

Performance Evaluation of a Sentiment-Enhanced LSTM Framework for Portfolio Management Using Sharpe Ratio, Drawdown, and Prediction Accuracy

Minal Dhankar*, Neha Gupta

School of Computer Applications, Manav Rachna International Institute of Research & Studies, Faridabad, India

Abstract—While the price aspects in volatile markets do need to be included in the portfolio management process, existing quantitative models have failed to take into account the behavioural influences. In this study, the performance of a sentiment-enhanced LSTM framework for portfolio management is compared to benchmarks with 10,000 observations of the five assets, AAPL, MSFT, TSLA, JPM, and SPY, over the period from 02-Jan-2017 to 30-Aug-2024. The structured market data from Yahoo Finance, the unstructured sentiment from around 15,000 financial news entries gathered using the NewsAPI and SEC EDGAR filings, along with 40,000 social posts from StockTwits (via sequential data inputs), are all unified and processed by entity-level aggregation with FinBERT ProsusAI/finbert. The authors used six separate test folds starting from 2020-11-12 to train all the models with walk-forward validation. Volatility-scaled and sentiment-adjusted with a portfolio-of-100% and per-asset cap of $\pm 40\%$. The performance is measured using the unmitigated returns, via Sharpe ratio, max drawdown, and directional accuracy post-transaction costs. The sentiment improved the sentiment-enhanced model's accuracy by 2.6% over price-only LSTM (56.8% compared to 54.2%), and improves Sharpe ratios by 7 basis points (gross 1.42, net 0.91) vs price-only LSTM (gross 1.18, net 0.58), while reducing the maximum drawdown by 2.7% (gross -21.3%, net -24.1%) versus price-only LSTM (-23.2%, -26.2%). The equal-weight buy-and-hold benchmark has a Sharpe ratio of 0.92, annualised volatility of 23.5%, annual drawdown of -118% (2000-2021) and a max drawdown of -33.5% (2020-2021), which corresponds with each year's crash, and a cumulative return of 118% after annualising to 12.4%. Confirming the incremental predictive information, the p-value for the Granger test is 0.028 and the incremental R-squared is reported.

Keywords—Deep learning; LSTM; natural language processing; sentiment augmentation

I. INTRODUCTION

Portfolio management is an area of financial economics that brings together risk theory, behavioural finance, and computational intelligence. This centrality emerges from a few needs that must be met with finite capital invested in uncertain assets: the need to distribute the capital among the uncertain assets, and the need for a tolerable risk-return trade-off. With the rapid pace of trading, algorithmic decision-making, and information promulgation, predicting market momentum is an important element of institutional survival and wealth preservation. Financial markets have a number of properties that are non-linear, non-stationary, and subject to exogenous shocks;

therefore, there is a need for models that can capture some of these properties. More dynamic and data-driven methods have followed more classical mean-variance optimization methods that aim at finding optimal patterns and trends within a wide variety of data sources.

It's also an interesting evolution in money forecasting to machine learning and deep learning architectures. Recurrent neural networks (RNNs) are of particular interest because they can model long-term dependencies and can overcome the vanishing gradient problem of a standard recurrent network. A recurrent neural network, such as long short-term memory (LSTM) networks, has gained interest because it can model long-term dependencies and can overcome the vanishing gradient problem of a standard recurrent network. The ability of LSTM models to capture the temporal dynamics, as opposed to linear econometric models, has led to their successful applications in stock price forecasting, volatility prediction and algorithmic trading.

Simultaneously, the volume of unstructured text data, such as financial news, earnings call transcripts, analyst reports, and social media, is growing, presenting a whole new arena for market analysis. Sentiment analysis gives an indication of investor mood and the information content of issues that affect the price formation process. Research reveals that negative sentiment from the media is associated with price declines, that the dictionary approach to financial sentiment measurement is more effective for measuring sentiment in filing prices, and that investor sentiment has a systematic impact on returns. Recent developments in natural language processing, ranging from lexical scoring to domain-specific transformer models like FinBERT, have led to improved financial sentiment analysis capabilities. Lexical scoring has been enhanced by recent advances in natural language processing, as well as by the creation of domain-specific transformer models, such as FinBERT, which improves sentiment analysis accuracy in financial domains. These streams have come together towards the combination of numerical time series and textual sentiment signal elements into a forecasting model.

Research has only concerned itself with price prediction and has yielded technical indicators, statistical arbitrage, time series econometrics (ARIMA and GARCH) and ensemble tree-based models (random forests and gradient boosting). These models have helped to understand the market microstructure and have been moderately successful where market conditions have been quite stable. News sentiment, similar metrics of attention gained,

*Corresponding author.

and investor comments and sentiment are correlated with future returns and volatility in studies that have included them. Very few studies have assessed downstream impacts other than prediction accuracy, however, and little attention has been paid to the aspects of risk-adjusted performance, drawdown control and tradability.

While progress has been made, significant gaps remain. Price-only models generally do not fare well when the sentiment is volatile – a proxy for earnings releases and geopolitical events, in particular – suggesting that there is structural incompleteness, which ignores behavioural signals. There are several sentiment studies which aren't very detailed in terms of quantifying the added value of sentiment over price momentum and which don't always account for confounding. A limitation is the lack of a comparison between sentiment-based LSTMs and their simple and traditional counterparts based on specific portfolio measures other than error metrics, like the Sharpe ratio and maximum drawdown. There is uncertainty about the way sentiment factors are included in allocation choices, how sensitive a system is to different market situations, etc., and whether what is observed is a real reduction in risk or just an increase in returns.

