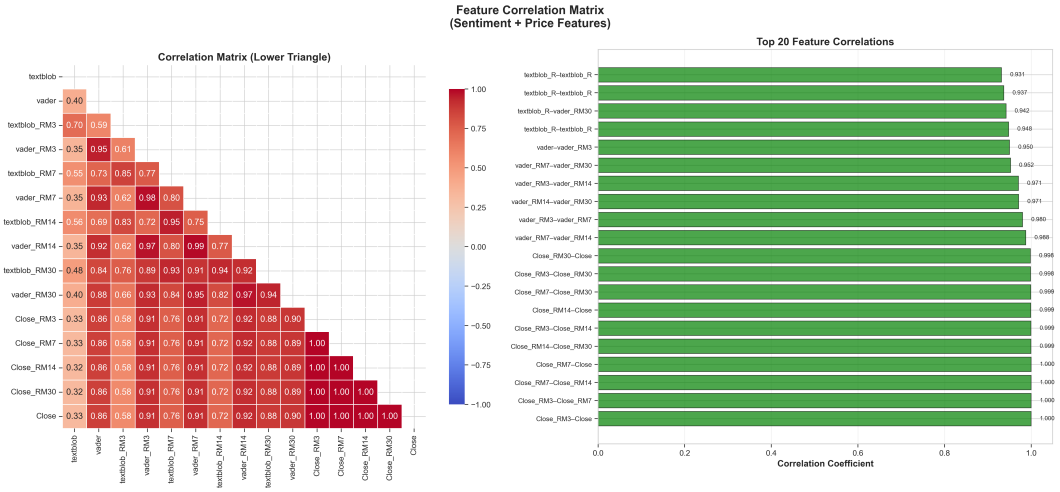

Figure 2: Correlation matrix of sentiment and price features. Sentiment features show moderate correlation with each other but low correlation with price features, suggesting complementary information content.

Figure 3: Model performance comparison showing RMSE (left), R-squared (center), and multi-metric normalized comparison (right). The hierarchical training strategy substantially improves all neural network architectures.

