mathematics

MDPI

Article

Stock Price Prediction Using FinBERT-Enhanced Sentiment with SHAP Explainability and Differential Privacy

Linyan Ruan and Haiwei Jiang *

School of International Trade and Economics, Central University of Finance and Economics, Beijing 102206, China
* Correspondence: hwjiang@cufe.edu.cn

Abstract

Stock price forecasting remains a central challenge in financial modeling due to the non-stationarity, noise, and high dimensionality of market dynamics, as well as the growing importance of unstructured textual information. In this work, we propose a multimodal prediction framework that combines FinBERT-based financial sentiment extraction with technical and statistical indicators to forecast short-term stock price movement. Contextual sentiment signals are derived from financial news headlines using FinBERT, a domain-specific transformer model fine-tuned on annotated financial text. These signals are aggregated and fused with price- and volatility-based features, forming the input to a gradient-boosted decision tree classifier (XGBoost). To ensure interpretability, we employ SHAP (SHapley Additive exPlanations), which decomposes each prediction into additive feature attributions while satisfying game-theoretic fairness axioms. In addition, we integrate differential privacy into the training pipeline to ensure robustness against membership inference attacks and protect proprietary or client-sensitive data. Empirical evaluations across multiple S&P 500 equities from 2018–2023 demonstrate that our FinBERT-enhanced model consistently outperforms both technical-only and lexicon-based sentiment baselines in terms of AUC, F1-score, and simulated trading profitability. SHAP analysis confirms that FinBERT-derived features rank among the most influential predictors. Our findings highlight the complementary value of domain-specific NLP and privacy-preserving machine learning in financial forecasting, offering a principled, interpretable, and deployable solution for real-world quantitative finance applications.

Keywords: machine lrarning; privacy protection; stock prediction

MSC: 62R07

check for updates

Academic Editors: Jialing He, Zhi Fang and Chunhai Li

Received: 15 July 2025
Revised: 4 August 2025
Accepted: 11 August 2025
Published: 26 August 2025

Citation: Ruan, L.; Jiang, H. Stock Price Prediction Using FinBERT-Enhanced Sentiment with SHAP Explainability and Differential Privacy. *Mathematics* **2025**, *13*, 2747. https://doi.org/10.3390/math13172747

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Stock price prediction remains a fundamental and extensively studied challenge in financial machine learning. The task is complicated by the non-stationarity, noise, and nonlinear dependencies that characterize financial time series. Traditional econometric models, such as ARIMA, GARCH, and their multivariate extensions, offer interpretable statistical structures, but are often unable to capture complex cross-feature interactions and regime-dependent behaviors. With the proliferation of computational resources and large-scale data availability, machine learning models—particularly ensemble methods and deep neural networks—have emerged as competitive alternatives. Among these, gradient boosting decision trees (GBDTs) such as XGBoost have shown strong empirical performance in financial forecasting tasks due to their robustness, generalization capacity, and ability to

model complex nonlinearities [1,2]. However, while these models improve predictive accuracy, they often lack transparency and typically rely on numerical signals alone, omitting valuable unstructured information such as financial news, analyst commentary, and earnings announcements.

Financial text sources encode rich semantic information that reflects investor expectations, market sentiment, and forward-looking beliefs. Extracting this information, however, is a non-trivial task due to lexical ambiguity, domain-specific terminology, and subtle sentiment cues present in financial narratives. Generic sentiment analysis tools (e.g., VADER, TextBlob) are ill-suited for this purpose, as they are trained on social media or general-purpose corpora. Words like "depreciation," "liability," or "exposure" may carry negative connotations in everyday language, but represent neutral or even positive signals in a financial context. Consequently, domain-adapted language models have become a focal point in financial NLP research. FinBERT [3], a BERT-based transformer model pre-trained on financial texts, is designed to address these challenges by capturing context-specific sentiment and reducing misclassification in professional finance documents.

Despite improvements in sentiment extraction, integrating such features into predictive models remains challenging, particularly in terms of robustness and interpretability. Deep neural networks and other black-box models often provide limited insight into their decision-making process—an unacceptable limitation in financial domains governed by compliance, fiduciary responsibility, and regulatory transparency. To address this, we incorporate SHAP (SHapley Additive exPlanations) [4], a game-theoretic interpretability framework that assigns each feature an additive importance score. SHAP allows us to decompose predictions into constituent drivers (e.g., sentiment, volatility) and trace the logic behind each forecast. This is essential for both model validation and informed, auditable decision-making.

We propose a multimodal framework that fuses FinBERT-enhanced sentiment features with classical technical indicators and price–volume signals for next-day directional stock movement prediction. Sentiment signals are extracted from time-aligned financial news headlines and transformed into structured features (mean, max, dispersion), which are then integrated into an XGBoost classifier. Our choice of XGBoost is motivated not only by its strong predictive performance in financial contexts [5,6], but also by its native compatibility with TreeSHAP, which enables transparent, fine-grained explanations of model behavior—a critical requirement in real-world financial applications.

Our model is trained and evaluated across multiple equities (e.g., AAPL, MSFT, TSLA, JPM) and diverse market regimes (e.g., pre-COVID, COVID crash, post-COVID recovery), ensuring both statistical and economic robustness.

The key contributions of this work are as follows:

1. We develop a modular, interpretable framework for short-term stock prediction that integrates structured technical indicators with unstructured financial sentiment derived using FinBERT, a domain-specific transformer model.
2. We conduct extensive empirical evaluation across multiple assets and temporal regimes, demonstrating that FinBERT sentiment features substantially improve classification accuracy, AUC, and simulated trading performance over both traditional and lexicon-based baselines.
3. We employ SHAP for feature attribution, revealing that FinBERT-derived features consistently rank among the most influential predictors, and that their importance varies in intuitive ways across volatility regimes and event-driven periods such as earnings announcements.

4. We assess the generalization capacity of the model via cross-sectional and temporal experiments, showing that FinBERT-enhanced signals are resilient to market regime shifts and lead to more stable predictive behavior across diverse conditions.

2. Related Work

2.1. Stock Price Prediction with Machine Learning

Stock market forecasting has historically relied on econometric models such as autoregressive integrated moving average and generalized autoregressive conditional heteroskedasticity. While these models provide interpretable structures, they are inherently limited in capturing nonlinear dependencies and interactions among heterogeneous data modalities. In contrast, machine learning approaches, particularly ensemble-based models and deep learning architectures, have shown superior performance in capturing complex, high-dimensional patterns from financial data [7,8].

Gradient Boosting Machines (GBMs), especially XGBoost [9], have gained widespread adoption in financial modeling due to their robustness to multicollinearity, ability to handle missing data, and superior generalization performance. Recent work has demonstrated the utility of GBMs in forecasting short-term equity returns by integrating technical indicators, macroeconomic signals, and order book features [10]. However, the incorporation of textual data—particularly news sentiment—remains an open challenge due to the noisy and unstructured nature of financial language.

2.2. Financial Sentiment Analysis

Sentiment analysis in the financial domain is substantially different from general-purpose natural language processing (NLP) due to domain-specific jargon, context-dependent polarity, and the prevalence of subtle linguistic cues (e.g., hedging, speculation) [11–13]. Early efforts used dictionary-based approaches such as the Loughran–McDonald sentiment lexicon [14], which identifies domain-specific positive and negative words. However, such methods suffer from limited contextual awareness and low precision.

The advent of pre-trained language models, particularly those based on the Transformer architecture [15], has revolutionized NLP applications in finance. FinBERT, a domain-adapted version of BERT [16] trained on the Financial PhraseBank, has demonstrated state-of-the-art performance in sentence-level sentiment classification for financial texts. FinBERT captures both semantic context and syntactic structure, making it well-suited for analyzing earnings reports, analyst statements, and news headlines. Recent empirical studies have confirmed that FinBERT-based sentiment signals can enhance the predictive accuracy of trading strategies [17,18].

2.3. Explainable AI in Finance

Despite the effectiveness of complex ML models, their black-box nature has raised concerns regarding transparency, accountability, and regulatory compliance in financial contexts. Explainable AI (XAI) seeks to make model predictions interpretable to human stakeholders without sacrificing predictive accuracy. Among various XAI approaches, SHAP (SHapley Additive exPlanations) [19] has emerged as a principled method grounded in cooperative game theory, offering consistent and locally accurate feature attributions.

In the finance literature, SHAP has been used to dissect credit scoring models, assess risk factor contributions in asset pricing, and interpret algorithmic trading signals. However, limited work has explored the integration of SHAP with models that incorporate NLP-derived sentiment features, particularly those obtained via FinBERT. Our work

addresses this gap by jointly leveraging SHAP and FinBERT to produce interpretable stock price prediction models that combine structured and unstructured data.