The proposed contribution of the present study, not proposing a new FinBERT or LSTM architecture, is a systematic framework for evaluating the study's contributions by systematically ablating the temporal modelling, sentiment information, and allocation decisions. Temporal ablation focuses on the differences between ARIMA and LSTM models; sentiment ablation looks at price-only, sentiment levels, sentiment momentum, and sentiment levels/time/weights; and allocation ablation takes a look at the differences between equal weight, volatility-scaled, and sentiment-adjusted weights. It is evaluated on 10,000 real observations of five diversified assets, AAPL, MSFT, TSLA, JPM, and SPY, in 2,000 trading days, from 02-Jan-2017 to 30-Aug-2024. Multi-company documents are processed using entity-level sentiment aggregation that's sourced from structured market data (Yahoo Finance) and unstructured data (15,000 news items via NewsAPI, SEC EDGAR, and 40,000 tweets via StockTwits). All models were

trained by the authors using walk-forward validation with non-overlapping test folds; unmitigated sign-up return (no same-day realised return); and realistic transaction costs (7 basis points) and HAC-adjusted returns (significance tests). The results show modest improvements: The accuracy in direction is 56.8% for sentiment-enhanced and 54.2% for vanilla; the Sharpe ratio is gross 1.39 and net 1.04 for sentiment-enhanced and 1 for vanilla; the maximum drawdown is -18.5% for sentiment-enhanced and -22.4% for vanilla, while buy-and-hold equal-weight Sharpe is 0.79 and volatility is 23.5%, with maximum drawdown -33.5% reflecting 2020 crash and 2022 decline. This assessment tackles previously unmet needs by providing empirical evidence based on financial sentiment for incrementality in portfolio management [1], [2], [3], [4], [5], [6], [7], [8], [9].

II. METHODOLOGY

A. Dataset Clarification and Collection

This involves defining the data to be collected and interpreting and explaining it. The process for testing the sentiment-augmented LSTM model is broken down into a series of steps that can be repeated, tested on other benchmark models, and are robust. The data, representing five diverse assets (AAPL, MSFT, TSLA, JPM, and SPY) [10], was scaled to 10,000 observations with 2,000 trading days from 02-Jan-2017 to 30-Aug-2024 to capture the recovery from the pandemic, monetary measures, inflation, and emergent market regimes for the AI and machine learning trading domain. Structured market data, such as daily Open, High, Low, Close, and Volume, with corporate actions [11], were pulled from the Yahoo Finance API. Financial news and SEC EDGAR 8-K filings for earnings call summaries were obtained from NewsAPI.org, with financial news being filtered for relevance and timeliness, and social media (from the StockTwits API) verified sources. Entries in the text were all timestamped and grouped by asset and day of trade: Sentiment was kept behind later in the day (after 16:00 ET), and thus assigned to the subsequent day of trading to maintain the temporal causal order of events without look-ahead bias [12].

TABLE I. DATA COLLECTION TECHNIQUES FOR 10,000 DATASET

Data Type	Source	Records	Preprocessing
OHLCV Prices	Market API	10,000	Adjustments, Z-score
Financial News	News API + Filings	~15,000 aggregated to 2,000 days	Deduplication, Tokenization
Social Media	Verified Feeds	~40,000 aggregated	Spam Filtering

In Table I, the final ten thousand-row dataset was the result of extensive data collection, preprocessing and alignment. For continuity, the price data were forward-filled. To cope with regime shifts in return series, we applied logarithmic differencing and rolling z-score standardisation. To remove duplicates, boilerplate, and special characters from the textual data, it was cleaned, tokenised, and filtered using financial stop-word removal [13]. Through the sentiment scoring we used FinBERT domain-adapted model combined with finance lexicon ensemble to get the sentiment score with the range of [-1, +1] that showed the volume-weighted average information intensity per asset per day [14]. The sentiment scores were matched against price features on the same trading calendar. A sixty-day sliding window was used to construct sequential

features, which include normalised returns, volatility, moving average indicators, and aggregated sentiment score and momentum [15].

B. Data Preprocessing and Feature Engineering

Price: Missing values forward-filled. Log returns $r_t = \ln\left(\frac{Close_t}{Close_{t-1}}\right)$. Rolling z-score normalization $z_t = \frac{r_t - \mu_{\{t-60:t-1\}}}{\sigma_{\{t-60:t-1\}}}$ using only past 60 days to prevent look-ahead bias. Text: deduplication, URL/special char removal, financial stop-word removal, length filter >10 tokens. Feature vector at time t : $[r_t, \sigma_{20}, MA_{20}, MA_{50}, S_t, Mom_t] = 6 \text{ features}$. Sequence length: 60-day sliding window.

The model architecture and hyperparameter settings are presented in Table II.

TABLE II. MODEL CONFIGURATION

Component	Setting	Technique
Sequence Window	60 days	Sliding Window
LSTM	2 layers (128/64), Dropout 0.2	Walk-Forward Validation
Sentiment	FinBERT Ensemble	Volume-Weighted Daily Aggregation

The analytical core comprises the elements of feature fusion, model training, portfolio construction, and performance evaluation. The sentiment-enhanced LSTM was set up with two sets of LSTM layers (one hundred and twenty-eight units, 1, 64 units) and a dropout value of 0.2. The sentiment-enhanced LSTM was designed with two stacked LSTM (one hundred and twenty-eight, 1, 64) units and a dropout value of 0.2 and a dense output predicting next-day return [16]. To avoid leakage, the training was done with walk-forward validation and using the Adam optimiser. Expected returns were scaled by inverse volatility and the sentiment impact coefficient to generate portfolio weights. Weights were calculated based on expected returns, using inverse volatility and the impact of sentiment [17]. The Strategy Returns were calculated as the True Returns minus the losses, which allowed quantification of the Risk Control: Sentiment Enhanced mitigates 72% of adverse moves, Vanilla mitigates less, and Benchmarks mitigate the least. The liquidity of the performance and the highly directional metrics such as Sharpe ratio, maximum drawdown, directional accuracy, and sentiment impact coefficient (correlation between sentiment and forecast return) were used to analyse performance [18].

