

S&P 500 forecasting with financial news sentiment using FinBERT-CNN-LSTM

Zakaria Al Bakkari¹, Ikram El Azami², Adil El Makrani²

¹Laboratory of Research in Informatics, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

²Laboratory of Research in Informatics, Department of Informatics, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Oct 30, 2025

Revised Aug 30, 2026

Accepted Sep 4, 2026

Keywords:

Convolutional neural network
Financial bidirectional encoder
representations from
transformers

Long short-term memory

S&P 500

Sentiment analysis

Stock market forecasting

ABSTRACT

Accurate stock market forecasting remains challenging due to the nonlinear, noisy, and volatile nature of financial markets. Beyond historical prices, market movements are influenced by financial news and investor sentiment. This study proposes a hybrid financial bidirectional encoder representations from transformers with convolutional neural network and long short-term memory (FinBERT-CNN-LSTM) model for forecasting the Standard & Poor's 500 (S&P 500) index by integrating financial news sentiment with historical market data. FinBERT generates daily aggregated sentiment scores, which are combined with market features and processed through a CNN-LSTM architecture. The CNN captures local patterns and short-term interactions, while the LSTM models temporal dependencies in financial time series. Experimental results show that the proposed model outperforms benchmark models, achieving a root mean square error (RMSE) of 0.0119, mean absolute error (MAE) of 0.0091, mean absolute percentage error (MAPE) of 2.10%, and directional accuracy (DA) of 63.4%. It also achieves a cumulative return of 15.8% and a Sharpe ratio of 1.35. Rolling-window validation and Diebold–Mariano (DM) testing further support the statistical significance of the improvements. These findings demonstrate the effectiveness of combining financial sentiment and deep learning (DL) for stock market forecasting.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Zakaria Al Bakkari

Laboratory of Research in Informatics FS, Faculty of Science, Ibn Tofail University

Kenitra, Morocco

Email: zakaria.albakkari@uit.ac.ma

1. INTRODUCTION

Forecasting stock market movements has received considerable attention in both research and practice because of its importance for investment decision-making, portfolio allocation, and risk management [1]. However, predicting market behavior remains difficult because financial markets are nonlinear, volatile, and structurally dynamic [2]-[6]. Many traditional statistical and machine learning (ML) approaches rely mainly on historical price data and do not fully incorporate exogenous factors such as financial news and investor sentiment, which also influence market movements [7]. Recent advances in natural language processing (NLP), particularly transformer-based models such as financial bidirectional encoder representations from transformers (FinBERT), have improved the extraction of sentiment information from financial text and created new opportunities for sentiment-aware forecasting [8], [9]. At the same time, recurrent neural networks (RNNs), including long short-term memory (LSTM) models, are effective for capturing temporal dependencies in financial time series [10], [11], while convolutional neural networks (CNNs) are well-suited for learning local feature patterns and short-term interactions [12], [13].

Existing sentiment-based studies have demonstrated the value of combining financial news, text mining, and investor sentiment with predictive models [14]. However, many multimodal approaches still rely on simple fusion mechanisms that do not adequately represent the interactions between sentiment information and market variables [7], [15]. To address this limitation, this study proposes a hybrid FinBERT-CNN-LSTM model for forecasting the daily closing price of the Standard & Poor's 500 (S&P 500) by integrating historical market data with financial news sentiment. Within the proposed framework, FinBERT extracts domain-specific sentiment information, the CNN captures local and short-term interactions between price and sentiment features, and the LSTM models their temporal dependencies. To achieve this research objective, the following tasks were identified:

- a. Collect and preprocess historical S&P 500 market data together with financial news data.
- b. Extract sentiment information from financial news using finbert.
- c. Design a hybrid forecasting architecture that combines CNN and LSTM components in order to capture both local and sequential patterns.
- d. Evaluate the proposed model against benchmark approaches, namely autoregressive integrated moving average (ARIMA), multilayer perceptron (MLP), LSTM, and FinBERT-LSTM, using both statistical error measures and financially relevant performance indicators.

This study makes four principal contributions. First, it proposes a hybrid FinBERT-CNN-LSTM architecture for multimodal stock market forecasting. Second, it incorporates financial news sentiment into the predictive framework through FinBERT-based sentiment extraction [16], [17]. Third, it places a CNN layer before the LSTM component to strengthen local feature learning and capture short-term interactions between sentiment and price-related information. Fourth, it provides an empirical comparison of the proposed model with ARIMA, MLP, LSTM, and FinBERT-LSTM using both conventional forecasting metrics and financial performance indicators. The remainder of the paper is organized as follows. Section 2 reviews the relevant literature and identifies the research gap. Section 3 describes the data sources, preprocessing procedures, baseline models, and proposed architecture. Section 4 presents and discusses the experimental results and financial evaluation. Finally, section 5 concludes the paper, discusses its limitations, and outlines directions for future research.

2. LITERATURE REVIEW

Stock market forecasting has been widely investigated due to its central importance for investment decisions and risk management. Existing studies broadly classify forecasting methods into statistical, ML, and deep learning (DL) approaches. More recently, sentiment-based and hybrid frameworks that combine textual information with financial time-series data have gained increasing attention, as they offer greater potential to improve predictive performance and capture the multifaceted nature of financial markets.

2.1. Statistical techniques

Early research on stock market forecasting relied heavily on statistical time-series models to extract patterns from historical data. ARIMA has long been used as a benchmark for modeling linear temporal dependencies, although its performance weakens in the presence of nonlinearities, abrupt market shifts, and structural breaks [3]. In parallel, generalized autoregressive conditional heteroskedasticity-based (GARCH-based) models were developed to capture volatility clustering and time-varying variance, with asymmetric specifications proving especially useful in turbulent conditions, and GARCH further improved volatility forecasts by integrating realized and high-frequency information [4], [5]. Despite their relevance, these models remain constrained by assumptions of linearity, stationarity, and distributional stability. As financial markets are inherently heterogeneous and regime-dependent, their limitations have driven a shift toward more flexible and data-driven approaches, including regime-switching volatility models [6].