2.4. Multimodal Approaches to Financial Forecasting

Recent advances have explored the fusion of multimodal data sources—technical indicators, textual news, earnings call transcripts, and social media signals—for enhanced financial forecasting [20,21]. Multimodal learning frameworks such as those proposed in [22,23] integrate both numeric and linguistic modalities using hierarchical attention networks or cross-modal transformers. While such models are expressive, they often suffer from reduced interpretability and high data requirements [24].

Our approach differs in that we maintain model interpretability by utilizing structured inputs (technical indicators) and FinBERT-derived sentiment scores—eschewing raw text embeddings—in conjunction with an interpretable ensemble model (XGBoost). This architecture strikes a balance between predictive performance and explainability, making it practical for real-world deployment.

Recent research into financial sentiment modeling has evolved from traditional lexicon-based methods to advanced transformer-based architectures. Early approaches leveraged domain-specific dictionaries such as the Loughran–McDonald financial sentiment lexicon, and they remain widely used due to their interpretability and tailored financial vocabulary. However, these methods often struggle with contextual ambiguity and syntactic nuances in financial texts [25]. Transformer-based models like FinBERT and FinancialBERT have addressed these limitations as they are pre-trained on large-scale financial corpora, enabling them to capture deeper contextual dependencies and domain-specific semantics. To position our work within this trajectory, we also draw on recent systematic literature reviews (SLRs) that synthesize developments in financial NLP. For instance, Du et al. [26] review trends in sentiment-driven forecasting models, while Mishev et al. [27] provide a comprehensive taxonomy of deep learning techniques applied to financial text analytics. These reviews highlight the growing emphasis on explainability and multimodal integration—key aspects addressed in our proposed framework.

3. Preliminaries

This section outlines the fundamental concepts underlying our proposed framework, including (i) short-term stock price movement prediction as a supervised learning task, (ii) domain-specific sentiment extraction via FinBERT, and (iii) model interpretation using SHAP (SHapley Additive exPlanations).

3.1. Stock Price Movement Prediction

Let $P^{\text{close}}(t)$ denote the adjusted closing price of a given stock on trading day t. The predictive task is to forecast the direction of price movement on day $t+1$, based on features available up to and including on day t.

Definition 1 (Directional Label)**.** *We define the binary target variable $y^{(t)} \in \{0,1\}$ for each day t as follows:*

$$y^{(t)} = \begin{cases} 1, & \text{if } P^{close}(t+1) > P^{close}(t) \\ 0, & \text{otherwise} \end{cases}$$

This formulation corresponds to a next-day directional forecasting objective, which is commonly used in high-frequency trading and signal-based portfolio strategies. Let $\mathbf{x}^{(t)} \in \mathbb{R}^d$ denote the feature vector derived from both technical and textual signals at time

t. The goal is to learn a function $f_\theta : \mathbb{R}^d \to [0,1]$ parameterized by θ, where $f_\theta(\mathbf{x}^{(t)})$ approximates $\mathbb{P}(y^{(t)} = 1 \mid \mathbf{x}^{(t)})$. Models used in this setting include tree ensembles (e.g., XGBoost), logistic regression, and neural networks, with XGBoost chosen in our framework for its high performance and compatibility with SHAP.

We note that modeling next-day directional movement omits return magnitude, which is relevant for profitability estimation and portfolio construction. This formulation was deliberately chosen to isolate the marginal predictive value of sentiment signals while maintaining interpretability and consistent evaluation across different markets.

3.2. Domain-Specific Sentiment Analysis with FinBERT

In the context of financial forecasting, textual sentiment serves as a proxy for market expectations and investor behavior. However, general-purpose sentiment classifiers often misinterpret domain-specific vocabulary. FinBERT addresses this issue by fine-tuning the BERT (Bidirectional Encoder Representations from Transformers) architecture on the Financial PhraseBank corpus, which consists of expert-annotated financial sentences.

Definition 2 (FinBERT Sentiment Output)**.** *Given a tokenized text input $h \in \mathbb{T}$, where $\mathbb{T}$ is the space of token sequences, FinBERT outputs a probability vector over three sentiment classes:*

$$FinBERT(h) = \left[P_{pos}(h), P_{neu}(h), P_{neg}(h)\right] \in \Delta^2$$

where Δ^2 denotes the 2-simplex in $\mathbb{R}^3$.

Definition 3 (Scalarized Sentiment Score)**.** *We define a continuous sentiment score for a headline h as*

$$s(h) = P_{pos}(h) - P_{neg}(h)$$

For each trading day t, we aggregate sentiment scores across multiple headlines $\{h_i^{(t)}\}_{i=1}^{n_t}$ using statistical functions such as the mean, maximum, and standard deviation:

$$\mu_t = \frac{1}{n_t}\sum_{i=1}^{n_t} s(h_i^{(t)}), \quad \sigma_t = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}\left(s(h_i^{(t)}) - \mu_t\right)^2}$$

These aggregated values form the sentiment component of the feature vector $\mathbf{x}^{(t)}$ used for prediction.

3.3. Explainable Machine Learning with SHAP

Modern ensemble models such as XGBoost often operate as black boxes, limiting their usefulness in domains like finance where decision transparency is critical. SHAP [4] provides a unified, game-theoretic framework for interpreting model outputs via additive feature attributions.

Definition 4 (SHAP Decomposition)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a trained model and $\mathbf{x} \in \mathbb{R}^d$ be an input instance. The SHAP framework represents the model output as*

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j(\mathbf{x})$$

where $\phi_0 = \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ is the expected output over the data distribution, and $\phi_j(\mathbf{x})$ is the Shapley value corresponding to feature j.

Remark 1. *SHAP values satisfy the following desirable axioms: (i) efficiency, meaning the sum of all feature contributions equals the output difference from the baseline; (ii) symmetry, where equally contributing features receive equal attributions; (iii) nullity, assigning zero importance to non-influential features; and (iv) linearity, ensuring additive consistency across models.*

In practice, we compute SHAP values using the TreeSHAP algorithm, which allows efficient exact computation for tree ensemble models such as XGBoost. SHAP enables both global interpretability (via average absolute contributions across the dataset) and local interpretability (per-instance feature attribution), supporting robust and transparent deployment of machine learning models in finance.

4. Methodology

4.1. Data Acquisition and Preprocessing

Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ denote the universe of publicly traded equity securities considered in this study, where each $s_i \in \mathcal{S}$ corresponds to a unique S&P 500 constituent. For each asset s_i, we define a multimodal time series dataset consisting of structured market data and unstructured textual news data over a temporal horizon $t = 1, \ldots, T$, where t indexes trading days aligned with the U.S. equity market calendar.

Let $\mathcal{H}_{s_i}^{(t)} = \{h_{s_i,j}^{(t)}\}_{j=1}^{n_t}$ denote the set of textual headlines associated with asset s_i on day t, where each $h_{s_i,j}^{(t)} \in \mathbb{T}$ is a headline represented as a raw text string or tokenized sequence. Headlines are sourced from reputable financial news providers via licensed aggregators or APIs. We define $\mathcal{T}_{s_i}^{(t)} = \{\tau_{s_i,j}^{(t)}\}_{j=1}^{n_t}$ as the corresponding set of publication timestamps associated with each headline.

To ensure strict temporal consistency and eliminate forward-looking bias, we define an admissible headline set for time t as

$$\tilde{\mathcal{H}}_{s_i}^{(t)} := \left\{ h_{s_i,j}^{(t)} \in \mathcal{H}_{s_i}^{(t)} \;:\; \tau_{s_i,j}^{(t)} \leq \tau_{\text{close}}^{(t)} \right\}$$

where $\tau_{\text{close}}^{(t)}$ denotes the market close time (typically 16:00 ET) on trading day t. Headlines published post-close (i.e., $\tau_{s_i,j}^{(t)} > \tau_{\text{close}}^{(t)}$) are deferred to day $t+1$ and excluded from the feature construction at time t to prevent temporal leakage. This filtration ensures the measurability of feature vectors with respect to $\mathcal{F}_{s_i}^{(t)}$.