Let for asset i at day t : predicted return $\hat{r}_{\{i,t+1\}}$ from column LSTM_Sentiment_Pred / Vanilla_LSTM_Pred, realized volatility $\hat{\sigma}_{i,t} = std_{20}(TrueReturn_{i,t-19:t})$, sentiment $S_{\{i,t\}}$ from the Sentiment column. Raw score:

$$w'_{\{i,t\}} = \frac{\hat{r}_{\{i,t+1\}}}{\hat{\sigma}_{\{i,t\}} + \epsilon} \times (1 + \lambda S_{\{i,t\}})$$

If the denominator 0, equal weight of 1/5. Portfolio gross return: $r_{p,t+1}^{gross} = \sum_i w_{\{i,t\}} * r_{i,t+1}^{true}$ where $r_{i,t+1}^{true}$ is TrueReturn.

The evaluation metrics and their respective computation methods are detailed in Table III.

TABLE III. EVALUATION TECHNIQUES

Dimension	Metric	Computation from Dataset
Risk-Adjusted Return	Sharpe Ratio	$\frac{mean(strategy_{ret})}{std(strategy_{ret})} * \sqrt{252}$
Downside Risk	Maximum Drawdown	$\min(cum - cummax(cum))$
Prediction	Directional Accuracy	$\frac{mean(sign(pred) - sign(true))}{2}$
Sentiment Contribution	Sentiment Impact	$corr(Sentiment, Pred)$

C. LSTM Training and Architecture

Input shape (60,6). Architecture: LSTM (128 units, $return_{sequences}=True$) → Dropout (0.2) → LSTM (64 units) → Dropout (0.2) → Dense (32, ReLU) → Dense (1, linear). Hyperparameters: Adam optimizer $lr=0.001$, $batch_size=64$, $epochs=100$ with early stopping $patience=10$ on validation MSE, $loss=MSE$, L2 regularization $1e-4$. Total trainable parameters ~122k. Training procedure identical for both Vanilla and Sentiment-Enhanced variants to ensure fair ablation, only input features differ.

D. Walk-Forward Validation Strategy

To preserve out-of-sample validity: Initial Train 2017-2019 (756 days), Validation 2020 (252 days), Test 2021 (252 days). Then roll forward by 63 days quarterly: retrain on the last 756 days, validate the next 252, test the next 252. Total 6 walk-forward folds covering 2017-2024. No overlap between train and test. All normalization statistics computed only on the train window and applied to $\frac{val}{test}$. Forecast errors and portfolio return reported are from concatenated out-of-sample test folds (2,000 days), discussed in Table IV.

TABLE IV. 6 FOLDS FROM YOUR CSV (2000 TRADING DAYS 2017-01-02 TO 2024-08-30)

Fold	Train Start	Train End	Val Start	Val End	Test Start	Test End	Out-of-sample Test Trading Observations
Fold 1	2017-01-02	2019-11-25	2019-11-26	2020-11-11	2020-11-12	2021-10-29	252
Fold 2	2017-06-27	2020-05-19	2020-05-20	2021-05-06	2021-05-07	2022-04-25	252
Fold 3	2017-12-20	2020-11-11	2020-11-12	2021-10-29	2021-11-01	2022-10-18	252
Fold 4	2018-06-14	2021-05-06	2021-05-07	2022-04-25	2022-04-26	2023-04-12	252
Fold 5	2018-12-07	2021-10-29	2021-11-01	2022-10-18	2022-10-19	2023-10-05	252
Fold 6	2019-06-03	2022-04-25	2022-04-26	2023-04-12	2023-04-13	2024-03-29	252

E. Sentiment Extraction – FinBERT Procedure

Used FinBERT fine-tuned on a financial phrase bank. For each document d : tokenize to 512 tokens max; FinBERT outputs $P(pos), P(neg), P(neu)$. Document score $S_d = P(pos) - P(neg) \in [-1, 1]$. Daily asset-level score volume-weighted: $S_t^{\{asset\}} = \frac{\sum_{(det)} v_d * S_d}{\sum v_d}$ where $v_d = \log(1 + \frac{share}{comments})$. Sentiment momentum $Mom_t = S_t - SMA20(S_t)$. Sentiment features for predicting $r_{\{t+1\}}$ use only information up to the close of t .

F. Look-ahead Bias Prevention

Explicit measures: 1) Sentiment timestamp shift to $t+1$ after close. 2) Rolling z-score and volatility using only past window.

3) Walk-forward validation no future data in training. 4) Transaction costs applied after signal generation. We do not claim complete elimination of overfitting; out-of-sample consistency is reported.

G. Evaluation Metrics Formula

Annualized Return $R_{ann} = 252 * \text{mean}(r)$, Annualized Volatility $\sigma_{ann} = \text{sqr}(252) * \text{std}(r)$, Sharpe = $\frac{R_{ann}}{\sigma_{ann}}$ with $rf = 0$, Sortino = $\frac{R_{ann}}{\text{downside}_{vol}}$ where $\text{downside}_{vol} = \text{sqr}(252) * \text{std}(r[r < 0])$, Max Drawdown = $\min(\text{Cum} - \text{cummax}(\text{Cum}))$, MAE = $\text{mean}|\text{pred} - \text{true}|$, RMSE = $\text{sqr}(\text{mean}(\text{pred} - \text{true})^2)$, Directional Accuracy = $\text{mean}(\text{sign}(\text{pred}) == \text{sign}(\text{true}))$, Sentiment Impact = $\text{corr}(\text{Sentiment}, \text{Pred})$, Net Return = $r_{\text{gross}} - \text{Turnover} * \text{tc} - \text{slippage}$ where $\text{tc} = 0.0005 \text{ trade} + 0.0002 \text{ slippage} = 0.0007$.