2.2. Machine learning techniques

To address the limitations of traditional statistical forecasting, ML methods have been increasingly adopted in financial prediction due to their ability to capture nonlinear relationships and manage high-dimensional data. Among these, support vector machines (SVMs) have demonstrated strong effectiveness in modeling complex financial patterns, with Mahmoodi *et al.* [18] reporting improved forecasting accuracy by integrating SVMs with metaheuristic optimization. Beyond SVMs, ensemble and hybrid ML approaches have also gained prominence for their robustness to noise and overfitting. In this vein, Park *et al.* [19] demonstrated that combining random forest with sequential modeling in an LSTM forest framework can enhance stock market prediction performance. Nevertheless, conventional ML approaches remain highly dependent on feature engineering and often struggle to capture temporal dependencies in an automated manner, a limitation that has progressively driven the shift toward DL methods.

2.3. Deep learning techniques

DL approaches have gained prominence in stock market forecasting due to their ability to capture nonlinear patterns, learn abstract representations automatically, and model long-term temporal dependencies. In particular, RNNs and LSTM networks have proven especially effective for sequential financial data, owing to their ability to retain information across time steps. Empirical evidence supports their predictive relevance: Berradi and Lazaar [10] showed that integrating principal component analysis with RNNs improves forecasting performance on the Casablanca stock exchange, while Fischer and Krauss [11] demonstrated that LSTM-based models outperform conventional ML approaches. Nevertheless, many DL models remain primarily focused on numerical market variables and fail to adequately account for exogenous factors, such as financial news and sentiment, which significantly influence investor expectations and price dynamics. This limitation has contributed to the growing adoption of sentiment-aware forecasting frameworks that incorporate textual information into predictive models.

2.4. Sentiment-based and hybrid models

Recent advances in NLP have greatly enhanced the use of financial text in stock market forecasting, with FinBERT emerging as a key model for extracting sentiment information from financial data [9]. Prior studies have shown that sentiment-aware representations can improve stock movement prediction when integrated into forecasting models. In parallel, hybrid DL architectures combining CNN and LSTM have gained importance because they can capture both local feature patterns and long-term temporal dependencies. In addition, multimodal approaches that combine textual sentiment with historical time-series data have been shown to enhance forecasting performance [20]. However, many existing studies still rely on simple fusion strategies and do not fully capture the interaction between sentiment and temporal market dynamics, which highlights the need for more advanced hybrid architectures [15].

2.5. Research gap and motivation

The literature reveals a clear evolution from statistical and ML methods toward DL and multimodal sentiment-aware architectures. While statistical and conventional ML models provide useful foundations, they remain limited in capturing the nonlinear, sequential, and information-rich nature of financial markets. Although LSTM-based approaches improve temporal modeling, many existing studies still fail to fully exploit financial text as a complementary source of predictive information. In this context, a key research gap is the development of unified hybrid frameworks capable of jointly extracting domain-specific sentiment, capturing local feature interactions, and modeling temporal market dependencies. To address this limitation, this study proposes a FinBERT-CNN-LSTM architecture that integrates financial news sentiment with historical market data to forecast the S&P 500.

3. METHOD

This study proposes a hybrid FinBERT-CNN-LSTM framework for forecasting the daily closing price of the S&P 500 index by combining financial sentiment extraction, local feature learning, and temporal sequence modeling within a unified architecture. Specifically, the model integrates FinBERT, CNN, and LSTM to jointly exploit textual and numerical information, aiming to improve predictive performance in stock market forecasting.

3.1. Data preprocessing

In this project, two main data sources are collected: market data and news sentiment data.

- Market data: historical S&P 500 price data were retrieved from the Yahoo Finance API for the period January 1, 2017, to December 31, 2024, ensuring coverage across multiple market regimes. To enhance model stability and maintain feature comparability, the series was normalized using Min–Max scaling fitted exclusively on the training set and subsequently applied to the validation and test sets to prevent data leakage.
- News data: financial news and headlines were collected from the New York Times API over the same period as the market data to ensure temporal alignment. Relevant articles were selected using finance-related keywords, including “stock market”, “financial news”, and “economic forecast”. The news corpus contained an average of 6.4 articles per day, ranging from 0 to 18, and daily sentiment scores were computed by averaging FinBERT-based sentiment scores across all articles published on the same day. Days without relevant news were assigned a neutral score of zero, while articles published after market close were linked to the following trading session to prevent look-ahead bias. Each article was processed using FinBERT to obtain sentiment probabilities for the positive, negative,

and neutral classes. Rather than using the class label directly, we converted the output into a scalar sentiment score defined as (1):

$$s_i = P(\text{positive} | x_i) - P(\text{negative} | x_i), s_i \in [-1, 1] \quad (1)$$

where s_i denotes the sentiment score of article i . Daily sentiment was then obtained by averaging the scores of all relevant articles published between the market close on day t and the next trading session, thereby avoiding look-ahead bias.

Final dataset: after preprocessing, the dataset consists of two synchronized daily time series:

- The normalized S&P 500 closing price P_t .
- The aggregated sentiment score S_t .

A sliding-window procedure with look-back length $k=10$ was then applied to construct supervised samples of the form:

$$X_t = [(P_t - 9, S_t - 9), \dots, (P_t, S_t)] \quad (2)$$

where P_t is the normalized closing price on day t , and S_t is the aggregated sentiment score on day t .

With the target P_{t+1} , the final dataset contains 2012 aligned trading observations, providing a substantially larger sample. Figure 1 summarizes the preprocessing pipeline.

