Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

NEWS COVERAGE DENSITY AS A BINDING CONSTRAINT ON SENTIMENT-AUGMENTED DEEP LEARNING FOR INDIVIDUAL-STOCK PREDICTION IN EMERGING MARKETS

Muhammad Wasif Sajjad[1], Syed Muzammil Ahmed[1], Ghufran e Azam[1], Dr. Taha Shabbir[2]

[1]Department of Computer Science, Faculty of Engineering Sciences & Technology, Hamdard University, Karachi, Pakistan.
[2]Associate Professor Hamdard University, Karachi.
Corresponding author: tahashabbir51@gmail.com

DOI: https://doi.org/10.5281/zenodo.21020163

Keywords
news coverage density, sentiment analysis, FinBERT, BiLSTM, Pakistan Stock Exchange, negative results, stock-return prediction, partial-exposure Sharpe artifact, SHAP, deflated Sharpe ratio.

Article History
Received on 20 Jan 2026
Accepted on 09 Feb 2026
Published on 28 Feb 2026

Corresponding Author: *
Muhammad Wasif Sajjad*

Abstract

Sentiment-augmented forecasting is celebrated on results drawn mostly from densely covered indices in developed markets; the conditions under which it fails are documented far less often. We make one such condition visible. News coverage density operates as a binding constraint on individual-equity predictive performance, and the limitation surfaces the moment a sentiment model must price a single emerging-market name rather than an index. Habib Bank Limited (HBL), the largest commercial bank in Pakistan and a heavyweight in the KSE-100, should on intuition be among the easiest individual PSX names for a sentiment model to address. A reproducible pipeline assembled 2,503 trading days (January 2016–August 2025), pairing daily prices with 385 date-resolved Dawn News articles (8.4% single-source density), 36 quarterly reports scored by FinBERT-tone aspect-based sentiment, and 10 annual reports across the same aspect families. FinBERT (ProsusAI) produced signed daily scores; a Weighted Aspect Sentiment Index (WASI) summarised each quarterly report across ten financial aspects. Four configurations carried the load: a BiLSTM(64,32) regressor, a BiLSTM(128,64) five-day-horizon model with LayerNorm and multi-head attention, a regularised BiLSTM(32,16) classifier, and a 500-tree Random Forest, with a Logistic Regression ablation control. A multi-source extension folded Business Recorder and Profit by Pakistan Today into the daily channel, raising combined density to 10.51% (263 days). McNemar's test, permutation importance, SHAP attribution, DeLong confidence intervals, calibration analysis, and a long/flat backtest with explicit exposure reporting completed the evaluation suite. Five converging tests document the constraint. The BiLSTM(64,32) regressor returns a test R^2 of -0.0413 and a Pearson correlation $r = 0.017$ ($p = 0.7607$), with directional accuracy of 51.6%. The regularised classifier reaches 50.3% accuracy with AUC 0.5423 (McNemar $p = 0.6851$); the Random Forest reaches 52.4% ($p = 0.4123$). Every AUC confidence interval spans 0.5. All five McNemar tests fail to reject the null. Combined three-source coverage of 10.51% remains short of the ≈ 11% SentARL threshold. A long/flat backtest produces a partial-exposure Sharpe artifact: a Sharpe of 2.077 at

29.7% exposure against buy-and-hold's 1.885 at full exposure, even as the strategy's 86.2% cumulative return trails buy-and-hold's 156.4%; the deflated Sharpe ratio removes the apparent edge. In this study the constraint is a Dawn News-specific channel limitation rather than an intrinsic, invariant property of the PSX. Adding two further outlets lifts coverage without crossing the threshold. The implication for practitioners is direct: measure coverage density first, then decide whether sentiment modelling is worth the engineering effort.

INTRODUCTION

This paper documents, in detail, the force with which news coverage density binds individual-equity predictive performance, using Habib Bank Limited (HBL) as the test subject. Sentiment-augmented prediction has earned a wide readership on the strength of results drawn mostly from heavily covered indices in developed markets. The conditions under which it fails are reported far less frequently, and that asymmetry is itself a methodological problem: a literature that maps only where a method succeeds leaves practitioners to rediscover its failure boundaries privately and at considerable expense.