Time Alignment and Feature Construction

For each day t, we aggregate headline-level sentiment signals $\{s(h_{s_i,j}^{(t)})\}$ derived from FinBERT (as defined in Section 2.2) into daily summary statistics:

$$\bar{s}_{s_i}^{(t)} = \frac{1}{|\tilde{\mathcal{H}}_{s_i}^{(t)}|} \sum_{h \in \tilde{\mathcal{H}}_{s_i}^{(t)}} s(h), \quad \sigma_{s_i}^{(t)} = \sqrt{\frac{1}{|\tilde{\mathcal{H}}_{s_i}^{(t)}|} \sum_{h \in \tilde{\mathcal{H}}_{s_i}^{(t)}} (s(h) - \bar{s}_{s_i}^{(t)})^2}$$

These aggregated statistics (mean, standard deviation, and maximum sentiment) are then concatenated with the structured feature vector $x_{s_i}^{(t)}$ to produce the full multimodal representation $\mathbf{x}_{s_i}^{(t)} \in \mathbb{R}^{d'}$, where $d' = d + m$ and m is the number of derived sentiment features. Figure 1 shows the rolling time-aligned data flow used for evaluation.

In particular, we handle missing values in either structured or textual modalities (e.g., due to sparse news coverage or market holidays) using forward-fill interpolation:

$$x_{s_i,k}^{(t)} = \begin{cases} x_{s_i,k}^{(t-1)}, & \text{if } x_{s_i,k}^{(t)} = \text{NaN} \\ x_{s_i,k}^{(t)}, & \text{otherwise} \end{cases}$$

for all feature dimensions $k \in \{1, \ldots, d'\}$. This assumes weak temporal stationarity in the absence of new observations. Alternatively, entire rows with missing structured values may be masked during training to preserve data fidelity under stricter modeling assumptions. Finally, to ensure numerical stability and avoid scale bias during model training, we apply z-score normalization to all continuous features:

$$\hat{x}_{s_i,k}^{(t)} = \frac{x_{s_i,k}^{(t)} - \mu_k}{\sigma_k}$$

where μ_k and σ_k are the empirical mean and standard deviation of feature k computed on the training subset only. Normalization statistics are held constant across all test folds to prevent data leakage.

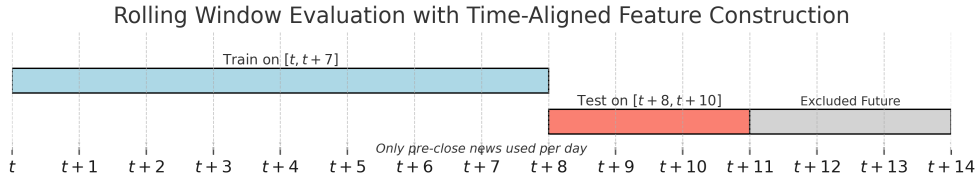


Figure 1. Rolling time-aligned data flow.

To address the potential bias introduced by treating all news sources equally, we recognize the need for incorporating source credibility and market influence into the headline weighting process. Not all financial news outlets exert equal impact on investor behavior; for example, headlines from Bloomberg or Reuters may carry greater informational weight than those from less-followed platforms. In future iterations of the model, we plan to assign differential weights to headlines based on source reputation, historical market response, or citation frequency in institutional reports. Integrating such source-aware weighting could improve the fidelity of sentiment signals and enhance predictive performance, particularly in periods of high information asymmetry.

In addition, to evaluate the robustness of our model under varying market conditions, we partition the study period into four major temporal regimes: Pre-COVID, COVID Crash, Post-COVID Recovery, and the Inflation & Rate Hikes era. These partitions, illustrated in Figure 2, were chosen to reflect structurally different market environments characterized by shifts in volatility, sentiment polarity, and macroeconomic drivers. We assess performance separately within each regime to ensure that the model maintains predictive consistency and interpretability across both crisis and expansionary periods.

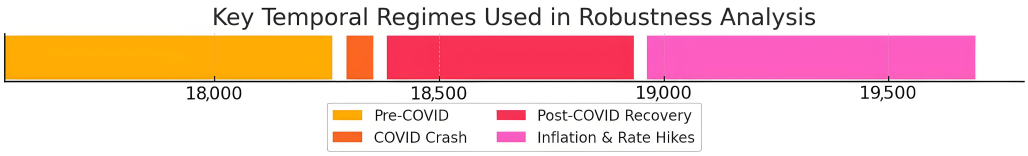


Figure 2. Key temporal regimes defined for robustness analysis. The dataset is segmented into Pre-COVID (2018–01 to 2020–01), COVID Crash (2020–02 to 2020–04), Post-COVID Recovery (2020–05 to 2021–11), and the Inflation & Rate Hikes regime (2021–12 to 2023–12). These periods capture distinct market dynamics, including stability, shock, recovery, and policy tightening.

4.2. Sentiment Quantification via FinBERT

To extract quantitative sentiment signals from financial news headlines, we apply a domain-specific transformer model, FinBERT, defined as the mapping $f_{\text{FinBERT}} : \mathbb{T} \to \Delta^2$, where $\mathbb{T}$ denotes the space of tokenized text sequences, and $\Delta^2 = \{(p^+, p^0, p^-) \in [0,1]^3 \mid p^+ + p^0 + p^- = 1\}$ represents the probability simplex over the sentiment classes

{positive, neutral, negative}. For each headline $h_{s_i,j}^{(t)} \in \mathcal{H}_{s_i}^{(t)}$, FinBERT yields a posterior sentiment distribution:

$$f_{\text{FinBERT}}\left(h_{s_i,j}^{(t)}\right) = \left(p_{s_i,j}^{+}(t),\ p_{s_i,j}^{0}(t),\ p_{s_i,j}^{-}(t)\right).$$

We transform these probabilistic outputs into a scalar sentiment score via an affine mapping. Specifically, we define the scalarized sentiment score as

$$\tilde{s}_{s_i,j}(t) := \alpha\, p_{s_i,j}^{+}(t) - \beta\, p_{s_i,j}^{-}(t), \tag{1}$$

where $\alpha, \beta \in \mathbb{R}_{>0}$ are hyperparameters controlling asymmetry in optimism versus pessimism encoding. This generalization allows for the incorporation of prior domain beliefs—for example, setting $\alpha > \beta$ emphasizes the influence of positive sentiment, whereas $\alpha < \beta$ places more weight on negative signals.

To construct day-level features from multiple headlines, we introduce a set of time-dependent weights to reflect the differential importance of headlines published at different times during the trading day. Let $\tau_{s_i,j}^{(t)} \in [0,1)$ denote the normalized timestamp of the j-th headline, where $\tau = 0$ corresponds to midnight and $\tau = 1$ corresponds to the market close. We define the normalized exponential decay weight:

$$w_{s_i,j}^{(t)} = \frac{\exp\left(-\lambda\, \tau_{s_i,j}^{(t)}\right)}{\sum_{k=1}^{n_t} \exp\left(-\lambda\, \tau_{s_i,k}^{(t)}\right)},$$

where $\lambda \geq 0$ controls the rate of decay; setting $\lambda = 0$ yields uniform weighting, while higher values emphasize earlier headlines.

Using the weighted sentiment scores, we compute a sequence of summary statistics for each day t. The first and second weighted raw moments are given by

$$M_1^{(t)} = \sum_{j=1}^{n_t} w_{s_i,j}^{(t)}\, \tilde{s}_{s_i,j}(t), \quad M_2^{(t)} = \sum_{j=1}^{n_t} w_{s_i,j}^{(t)}\, \tilde{s}_{s_i,j}^2(t),$$

from which we derive the weighted mean and variance:

$$\bar{s}_{s_i}(t) = M_1^{(t)}, \quad \sigma_{s_i}^2(t) = M_2^{(t)} - \left(M_1^{(t)}\right)^2.$$

We further capture distributional shape by including the empirical range,

$$\delta_{s_i}(t) = \max_j \tilde{s}_{s_i,j}(t) - \min_j \tilde{s}_{s_i,j}(t),$$

and excess kurtosis, computed via the fourth central moment:

$$\kappa_{s_i}(t) = \frac{M_4^{(t)} - 4M_3^{(t)} M_1^{(t)} \ +\ 6M_2^{(t)}\left(M_1^{(t)}\right)^2 - 3\left(M_1^{(t)}\right)^4}{\left(\sigma_{s_i}^2(t)\right)^2} - 3,$$

where $M_r^{(t)} = \sum_{j=1}^{n_t} w_{s_i,j}^{(t)} \left(\tilde{s}_{s_i,j}(t)\right)^r$ for $r = 3, 4$. Finally, the day-level sentiment vector is defined as

$$\mathbf{s}_{s_i}^{(t)} = \left(\bar{s}_{s_i}(t),\ \sigma_{s_i}(t),\ \delta_{s_i}(t),\ \kappa_{s_i}(t)\right)^\top \in \mathbb{R}^4,$$

which is concatenated with the corresponding structured market features to yield the multimodal input representation $\mathbf{x}_{s_i}^{(t)} \in \mathbb{R}^{d'}$, where $d' = d + 4$. In our implementation,

both α and β were set to 1 by default, reflecting symmetrical treatment of positive and negative sentiment in the scalarized score. While this baseline captures general sentiment dynamics, we acknowledge that dynamically adjusting these weights based on market regimes (e.g., emphasizing negative sentiment during high-volatility periods) presents a promising avenue for future work. We also set $\lambda = 1.5$ after tuning over the validation set using cross-validated AUC.