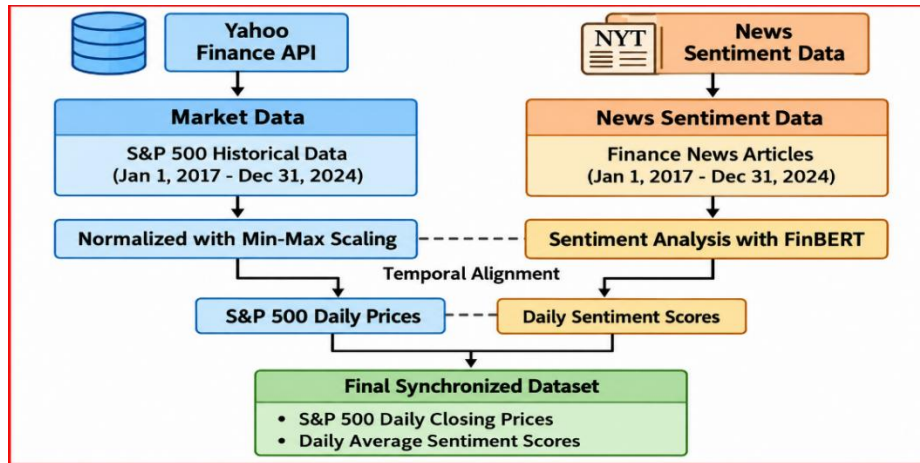


Figure 1. Data preprocessing

Figure 1 illustrates the construction of the multimodal dataset by aligning S&P 500 price data with FinBERT-derived sentiment scores on a daily basis, with missing sentiment values set to 0. The resulting synchronized time series was then used as input to the predictive model. To ensure a rigorous evaluation, the dataset was split chronologically into training, validation, and test subsets, as summarized in Table 1.

Table 1. Dataset split

Dataset	Percentage (%)	Period
Training	63	2017-2021
Validation	25	2022-2023
Test	12	2024

This chronological split prevents information leakage and enables evaluation under a realistic out-of-sample forecasting setting. Descriptive statistics of the dataset are reported in Table 2. Table 2 indicates substantial variation in S&P 500 prices over the study period, reflecting shifting market conditions, while FinBERT-based sentiment scores remain centered around zero, suggesting overall balanced news sentiment. The dataset also shows an average of approximately six financial articles per day. To further examine the link between market behavior and investor sentiment, Figure 2 provides a time-series visualization of closing prices and daily sentiment scores, offering an intuitive view of their joint evolution.

Table 2. Descriptive statistics of the dataset

Variable	Mean	Std. Dev	Min	Max
S&P 500 close	3675.2	976.5	2237.4	6090.3
Sentiment score	0.04	0.30	-0.91	0.86
Articles per day	6.4	3.1	0	18

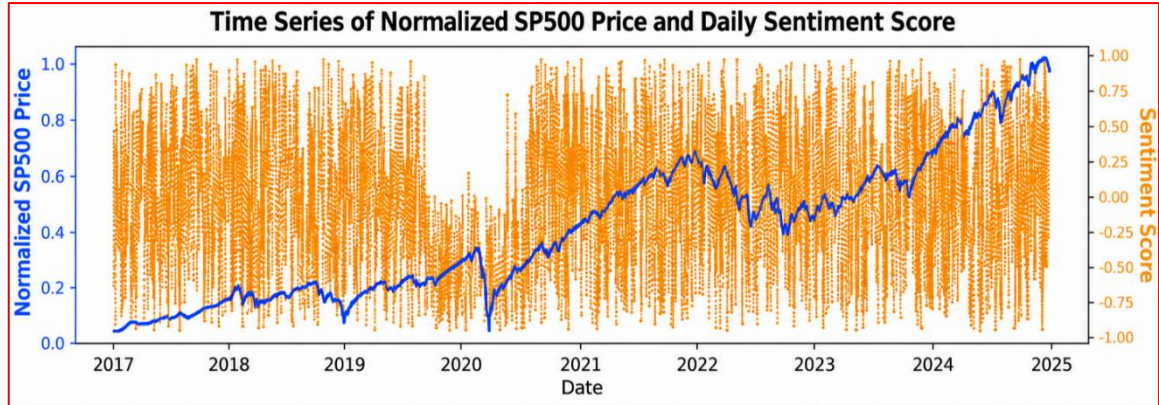


Figure 2. Time-series visualization of S&P 500 prices and sentiment scores

The dual-axis plot highlights the temporal co-evolution of market performance and investor sentiment, suggesting a potential association between price movements and sentiment fluctuations over the study period. It notably reflects the sharp decline in the S&P 500 during the first half of 2020 following the coronavirus disease 2019 (COVID-19) shock, followed by a market rebound as restrictions were gradually lifted, despite persistently negative sentiment.

3.2. Problem formulation

Based on the preprocessed dataset, the following section presents the predictive models used in this study to address the forecasting problem:

$$\hat{P}_{t+1} = f_{\theta}(X_t) \quad (3)$$

With:

- X_t is the input sequence defined in (2).
- θ represents the learnable model parameters.
- $\hat{P}_{t+1}$ is the predicted closing price for day $t+1$.

Each article i is processed using FinBERT to obtain a sentiment score: s_i . The sentiment aggregation formula is (4):

$$S_t = \frac{1}{N_t} \sum_{i=1}^{N_t} s_{i,t} \quad (4)$$

With:

- N_t is the number of relevant financial news articles published on day t .
- $s_{i,t}$ is the FinBERT sentiment score assigned to article i on day t .

In the absence of relevant news, $S_t=0$.

3.3. Baseline models and proposed architecture

To evaluate the effectiveness of multimodal learning for stock-market forecasting, the study implemented a statistical baseline (ARIMA) and four neural architectures: MLP, LSTM, FinBERT-LSTM, and the proposed FinBERT-CNN-LSTM. These models progressively incorporate nonlinear relationships, temporal dependencies, and sentiment information.