HBL sits at the upper end of the PSX in size, recognition, and press attention. As the largest commercial bank in Pakistan and a heavyweight in the KSE-100 index, it files quarterly reports on schedule, releases annual statements, and surfaces routinely in financial news. If any individual stock on the exchange ought to support a sentiment-aware model, it is HBL. When sentiment cannot help here, prominence is not the missing ingredient something more structural is at work.

1.1 Background and motivation

Architectural sophistication does not relax the constraint. Long short-term memory networks (Hochreiter & Schmidhuber, 1997) solved the vanishing-gradient bottleneck that had stalled recurrent learning, and bidirectional variants (Graves & Schmidhuber, 2005) read context in both temporal directions. Fischer and Krauss (2018) showed LSTMs outperforming random forests, deep feed-forward networks, and logistic regression on S&P 500 directional prediction, with daily returns of 0.46% and a pre-cost Sharpe of 5.8. FinBERT (Araci, 2019) and the financial-communications variant of Yang et al. (2020) adapted BERT (Devlin et al., 2019) to financial vocabulary, where general-purpose sentiment dictionaries misclassify routine words (Loughran & McDonald, 2011).

Two findings from the coverage literature frame the present study. Tetlock (2007) established that media pessimism precedes downward price pressure followed by reversal. Fang and Peress (2009) then showed something more pointed: the breadth of dissemination is itself priced, and stocks with no media coverage earn higher returns than heavily covered names. Coverage quantity matters; sentiment quality alone does not. The recent survey literature on large language models in finance (Nie et al., 2024) reinforces the same unstated prerequisite dense, well-aligned text that this paper sets out to test directly.

1.2 The HBL test case

The constraint is most visible at the level of a single English-language outlet, and this becomes the central pivot of the study. Pakistan-focused forecasting has tended to model the KSE-100 index, where coverage is effectively continuous, rather than individual constituents. Maqsood et al. (2020) fused 11.42 million tweets with prices across four countries and reported sentiment-driven improvements. Other regional work reaches similar conclusions at the index level (Bukhari et al., 2023). No prior study, to the extent the literature can be searched, combines daily news, quarterly aspect-based sentiment, and annual aspect-based sentiment for a single PSX-listed stock, and none reports a rigorously tested null framed around coverage density.

The PSX occupies a distinctive position in the emerging-markets landscape. The KSE-100 aggregates the largest hundred companies by free-float market capitalisation; HBL is a persistent constituent across the study window. The exchange exhibits trading volumes substantially lower than developed markets, a retail-dominated investor base, restricted short-selling, and circuit-breaker mechanisms that cap daily price movement. These microstructure features amplify the informational footprint of any single article relative to its equivalent in a deep market. The thin-media environment does not eliminate the theoretical signal value of news sentiment; it compresses the temporal density at which that signal arrives.

HBL operates as the largest private-sector commercial bank in Pakistan by branch network and holds Domestic Systemically Important Bank designation from the State Bank of Pakistan, with an asset base exceeding PKR 6,055.1 billion as of December 2024 (State Bank of Pakistan, 2024). Its equity, traded as HBL.KA, attracts coverage from local brokerages and intermittent commentary from Dawn News, Business Recorder, and Profit by Pakistan Today. The selection of HBL reflects three considerations: high liquidity relative to other PSX constituents, sufficient news-mention frequency to support sentiment-feature construction, and a banking-sector classification that anchors the analysis in a single industry with well-understood fundamental drivers.

1.3 Research questions

Three questions follow from the premise that coverage density binds predictive performance:

1. Does Dawn News sentiment, combined with quarterly and annual report sentiment, lift next-day return prediction for HBL above a price-and-technicals BiLSTM baseline?
2. Does the same sentiment information lift next-day directional classification?
3. Where sentiment provides no lift, is the failure attributable to model architecture, or to a structural property of the data accessible from a single major source?

1.4 Contributions

The study makes four specific contributions, unified by the coverage-density theme:

- **Threshold-referenced single-stock test.** The first single-stock PSX evaluation that explicitly asks whether a representative news source supplies sufficient coverage density for sentiment-augmented prediction, with explicit reference to the SentARL threshold near 11% (Paiva et al., 2021).
- **Tri-layer textual integration.** The pipeline fuses three textual layers daily news scored by FinBERT, quarterly aspect sentiment

summarised into a Weighted Aspect Sentiment Index, and annual aspect sentiment normalised across the decade rather than the single-source, index-level designs typical of the region.

