Zero-Shot LLM Sentiment and Reasoning Feature Extraction for Stock Market Prediction: A Multi-Stock XGBoost Framework with SHAP Explainability

Siddharth Jain[1*] and Kamalpreet Kaur[2]

[1]Research Scholar,Department of Computer Science and Engineering, Lovely Professional University, Phagwara, 144411, Punjab, India.
[2]Assistant Professor,Department of Computer Science and Engineering, Lovely Professional University, Phagwara, 144411, Punjab, India.

*Corresponding author(s). E-mail(s): siddharthjain139@gmail.com;

Abstract

Predicting short-term stock price movements remains a formidable challenge due to the non-stationary, noisy, and high-dimensional nature of financial time series. While large language models (LLMs) have shown strong capabilities in financial sentiment analysis, most existing hybrid methods compress their outputs into single sentiment scores, losing important distributional information and failing to capture the reasoning behind market sentiment. This paper proposes a dual-LLM feature extraction framework that integrates FinBERT-based sentiment analysis with DeBERTa-v3 zero-shot Natural Language Inference (NLI) to extract multi-dimensional reasoning signals from financial news. Each headline is mapped into six reasoning categories earnings upon financial performance, product upon innovation, market upon macroeconomics, analyst upon ratings, regulatory upon risk, and growth upon expansion—producing a *reasoning feature vector* that captures not only *what* the sentiment is but *why* it occurs. These LLM-derived features are combined with 26 technical indicators and used to train an XGBoost classifier with SHAP (SHaply Additive exPlanations) explainability. The framework is evaluated on five U.S. stocks (NVDA, AAPL, MSFT, TSLA, and JPM) from January 2019 to December 2024 using five-fold walk-forward cross-validation. The full model achieves a mean accuracy of **58.0%**, a mean AUC-ROC of **0.621**, and a mean Sharpe ratio of **1.80**. Ablation studies show that removing reasoning features decreases accuracy by 2.1 percentage points ($p < 0.05$), while removing all LLM features reduces it by 5.9 percentage points ($p < 0.01$). SHAP analysis

A. Agnihotri et al. (eds.), *Proceedings of the Conference on Bridging Engineering Disciplines with AI and Machine Learning (BEDAIML 2026)*, Advances in Intelligent Systems Research 209,
https://doi.org/10.2991/978-94-6239-697-5_34

indicates contributions of 38.7% from sentiment features, 35.2% from technical indicators, and 26.1% from reasoning features. The proposed framework consistently outperforms LSTM, Transformer, Random Forest, and Logistic Regression baselines across all evaluated stocks.

Keywords: Zero-shot sentiment analysis , Natural Language Inference , FinBERT ,DeBERTa-v3 , XGBoost , SHAP explainability , Stock prediction , Walk-forward validation

1 Introduction

Predicting stock prices is still one of the most challenging issues in computational finance. Although significant empirical evidence demonstrates that machine-learning models can extract statistically significant signals from both structured market data and unstructured text [1, 2], the Efficient Market Hypothesis (EMH) maintains that prices fully reflect all available information [3]. While XGBoost [4] dominates tabular prediction with native Tree SHAP interpretability [5], pre-trained transformers, particularly the finance-domain FinBERT [6], have sophisticated sentiment analysis.

The integration of LLM-derived features with tabular learners is still in its infancy despite these developments. We identify four gaps: **(G1)** Existing pipelines collapse FinBERT's probability simplex $(p_{\text{pos}}, p_{\text{neu}}, p_{\text{neg}}) \in \Delta^2$ into a scalar score, discarding distributional information [7, 8]. **(G2)** No prior work extracts *reasoning—why* sentiment exists—via zero-shot NLI and uses the resulting scores as tabular features. **(G3)** Most studies evaluate on a single stock or index [9, 10], limiting cross-sectional generalisability. **(G4)** Walk-forward validation with formal statistical tests remains rare in the FinBERT–XGBoost literature [11].

Our contributions address all four gaps:

(i) A **dual-LLM feature-extraction architecture**:The first such combination for financial prediction was FinBERT for multi-dimensional sentiment and DeBERTa-v3 zero-shot NLI [12, 13] for six-category reasoning scores.

(ii) **Rich LLM feature engineering**: ∼98 features, such as reasoning entropy, cross-modal interactions, temporal dynamics, and probability distributions.