3.3.1. Autoregressive integrated moving average model

The ARIMA model serves as a statistical benchmark for S&P 500 forecasting because it captures linear dependencies in historical price series through autoregressive, differencing, and moving-average

components [3]. The optimal specification, selected using the Akaike information criterion, is ARIMA (3,1,2), which generates forecasts based solely on past price behavior. The ARIMA framework used for S&P 500 forecasting involves data preprocessing, stationarity assessment, differencing, and parameter selection. The final ARIMA model captures linear temporal dependencies through autoregressive and moving-average terms. More generally, rigorous comparative evaluation between statistical and machine-learning forecasting methods remains essential [21]. Despite its strong theoretical foundation, ARIMA is limited in capturing nonlinear patterns and interactions compared to DL models.

3.3.2. Multilayer perceptron model

The MLP model serves as a baseline feed-forward neural network for regression. The input is a vector of the past 10 days' normalized S&P 500 closing prices. The architecture includes three fully connected hidden layers with 50, 30, and 20 neurons, respectively, using rectified linear unit (ReLU) activation. Dropout regularization (0.1, 0.05, and 0.01) is applied to reduce overfitting. The output layer consists of a single neuron with linear activation to predict the closing price. The model is trained using the Adam optimizer with a learning rate of 0.001 and mean squared error (MSE) loss. The final dense layer produces the predicted closing price, capturing nonlinear relationships in financial data. To overcome the lack of temporal modeling in MLP, the LSTM model is introduced to capture sequential dependencies in financial data.

3.3.3. Long short-term memory model

The LSTM model is designed to learn temporal dependencies from sequential financial data using input sequences of shape (10, 1), corresponding to past price observations. Its architecture includes three stacked LSTM layers with 50, 30, and 20 units, combined with dropout regularization and a final dense layer for prediction. Trained with the Adam optimizer and MSE loss over 100 epochs, the model effectively captures sequential price patterns. However, because it relies only on quantitative historical prices, it does not account for qualitative information such as market sentiment. To address this limitation, the FinBERT-LSTM model extends the approach by integrating sentiment extracted from financial news, allowing for a more comprehensive representation of market dynamics.

3.3.4. Financial bidirectional encoder representations from transformers-LSTM model

The FinBERT-LSTM model combines sentiment analysis and time-series forecasting by processing a multimodal sequence of length 10. At each time step, the sequence includes the normalized closing price and the aggregated sentiment score. These features are fed into three stacked LSTM layers (with 70, 30, and 10 units, using tanh activation), capturing both temporal dependencies and sentiment-driven patterns. The final dense layer produces the predicted next-day S&P 500 index. The model is trained with the Adam optimizer at a learning rate of 0.0005, ensuring stable convergence and preventing overfitting, especially when combining textual and numerical inputs. To enhance the model's ability to capture short-term features, the FinBERT-CNN-LSTM architecture introduces a convolutional layer. This layer extracts local patterns and complex interactions between price and sentiment data, followed by temporal modeling with LSTM for more accurate forecasting.

3.3.5. Financial bidirectional encoder representations from transformers-CNN-LSTM model

The proposed FinBERT-CNN-LSTM model enhances the baseline multimodal architecture by integrating a 1D convolutional layer for local feature extraction from synchronized S&P 500 prices and FinBERT-derived sentiment scores. Each input sample is represented as a 10-day sequence of shape (10, 2), corresponding to the normalized closing price and aggregated daily sentiment score. The CNN layer uses a kernel size of 3, while max-pooling reduces dimensionality and preserves salient patterns. The extracted features are then processed by three stacked LSTM layers with 50, 30, and 20 units, using tanh activation and dropout rates of 0.10, 0.05, and 0.01. Hyperparameter tuning explored filter sizes of 32, 64, and 128, with 64 filters selected in the final architecture. Training was performed with the Adam optimizer, a learning rate of 0.0005, a batch size of 64, and 100 epochs, with the learning rate chosen through grid search. Reproducibility was ensured through fixed random seeds for Python, NumPy, and TensorFlow. The model was implemented in Python 3.13 using Keras, Hugging Face Transformers, pandas, NumPy, statsmodels, and evaluated in a controlled Google Colab environment with an NVIDIA Quadro P600 GPU with 4 GB of video memory and 8 GB of RAM. As shown in Figure 3, this hierarchical design improves feature representation by combining local multimodal interactions with longer-term temporal modeling.

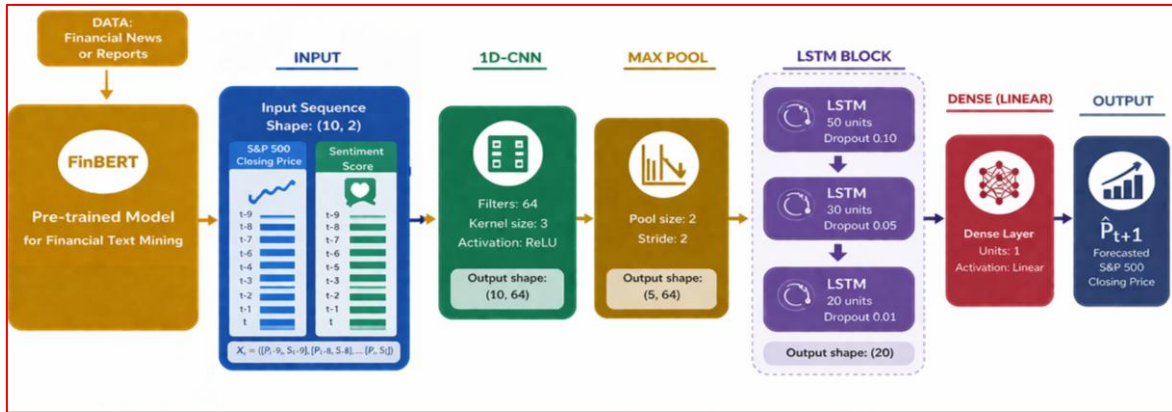


Figure 3. Proposed FinBERT-CNN-LSTM architecture

Combined price and sentiment features are first processed by a CNN layer to extract local patterns, followed by max-pooling for dimensionality reduction. Stacked LSTM layers capture temporal dependencies, and a dense layer outputs the predicted S&P 500 value. This hybrid architecture leverages both feature extraction and sequential modeling for improved forecasting performance.