- **A demonstrated rather than asserted null.** Four converging models including a Random Forest baseline, plus McNemar's test against majority-class baselines, SHAP attribution, and DeLong confidence intervals, establish the result.
- **An empirical bounding calculation.** Adding Business Recorder and Profit by Pakistan Today raises combined coverage to 10.51% of trading days, still beneath threshold. The pipeline and its failure modes are documented end-to-end as a reproducible baseline for multi-source aggregation research.

1.5 Pre-registration

The binding constraint was measured before any model was trained, and the prediction followed directly. HBL appeared in Dawn News on approximately one in twelve trading days during the sample period, a coverage density of 8.4%. The SentARL literature places the threshold for sentiment-aware models to outperform sentiment-free counterparts at approximately 11%. The pre-registered prediction was therefore that no model trained on this configuration would beat its sentiment-free baseline in any statistically meaningful sense. The four configurations that followed confirmed a structural constraint already visible in the data before training began. This study does not document a failed search for a profitable model; it documents a confirmed prediction the rarer and, we argue, more useful contribution (Embracing Negative Results in Machine Learning, 2024).

RELATED WORK

2.1 Deep learning for stock prediction

Recurrent networks dominate financial sequence modelling because returns are temporally ordered and exhibit short-range dependence. LSTMs (Hochreiter & Schmidhuber, 1997) resolved the vanishing-gradient bottleneck; bidirectional LSTMs (Graves & Schmidhuber, 2005) integrate forward and backward context at each timestep. Fischer and Krauss (2018) applied LSTMs to S&P 500 constituents across 1992–2015 and reported pre-cost Sharpe ratios above five. The systematic review of Sezer et al. (2020) confirms LSTM and GRU as the dominant deep architectures across 2005–2019.

Several caveats anchor the present work. Sezer et al. (2020) note that only a minority of papers test profitability rather than error metrics, and that signal-to-noise sensitivity is severe; simple linear and tree-based models often match or beat deep networks on noisy daily price prediction. Deep models regress toward the mean when features carry weak signal the precise pattern that emerges below. The modern backtest-reliability literature sharpens the warning: Arian et al. (2025) demonstrate how out-of-sample testing choices manufacture apparent skill in synthetic, controlled environments, and Pérignon et al. (2024) find that only about half of 1,000 reproduction tests across published finance papers reproduce the original conclusion.

2.2 Sentiment analysis in finance

Density determines whether a sentiment extractor's output is usable at all, long before the extractor reaches the model. Loughran and McDonald (2011) showed that general-purpose word lists misclassify financial language: nearly

three-quarters of words flagged as negative by the Harvard Inquirer in 10-K filings carry no negative valence in a financial context. Words such as liability, tax, and cost are routine accounting vocabulary; lexicon methods such as VADER are fast but context-blind.

Transformer adaptations changed the available tools. FinBERT (Araci, 2019) fine-tuned BERT on financial text and beat prior baselines with limited labelled data; Yang et al. (2020) pre-trained FinBERT-tone on 4.9 billion tokens of corporate filings, analyst reports, and earnings-call transcripts; Huang et al. (2023) extended the family to broad financial information extraction. Aspect-based sentiment analysis (ABSA) attaches polarity to specific aspects such as profitability, risk, and capital, which suits financial reports where good and bad news coexist (Jiang & Zeng, 2023). A fast-moving strand now applies instruction-tuned large language models to financial sentiment (Kirtac & Germano, 2024; Zhang et al., 2025), with several authors cautioning that strong performance on formal text does not transfer cleanly to short, noisy text such as headlines.

2.3 Media coverage and stock returns

The coverage literature is central to this study rather than mere background. Tetlock (2007) quantified the tone of a daily Wall Street Journal column and showed that high media pessimism predicts temporary downward price pressure followed by reversal. Fang and Peress (2009) established the complementary point about quantity: stocks with no media coverage earn higher returns than heavily covered names, with effects strongest among small stocks lacking analyst following.