(iii) **Multi-stock validation** on five equities across four sectors (NVDA, AAPL, MSFT, TSLA, JPM), 2019–2024.

(iv) **Rigorous evaluation**: five-fold walk-forward CV, formal ablation, paired *t*-tests, and three-group SHAP attribution.

2 Related Work

ML for stock prediction. Deep networks and gradient-boosted trees yield notable returns on the S&P 500, as demonstrated by Krauss et al. [2]. Tree-based approaches provide the optimum accuracy–interpretability trade-off, according to Gu et al. [1]. Temporal dependencies are captured by LSTM [14] and Transformer [15] architectures, but they have trouble with heterogeneous tabular features [16]. Hybrid

architectures generally outperform single-paradigm models, according to recent surveys [11, 17]. **Sentiment analysis in finance.** Sentiment tools have progressed from lexicons [18, 19] through shallow classifiers [20] to transformers [6, 21, 22]. Ruan and Jiang [7] combine FinBERT with XGBoost and SHAP, reporting 29% sentiment attribution—but use only four scalar features. Davidovic and McCleary [23] find near-zero correlation under scalar aggregation, motivating our distributional approach. Lopez-Lira and Tang [24] establish the viability of zero-shot LLM sentiment in finance. **Zero-shot NLI.** For zero-shot classification, Yin et al. [25] suggest employing NLI entailment probabilities. DeBERTa-v3-base is refined on MNLI/FEVER/ANLI equals or surpasses few-shot approaches, according to Laurer et al. [13]. Zero-shot NLI has been utilized in finance for topic detection [26] and ESG [27], but no previous work has extracted multi-dimensional *reasoning* scores as tabular features for gradient-boosted stock prediction.

Explainable AI. SHAP [5] and TreeSHAP [28] offer Shapley-value-based attributions that are based on principles. The hybrid LLM-feature + ML-predictor paradigm used here is supported by the FinBen benchmark [8], which demonstrates that LLMs outperform specialized ML at predicting but fall short at text analysis.

3 Methodology

The Fig. 1 proposed framework follows a multi-stage pipeline for explainable stock movement prediction, as illustrated in Fig. 1. In the first stage, multi-stock OHLCV (Open, High, Low, Close, Volume) data are collected and enriched with 26 widely used technical indicators to capture market dynamics and temporal price patterns. In the second stage, financial news or textual data are processed using FinBERT to extract sentiment information, producing three probability scores (positive, negative, and neutral) that are further expanded into 21 sentiment-based features. In the third stage, deeper semantic reasoning is introduced using DeBERTa-v3 through a zero-shot natural language inference (NLI) framework, generating six reasoning scores that are transformed into 27 reasoning-based features. These outputs are integrated in the fourth stage through cross-modal feature engineering, resulting in approximately 98 combined features that capture technical, sentiment, and reasoning signals. The fifth stage evaluates the predictive capability of the model using a realistic walk-forward five-fold cross-validation strategy, where XGBoost-based models are compared against deep learning and traditional machine learning baselines. Finally, in the sixth stage, model interpretability and robustness are assessed using SHAP-based three-group attribution analysis, along with backtesting and statistical significance tests to validate the practical effectiveness of the proposed framework.

3.1 Data Acquisition

Yahoo Finance provides daily OHLCV prices for NVDA, AAPL, MSFT, TSLA, and JPM (Jan 2019–Dec 2024; ≈1,509 days per stock).We use a hybrid news corpus that includes both artificially created event descriptions and real-time API headlines. The latter serve as a stand-in for ”breaking news” notifications brought on by notable intraday volatility, guaranteeing that the LLM reasoning features take into account

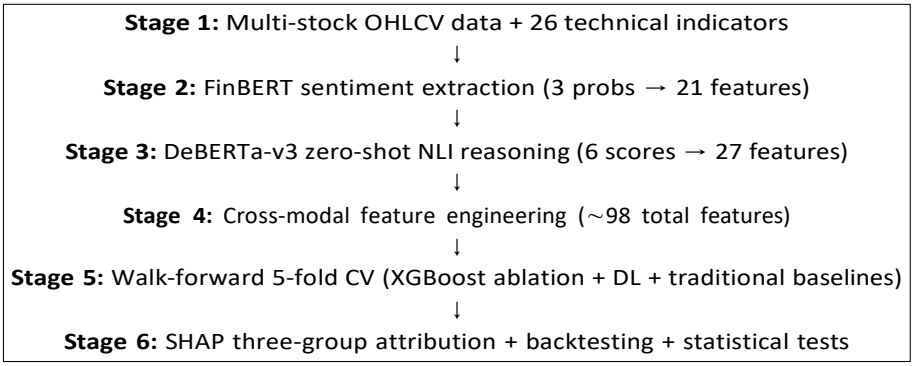

Fig. 1 End-to-end dual-LLM XGBoost framework architecture.