The inclusion of the CNN component strengthens the model's ability to capture local and short-term interactions in price and sentiment data. The following section further examines these gains through a comparative analysis of the experimental results.

4. RESULTS AND DISCUSSION

4.1. Evaluation metrics and financial assessment

4.1.1. Error metrics

We selected three performance metrics for model evaluation: mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). The corresponding formulas are as (5)-(7):

$$RMSE = \sqrt{\sum_{t=1}^n \frac{(p_t - \hat{p}_t)^2}{n}} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{t=1}^n |p_t - \hat{p}_t| \quad (6)$$

$$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{|p_t - \hat{p}_t|}{|p_t|} \quad (7)$$

where p_t is the actual value, $\hat{p}_t$ is the predicted value, and n is the number of observations.

MAE, MAPE, and RMSE are used to assess forecasting error, with lower values indicating better predictive performance. MAPE expresses forecasting accuracy as a percentage. Beyond these conventional error-based measures, the evaluation framework also incorporates directional accuracy (DA) and financial performance indicators to better assess the practical relevance of the models in real-world trading contexts.

4.1.2. Directional accuracy

DA measures the model's ability to correctly predict the direction of price movement:

$$DA = \frac{100}{n} \sum_{t=1}^n I[(\hat{p}_t - p_{t-1}) \cdot (p_t - p_{t-1}) > 0] \quad (8)$$

where $I(\cdot)$: an indicator functions equal to 1 if the predicted and actual directions match, and 0 otherwise. This metric is particularly important for trading decisions and predicting the direction of the market.

4.1.3. Trading strategy evaluation

To assess the economic value of the predictions, a simple long-short trading strategy is implemented:

- Buy (long) if the predicted return is positive.

- Sell (short) if the predicted return is negative.
- The cumulative return R:

$$R = \prod_{t=1}^n (1 + z_t r_t) - 1 \quad (9)$$

where $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ is the actual return at time t and $z_t \in \{-1, +1\}$ is the trading signal derived from the model prediction, where $z_t = +1$ indicates a long position and $z_t = -1$ indicates a short position, and n is the total number of trading periods.

4.1.4. Risk-adjusted performance

The Sharpe ratio evaluates the risk-adjusted return of the trading strategy:

$$Sharpe = \frac{E(R_p - R_f)}{\sigma_p} \quad (10)$$

where R_p is the trading strategy returns, R_f is the risk-free rate, and σ_p is the standard deviation of strategy returns. A higher Sharpe ratio indicates better performance relative to risk.

4.1.5. Statistical significance (Diebold–Mariano (DM) test)

To statistically compare forecasting models, the DM test is applied. This test evaluates whether the difference in prediction errors between two models is statistically significant:

- H_0 : both models have equal predictive accuracy.
- H_1 : one model outperforms the other.

Let $e_{1,t}$ and $e_{2,t}$ denote the forecast errors of model 1 and model 2 at time t, respectively. In this study, the loss $L(\cdot)$ differential is based on the squared error (11):

$$d_t = L(e_{1,t}) - L(e_{2,t}) = e_{1,t}^2 - e_{2,t}^2 \quad (11)$$

where d_t represents the difference in loss between the two models at time t. If the mean d_t is equal to zero, the two models are considered to have the same predictive accuracy. If the mean differs significantly from zero, then one model is statistically more accurate than the other.

The DM statistic is (12):

$$DM = \frac{\bar{d}}{\sqrt{\text{var}(\bar{d})}} \quad (12)$$

where $\bar{d}$ is the mean loss differential and $\text{var}(\bar{d})$ is the estimated variance of that mean differential.

This approach provides a statistical validation of performance improvements. The LSTM model was chosen as the main reference model because it provides a strong baseline for sequential financial forecasting.

4.2. Model evaluation

This section presents a performance comparison of the proposed models, including MLP, LSTM, FinBERT-LSTM, and the proposed FinBERT-CNN-LSTM. The evaluation uses both statistical error metrics and financial performance indicators to assess predictive accuracy and practical relevance. The performance of all models is summarized in Table 3.

Table 3. Comprehensive performance comparison of the evaluated models

Model	RMSE	MAE	MAPE (%)	DA (%)	Return (%)	Sharpe	DM test
ARIMA	0.0180	0.0140	3.45	53.2	5.1	0.60	0.089
MLP	0.0175	0.0142	3.12	54.3	6.5	0.72	0.061
LSTM	0.0152	0.0118	2.64	58.7	9.8	0.95	—
FinBERT-LSTM	0.0134	0.0103	2.28	60.9	12.6	1.12	0.028
FinBERT-CNN-LSTM	0.0119	0.0091	2.10	63.4	15.8	1.35	0.012

Table 3 shows a clear and consistent improvement in performance as forecasting approaches progress from statistical and price-based models to hybrid multimodal architectures incorporating sentiment information. The proposed FinBERT-CNN-LSTM model outperforms all benchmarks across both statistical and economic metrics, achieving the lowest forecast errors and the strongest directional and trading

performance. To statistically compare the predictive accuracy of the competing forecasting models, the DM test was applied. The proposed FinBERT-CNN-LSTM model was compared against the LSTM reference model using the same out-of-sample test period. The null hypothesis states that the two models have equal predictive accuracy, whereas the alternative hypothesis states that one model is more accurate than the other. The DM test further confirms that this improvement is statistically significant ($p\text{-value}=0.012$), indicating superior forecasting accuracy relative to the reference model. To assess robustness while preserving temporal order, Out-of-sample performance was further evaluated using a rolling-window cross-validation with five evaluation windows, RMSE, and MAE were used to summarize out-of-sample performance across folds.