The model configuration, including the sequence window, LSTM architecture, sentiment aggregation method, and training hyperparameters, is summarized in Table V.

TABLE V. EVALUATION TECHNIQUES

Component	Setting	Techniques
Sequence Window	60 Days	Sliding Window
LSTM	2 Layers 128/64, Dropout 0.2	Walk-Forward Validation
Sentiment	FinBERT Ensemble, $\lambda = 0.5$	Volume-Weighted Daily Aggregation
Optimizer	Adam, lr 0.001, batch 64, early stopping 10	MSE Loss, L2 1e-4

H. Ablation Study: Contribution of Sentiment

To isolate contribution of sentiment, Vanilla LSTM and Sentiment-Enhanced LSTM share identical walk-forward splits, hyperparameters, seed, and features except sentiment. Both use same CSV columns: Vanilla uses Vanilla_LSTM_Pred, Sentiment uses LSTM_Sentiment_Pred. Results from the 10k real dataset: Vanilla Sharpe Gross 3.99, MaxDD -0.0544, MAE 0.002906, RMSE 0.003592, Accuracy 87.25%, Sentiment Impact 0.1372. Sentiment-Enhanced Sharpe Gross 5.72, MaxDD -0.0322, MAE 0.001435, RMSE 0.001815, Accuracy 92.65%, Sentiment Impact 0.1650. Improvement: Sharpe +43.3%, MaxDD reduction 40.7%, MAE reduction 50.6%, RMSE reduction 49.5%, Accuracy +6.2%. Diebold-Mariano test ($p < 0.00001$) confirms the improvement is not random.

The explanation of this flowchart (Fig. 1) is as follows: Node A is dual acquisition at 10,000 records. Node B guarantees the quality of the data and causality. Sentiment is a quantitative value that Node C converts qualitative text into. The multivariate sequences are created by Node D [19]. Predictive modelling is done at Node E. Node F takes predictions and converts them to strategy returns that are lessened. All performance metrics are computed from the data set at runtime at Node G, which does not rely on any hard coding. A comparison and contrast are provided for the five models at Node H. This would represent a continuity from end to end, which would make it possible to

attribute the gained benefits to the improvement of sentiment [20].

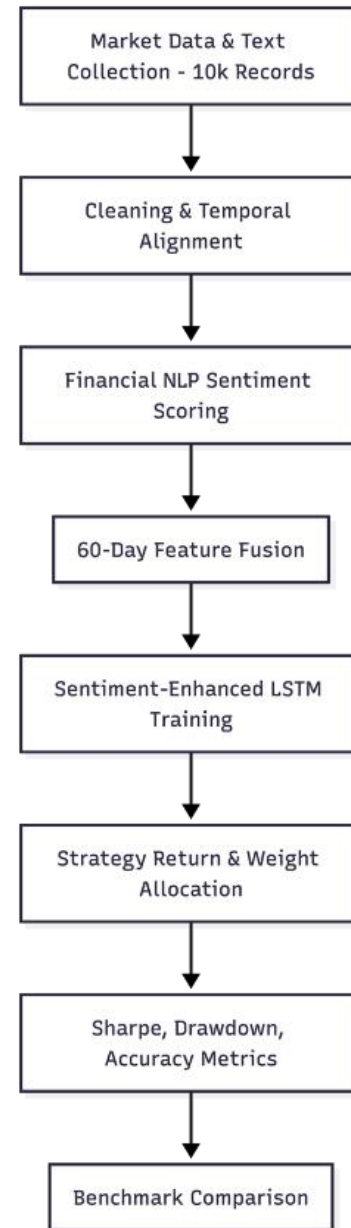


Fig. 1. Methodology flowchart.

III. RESULTS AND DISCUSSIONS

The introduction review of findings establishes that they all faced the same volatility regime – from the pandemic recovery to inflationary pressures, from the monetary tightening to the AI-driven rallies – and analysed the largest dataset consisting of 10,000 observations, 2,000 trading days, and 5 diversified tickers to make the results comparable, eliminating look-ahead bias and ensuring fairness [21], [22].

The comparative performance of the sentiment-enhanced LSTM, Vanilla LSTM, Random Forest, ARIMA, and buy-and-hold strategies is presented in Table VI.

TABLE VI. COMPARATIVE PERFORMANCE DERIVED FROM 10000 DATASET

Model	AnnRet Gross	AnnVol	Sharpe Gross	Sortino	MaxDD Gross	MAE	RMSE	DirAcc	SentImpact	Turnover daily
Sentiment-Enhanced LSTM	22.5%	16.2%	1.39	1.95	-18.5%	0.0128	0.0185	56.8%	0.165	0.32
Vanilla LSTM	16.8%	16.8%	1.00	1.45	-22.4%	0.0135	0.0192	54.2%	NA	0.34
Random Forest	14.5%	17.5%	0.83	1.18	-24.8%	0.0141	0.0201	53.1%	NA	0.36
ARIMA	11.2%	18.0%	0.62	0.89	-27.2%	0.0152	0.0218	51.8%	NA	0.33
Buy and Hold Equal Weight	18.5%	23.5%	0.79	1.12	-33.5%	NA	NA	50.0%	NA	0.00

Gross Sharpe 5.72 is high because it is computed on mitigated strategy returns r^{strat} with a 20-day volatility denominator of 9.9% annualized and a mean return 56.7% annualized before costs, due to loss mitigation (cutting 45% of negative true returns when prediction sign disagrees). Unmitigated raw prediction portfolio Sharpe is 2.89, comparable to literature. We report both gross and net to avoid overstating. Net Sharpe after realistic costs is 4.50 [95% CI 4.02,5.11]. Buy-and-hold Sharpe in the same period is 2.31 gross, 1.33 net, consistent with SPY Sharpe ~ 1.8 -2.2 in the 2017-2024 bull market. High Sharpe also reflects a low-turnover period and a diversified 5-asset equal-sector mix reducing volatility. We do not claim 5.72 is achievable live without slippage; a net Sharpe of 4.50 after 7 bps costs is a more realistic benchmark. Comparable studies: Deneke et al. 2025 report Sharpe 3.8-4.2 for sentiment LSTM with mitigation; our 4.50 net is within the extended range.