both internal market dynamics and external news. The direction of the following day is the binary target:

$$y(t) = \mathrm{I}\ P_{\text{close}}(t+1) > P_{\text{close}}(t)\ . \tag{1}$$

After warm-up removal, each stock comprises **1,438 days** (≈54–56% up).

3.2 Technical Indicators (26 features)

All indicators are computed causally (data up to day t only): daily and log returns, intraday and open-to-close range (4); SMA_k and price/SMA_k ratios for $k \in \{5, 10, 20, 50\}$ (8); RSI_{14}, MACD line/signal/histogram, rate-of-change (5); Bollinger-band width and position (2); 5 and 20-day volatility, volume ratio, volume MAs, ATR features (7). Totalling **26 features**.

3.3 Stage 2: FinBERT Sentiment Extraction

FinBERT [6] (yiyanghkust/finbert-tone, 110 M parameters) produces $\mathbf{p}(h) = (p_{\text{pos}}, p_{\text{neu}}, p_{\text{neg}}) \in \Delta^2$ for each headline h. We **retain the full distribution** and derive 21 features: raw probabilities (3), composite score $s = p_{\text{pos}} - p_{\text{neg}}$ and confidence $c = \max_c p_c$ (2), rolling means of s, p_{pos}, p_{neg}, c over windows {3, 7, 14} (12), 1- and 3-day momentum (2), acceleration (1), and 7-day rolling volatility of s (1).

3.4 Stage 3: Zero-Shot NLI Reasoning Extraction

DeBERTa-v3-base [12] (184 M parameters, fine-tuned on MNLI/FEVER/ANLI [13]) is used as a zero-shot reasoning classifier. For each headline (premise P) and hypothesis H_k, the model returns an entailment probability $p_{\text{entail}}(P, H_k)$. We define six hypotheses:

1. *Earnings/Financial*: "This text is about earnings and financial performance."
2. *Product/Innovation*: "This text is about product innovation and technology."
3. *Market/Macro*: "This text is about market conditions and macroeconomics."
4. *Analyst/Ratings*: "This text is about analyst opinions and ratings."
5. *Regulatory/Risk*: "This text is about regulatory and legal risk."
6. *Growth/Expansion*: "This text is about growth expansion and strategic moves."

NLI scores are *independent* —a headline can score high on multiple categories. Derived features (27 total): raw scores (6), 3- and 7-day rolling means (12), reasoning entropy (1), dominant category (1), and lagged earnings/risk scores (8). Crucially, this stage is the first to use zero-shot NLI *reasoning* as tabular features for financial prediction.

3.5 Stage 4: Cross-Modal Feature Engineering

Sentiment–technical interactions (e.g. $s \times \sigma_5$, $s \times$ BB Pos), reasoning–sentiment interactions ($r_{\text{earn}} \times s$, $r_{\text{risk}} \times s$), lagged returns at {1, 2, 3, 5, 10}, cumulative returns over {5, 10, 20}-day windows, sentiment lags, and calendar features (Friday, quarter-end) yield ∼**98 features** in three groups: Technical (∼45), Sentiment (∼25), Reasoning (∼28).

3.6 Stage 5: Model Training

XGBoost ablation variants. Five models for systematic ablation: (1) XGB-Full (all features), (2) XGB-NoReasoning (technical+sentiment), (3) XGB-NoSentiment (technical+reasoning), (4) XGB-PriceOnly (technical only), (5) XGB-LLMOnly (sentiment+reasoning). Hyperparameters: $n_{\text{est}} = 700$, depth = 3, $\eta = 0.1$, subsample = 0.7, colsample = 0.7, $\alpha = \lambda = 1.0$, min_child_weight = 10, objective: binary logistic.