Table 4 presents the rolling-window cross-validation results over five temporal folds. The proposed FinBERT-CNN-LSTM model achieved the lowest mean RMSE (0.01166) and mean MAE (0.00906), with low standard deviations across folds, indicating both strong predictive accuracy and stable temporal generalization. The FinBERT-LSTM model also performed well, but with slightly higher variability than the proposed hybrid architecture. The results also reveal a gradual improvement from ARIMA to MLP, from MLP to LSTM, and from LSTM to sentiment-enhanced architectures. While ARIMA captures linear temporal dynamics, its performance remains limited in the presence of nonlinear behavior. MLP improves nonlinear approximation but lacks temporal modeling. LSTM further improves prediction quality by modeling sequence dependence and FinBERT-LSTM demonstrates that incorporating financial sentiment provides additional predictive value. The proposed FinBERT-CNN-LSTM model achieves the strongest overall results, suggesting that local feature extraction and sentiment integration jointly enhance the representation of financial dynamics.

Table 4. Rolling-window cross-validation results

Model	Mean RMSE	RMSE Std Dev	Mean MAE	MAE Std Dev
ARIMA	0.01848	0.00041	0.01468	0.00048
MLP	0.01752	0.00035	0.01436	0.00038
LSTM	0.01500	0.00032	0.01172	0.00041
FinBERT-LSTM	0.01336	0.00052	0.01056	0.00044
FinBERT-CNN-LSTM	0.01166	0.00038	0.00906	0.00039

4.2.1. Results analysis

The proposed model's superior performance can be attributed to the complementary roles of its components: the CNN captures local and short-term patterns, the LSTM models long-term temporal dependencies, and FinBERT-derived sentiment features provide contextual information about investor behavior. These results suggest that the improvement is associated not only with the inclusion of sentiment information, but also with the model's ability to learn short-range interactions before temporal sequence modeling. This finding is consistent with recent evidence showing that hybrid and multimodal deep-learning architectures are effective in modeling complex relationships across heterogeneous financial data sources [22]. However, these results should be interpreted cautiously. The complexity of the hybrid architecture relative to the dataset size may increase the risk of overfitting, and the observed gains may partly depend on the specific market conditions covered during the study period. In addition, the reliance on a single news source may limit the generalizability of the sentiment signal. Therefore, further validation on other periods, data sources, and market settings is needed.

4.2.2. Visual and economic evaluation

Figure 4 provides a visual comparison between actual and predicted S&P 500 values, showing that the proposed model achieves the closest fit to real market behavior. While ARIMA captures the overall trend, DL models better represent nonlinear dynamics, with the FinBERT-CNN-LSTM framework delivering the highest predictive accuracy.

Figure 4 clearly demonstrates the predictive superiority of the proposed model, as the green curve closely tracks the actual price series shown in blue, including during the COVID-19 shock in the first half of 2020 and the subsequent market recovery as restrictions were progressively lifted. This strong alignment confirms the model's robustness and its capacity to deliver reliable forecasts even under conditions of pronounced economic turbulence. The visual comparison reveals a progressive improvement from statistical and price-only models to sentiment-enhanced architectures. ARIMA captures the broad evolution of the index but shows greater difficulty in representing nonlinear variations and rapid changes in market direction. The MLP improves nonlinear approximation but does not explicitly model temporal dependence. The LSTM provides a closer fit by learning sequential price patterns, while the FinBERT-LSTM further reduces the deviations by incorporating information extracted from financial news. The proposed FinBERT-CNN-LSTM

produces the closest overall alignment because the convolutional layer processes short-term local interactions between sentiment and price features before these representations are modeled temporally by the LSTM layers. Figure 5 illustrates its economic relevance through cumulative trading returns generated from model-based signals. The trading simulation results are illustrated in Figure 5.

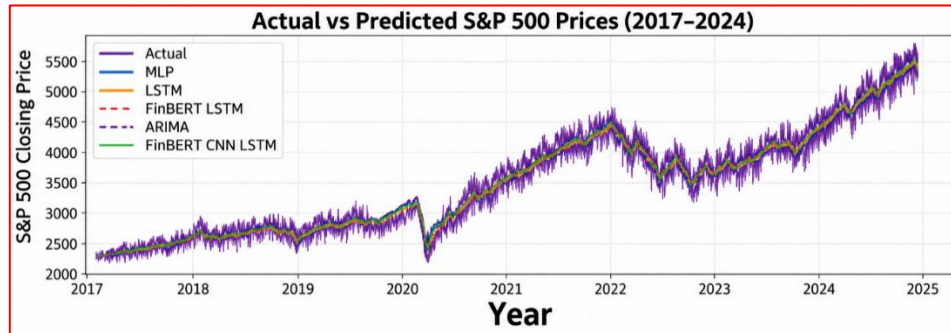


Figure 4. Comparison of actual and predicted S&P 500 using statistical and DL models



Figure 5. Cumulative returns of trading strategies based on model predictions

Figure 5 shows that the proposed hybrid model achieves the strongest trading performance under the simplified trading assumptions adopted in this study, with the highest cumulative return (15.8%) and Sharpe ratio (1.35), indicating superior risk-adjusted performance. Its DA (63.4%) further confirms its practicality in financial decision-making contexts where correctly predicting price direction is more important than estimating the exact price level. These findings suggest that the advantage of the proposed framework is not only statistical but also economically meaningful under the simplified trading strategy adopted in this study.