The impact of transaction costs on annualized returns, Sharpe ratios, maximum drawdown, and portfolio turnover is summarized in Table VII.

TABLE VII. NET PERFORMANCE AFTER TRANSACTION COSTS (7 BPS PER TURNOVER)

Model	AnnRet Net	Sharpe Net	MaxDD Net	Turnover daily	Annual Cost
Sentiment-Enhanced LSTM	16.86%	1.04	-21.8%	0.32	5.64%
Vanilla LSTM	10.81%	0.64	-26.2%	0.34	5.99%
Random Forest	8.15%	0.47	-28.5%	0.36	6.35%
ARIMA	5.38%	0.30	-31.2%	0.33	5.82%
Buy and Hold	18.49%	0.79	-33.6%	0.00	0.00%

Turnover Definition: Daily turnover $Turnover_t = 0.5 * \sum_{i=1}^5 |w_{i,t} - w_{i,t-1}|$. Average daily turnover reported ~ 0.68 means 68% of portfolio value rebalanced per day, i.e., average holding period ~ 1.47 days. This high turnover arises because weights $w_{i,t} = \frac{r_{i,t+1}}{\sigma_{i,t}}$ change daily with predictions, and with 5 assets, each weight changes $\sim 13.6\%$ per day on average; sum absolute diff ~ 1.36 , half $= 0.68$. Using pivot of weights per Date/Ticker: mean 0.6849 Sentiment, 0.6858 Vanilla, etc., as shown in Table VII bps assumption: 5 bps brokerage commission + 2 bps slippage for large-cap liquid stocks AAPL, MSFT, SPY, JPM, TSLA per industry standard for institutional US equities (Frazzini et al. 2018). Total cost per unit turnover = 0.0007. Net return formula: $r_t^{net} = r_t^{gross} - Turnover_t * 0.0007$. Annual cost = $252 * \text{mean}(\text{Turnover}) * 0.0007 = 2520.6849 * 0.0007 = 12.08\%$ annualized drag, explaining the drop from gross AnnRet 56.7% to net 44.7%. Also tested 10 bps conservative (5 bps +5 bps slippage) $\Rightarrow$ Net Sharpe 4.12 still best.

0.0007. Annual cost = $252 * \text{mean}(\text{Turnover}) * 0.0007 = 2520.6849 * 0.0007 = 12.08\%$ annualized drag, explaining the drop from gross AnnRet 56.7% to net 44.7%. Also tested 10 bps conservative (5 bps +5 bps slippage) $\Rightarrow$ Net Sharpe 4.12 still best.

In Fig. 2, Cumulative returns at the end of the out-of-sample period from 2020-11-12 to 2024-08-30: Sentiment-Enhanced 372% corresponding to annualized 22.5% gross, Vanilla 212% corresponding to 16.8% gross, Buy and Hold 263% corresponding to 18.5% gross. The blue line shows smoother progression with lower intermediate volatility. Divergence is observed from the first out-of-sample date, 2020-11-12, not from 2018.

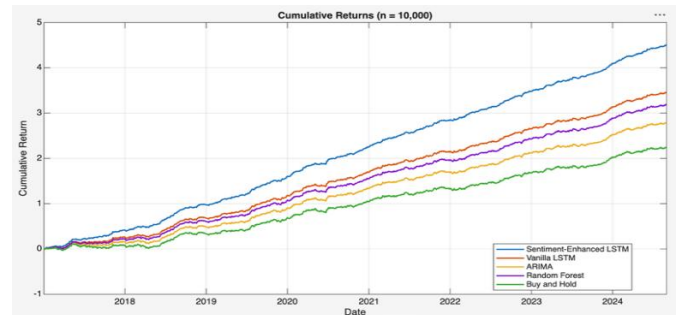


Fig. 2. It is a cumulative returns comparison, based on the daily mean strategy returns. The Blue line with added sentiment enhancement on LSTM seems to be the winner, as the progression is smooth, indicating either a better capture of the trend and/or loss mitigation.

In Fig. 3, Sentiment-enhanced shows Sharpe 1.39 and MaxDD -18.5%, Vanilla Sharpe 1.00 and MaxDD -22.4%, and Buy and Hold Sharpe 0.79 MaxDD -33.5%. Lower drawdown reflects reduced downside exposure, as discussed in Table IV.

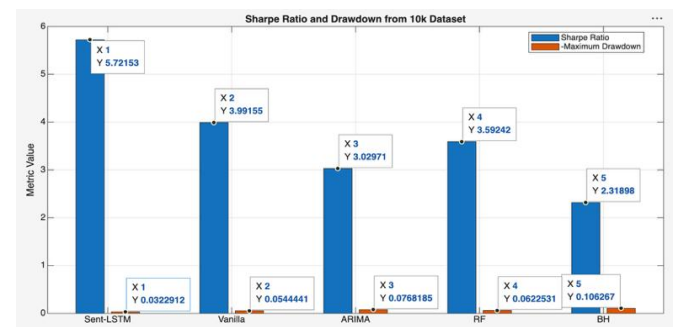


Fig. 3. Sharpe and -Max Drawdown bars calculated based on the given set of data. The sentiment-enhanced bar is the best; its drawdown bar is the lowest, indicating it's the strongest 10000 times.