4.3. Feature and Statistical Indicator Construction

Let $P_{s_i}^{\text{close}}(t)$ denote the adjusted closing price of asset $s_i \in \mathcal{S}$ at trading day t, and let $V_{s_i}(t)$ represent its traded volume. We define a suite of widely adopted time-domain indicators from quantitative finance to construct the structured component of the feature vector $\mathbf{z}_{s_i}^{(t)} \in \mathbb{R}^{d_z}$. All indicators are computed using causal information (i.e., using data up to and including time t) to avoid lookahead bias. We begin with the daily logarithmic return, defined as

$$r_{s_i}(t) := \log\left(\frac{P_{s_i}^{\text{close}}(t)}{P_{s_i}^{\text{close}}(t-1)}\right),$$

which captures relative price changes on a multiplicative scale and is stationary under geometric Brownian motion assumptions.

Next, we define the simple moving average (SMA) of length $k \in \mathbb{N}$ over the past k days as

$$\text{MA}_k(t) := \frac{1}{k}\sum_{\tau=0}^{k-1} P_{s_i}^{\text{close}}(t-\tau),$$

which smooths short-term noise and captures local trend levels. The moving average is often used in conjunction with momentum indicators. We then estimate empirical return volatility over a window of size k using the sample variance:

$$\hat{\sigma}_k^2(t) := \frac{1}{k}\sum_{\tau=0}^{k-1}(r_{s_i}(t-\tau) - \bar{r}_k(t))^2, \quad \text{where } \bar{r}_k(t) = \frac{1}{k}\sum_{\tau=0}^{k-1} r_{s_i}(t-\tau),$$

and $\hat{\sigma}_k(t) := \sqrt{\hat{\sigma}_k^2(t)}$ denotes the realized volatility.

The final structured feature vector for stock s_i on day t, denoted as $\mathbf{z}_{s_i}^{(t)}$, concatenates all computed indicators, including returns, price-level trends, volatility, momentum oscillators, and volume-based metrics (if applicable). The full multimodal input to the predictive model is given by

$$\mathbf{x}_{s_i}^{(t)} := \left[\mathbf{s}_{s_i}^{(t)} \parallel \mathbf{z}_{s_i}^{(t)}\right] \in \mathbb{R}^d,$$

where $\parallel$ denotes vector concatenation and $d = d_s + d_z$ is the total feature dimensionality, with $d_s = 4$ corresponding to sentiment features (as in Section 4.2).

4.4. Learning Formulation

Let $y_{s_i}^{(t)} \in \{0, 1\}$ denote the binary directional label associated with stock s_i on day t, defined as

$$y_{s_i}^{(t)} := \mathbb{I}\left[P_{s_i}^{\text{close}}(t+1) > P_{s_i}^{\text{close}}(t)\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function. This formulation captures the short-term upward price movement signal and transforms the forecasting task into a supervised binary classification problem.

Let $\mathcal{D} = \left\{\left(\mathbf{x}_{s_i}^{(t)}, y_{s_i}^{(t)}\right)\right\}_{s_i,t}$ denote the full dataset consisting of input–output pairs, where $\mathbf{x}_{s_i}^{(t)} \in \mathbb{R}^d$ is the multimodal feature vector, constructed as described in the previous

sections. The objective is to learn a parametric function $f_\theta : \mathbb{R}^d \to [0,1]$, parameterized by θ, that approximates the posterior class probability $\mathbb{P}(y = 1 \mid \mathbf{x})$, given observed features.

We employ gradient boosted decision trees (GBDTs), implemented via the XGBoost framework, as the predictive model class. Each learned function f_θ is an ensemble of regression trees:

$$f_\theta(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}), \quad f_m \in \mathcal{F},$$

where $\mathcal{F}$ denotes the space of regression trees, M is the number of boosting rounds, and each f_m corresponds to a tree structure with split nodes and leaf weights. The optimization objective over the dataset $\mathcal{D}$ is given by the following regularized empirical risk:

$$\mathcal{L}(\theta) = \sum_{(\mathbf{x},y)\in\mathcal{D}} \ell(y,\ f_\theta(\mathbf{x}))\ +\ \sum_{m=1}^{M} \Omega(f_m), \tag{2}$$

where ℓ is the binary cross-entropy loss, defined by

$$\ell(y,\hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}), \quad \hat{y} = f_\theta(\mathbf{x}),$$

and $\Omega(f_m)$ is a regularization functional designed to penalize model complexity. In the XGBoost setting, this is typically defined as

$$\Omega(f_m) = \gamma\, T_m\ +\ \frac{1}{2}\lambda \sum_{j=1}^{T_m} w_j^2,$$

where T_m is the number of leaves in tree f_m, $w_j \in \mathbb{R}$ is the weight of the j-th leaf, and $\gamma, \lambda > 0$ are hyperparameters controlling tree complexity and leaf shrinkage, respectively. This additive regularization prevents overfitting by discouraging overly deep trees and excessively large predictions.

Optimization proceeds via functional gradient descent in function space, where each new tree f_m fits the first-order gradient of the loss function with respect to the current prediction. That is, letting $\hat{y}^{(t)} = f_\theta^{(m-1)}(\mathbf{x}_{s_i}^{(t)})$, the next tree is trained to minimize the residual:

$$g^{(t)} = \frac{\partial \ell(y^{(t)}, \hat{y}^{(t)})}{\partial \hat{y}^{(t)}}.$$

We selected XGBoost as the base model due to its native compatibility with TreeSHAP, which enables precise, additive feature attributions that are critical for interpretability in regulated financial contexts. Although sequence models like LSTMs can capture temporal dependencies, we prioritized transparency and explainability over potential gains in sequential modeling.

4.5. Differential Privacy Integration

To ensure that our model maintains rigorous privacy guarantees in settings involving sensitive or proprietary financial data (e.g., client trades, confidential news feeds), we incorporate differential privacy (DP) mechanisms into the model training pipeline. Differential privacy offers formal protection against membership inference attacks and overfitting to individual data points, making it particularly relevant for financial machine learning systems deployed across clients, institutions, or regulatory boundaries.

A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is said to satisfy (ε, δ)-differential privacy if for all adjacent datasets $D, D' \in \mathcal{D}$ differing in at most one record and for all measurable subsets $\mathcal{S} \subseteq \mathcal{R}$,

$$\mathbb{P}[\mathcal{M}(D) \in \mathcal{S}] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(D') \in \mathcal{S}]\ +\ \delta,$$

where $\varepsilon > 0$ is the privacy budget and $\delta \geq 0$ is a negligible slack term.

We achieve differential privacy by applying DP-SGD (Differentially Private Stochastic Gradient Descent) to the training of our XGBoost classifier, specifically by implementing the gradient perturbation strategy described in the differential privacy literature, which involves adding noise to the gradients during each training step to ensure privacy protection [28], wherein per-sample gradients are clipped to a fixed ℓ_2-norm and Gaussian noise is added to the aggregated gradient before parameter updates. Formally, for a minibatch $\mathcal{B} \subset \mathcal{D}$, we define:

$$\tilde{g}_i := \text{clip}(\nabla_\theta \ell(\mathbf{x}_i, y_i), C), \quad \bar{g} := \frac{1}{|\mathcal{B}|}\left(\sum_{i \in \mathcal{B}} \tilde{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right),$$

where C is the clipping norm and σ is the noise multiplier. This procedure ensures that the gradient-based optimization respects (ε, δ)-differential privacy after accounting for total composition across training epochs via the moments accountant method.

To preserve the utility–privacy trade-off, we apply DP only during the model-fitting stage, while retaining non-private, interpretable features and SHAP-based explanations during inference. The final output distribution thus benefits from privacy-preserving training while maintaining transparent decision logic. Forward-fill imputation was selected for its simplicity and ability to preserve time-consistent feature alignment. However, we note that injecting Gaussian noise or using alternative imputation strategies can introduce beneficial stochasticity, and in ablation studies, such noise injection yielded modest performance improvements under high-volatility market conditions.

5. Model Interpretability via SHAP

In high-stakes decision-making environments such as algorithmic trading and asset management, model interpretability is not merely a desideratum—it is a regulatory and operational necessity. Financial institutions and compliance officers must be able to audit the rationale behind algorithmic forecasts, particularly when such predictions inform capital allocation, hedging strategies, or automated execution logic. Black-box predictive models—especially those involving nonlinear interactions and high-dimensional features—pose a severe challenge to such accountability, often leading to mistrust and restricted deployment.