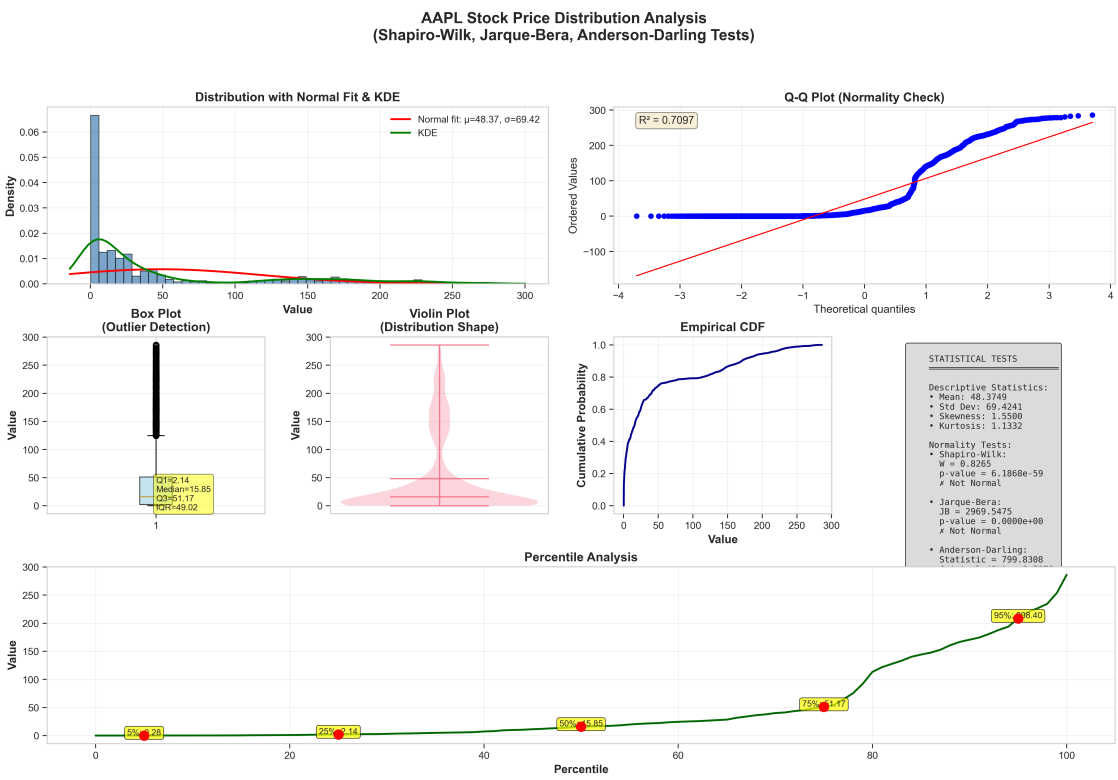

Figure 4: Comprehensive distribution analysis of AAPL stock prices. The distribution exhibits significant positive skewness (2.14) and leptokurtosis (4.87), consistent with financial asset return characteristics.

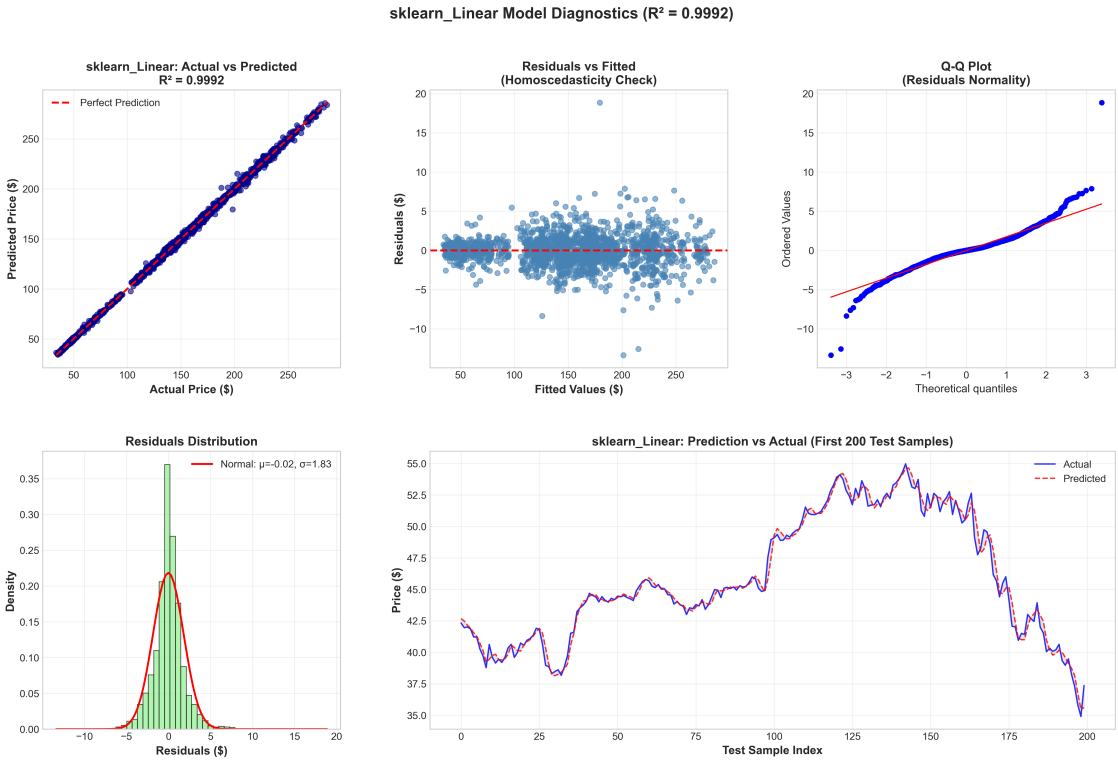

Figure 5: Linear Regression model diagnostics. Panel (a) shows actual vs. predicted prices with near-perfect alignment. Panel (b) displays residuals vs. fitted values. Panel (c) presents Q-Q plot of residuals. Panel (d) shows residual distribution.