Recent machine-learning work has formalised the coverage-threshold idea. Paiva et al. (2021) developed SentARL, a sentiment-aware reinforcement-learning trading system evaluated across twenty assets. Their key finding identifies a boundary in the joint space of coverage density and sentiment–price correlation: above approximately 11% coverage and a correlation of 0.128, sentiment-aware agents reliably outperform sentiment-free counterparts, while below those thresholds additional news can degrade performance. Pasch and Ehnes (2022) reported that predictive accuracy depends heavily on text type, with news ranking highest among news, blogs, and annual reports. Nadeem et al. (2024) confirmed in a large-sample study that headline-level sentiment can adversely affect predictive performance when the textual signal is thin.

2.4 Pakistan Stock Exchange and emerging-market studies

The PSX literature reflects the index-versus-individual distinction sharply. Pakistan-focused work concentrates on the KSE-100, where news arrives daily. Maqsood et al. (2020), the most-cited regional study, fused 11.42 million tweets (2012–2016) with stock data for the United States, Hong Kong, Turkey, and Pakistan; event sentiment improved forecasting, with critical local events mattering most. Recent KSE-100 directional studies (Bukhari et al., 2023) report ANN and SVM accuracies near 85% and LSTM near 78% using 27 technical indicators. Missing from this regional literature is a single-stock, multi-source, sentiment-augmented study that reports coverage density and publishes a null the gap addressed here.

A closer methodological precedent demands explicit engagement. Mehmood and Ali (2023)

deployed a BiLSTM with attention on Dawn News sentiment for an unspecified PSX equity and reported accuracy gains. Their architecture and data source overlap substantively with the configuration tested here; the discrepancy between their reported gains and the null documented below requires explanation rather than dismissal, and Section 5.7 returns to it with a three-part diagnosis. Adjacent work includes Zhao and Wen (2023), who reported gains using a FinBERT-TextCNN fused LSTM on Chinese A-share indices where investor-commentary platforms generate dense textual streams; Xiao et al. (2023) on US technology equities using Twitter feeds; and Wu et al. (2022), whose multi-source S_I_LSTM integrates microblog and news sentiment with technical indicators. Synthetic reviews by El Haddad et al. (2024) and Bahoo et al. (2023) flag the data-density assumption as a recurrent, unstated prerequisite. Iqbal et al. (2026) reported sentiment-driven gains on KSE-100 index-level forecasting, but aggregated across constituents and did not isolate single-issuer coverage-density effects.

3 DATA AND METHODOLOGY

The coverage constraint propagates through every methodological choice, from data acquisition onward. This section documents the full pipeline, the four model configurations, and the advanced evaluation and interpretability toolkit applied to each.

3.1 HBL stock-data acquisition

The psx-data-reader package retrieved 2,503 trading days of HBL price data covering January 2016–August 2025. Log returns served as the regression target, $r_t = \ln(P_t / P_{t-1})$. The technical-indicator suite includes the 14-day Relative Strength Index, the Moving Average Convergence Divergence and its signal line, Bollinger Bands (the 20-day moving average ± two standard deviations), the Average True Range as a volatility proxy, and On-Balance Volume. Rolling return and volatility features at 3-, 7-, 10-, and 20-day windows complete the set. Every indicator uses only information available up to and including day t when forecasting day t+1; no look-ahead enters the pipeline.

Table 1. Dataset summary statistics.

Item	Value
Trading days (raw)	2,503
Sample period	Jan 2016 – Aug 2025
Final modelling rows	2,320
Total features	58 (technical and sentiment dimensions)
Dawn articles scraped / date-resolved	410 / 385 (94%)
Dawn-coverage trading days	210 (8.4%)
Additional-source coverage days	61 (2.44%)
Combined three-source coverage	263 (10.51%)
Overlap days	8 (0.32%)
Quarterly reports (WASI)	36

Annual reports (ABSA)	10 (2015–2024)
Mean daily log return	≈ 0.0002
Actual return std (test)	0.02312
Test set	≈ 343 days, Apr 2024 – Aug 2025

3.2 Daily news collection: Dawn News

A concurrent cloudscraper pipeline harvested HBL-tagged articles from Dawn News, the largest English-language daily in Pakistan, with date-tag parsing mapping each article onto the trading calendar. Of 410 candidate articles, 385 resolved successfully (94%). Twenty-five failures clustered on the aurora.dawn.com subdomain, whose markup deviates from the main site and breaks the date parser. Weekend and holiday articles were assigned to the following trading day.