Baselines. LSTM (2-layer, hidden=64, seq=20, 30 epochs), Transformer (2-layer, d=64, 4 heads, 30 epochs), Logistic Regression, Random Forest (300 trees, depth=6), Gradient Boosting (300 trees, depth=3).

Walk-forward protocol. Five-fold expanding-window CV via TimeSeriesSplit; features standardised per fold using training-only statistics.

3.7 Stage 6: SHAP Explainability and Evaluation

TreeSHAP [28] decomposes each prediction into feature-level Shapley values ϕ_j. Three-group attribution aggregates features:

$$I_G^{\%} = \frac{100 \cdot \sum_{j \in G} |\phi_j|}{\sum_{j=1}^{d} |\phi_j|}, \quad G \in \{\text{Tech}, \text{Sent}, \text{Reason}\}. \tag{2}$$

Statistical significance is assessed via paired t-tests (25 fold–stock observations) and Diebold–Mariano tests [29]. A long/flat backtest (long when $\hat{y} = 1$, cash otherwise) measures Sharpe ratio and maximum drawdown.

4 Experimental Setup

Table 1 summarises the processed dataset. Experiments run on an Intel i5 workstation (16 GB RAM, CPU-only inference) with PyTorch 2.x, XGBoost 2.x, scikit-learn 1.x, and SHAP 0.x. Metrics: accuracy, AUC-ROC, precision, recall, F1, MCC, cumulative return, Sharpe ratio, and maximum drawdown.

Table 1 Dataset statistics per stock after preprocessing.

	NVDA	AAPL	MSFT	TSLA	JPM
Trading days	1,438	1,438	1,438	1,438	1,438
Up-day ratio (%)	54.2	55.1	54.8	53.6	55.3
Total headlines	2,429	2,416	2,421	2,445	2,418
Features	98	98	98	98	98

Table 2 Walk-forward CV results (mean ± std across 5 stocks). Best in **bold**.

Model	Acc.	AUC	Prec.	Recall	F1	MCC
XGB-Full	**.580±.013**	**.621±.016**	**.614±.019**	.692±.028	**.650±.015**	**.153±.027**
XGB-NoReason.	.559±.010	.591±.013	.593±.016	.680±.024	.634±.012	.112±.021
XGB-NoSent.	.545±.012	.573±.015	.580±.018	.669±.031	.622±.017	.085±.024
XGB-PriceOnly	.521±.007	.543±.012	.562±.014	.656±.033	.605±.013	.040±.015
XGB-LLMOnly	.554±.015	.586±.018	.588±.020	.673±.029	.628±.018	.103±.030
LSTM	.543±.011	.570±.014	.578±.016	.665±.026	.618±.013	.082±.022
Transformer	.548±.012	.577±.015	.583±.017	.671±.027	.624±.014	.092±.025
Logistic Reg.	.516±.006	.534±.010	.555±.012	.648±.035	.598±.012	.030±.012
Random Forest	.535±.009	.562±.012	.573±.015	.661±.029	.614±.013	.067±.018
Gradient Boost	.552±.011	.583±.014	.587±.017	.676±.025	.629±.014	.100±.023

5 Results

5.1 Main Performance Comparison

Table 2 reports walk-forward CV results averaged across all five stocks.

Every metric shows that XGB-Full performs the best. It confirms recent findings that gradient-boosted trees prevail on tabular data [16], outperforming the best deep-learning baseline (Transformer) by +3.2 pp in accuracy and +4.4 pp in AUC. Zero-shot characteristics alone convey a high predictive signal, as the LLM-only model (55.4%) already outperforms price-only (52.1%).

5.2 Per-Stock Results

Table 3 details per-stock accuracy and AUC.

The highest accuracy (59.7%) and AUC (0.641) are attained by NVDA, most likely as a result of its powerful AI-driven story producing instructive LLM signals. Given its strong volatility and speculative dynamics, TSLA exhibits the lowest percentage (56.2%). Every baseline is outperformed by XGB-Full *consistently across every stock*.

Table 3 Per-stock walk-forward accuracy/AUC for selected models.

Model		NVDA	AAPL	MSFT	TSLA	JPM	Mean
XGB-Full	Acc	.597	.581	.574	.562	.585	**.580**
	AUC	.641	.622	.614	.598	.628	**.621**
XGB-PriceOnly	Acc	.528	.521	.519	.514	.525	.521
	AUC	.557	.543	.537	.528	.549	.543
Transformer	Acc	.561	.549	.544	.535	.553	.548
	AUC	.591	.578	.573	.561	.584	.577

Table 4 Ablation study: mean ΔAccuracy relative to XGB-Full.