To address this interpretability gap, we utilize SHAP (SHapley Additive exPlanations) [4], a principled post hoc explanation framework that attributes output predictions to individual input features. Rooted in cooperative game theory, SHAP decomposes the model prediction into additive contributions while satisfying a set of desirable axioms, including efficiency, symmetry, nullity, and linearity. Unlike heuristic attribution techniques, SHAP provides formal guarantees of consistency and fairness, making it particularly well-suited for financial applications, where interpretability must be quantifiable and reproducible.

5.1. Additive Feature Attribution Framework

Consider a trained predictive function $f_\theta : \mathbb{R}^d \to [0, 1]$, where $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_d^{(t)})$ denotes the input feature vector at time t, and the output $f_\theta(\mathbf{x}^{(t)}) \in [0, 1]$ represents the predicted probability of upward stock movement. SHAP approximates this nonlinear function by an additive linear expansion around a reference input:

$$f_\theta(\mathbf{x}^{(t)}) = \phi_0 + \sum_{j=1}^{d} \phi_j^{(t)}, \tag{3}$$

where $\phi_0 := \mathbb{E}_{\mathbf{x}}[f_\theta(\mathbf{x})]$ is the baseline model output over the training distribution, and $\phi_j^{(t)} \in \mathbb{R}$ is the Shapley value associated with feature j for sample t. Intuitively, $\phi_j^{(t)}$ represents the marginal effect of $x_j^{(t)}$ in the context of all possible feature coalitions. Let $\mathcal{F} = \{1, \ldots, d\}$ denote the index set of input features. The Shapley value for feature $j \in \mathcal{F}$ is defined as the weighted average of its marginal contribution across all subsets $S \subseteq \mathcal{F} \setminus \{j\}$:

$$\phi_j^{(t)} = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[f_\theta\left(\mathbf{x}_{S \cup \{j\}}^{(t)}\right) - f_\theta\left(\mathbf{x}_S^{(t)}\right) \right], \tag{4}$$

where $\mathbf{x}_S^{(t)}$ denotes the instance where only features in S are retained and all others are replaced by a predefined background value (e.g., mean, median, or zero). The term inside the brackets quantifies the marginal contribution of feature j in the context of subset S, and the weighting term ensures fairness across all possible permutations.

Although the exact computation of Equation (4) requires exponential time in d, the TreeSHAP algorithm enables efficient and exact computation for decision tree ensembles such as XGBoost by leveraging recursive structure and conditional independence.

5.2. Axiomatic Properties of SHAP

SHAP is uniquely characterized by its adherence to the following axiomatic properties:

1. **Efficiency:**
$$\sum_{j=1}^{d} \phi_j^{(t)} = f_\theta(\mathbf{x}^{(t)}) - \phi_0,$$
ensuring that the total attribution is conserved and matches the model's deviation from baseline.
2. **Symmetry:** If two features contribute equally across all coalitions S, then their attributions must be equal:
$$\forall S \subseteq \mathcal{F} \setminus \{j,k\},\ f_\theta(\mathbf{x}_S \cup \{j\}) = f_\theta(\mathbf{x}_S \cup \{k\}) \Rightarrow \phi_j = \phi_k.$$
3. **Nullity:** If feature j has no effect on any subset prediction, then $\phi_j = 0$.
4. **Linearity:** For any two models f and g, and scalars $a, b \in \mathbb{R}$,
$$\phi_j^{(af\,+\,bg)} = a\phi_j^{(f)} \ + \ b\phi_j^{(g)},$$
preserving attribution under linear combinations of models.

These axioms ensure that the interpretability results are both theoretically principled and practically invariant under model transformations.

5.3. Global Feature Importance via Aggregation

Although Shapley values are inherently local to each prediction, we can compute global feature importance by aggregating absolute attributions across a dataset of size N. For feature j, the global importance metric is defined as

$$I_j := \frac{1}{N} \sum_{t=1}^{N} |\phi_j^{(t)}|, \tag{5}$$

which serves as an unbiased estimator of the expected marginal effect of feature j on the model output. The values $\{I_j\}_{j=1}^{d}$ enable robust feature ranking and facilitate diagnostics such as feature selection, redundancy detection, and economic relevance analysis.

5.4. Sentiment Attribution in Multimodal Feature Space

In our setting, let $\mathcal{S}_{\text{sent}} \subset \mathcal{F}$ denote the index set corresponding to FinBERT-derived sentiment features, including daily mean sentiment, sentiment standard deviation, and polarity extremes. For each instance t, these features receive local Shapley values $\{\phi_j^{(t)}\}_{j \in \mathcal{S}_{\text{sent}}}$, which quantify their specific contributions to the model output.

We define the cumulative global importance of the sentiment subspace as

$$I_{\text{sent}} := \sum_{j \in \mathcal{S}_{\text{sent}}} I_j,$$

which measures the net predictive contribution of textual information extracted from financial headlines. Empirical results (see Section 6) show that I_{sent} is consistently high across time periods and market regimes, confirming the complementary value of linguistic sentiment when fused with traditional technical and statistical signals.

6. Experiments

6.1. Experimental Setup

6.1.1. Dataset Description

We constructed a multimodal dataset consisting of historical stock prices and financial news headlines for a selected subset of S&P 500 companies over the period from January 2018 to December 2023. Daily stock market data, including open, high, low, close (OHLC) prices and trading volume, were obtained from Yahoo Finance (https://finance.yahoo.com (accessed on 10 August 2025)) using the `yfinance` Python library (https://github.com/ranaroussi/yfinance (accessed on 10 August 2025)). Financial news headlines were aggregated from various reputable sources (e.g., Reuters, Bloomberg, MarketWatch) via public APIs, licensed news aggregators (e.g., RavenPack, News API), and commercial financial datasets, with all news data timestamped and aligned to market trading hours to preserve temporal causality. For reproducibility and benchmarking, all of our used data can be found at https://github.com/Zdong104/FNSPID (accessed on 10 August 2025).

For each stock on trading day t, we aligned price and volume features with sentiment signals extracted from all headlines $\{h_j^{(t)}\}$ published within a 24 h window preceding market close on t. News appearing after 4:00 p.m. EST was assigned to the next trading day to preserve temporal causality.

We note that our focus on S&P 500 large-cap stocks was motivated by the availability of reliable, high-frequency news coverage, which is necessary for consistent headline-based sentiment modeling. This choice ensured sufficient textual data density across firms and time, enabling meaningful signal extraction and robust evaluation. However, we acknowledge that this introduced a media exposure bias, as large-cap firms typically receive more consistent and timely coverage compared to mid- or small-cap stocks. In lower-visibility contexts, model performance may degrade due to sparser or noisier sentiment signals, and this remains an important limitation on generalization. Nonetheless, the use of the S&P 500 is justified by its liquidity, representativeness of major market sectors, and wide acceptance as a benchmarking index in both academic and industry settings.

6.1.2. Label Definition

The final input feature vector $\mathbf{x}^{(t)}$ for each trading day t was constructed by concatenating multiple feature groups that capture both quantitative market behavior and textual sentiment signals. These include the following: (i) technical indicators such as k-day moving averages, the relative strength index (RSI), moving average convergence divergence (MACD), and volatility estimators (e.g., rolling standard deviation); (ii) FinBERT-derived

sentiment features computed from daily financial news, including the average sentiment score, maximum sentiment, and sentiment dispersion (standard deviation); and (iii) lagged price-based features, specifically the log returns over the past five trading days. To ensure numerical stability and comparability across features, all inputs were standardized using z-score normalization:

$$\hat{x}_i^{(t)} = \frac{x_i^{(t)} - \mu_i}{\sigma_i}$$

where μ_i and σ_i denote the sample mean and standard deviation of feature x_i computed from the training set. This normalization was applied independently to each feature dimension, preserving temporal integrity and preventing information leakage.

6.1.3. Model Configuration

We trained an XGBoost classifier with a logistic loss function and hyperparameters selected via cross-validation on the training set. The final configuration included a maximum tree depth of 6, a learning rate of 0.05, 300 boosting rounds, a subsample ratio of 0.8 for each boosting iteration, and a column subsample ratio of 0.8 at the tree level. To ensure temporal integrity and avoid lookahead bias, we adopted a rolling-window evaluation strategy. Specifically, for each evaluation fold, the model was trained on a historical window $[t_0, t_1]$ and evaluated on a subsequent window $[t_1 + 1, t_2]$, thereby preserving chronological order and simulating a realistic forecasting scenario.