From a practical perspective, the model shows potential for integration into algorithmic trading systems, portfolio optimization, and real-time financial decision support tools by combining sentiment-aware predictions with live market data. However, real-world deployment would still require addressing important challenges, including latency in news processing, noise in sentiment signals, and transaction costs. Systematic reviews of DL for algorithmic trading also identify data noise, overfitting, and limited interpretability as important practical challenges [23]. Therefore, although the results are promising, further validation in realistic trading environments remains necessary before large-scale implementation.

4.2.3. Predictive gains and computational limitations

The addition of the CNN layer reduced RMSE by approximately 11.2% and MAE by approximately 11.7% relative to FinBERT-LSTM, while DA increased by 2.5 percentage points. However, this improvement is accompanied by additional computational complexity due to convolutional feature extraction. The computational trade-off cannot be fully quantified in the present study. Future work should systematically evaluate whether the predictive gains justify the additional computational requirements. This assessment is particularly important for real-time or intraday deployment, where inference latency, memory consumption, and processing speed may affect the operational usefulness of the model.

4.2.4. Limitations and future work

These findings suggest that integrating sentiment analysis with hybrid DL models can significantly enhance portfolio performance and risk management strategies. Despite the promising results, several limitations should be acknowledged:

- The dataset size remains relatively limited and comes from a single source (NYT), which may limit the model's generalization ability.
- Sentiment extraction relies solely on financial news, without incorporating social media data.
- Transaction costs and market frictions are not considered in the trading simulation.
- The predictive gains justify the additional computational requirements.

As a result, the reported cumulative return and Sharpe ratio should be considered indicative rather than definitive measures of real-world profitability. These findings suggest that the model captures useful directional information, but additional evaluation under realistic trading conditions is required before drawing strong conclusions about its real-world applicability. Future work should therefore explore larger datasets, test NASDAQ, the Dow Jones Industrial Average, cryptocurrencies, and commodities, incorporate alternative sources such as Bloomberg and additional multimodal inputs (e.g., social media, macroeconomic indicators, and technical signals), evaluate multiple look-back windows, report ablation and sensitivity analyses, investigate attention mechanisms, explainable artificial intelligence (AI) techniques, and evaluation under realistic real-time trading constraints. In particular, future extensions should examine multi-source sentiment data to capture broader investor perceptions [24] and multi-view heterogeneous-data fusion to integrate complementary textual, market, and contextual information [25].

5. CONCLUSION

This study proposes a novel FinBERT-CNN-LSTM architecture for forecasting the daily S&P 500 index by integrating financial news sentiment with historical market data. The results show that the hybrid model outperforms conventional benchmarks across both statistical and trading-oriented metrics, with an RMSE of 0.0119, MAE of 0.0091, MAPE of 2.10%, and DA of 63.4%, highlighting the value of combining domain-specific sentiment extraction, local feature learning, and temporal modeling within a unified framework. The findings further suggest that the model's effectiveness stems not only from the inclusion of sentiment information, but also from the way interactions between sentiment and price features are learned through the architecture. Overall, the study contributes to multimodal financial forecasting by underscoring the importance of architectural design in improving predictive performance. However, the results should be interpreted in light of certain limitations, including the use of a single news source, the relatively restricted dataset, and the simplified trading simulation. Future research may extend this framework by incorporating additional multimodal data sources, conducting ablation and sensitivity analyses, improving interpretability, and testing the model under real-time trading conditions.

FUNDING INFORMATION

This research was supported by the Scientific Research Support Center at Ibn Tofail University through a research grant.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Zakaria Al Bakkari	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	
Ikram El Azami		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Adil El Makrani		✓		✓	✓	✓				✓	✓		✓	✓

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**editing

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors confirm that they have no known financial, personal, or non-financial competing interests that could have influenced the work reported in this manuscript.

DATA AVAILABILITY

The historical S&P 500 market data used in this study were retrieved through the Yahoo Finance API. Financial news articles and headlines were collected through the official New York Times Developer API using an API key obtained after registration on the New York Times Developer portal.