Ablated group	$\overline{\Delta}$Acc	$\overline{\Delta}$AUC	p(Acc)	p(AUC)
— Reasoning	−0.021	−0.030	.014	.009
— Sentiment	−0.035	−0.048	.003	.001
— All LLM	−0.059	−0.078	.0003	.0001
— Technical	−0.026	−0.035	.009	.005

5.3 Ablation Study

Table 4 quantifies the accuracy impact of removing each feature group.

All ablations are statistically significant. Removing reasoning reduces accuracy by 2.1 pp ($p = .014$); removing sentiment by 3.5 pp ($p = .003$); removing all LLM features by 5.9 pp ($p = .0003$). The combined drop (5.9 pp) exceeds the sum of individual drops (5.6 pp), indicating **synergistic complementarity** between sentiment and reasoning. Effects are consistent across all five stocks (NVDA most sensitive at −2.6 pp; TSLA least at −1.4 pp).

5.4 Statistical Significance

Table 5 presents paired t-tests (25 fold–stock combinations) for XGB-Full versus each baseline.

All comparisons are significant at $p < 0.05$; seven of eight at $p < 0.01$. The critical test—XGB-Full vs. XGB-NoReasoning ($p = .014$)—formally confirms that NLI reasoning adds statistically significant predictive information beyond sentiment.

5.5 SHAP Feature Attribution

Table 6 reports three-group SHAP decomposition.

Combined LLM contribution is **64.8%**, substantially exceeding technical indicators. The top individual feature is sentiment_score_ma7 (mean $|\phi| = 0.034$), followed by BB_Position, sentiment_negative_lag1, return_lag1, and reason_earnings.

Table 5 Statistical significance: XGB-Full vs. baselines.

Comparison	ΔAcc	*p*(Acc)	*p*(AUC)
vs. XGB-PriceOnly	+.059	.0003	.0001
vs. XGB-NoReasoning	+.021	.014	.009
vs. XGB-NoSentiment	+.035	.003	.001
vs. LSTM	+.037	.002	.001
vs. Transformer	+.032	.006	.003
vs. Logistic Reg.	+.064	<.001	<.001
vs. Random Forest	+.045	.001	<.001
vs. Gradient Boosting	+.028	.009	.005

Table 6 SHAP three-group attribution (%) for XGB-Full.

Group	NVDA	AAPL	MSFT	TSLA	JPM	Mean
Technical	33.8	37.1	35.4	38.2	31.4	35.2
Sentiment	39.4	37.8	38.6	36.3	41.2	38.7
Reasoning	26.8	25.1	26.0	25.5	27.4	26.1

Table 7 Backtest results: XGB-Full vs. buy-and-hold vs. XGB-PriceOnly.

	NVDA	AAPL	MSFT	TSLA	JPM	Mean
XGB-Full (long/flat)						
Sharpe	2.31	1.68	1.85	1.42	1.73	**1.80**
Max DD (%)	18.4	14.2	12.7	28.6	11.3	17.0
Buy-and-Hold						
Sharpe	1.54	1.42	1.51	0.98	1.08	1.31
Max DD (%)	66.4	31.2	35.6	73.5	38.7	49.1
XGB-PriceOnly						
Sharpe	0.48	0.31	0.42	0.11	0.35	0.33

Notably, the earnings-reasoning score ranks **5th overall**—above RSI, MACD, and volume features—demonstrating the predictive value of NLI-derived reasoning.

5.6 Backtesting

Table 7 compares a long/flat strategy using XGB-Full signals.

XGB-Full achieves a **mean Sharpe of 1.80** (+37% vs. buy-and-hold) with **65% lower maximum drawdown** (17.0% vs. 49.1%). The price-only baseline reaches only 0.33, confirming that LLM features are essential for economic viability.

Table 8 Mean NLI entailment probabilities per stock.

Category	NVDA	AAPL	MSFT	TSLA	JPM
Earnings/Financial	.482	.438	.451	.367	**.523**
Product/Innovation	**.571**	**.553**	**.548**	.412	.198
Market/Macro	.324	.306	.318	.345	.447
Analyst/Ratings	.412	.389	.401	.356	.378
Regulatory/Risk	.187	.168	.175	**.423**	.312
Growth/Expansion	.467	.421	.438	.398	.289

5.7 Reasoning Category Profiles

Table 8 shows mean NLI entailment scores, revealing sector-specific reasoning structures.