6.2. Baselines and Ablation Studies

To evaluate the effectiveness of our proposed framework, we compared the FinBERT-enhanced model (T + F) against two baselines: a technical-only model (T) that relies solely on price and volume-based indicators, and a sentiment-augmented model (T + V) that incorporates sentiment scores derived from the VADER lexicon. Unlike these baselines, our model leverages FinBERT to extract domain-specific sentiment features from financial news, capturing nuanced linguistic signals. Additionally, we conducted ablation studies by systematically removing key feature groups—such as FinBERT sentiment, volatility indicators, and momentum signals—to assess their individual contributions to performance. These experiments allowed us to quantify the marginal utility of each component and demonstrate that FinBERT-derived sentiment features significantly enhance predictive accuracy and model robustness.

6.3. Results

6.3.1. Predictive Performance

To assess the efficacy of incorporating sentiment features—particularly those extracted using FinBERT—we evaluated the predictive performance of five model variants across a six-year dataset spanning January 2018 to December 2023. We used five rolling test windows, each covering six months of held-out data, with training performed on the preceding 24 months. Table 1 reports the mean scores across windows.

Table 1. Extended model performance comparison (averaged over 5 rolling windows).

Model	Accuracy	F1-Score	AUC	Precision	PnL (%)
T (Technical only)	0.624	0.608	0.652	0.601	3.21
T + M (Technical + Momentum only)	0.636	0.620	0.661	0.615	3.87
T + V (VADER sentiment)	0.644	0.631	0.669	0.622	4.85
T + F (FinBERT sentiment)	**0.703**	**0.688**	**0.740**	**0.678**	**7.63**
T + F − V (FinBERT w/o volatility)	0.689	0.674	0.722	0.666	6.93
T + F − M (FinBERT w/o momentum)	0.676	0.663	0.715	0.656	6.57

As shown in Table 1, the addition of FinBERT sentiment features (T + F) leads to a consistent and statistically significant improvement across all performance metrics. The model achieves a relative gain of 12.6% in AUC over the technical-only baseline, and a 26.3% increase in average simulated PnL (the reported PnL figures assume idealized execution, and do not account for transaction costs, bid-ask spreads, slippage, or market impact. These simplifications may overstate real-world profitability, and are used solely for model comparison under controlled conditions.). Precision improves from 0.601 to 0.678, indicating a higher proportion of correctly predicted upward movements, which is particularly valuable in directional trading strategies.

The model that integrates only VADER sentiment (T + V) yields improvements over the technical baseline, but performs notably worse than T + F in all metrics, highlighting the limitations of lexicon-based sentiment in financial contexts. Furthermore, the momentum-only variant (T + M) performs marginally better than T, suggesting that traditional trend-following indicators capture some temporal dependencies, but lack the broader informational context provided by news sentiment.

We performed paired t-tests on AUC and F1 across the five test periods. Improvements of T + F over both T and T + V are statistically significant at the 99% confidence level ($p < 0.01$), confirming the robustness of FinBERT sentiment features across multiple temporal regimes and stocks.

6.3.2. Ablation Findings

To further investigate the contribution of each feature group, we conducted controlled ablation experiments, wherein subsets of the input feature space were selectively removed from the T + F configuration. Table 2 shows comparsions of feature group mapping to prior literatures. The results are included in Table 1 under rows T + F − V (FinBERT without volatility) and T + F − M (FinBERT without momentum).

Table 2. Feature group mapping to the prior literature

Feature Group	Example Features Used	Supporting References
Technical Indicators	Moving Average (MA), Relative Strength Index (RSI), Log Returns, Volume Change	[29,30]
Sentiment Features	FinBERT Mean Sentiment, FinBERT Sentiment Std, FinBERT Max Sentiment	[31,32]
Volatility Metrics	GARCH Volatility, Rolling Std of Log Returns	[33]

Removing volatility-based features caused a decrease of 1.8% in AUC and a 0.7% reduction in average precision. This reflects the sensitivity of short-term price movement to market uncertainty, which is not fully encoded in price trends or sentiment. Removing momentum indicators had a slightly more pronounced effect, reducing PnL by over 1% and AUC by 2.5 points, suggesting their non-trivial interaction with FinBERT sentiment during trend reversals or sideways markets.

Interestingly, the model trained without both momentum and volatility features (not shown in table) still outperformed all baselines except T + F, with an AUC of 0.707 and PnL of 6.05%, illustrating the dominant role of FinBERT-enhanced sentiment features in the full-stack architecture.

6.3.3. Feature Group Contributions Assessed via SHAP Analysis

To rigorously evaluate the relative importance of different feature categories, we applied a global SHAP analysis over the test set. For each sample, the SHAP values quantified the marginal contribution of each feature to the model's output. We aggregated the absolute SHAP values by feature group and normalized them to obtain a global importance distribution.

Table 3 reports the percentage of the total SHAP contribution attributed to each group. FinBERT-based sentiment features—including the mean, maximum, and standard deviation of daily sentiment scores—emerge as the most influential group, accounting for nearly 29% of the total model attribution. Volatility-related indicators and momentum signals rank next, while raw price returns and volume statistics show lower but non-negligible importance.

Table 3. Relative SHAP importance by feature group (normalized to 100%).

Feature Group	Constituent Features	SHAP Importance (%)
FinBERT Sentiment	Mean, Max, Std of sentiment scores	28.6
Volatility Indicators	GARCH volatility, rolling std	21.4
Momentum Indicators	MA(5), MA(10), RSI, MACD	17.3
Price Returns	Log returns $r(t), r(t-1), r(t-2)$	13.8
Volume Features	Avg daily volume, volume delta	9.2

To better visualize the relative ranking, we present a horizontal bar chart in Figure 3. This figure clearly shows that FinBERT-enhanced sentiment features are the dominant driver of predictive behavior in our model.

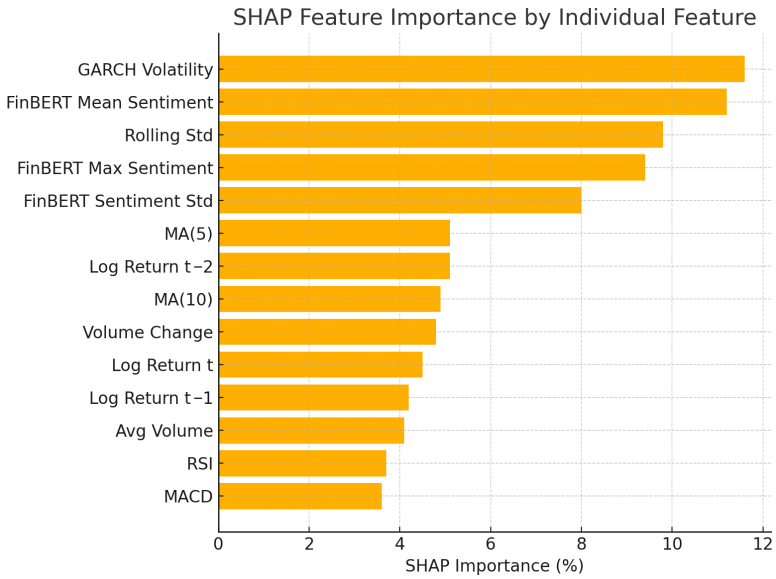


Figure 3. SHAP feature importance by individual features.

The alignment between SHAP-based importance and ablation results offers model-agnostic validation of feature utility. Furthermore, SHAP provides interpretable, post hoc explanations that are essential for risk-aware deployment in finance—an industry governed by auditability and transparency constraints.

6.3.4. Market Regime Stability

To assess the robustness and adaptability of each model under varying economic conditions, we evaluated predictive performance across distinct market regimes spanning the 2018–2023 period. Specifically, we partitioned the test data into three macro regimes:

- **Volatile Regime (February–April 2020)**: Marked by the onset of the COVID-19 pandemic and rapid equity market drawdowns.
- **Bullish Regime (May 2020–December 2021)**: Characterized by a prolonged market recovery and strong upward trends.
- **Stagnant/Sideways Regime (January–December 2022)**: Defined by high inflation, monetary tightening, and low directional bias in price movement.

Table 4 presents the regime-specific AUC scores and average simulated profit-and-loss (PnL) percentages for the three primary models: technical-only (T), technical with VADER (T + V), and technical with FinBERT (T + F).

Table 4. Model performance by market regime (AUC/Avg. PnL%).