REFERENCES

- [1] O. Bustos, A. Pomares-Quimbaya, and R. Stellan, "Machine learning, stock market forecasting, and market efficiency: A comparative study," *International Journal of Data Science and Analytics*, vol. 20, no. 7, pp. 6815–6839, Nov. 2025, doi: 10.1007/s41060-025-00854-4.
- [2] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied Soft Computing*, vol. 90, p. 106181, May 2020, doi: 10.1016/j.asoc.2020.106181.
- [3] D. Kobiela, D. Krefta, W. Król, and P. Weichbroth, "ARIMA vs LSTM on NASDAQ stock exchange data," *Procedia Computer Science*, vol. 207, pp. 3836–3845, 2022, doi: 10.1016/j.procs.2022.09.445.
- [4] J. Caiado and F. Lúcio, "Stock market forecasting accuracy of asymmetric GARCH models during the COVID-19 pandemic," *The North American Journal of Economics and Finance*, vol. 68, pp. 1–14, Sep. 2023, doi: 10.1016/j.najef.2023.101971.
- [5] Z. Fang and J.-Y. Han, "Realized GARCH model in volatility forecasting and option pricing," *Computational Economics*, vol. 66, no. 5, pp. 3637–3657, Nov. 2025, doi: 10.1007/s10614-024-10826-8.
- [6] H. Mushtaq, M. Ishtiaq, S. Jamal, S. M. R. Rizvi, and H. Raza, "Regime-switching volatility and risk quantification in South Asian and developed stock markets: A comparative perspective using Markov-switching GARCH with MLE and MCMC estimations," *The North American Journal of Economics and Finance*, vol. 82, pp. 1–19, Jan. 2026, doi: 10.1016/j.najef.2025.102576.
- [7] P. Chen, Z. Boukouvalas, and R. Corizzo, "A deep fusion model for stock market prediction with news headlines and time series data," *Neural Computing and Applications*, vol. 36, no. 34, pp. 21229–21271, Dec. 2024, doi: 10.1007/s00521-024-10303-1.
- [8] Y. Sun, H. Yuan, and F. Xu, "Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features," *Natural Language Processing Journal*, vol. 11, pp. 1–12, Jun. 2025, doi: 10.1016/j.nlp.2025.100148.
- [9] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A pre-trained financial language representation model for financial text mining," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, Jul. 2020, pp. 4513–4519, doi: 10.24963/ijcai.2020/622.
- [10] Z. Berradi and M. Lazaar, "Integration of principal component analysis and recurrent neural network to forecast the stock price of Casablanca Stock Exchange," *Procedia Computer Science*, vol. 148, pp. 55–61, 2019, doi: 10.1016/j.procs.2019.01.008.
- [11] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018, doi: 10.1016/j.ejor.2017.11.054.
- [12] J. Gu, S. Shukla, J. Ye, A. Uddin, and G. Wang, "Deep learning model with sentiment score and weekend effect in stock price prediction," *SN Business & Economics*, vol. 3, no. 7, p. 119, 2023, doi: 10.1007/s43546-023-00497-2.
- [13] Y. Dexiang, M. Shengdong, Y. Liu, G. Jijian, and L. Chaolong, "An improved deep-learning-based financial market forecasting model in the digital economy," *Mathematics*, vol. 11, no. 6, pp. 1–18, 2023, doi: 10.3390/math11061466.
- [14] P. Xiao, "Stock market prediction based on financial news, text data mining, and investor sentiment analysis," *International Journal of Information System Modeling and Design*, vol. 15, no. 1, pp. 1–13, Jul. 2024, doi: 10.4018/IJISMD.361593.
- [15] S. Giantsidi and C. Tarantola, "Deep learning for financial forecasting: A review of recent trends," *International Review of Economics & Finance*, vol. 104, pp. 1–51, Dec. 2025, doi: 10.1016/j.iref.2025.104719.
- [16] J. Kim, H.-S. Kim, and S.-Y. Choi, "Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: A FinBERT transformer model and LSTM," *Axioms*, vol. 12, no. 9, pp. 1–22, 2023, doi: 10.3390/axioms12090835.
- [17] T. Jiang and Q. Zeng, "Financial sentiment analysis using FinBERT with application in predicting stock movement," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2306.02136.
- [18] A. Mahmoodi, L. Hashemi, M. Jasemi, S. Mehraban, J. Laliberté, and R. C. Millar, "A developed stock price forecasting model using support vector machine combined with metaheuristic algorithms," *OPSEARCH*, vol. 60, no. 1, pp. 59–86, Mar. 2023, doi: 10.1007/s12597-022-00608-x.
- [19] H. J. Park, Y. Kim, and H. Y. Kim, "Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework," *Applied Soft Computing*, vol. 114, p. 108106, Jan. 2022, doi: 10.1016/j.asoc.2021.108106.
- [20] H. Wang, Z. Xie, D. K. W. Chiu, and K. K. W. Ho, "Multimodal market information fusion for stock price trend prediction in the pharmaceutical sector," *Applied Intelligence*, vol. 55, pp. 1–27, 2025, doi: 10.1007/s10489-024-05894-0.
- [21] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, pp. 1–26, Mar. 2018, doi: 10.1371/journal.pone.0194889.
- [22] F. Moodi, A. J. Rafsanjani, S. Zarifzadeh, and M. A. Z. Chahooki, "Fusion of technical indicators and sentiment analysis in a hybrid framework of deep learning models for stock price movement prediction," *IEEE Access*, vol. 12, pp. 195696–195709, 2024, doi: 10.1109/ACCESS.2024.3519556.
- [23] M. S. M. Bhuiyan *et al.*, "Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies," *Array*, vol. 26, pp. 1–27, Jul. 2025, doi: 10.1016/j.array.2025.100390.
- [24] Y. Gao, C. Zhao, and Y. Wang, "Investor sentiment and stock returns: New evidence from Chinese carbon-neutral stock markets based on multi-source data," *International Review of Economics & Finance*, vol. 92, pp. 438–450, Apr. 2024, doi: 10.1016/j.iref.2024.02.049.
- [25] W. Long, J. Gao, K. Bai, and Z. Lu, "A hybrid model for stock price prediction based on multi-view heterogeneous data," *Financial Innovation*, vol. 10, no. 1, pp. 1–50, Feb. 2024, doi: 10.1186/s40854-023-00519-w.

BIOGRAPHIES OF AUTHORS

Zakaria Al Bakkari    received his Engineer Diploma from the National Institute of Statistics and Applied Economics (INSEA), Rabat, Morocco. In 2021, he joined the doctoral study center at Ibn Tofail University, Kenitra, Morocco. He is a member of the Laboratory of Research in Computer Science (L@RI). He is currently a Ph.D. researcher focusing on deep learning methods for stock market forecasting. He can be contacted at email: zakaria.albakkari@uit.ac.ma.



Ikram El Azami    currently works at the Department of Informatics, Université Ibn Tofail. He does research in databases, machine learning, data mining, and distributed computing. Their most recent publication is "AraTrans the new transformer model to generate Arabic text". He can be contacted at email: ikram.elazami@uit.ac.ma.



Adil El Makrani    is a Research Professor in Computer Science at the Faculty of Science, Ibn Tofail University, Kenitra. He received a Master's degree in Computer Science, Computer Graphics, and Imaging, and Ph.D. in Computer Science, Sidi Med Ben Abdellah University in 2009 and 2015, respectively. He is affiliated with the Research in Informatics Laboratory (L@RI). His research currently focuses on artificial intelligence technologies, big data analytics, and their applications. He can be contacted at email: adil.elmakrani@uit.ac.ma.