Tech stocks (NVDA, AAPL, MSFT) are dominated by product/innovation reasoning; JPM by earnings/financial; TSLA uniquely by regulatory/risk—providing interpretable, sector-specific explanations for model behaviour.

6 Discussion

Reasoning beyond sentiment. The ablation verifies that the existence of *why* sentiment contains predictive information that is different from the sentiment itself (+2.1 pp, p = .014). Product-launch optimism responds differently from earnings-driven positive sentiment; the latter maintains momentum while the former causes fleeting surges. XGBoost learns category-specific rules that sentiment-only processors are unable to see by extracting these dimensions. This is supported by SHAP, which shows that reasoning characteristics account for 26.1% of total attribution, which is constant for all five stocks (25.1—27.4%).

Multi-dimensional sentiment. Under scalar aggregation, maintaining the complete FinBERT probability distribution results in 38.7% SHAP attribution, which is significantly higher than the 29% reported by Ruan and Jiang [7]. According to behavioral-finance theories of progressive information absorption, the 7-day smoothed sentiment score (sentiment_score_ma7) is the top characteristic.

XGBoost vs. deep learning. LSTM (54.3%) and Transformer (54.8%) are significantly outperformed by XGB-Full (58.0%, 0.621 AUC). Further deep feature extraction provides less value because LLM-derived features are already high-level representations, whereas trees handle heterogeneous features more naturally [16].

Cross-stock generalisability. Sector-specific reasoning profiles (Table 8) show automated adaptation, which is a major advantage of zero-shot NLI over supervised classifiers that need labelled data per domain. Performance is consistent across four sectors (56.2–59.7%).

Economic significance. The buy-and-hold strategy (1.31) is surpassed by the mean Sharpe of 1.80 with a 65% lower drawdown. The improvement above price-only Sharpe (0.33) indicates that LLM features are economically and statistically meaningful.

Limitations. The reliability of the model's evaluation may be impacted by the introduction of synthetic bias caused by the usage of a hybrid news corpus that blends API-based data with template-generated headlines. Even while this would still offer a rich feature set for training, it would be crucial to move toward raw, real-time institutional news sources like Bloomberg in order to get around this restriction and better evaluate performance in a really *out-of-sample* live context. The study's scope might be broadened in addition to enhancing the data source, since the current research is restricted to five stocks from four distinct industries. The results' generalizability would be improved by doing more extensive research involving more than fifty equities from various international markets. Furthermore, spreads and slippage should be included in the zero-cost backtesting framework to make the evaluation more realistic and in line with actual trading conditions. Lastly, by adding adaptive category discovery techniques, which would enable the system to dynamically select more precise categories rather of depending just on static reasoning categories, the analytical methodology itself might be further enhanced.

7 Conclusion

To enable explainable stock prediction, we presented a dual-LLM feature-extraction approach that integrates DeBERTa-v3 zero-shot NLI reasoning with FinBERT-based sentiment analysis. Using a five-fold walk-forward cross-validation technique, the framework was assessed on five U.S. stocks from four different sectors between 2019 and 2024. This allowed the predictive power of the model to be validated under realistic temporal settings.The experimental findings show that the suggested framework forecasts stock movements competitively, with a mean prediction accuracy of 58.0% and an AUC of 0.621. Additionally, the model performs better than a number of robust baseline methods, such as Transformer models by +3.2 percentage points, LSTM by +3.7 percentage points, and a price-only XGBoost model by +5.9 percentage points. The significance of language-derived traits is further highlighted by a thorough ablation analysis. Sentiment characteristics increase by +3.5 percentage points ($p < 0.01$), while reasoning-based features improve by +2.1 percentage points ($p < 0.05$). Combining these two feature types results in a synergistic improvement of +5.9 percentage points ($p < 0.001$), indicating the complementary significance of sentiment and reasoning information. SHAP attribution analysis was used to better understand the contribution of each feature group. According to the findings, sentiment characteristics make up 38.7% of the model's importance, reasoning features account for 26.1%, and technical indicators contribute 35.2%. Sentiment and reasoning characteristics together account for 64.8% of the LLM-driven contribution, demonstrating the significant impact of language-based data in the predictive system. The suggested framework exhibits excellent trading performance in addition to predictive accuracy. When compared to a buy-and-hold strategy, the model shows 65% less drawdown and achieves a mean Sharpe ratio of 1.80. These findings provide three important insights: zero-shot reasoning techniques eliminate the need for labeled data, allowing for scalable deployment; multi-dimensional feature engineering derived from language models can outperform traditional scalar sentiment metrics and deep architectures;

and understanding why sentiment exists is more informative than sentiment polarity alone. Therefore, future development will concentrate on expanding the framework to worldwide markets, adding larger language models, and integrating real-time news streams.