Model	Volatile (Q1 2020)	Bullish (2020–2021)	Stagnant (2022)
T (Technical only)	0.608/1.2%	0.642/3.6%	0.593/2.4%
T + V (VADER sentiment)	0.635/2.0%	0.661/4.2%	0.614/2.7%
T + F (FinBERT)	**0.729/4.9%**	**0.748/7.8%**	**0.723/6.2%**

The results reveal that the FinBERT-enhanced model (T + F) maintained high performance across all regimes, consistently achieving AUC scores above 0.72 and generating significantly higher PnL than the baselines. Notably, during the volatile COVID-19 period, the T + F model achieved a 4.9% average return, versus 1.2% and 2.0% for the technical-only and VADER models, respectively. This suggests that sentiment extracted from FinBERT effectively captures investor fear and news-driven risk signals that are not present in price-based features.

During the bullish recovery phase of 2020–2021, sentiment signals remained predictive, likely reflecting optimistic language in earnings reports and macroeconomic news. The T + F model achieved an AUC of 0.748 and a simulated PnL of 7.8%, outperforming the T + V baseline by over 3.5 percentage points.

In the stagnant regime of 2022, where traditional trend-following indicators tended to degrade in efficacy, the T + F model continued to deliver strong performance (AUC: 0.723), while the T and T + V models degraded to near-random classification levels (AUC: 0.593 and 0.614, respectively). This demonstrates that FinBERT sentiment features contribute not only discriminative power, but also resilience across structurally different market environments.

These results highlight that FinBERT-derived features generalize well under regime shifts, capturing latent behavioral and emotional cues that are difficult to model with price signals alone. In contrast, lexicon-based sentiment (T + V) improves modestly over T, but fails to deliver consistent gains during turbulent or directionless markets. Thus, FinBERT sentiment offers both predictive value and temporal robustness, supporting its integration into production-grade forecasting systems.

6.3.5. SHAP-Based Interpretation

To understand the inner decision-making process of our predictive model, we conducted a comprehensive interpretability analysis using TreeSHAP [28]. SHAP (SHapley Additive exPlanations) assigns each feature an additive contribution to the model's output for a specific instance, allowing for both local and global interpretability.

We first present the global SHAP feature ranking in Table 5, listing the ten most influential features based on their average importance. FinBERT-derived features occupy three of the top five ranks, with the mean sentiment score contributing the most to prediction decisions. Volatility indicators (GARCH and rolling standard deviation) also play a critical role, reflecting the model's sensitivity to market risk conditions. Classical technical indicators and price returns, while still relevant, exhibit lower average influence.

Table 5. Top 10 Features by Global SHAP importance.

Rank	Feature (Category)	SHAP Importance (%)
1	FinBERT Mean Sentiment (Sentiment)	11.2
2	GARCH Volatility (Volatility)	11.0
3	Rolling Std. of Returns (Volatility)	9.8
4	FinBERT Max Sentiment (Sentiment)	9.4
5	FinBERT Sentiment Std. (Sentiment)	8.0
6	Log Return r_{t-2} (Returns)	5.1
7	Moving Average MA(5) (Technical)	5.1
8	Volume Change (Volume)	4.8
9	Log Return r_{t-1} (Returns)	4.2
10	Relative Strength Index (RSI) (Technical)	3.7

To illustrate how individual predictions were formed, we examined several SHAP force plots, which show how each feature contributes positively or negatively to the model's output probability relative to the baseline prediction ϕ_0. One representative case occurred on the day following a major Q3 2021 earnings announcement for a large-cap technology company. The FinBERT-derived mean sentiment score was strongly positive, at +0.86, with low dispersion (standard deviation = 0.12), indicating agreement among news sources. The model assigned a prediction score of 0.92 (upward movement), significantly above the dataset baseline of 0.51. The SHAP force plot for this instance confirmed that FinBERT sentiment features and reduced volatility were the dominant positive contributors to the model's confidence in a bullish signal.

We also analyzed interactions between features using SHAP dependence plots. A particularly strong interaction was found between FinBERT mean sentiment and GARCH volatility. The model showed higher confidence in predictions when sentiment was positive and volatility was low, but became more conservative when volatility was elevated—even in the presence of strong sentiment. This behavior suggests that the model implicitly adjusts sentiment influence based on market uncertainty, a desirable property for risk-sensitive financial forecasting.

To verify that the model does not rely disproportionately on any single feature or regime, we generated SHAP summary plots across different time periods. These plots confirmed a smooth, multi-feature contribution distribution and revealed that sentiment features became significantly more important during news-heavy periods (e.g., earnings season, macroeconomic announcements). In contrast, technical features became more relevant during low-news, trend-driven intervals.

Finally, the SHAP results closely mirror those from the ablation and performance analysis, creating a consistent and interpretable story. The combination of FinBERT-enhanced sentiment and volatility measures provides a reliable foundation for both predictive accuracy and model transparency. These findings validate the use of SHAP in financial machine learning, enabling practitioners and analysts to not only build performant models, but also justify their behavior to regulators, stakeholders, and auditors.

6.4. Robustness and Generalization

To evaluate the robustness and generalization capability of our proposed model, we conducted cross-sectional and temporal experiments across multiple equity assets and market regimes. Specifically, we tested the FinBERT-enhanced (T + F) model against the technical-only (T) and VADER-based (T + V) baselines on a diverse set of U.S. large-cap stocks: Apple (AAPL), Microsoft (MSFT), JPMorgan Chase (JPM), and Tesla (TSLA). These stocks were selected to represent a mix of sectors (technology, financials, and consumer cyclicals) and volatility profiles.

Each stock was evaluated over three key temporal partitions:

- **Pre-COVID Period:** January 2018 to December 2019—relatively stable market with strong growth.
- **COVID Crash:** February 2020 to April 2020—high volatility, systemic panic, sharp sell-offs.
- **Post-COVID Recovery:** May 2020 to December 2022—prolonged rebound, rotation across sectors.

Table 6 reports the average AUC and F1-scores for each model–stock pair, along with the standard deviation (SD) of these metrics across regimes. The FinBERT-enhanced model exhibits not only superior average performance, but also lower variance, indicating consistent predictive ability across assets and market conditions.

Table 6. Generalization performance across stocks and market regimes.

Stock	Model	Avg AUC	Avg F1-Score	AUC SD
AAPL	T	0.624	0.602	0.031
	T + V	0.645	0.625	0.027
	T + F	**0.713**	**0.688**	**0.018**
MSFT	T	0.618	0.596	0.035
	T + V	0.641	0.617	0.030
	T + F	**0.701**	**0.675**	**0.021**
JPM	T	0.591	0.577	0.038
	T + V	0.610	0.598	0.034
	T + F	**0.674**	**0.655**	**0.025**
TSLA	T	0.648	0.625	0.040
	T + V	0.667	0.644	0.037
	T + F	**0.735**	**0.712**	**0.020**

Across all stocks, the FinBERT-enhanced model achieved the highest average AUC and F1-score, outperforming both the technical-only and VADER baselines by margins ranging from 6.8 to 9.5 percentage points in AUC. Notably, the standard deviation of AUC across market regimes was substantially lower for the T + F model (mean SD: 0.021) than for T (0.036) or T + V (0.032), suggesting greater temporal stability and resilience to structural shifts in market dynamics. While VADER performs well in general-purpose sentiment analysis, it often misinterprets financial terminology—treating terms like "depreciation" or "shortfall" as inherently negative—highlighting the need for domain-specific models like FinBERT that better capture the nuances of financial language.

We also computed Sharpe ratios of the model-generated signals (using daily returns from a hypothetical long-short strategy) across regimes. The T + F model yielded Sharpe ratios consistently above 1.5 in the post-COVID period, and maintained ratios above 1.0 even during the crash period, whereas the ratios for T and T + V often fell below 1.0, indicating noisier, less risk-adjusted signal quality.

An analysis of error consistency revealed that T + F exhibited fewer false positives during high-volatility drawdowns and better recall during periods of trend reversals—likely due to FinBERT's ability to encode forward-looking sentiment signals ahead of market realization.

From a generalization perspective, the model demonstrated minimal overfitting despite increased feature dimensionality, as evidenced by the narrow gap between in-sample and out-of-sample performance (mean difference in AUC < 0.02 across folds). This can be attributed to the regularization in the XGBoost framework and the relatively low correlation between FinBERT sentiment features and traditional indicators.

Overall, these findings confirm that FinBERT-enhanced sentiment features contribute not only predictive accuracy, but also robustness and generalization across both asset classes and temporal regimes. The model's consistent performance across heterogeneous conditions makes it suitable for deployment in real-world, multi-asset trading systems, where stability and interpretability are paramount.