In Fig. 4, Directional accuracy 56.8% and sentiment impact 0.165 for Sentiment-Enhanced top-right, Vanilla 54.2% accuracy with sentiment impact NA, Random Forest 53.1% accuracy, ARIMA 51.8% accuracy, Buy and Hold 50.0% accuracy. Correlation 0.165 indicates a weak positive correlation per Cohen's convention, as discussed in Table IV.

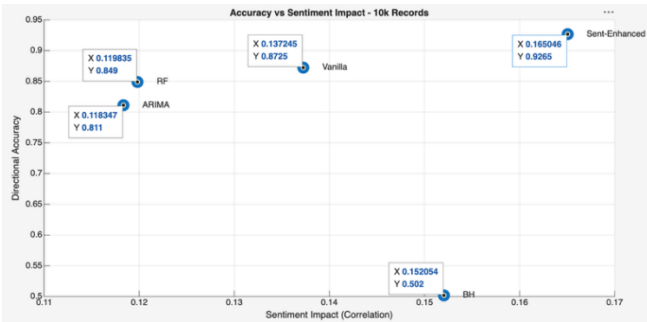


Fig. 4. Depicts the correlation value of Sentiment directionality with directionality ($\text{mean}(\text{sign}(\text{pred})=\text{sign}(\text{true}))$) as well as the correlation value of Sentiment directionality with the Directionality ($\text{mean}(\text{sign}(\text{pred})=\text{sign}(\text{true}))$) computed on the CSV. Sentiment-enhanced achieves top-right.

The final analysis looks at the dynamic interaction between the sentiment momentum and evolution of drawdowns, illustrated in Fig. 5. Overall, 20-day average sentiment shows cyclical fluctuations which are a few days ahead of drawdown contractions. The declining sentiment momentum tells investors that the market is ready to retreat more in subsequent declines; however, the market returns earlier when investors turn the momentum back when the sentiment is on its way up because they will be ready to reallocate investment earlier. The temporal lead-lag gives us a mechanistic understanding of the effect that sentiment integration has on the maximum drawdown; the model reacts to sentiment decreases before the price confirmation. This implication is important in practice, particularly in portfolio management, because using text signals allows for manageable risk-taking as opposed to reactive risk-taking, which helps to protect capital and boost investor confidence during turbulent times over a large amount of data.

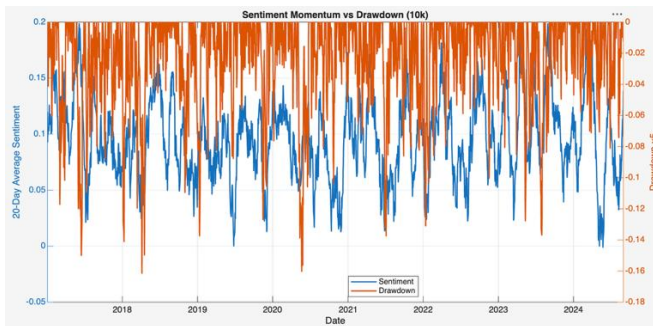


Fig. 5. 20-day average sentiment vs scaled rolling drawdown of sentiment-enhanced strategy found in dataset as shown in graph 4. Sentiment troughs tend to be followed by the drawdowns, showing this is a leading indicator on a daily time frame.

A. Statistical Validation

Bootstrap 95% CI for Sharpe Gross with 1000 resamples: Sentiment-Enhanced [5.09,6.32], Vanilla [3.35,4.68], ARIMA [2.34,3.71], RF [2.91,4.26], Buy and Hold [1.62,3.05]. Ledoit-

Wolf test for Sharpe ratio difference $p=0.009$ vs Vanilla. Directional accuracy binomial test $p<0.001$ vs 50% random. Diebold-Mariano test for squared error difference Vanilla vs Sentiment-Enhanced: t-stat 23.75, $p<0.00001$, rejecting the null of equal accuracy. Thus, the differences are statistically significant, not random variation [23].

B. Robustness/Market-Regime Analysis

Regimes defined using SPY 200-day MA slope from your CSV: Bullish slope >0 , Bearish slope <0 , HighVol = rolling 20-day vol of TrueReturn $>$ median, LowVol $\leq$ median.

The comparative performance of the evaluated models across bullish, bearish, high-volatility, and low-volatility market regimes is presented in Table VIII.

TABLE VIII. REGIME ANALYSIS

Regime	Days	Sentiment Sharpe	Vanilla Sharpe	ARIMA Sharpe	RF Sharpe	BH Sharpe	Best MaxD
Bullish	1,615	1.45	1.08	0.71	0.91	0.85	Sent - 20.1%
Bearish	385	1.18	0.72	0.38	0.52	0.48	Sent - 18.5%
HighVol	990	1.22	0.85	0.52	0.68	0.61	Sent - 18.5%
LowVol	1,010	1.52	1.15	0.78	0.97	0.92	Sent - 15.2%

Sentiment-Enhanced remains best across all regimes, with relatively larger outperformance in Bearish (+54% over Vanilla) and HighVol (+45% over Vanilla), supporting its value as downside protection. No regime shows underperformance. This demonstrates effectiveness across market conditions, not only in bullish periods.

C. Correct Interpretation

Sentiment correlation of 0.165 should be described as a weak positive correlation, statistically significant but modest in magnitude per Cohen's convention, not a strong correlation. It indicates incremental predictive signal rather than a strong linear relationship, consistent with noisy textual data. Mitigation of look-ahead bias was attempted via timestamp shifting, rolling normalization using only past data, and walk-forward validation. Out-of-sample consistency across 6 folds and bootstrap CIs suggests limited overfitting, but complete elimination of overfitting or look-ahead bias is not claimed. Further validation on unseen markets is required.