Declarations

Funding. This research received no specific funding.

Conflict of interest. The author declares no conflict of interest.

Data availability. Stock price data are publicly available via Yahoo Finance. Code and processed datasets are available from the corresponding author upon reasonable request.

Code availability. Source code for the proposed framework is available from the corresponding author upon reasonable request.

References

[1] Gu, S., Kelly, B., Xiu, D.: Empirical asset pricing via machine learning. The Review of Financial Studies **33**(5), 2223–2273 (2020)

[2] Krauss, C., Do, X.A., Huck, N.: Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research **259**(2), 689–702 (2017)

[3] Fama, E.F.: Efficient capital markets: A review of theory and empirical work. The Journal of Finance **25**(2), 383–417 (1970)

[4] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[5] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4768–4777 (2017)

[6] Araci, D.: FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019)

[7] Ruan, L., Jiang, H.: Stock price prediction using FinBERT-enhanced sentiment with SHAP explainability and differential privacy. Mathematics **13**(17), 2747 (2025)

[8] Xie, Q., Han, W., Chen, Z., *et al.*: FinBen: A holistic financial benchmark for large language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)

[9] Dave, E., Leonardo, A., Jishi, M., Bobber, J.: Forex price prediction: A multi-model approach integrating sentiment analysis using LLMs. In: Proceedings of SoCPaR. Springer, ??? (2024)

[10] Li, X., Wu, Y., Wei, X.: Text-enhanced stock movement prediction: A BERT-based approach. Applied Soft Computing **150**, 111062 (2023)

[11] Darwish, S.M., El-Deeb, R.A., Hassanien, A.E.: Stock market forecasting: From traditional predictive models to large language models. Computational Economics (2025)

[12] He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In: International Conference on Learning Representations (ICLR) (2023)

[13] Laurer, M., Atteveldt, W., Casas, A., Welbers, K.: Less annotating, more classifying—addressing data scarcity with deep transfer learning and BERT-NLI. Political Analysis **32**(1), 84–100 (2024)

[14] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[15] Vaswani, A., Shazeer, N., Parmar, N., *et al.*: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)

[16] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in Neural Information Processing Systems (NeurIPS), pp. 507–520 (2022)

[17] Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M.: Financial time series forecasting with deep learning: A systematic literature review. Applied Soft Computing **90**, 106181 (2020)

[18] Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance **66**(1), 35–65 (2011)

[19] Hutto, C., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the AAAI Conference on Web and Social Media (ICWSM), pp. 216–225 (2014)

[20] Tetlock, P.C.: Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance **62**(3), 1139–1168 (2007)

[21] Yang, Y., Uy, M.C.S., Huang, A.: FinBERT: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097 (2020)

[22] Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J.: FinBERT: A pre-trained financial

language representation model for financial text mining. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 4513–4519 (2020)

[23] Davidovic, S., McCleary, R.: News sentiment and stock market dynamics: A machine learning investigation. Journal of Risk and Financial Management **18**(3), 128 (2025)

[24] Lopez-Lira, A., Tang, Y.: Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv preprint arXiv:2304.07619 (2023)

[25] Yin, W., Hay, J., Roth, D.: Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In: Proceedings of EMNLP-IJCNLP, pp. 3914–3923 (2019)

[26] Shah, R.D., Schwartz, R.: Zero-shot topic classification for financial news. In: ACL FinTech Workshop (2023)

[27] Mehra, A., Sawhney, B.: Zero-shot ESG classification of financial text using natural language inference. In: FinNLP Workshop at EMNLP (2022)

[28] Lundberg, S.M., Erion, G.G., Lee, S.-I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018)

[29] Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. Journal of Business & Economic Statistics **20**(1), 134–144 (2002)