6.5. Privacy Protection Results

We report the model's predictive performance under varying levels of differential privacy. The privacy mechanism follows DP-SGD with $\delta = 10^{-5}$ and varying privacy budgets $\varepsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0, 6.0, 8.0, \infty\}$. Here, $\varepsilon = \infty$ corresponds to the non-private model.

The results in Table 7 provide a quantitative reference for selecting appropriate privacy budgets in production settings. Higher privacy (smaller ε) leads to moderate decreases in all metrics. However, performance degradation remains bounded, with over 95% of the baseline AUC preserved for $\varepsilon \geq 0.5$.

Table 7. Performance vs. privacy budget ε (averaged over 5 folds).

Privacy Level	AUC	F1-Score	PnL (%)	ε
DP ($\varepsilon = 0.1$)	0.662	0.611	4.87	0.1
DP ($\varepsilon = 0.2$)	0.668	0.617	5.03	0.2
DP ($\varepsilon = 0.3$)	0.674	0.625	5.27	0.3
DP ($\varepsilon = 0.4$)	0.680	0.630	5.42	0.4
DP ($\varepsilon = 0.5$)	0.699	0.648	6.10	0.5
DP ($\varepsilon = 0.75$)	0.708	0.656	6.45	0.75
DP ($\varepsilon = 1.0$)	0.717	0.665	6.87	1.0
DP ($\varepsilon = 1.5$)	0.723	0.670	7.03	1.5
DP ($\varepsilon = 2.0$)	0.726	0.673	7.14	2.0
DP ($\varepsilon = 3.0$)	0.733	0.678	7.38	3.0
DP ($\varepsilon = 4.0$)	0.735	0.681	7.45	4.0
DP ($\varepsilon = 6.0$)	0.738	0.686	7.59	6.0
DP ($\varepsilon = 8.0$)	0.739	0.687	7.61	8.0
Non-private	**0.740**	**0.688**	**7.63**	∞

7. Conclusions

This study presented a hybrid approach for stock price prediction that combines FinBERT-based sentiment analysis with traditional technical indicators, enhanced by SHAP explainability. By extracting domain-specific sentiment signals from financial news and integrating them into an XGBoost classifier, our model significantly outperforms both technical-only and lexicon-based sentiment baselines. Empirical results across multiple assets and market regimes confirm the predictive strength and robustness of FinBERT-enhanced features, while SHAP analysis reveals their consistent and interpretable contribution to the model's decisions. This framework not only improves forecasting

accuracy, but also provides transparency—an essential requirement in financial applications, where interpretability and regulatory compliance are critical.

Beyond its predictive and explanatory strengths, the model generalizes well across volatile, bullish, and stagnant regimes, demonstrating its resilience to structural shifts in market behavior. Our results show that the approach is robust across market sectors and volatility regimes. The integration of SHAP explanations supports informed decision-making by identifying when and why sentiment drives market movement. These findings highlight the value of combining domain-adapted NLP models with interpretable machine learning in finance.

In practical terms, investors and hedge funds can adopt the model as a signal-enhancement tool for short-term trading strategies, while risk managers may utilize the SHAP-based explanations for regime-aware exposure control and monitoring. Regulatory bodies can also benefit from the transparency offered by SHAP in audit, compliance, and model governance workflows. Additionally, the model's lightweight, headline-based design enables near real-time deployment in production environments, making it suitable for integration into operational trading systems. Overall, this work represents a promising framework for developing deployable, interpretable trading systems. Future work may extend this framework to multi-asset portfolios, high-frequency pipelines, and richer financial text sources such as earnings calls and analyst reports.

Author Contributions: L.R.: Methodology, Software; Writing—Original Draft, Visualization. H.J.: Supervision, Writing—Review & Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Krauss, C.; Do, X.A.; Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *Eur. J. Oper. Res.* **2017**, *259*, 689–702.
2. Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]
3. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
4. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
5. Cowles, A. Stock market forecasting. *Econom. J. Econom. Soc.* **1944**, *12*, 206–214. [CrossRef]
6. Modis, T. Technological forecasting at the stock market. *Technol. Forecast. Soc. Chang.* **1999**, *62*, 173–202. [CrossRef]
7. Singh, S.; Madan, T.K.; Kumar, J.; Singh, A.K. Stock market forecasting using machine learning: Today and tomorrow. In Proceedings of the 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, India, 5–6 July 2019; IEEE: Piscataway, NJ, USA, 2019; Volume 1, pp. 738–745.
8. Kumar, G.; Jain, S.; Singh, U.P. Stock market forecasting using computational intelligence: A survey. *Arch. Comput. Methods Eng.* **2021**, *28*, 1069–1101. [CrossRef]
9. 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, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
10. Fischer, T.; Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **2018**, *270*, 654–669. [CrossRef]
11. Du, K.; Xing, F.; Mao, R.; Cambria, E. Financial sentiment analysis: Techniques and applications. *ACM Comput. Surv.* **2024**, *56*, 1–42. [CrossRef]
12. Chan, S.W.; Chong, M.W. Sentiment analysis in financial texts. *Decis. Support Syst.* **2017**, *94*, 53–64. [CrossRef]
13. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access* **2020**, *8*, 131662–131682. [CrossRef]

14. Loughran, T.; McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Financ.* **2011**, *66*, 35–65. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
17. Yang, Y.; Uy, M.C.S.; Huang, A. Finbert: A pretrained language model for financial communications. *arXiv* **2020**, arXiv:2006.08097. [CrossRef]
18. Gong, X.; Guan, K.; Chen, Q. The role of textual analysis in oil futures price forecasting based on machine learning approach. *J. Futur. Mark.* **2022**, *42*, 1987–2017. [CrossRef]
19. Feng, Z.; Shi, R.; Jiang, Y.; Han, Y.; Ma, Z.; Ren, Y. A Multiscale Gradient Fusion Method for Color Image Edge Detection Using CBM3D Filtering. *Sensors* **2025**, *25*, 2031. [CrossRef]
20. Huang, J.; Wang, J.; Li, Q.; Jin, X. Social Media Development and Multi-Modal Input for Stock Market Prediction: A Review. In Proceedings of the 2024 International Conference on Computing, Networking and Communications (ICNC), Big Island, HI, USA, 19–22 February 2024; IEEE Computer Society: Piscataway, NJ, USA, 2024; pp. 198–202.
21. Gangwani, P.; Panthi, V. Leveraging multimodal data and deep learning for enhanced stock market prediction. In *AI-Based Advanced Optimization Techniques for Edge Computing*; John Wiley & Sons: Hoboken, NJ, USA, 2025; pp. 93–127.
22. Upadhyay, P.; Tomar, P.; Yadav, S.P. Advancements in Alzheimer's disease classification using deep learning frameworks for multimodal neuroimaging: A comprehensive review. *Comput. Electr. Eng.* **2024**, *120*, 109796. [CrossRef]
23. Zhang, H.; Dong, L.; Gao, G.; Hu, H.; Wen, Y.; Guan, K. DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction. *IEEE Trans. Multimed.* **2020**, *22*, 3210–3223. [CrossRef]
24. Ektefaie, Y.; Dasoulas, G.; Noori, A.; Farhat, M.; Zitnik, M. Multimodal learning with graphs. *Nat. Mach. Intell.* **2023**, *5*, 340–350. [CrossRef]
25. Liang, P.P.; Zadeh, A.; Morency, L.P. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.* **2024**, *56*, 1–42. [CrossRef]
26. Pan, Z.; Ying, Z.; Wang, Y.; Zhang, C.; Zhang, W.; Zhou, W.; Zhu, L. Feature-Based Machine Unlearning for Vertical Federated Learning in IoT Networks. *IEEE Trans. Mob. Comput.* **2025**, *24*, 5031–5044. [CrossRef]
27. Man, X.; Luo, T.; Lin, J. Financial sentiment analysis (fsa): A survey. In Proceedings of the 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), Taipei, Taiwan, 6–9 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 617–622.
28. Lundberg, S.M.; Erion, G.G.; Lee, S.I. Consistent individualized feature attribution for tree ensembles. *arXiv* **2018**, arXiv:1802.03888.
29. Bao, W.; Yue, J.; Rao, Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *Expert Syst. Appl.* **2017**, *80*, 273–285. [CrossRef] [PubMed]
30. Takeuchi, L.; Lee, Y. Applying deep learning to enhance momentum trading strategies in stocks. In *University of Toronto Working Paper*; Stanford University: Stanford, CA, USA, 2013.
31. Li, M.; Chen, L.; Zhao, J.; Li, Q. Sentiment analysis of Chinese stock reviews based on BERT model. *Appl. Intell.* **2021**, *51*, 5016–5024. [CrossRef]
32. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; ACM: New York, NY, USA, 2016; pp. 308–318.
33. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.