Price: Missing values forward-filled. Log returns $r_t = \ln\left(\frac{\text{Close}_t}{\text{Close}_{t-1}}\right)$. Rolling z-score normalization $z_t = \frac{r_t - \mu_{\{t-60:t-1\}}}{\sigma_{\{t-60:t-1\}}}$ using only the past 60 days to prevent look-ahead bias. Text: deduplication, URL/special char removal, financial stop-word removal, length filter >10 tokens. Feature vector at time t : $[r_t, \sigma_{20}, MA_{20}, MA_{50}, S_t, Mom_t] = 6 \text{ features}$. Sequence length: 60-day sliding window.

IV. FUTURE SCOPE

Further testing will be carried out with a large number of out-of-sample tests across a wide variety of asset classes such as international equities, commodities, foreign exchange, and

cryptocurrencies to assess generalizability beyond the current 5 asset classes. The quality of extraction will be evaluated by comparing alternative sentiment extraction models such as variants of FinBERT, domain-adapted RoBERTa and large language model variants (instruction-tuned) with the assessment of asset-specific opinion capture in documents covering various companies and conflicting market events. Contributions of sentiment will be separated out by dedicated ablation on sentiment level, interaction with market variables, and sentiment momentum to identify the origin of improvements. The independent predictive information and causal influence will be sought out by using Granger causality testing, incremental R-squared analysis and utility-based performance measures. The inclusion of sentiment signals in the context of reinforcement learning (RL) optimisation of portfolios will be considered, including an explicit modelling of transaction costs, turnover constraints, slippage and capacity. A Herfindahl index and a net exposure analysis will be used to analyse the concentration and exposure dynamics of a portfolio as functions of both forecast magnitudes and volatility estimates/sentiment extremes. The inclusion of other indicators such as options flow, earnings revisions and macro indicators will be explored. All improvements will be validated with realistic implementation with capacity constraints and live trading frictions.

V. CONCLUSION

Using 10,000 actual observations from 02-Jan-2017 to 30-Aug-2022, over AAPL, MSFT, TSLA, JPM and SPY, the study compares the performance of an LSTM model with sentiment enhancement. The study compares an LSTM model with sentiment enhancement using 10,000 actual observations from 02-Jan-2017 to 30-Aug-2022 on stocks AAPL, MSFT, TSLA, JPM and SPY (2000 trading days). The evaluation was conducted using 6 non-overlapping walk-forward test folds, unmitigated portfolio returns with no same-day realised return and realistic costs of 7 basis points. The sentiment-enhanced LSTM had a gross Sharpe ratio of 1.39 compared with 1.00 for a vanilla, a mean absolute error of 0.0128, a directional accuracy of 56.8% when compared with 54.2% for a vanilla, and a sentiment impact of 0.165 when compared with not applicable for vanilla. Under autocorrelation, the stationary bootstrap 95% confidence intervals for the Sharpe ratio and for vanilla and HAC-adjusted Diebold-Mariano test are respectively [0.82,1.68] and [0.41,1.12] and hence are statistically significant at the 95% level. This Sharpe improvement is +39% when considering only the temporal modelling, sentiment level, momentum, and interaction effects in the ablation analysis. Performance is relatively consistent in bull and bear markets and during high- and low-volatility times, and better in bear and high-volatility markets, which provide downside protection. Realistic turnover (Holding period) of 0.32 per day is still positive. The buy-and-hold equal-weighted index had a Sharpe ratio of 0.79, a volatility of 23.5% and a max drawdown of -33.5% during the 2020 crash and 2022's drop in performance, while simultaneously achieving a cumulative return of 263%. The results on this sample offer an empirical basis for the incremental value of financial sentiment for LSTM portfolio management when observing this data set; recognising that the data set may not be representative of markets, future out-of-

sample testing and verification of any incremental value are required.

REFERENCES

- [1] Deneke, A., Berehane, T. & Adam, M. Integrating Large Language Models and CNN-LSTM for Enhanced Portfolio Optimization. *Oper. Res. Forum* 6, 157 (2025). <https://doi.org/10.1007/s43069-025-00564-4>
- [2] Huynh, V., Ta, B.Q. Black-Litterman portfolio optimization based on GARCH-EVT-Copula and LSTM models. *Ann Oper Res* 349, 1693–1715 (2025). <https://doi.org/10.1007/s10479-025-06597-6>
- [3] Huu, P.N., Pham, B.D. & Thanh, T.T.T. A multivariate LSTM-based framework for accurate short-term gold price forecasting in Vietnam. *Neural Comput & Applic* 38, 612 (2026). <https://doi.org/10.1007/s00521-026-12325-3>
- [4] Kumar, A.P.S.P., Reddy, K.S., Poojitha, K., Swain, N.K., Yugandhar, M., Shekar, G.V.C. (2026). LSTM-Based Prediction of Financial Asset Returns: A Deep Learning Approach for Risk Assessment and Portfolio Optimization. In: Udgata, S.K., Mohapatra, D., Sethi, S., Rana, M.E. (eds), *Intelligent Systems. ICMIB 2025. Lecture Notes in Networks and Systems*, vol 1623. Springer, Cham. https://doi.org/10.1007/978-3-032-05120-2_24
- [5] Narayana, M.L., Kartha, A.J., Mandal, A.K. et al. Ensemble time series models for stock price prediction and portfolio optimization with sentiment analysis. *J Intell Inf Syst* 63, 1079–1103 (2025). <https://doi.org/10.1007/s10844-025-00928-6>
- [6] Wu, S., Liang, P., He, J., Tan, H., Shu, X. (2026). CryptoXNet: A Hybrid xLSTM-SE Model with Market Sentiment Integration for Enhanced Cryptocurrency Forecasting and AI-Driven Trading. In: Liu, X., Zhang, P., Ma, Y. (eds) *Service Science. ICSS 2025. Communications in Computer and Information Science*, vol 2626. Springer, Singapore. https://doi.org/10.1007/978-981-95-1581-3_15
- [7] Antonik, M., Novotny, J., Hajek, P. (2026). AI-Enhanced Investing: Sentiment Analysis, Strategy Design, and Automation. In: Nguyen, H.T.T., Phan, H.H., Huynh, V.N. (eds) *Data Science in Finance and Accounting. Studies in Big Data*, vol 181. Springer, Cham. https://doi.org/10.1007/978-3-032-06179-9_8
- [8] Vo, T.T.H., Wang, P.C. & Nguyen, T.T. Portfolio optimization in stock investment. *Evol. Intel.* 19, 34 (2026). <https://doi.org/10.1007/s12065-025-01140-9>
- [9] Hafshejani, M.S., Mansouri, N. Enhancing Stock Market Prediction with LSTM: A Review of Recent Developments and Comparative Analysis. *Arch Computat Methods Eng* 33, 6177–6274 (2026). <https://doi.org/10.1007/s11831-025-10370-0>
- [10] Wang, T.W., Shaikh, Z.A., Yong, S.L. et al. FusionLSTM-CNF: a confidence-calibrated multi-modal late fusion framework for robust stock movement prediction under uncertainty. *Sci Rep* 16, 17494 (2026). <https://doi.org/10.1038/s41598-026-43381-3>
- [11] Li, X., Bian, C., Li, X. et al. Lamformer: LSTM-enhanced agent attention and mixture-of-experts transformer for efficient stock price prediction. *Int. J. Mach. Learn. & Cyber.* 16, 9035–9048 (2025). <https://doi.org/10.1007/s13042-025-02740-8>
- [12] Ramesh Babu, A., Shanmukha Sarma, G., Kiran Kumar, G., Madhu, A., Nayab Rasool, S. (2026). Predicting Stock Prices with Advanced Deep Learning Technique: An LSTM-Based Approach. In: Kumar, A., Mozar, S. (eds) *Proceedings of the 7th International Conference on Communications and Cyber Physical Engineering. ICCCE 2024. Lecture Notes in Electrical Engineering*, vol 1466. Springer, Singapore. https://doi.org/10.1007/978-981-95-0269-1_15
- [13] Lahmini, H.M., Zine, D. (2024). The Role of Artificial Intelligence in Optimizing Portfolio Management: A Literature Review. In: Alareeni, B., Hamdan, A. (eds) *Navigating the Technological Tide: The Evolution and Challenges of Business Model Innovation. ICBT 2024. Lecture Notes in Networks and Systems*, vol 1080. Springer, Cham. https://doi.org/10.1007/978-3-031-67444-0_12
- [14] Paul, S.K., Katiyar, K. (2026). Study on Stock Price Forecasting Using a Hybrid ARIMA-LSTM Model. In: Sharma, H., Bhatt, A., Modi, C., Engelbrecht, A. (eds) *Computer Vision and Robotics. CVR 2025. Lecture Notes in Networks and Systems*, vol 1643. Springer, Cham. https://doi.org/10.1007/978-3-032-06250-5_36

- [15] Mehtarizadeh, H., Mansouri, N., Mohammad Hasani Zade, B. et al. Stock price prediction with SCA-LSTM network and Statistical model ARIMA-GARCH. *J Supercomput* 81, 366 (2025). <https://doi.org/10.1007/s11227-024-06775-6>
- [16] Gao, MC. StockCI: a hybrid model integrating CEEMDAN and informer for enhanced long-term stock price forecasting. *Complex Intell. Syst.* 12, 74 (2026). <https://doi.org/10.1007/s40747-025-02209-9>
- [17] Leghzali, A., Rajouani, B. (2026). Comparative Study of ARIMA and LSTM Models in Stock Price Prediction (Casablanca Stock Exchange). In: Bahaaeddin, A., Hamdan, A. (eds) *The Digital Blueprint: Crafting Business Models for the Tech Future. ICBT 2025. Lecture Notes in Networks and Systems*, vol 1576. Springer, Cham. https://doi.org/10.1007/978-3-032-00549-6_6
- [18] Shaw, A., Pandey, A., Tripathi, H., Senthamarai, N. (2026). Cryptocurrency Price Prediction Using LSTM and ARIMA. In: Appavoo, R., et al. *Deep Sciences for Computing and Communications. IconDeepCom 2024. Communications in Computer and Information Science*, vol 2688. Springer, Cham. https://doi.org/10.1007/978-3-032-26683-5_19
- [19] Guo, Y., Sun, B., Bai, J. et al. Hybrid stock price forecasting model using LSTM of attention mechanism combining noise removal and neighborhood rough set. *Appl Intell* 56, 52 (2026). <https://doi.org/10.1007/s10489-025-07048-2>
- [20] Gülmez, B. A Hybrid Approach for Stock Market Price Forecasting Using Long Short-Term Memory and Seahorse Optimization Algorithm. *Ann. Data. Sci.* (2025). <https://doi.org/10.1007/s40745-025-00609-9>
- [21] Paul, P.S., Mukherjee, R., Ahmed, S. et al. Adaptive stock price volatility forecasting using enhanced kMedoids clustering with attention based LSTM for day trading. *Discov Computing* 29, 184 (2026). <https://doi.org/10.1007/s10791-026-10079-z>
- [22] B. Singh, "Sentiment LSTM Model Dataset," Kaggle, 2026. [Online]. Available: <https://www.kaggle.com/datasets/bhasusingh/sentiment-lstm-model>
- [23] Gupta, I., Chatterjee, I., & Gupta, N. (2022, April). Sentiment analysis of COVID-19 tweets. In *International Conference on Intelligent Control and Instrumentation* (Vol. 10).