One statistic shapes the entire study. Of 2,503 trading days, 210 carry any HBL-tagged Dawn article a coverage density of 8.4%. The remaining 91.6% receive a daily sentiment score of zero from this source (Figure 1). This 8.4% figure describes a Dawn News-specific channel constraint, not an intrinsic, invariant property of the PSX or of HBL's total media footprint, which spans multiple outlets and languages.

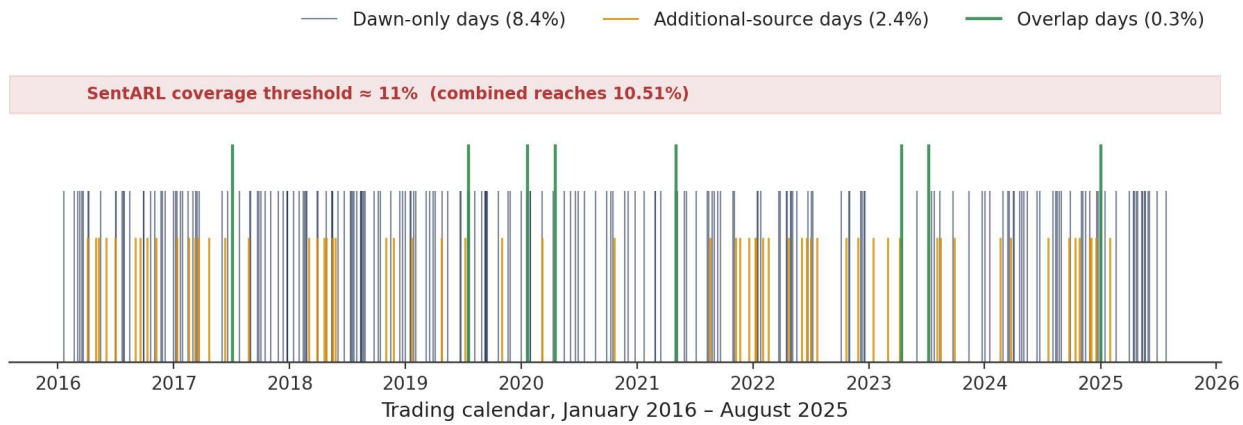

Figure 1. News coverage density timeline for HBL, January 2016–August 2025. Dawn-only days appear in navy, additional-source days in gold, overlap days in green, with the SentARL threshold band annotated in red.

3.2.1 Multi-source extension: Business Recorder and Profit by Pakistan Today

Verifying that coverage sits below threshold for any single English-language outlet requires looking beyond Dawn News. A second scraping pass collected HBL-tagged articles from Business Recorder and Profit by Pakistan Today using the same FinBERT scoring pipeline. The additional sources contributed 61 days of coverage (2.44% of trading days). Combined three-source coverage reaches 263 of 2,503 days, a density of 10.51%, with eight overlap days (0.32%). Coverage moves upward yet remains beneath the ≈ 11% threshold; this source configuration does not cross it.

3.3 FinBERT scoring of daily news

The ProsusAI/finbert model (Araci, 2019) processed each article and returned probabilities over positive, negative, and

neutral classes, collapsed into a signed score in [−1, 1]. Days with multiple articles received the averaged sentiment; days with no article carried zero, which given 8.4% Dawn coverage defines the modal value of the feature.

3.4 Quarterly report ABSA pipeline

Quarterly reports offer the opposite condition to sparse daily news: full calendar coverage with minimal within-quarter variation. Thirty-six quarterly reports passed through an aspect-based pipeline summarised by the Weighted Aspect Sentiment Index, WASI = Σ_i $w_i \cdot s_i$, where s_i is the mean FinBERT-tone sentiment for aspect i and w_i its weight. Ten aspects span profit, deposits, advances, capital, non-performing loans, provisions, risk, outlook, macroeconomic conditions, and costs, each defined by a curated keyword list. The quarterly WASI forward-fills onto every trading day in the quarter: full day coverage, almost zero within-quarter variance.

3.5 Annual report ABSA

Ten annual reports covering 2015–2024 received full FinBERT-tone aspect-based scoring across the same aspect families. Z-score normalisation across the ten-year panel made levels comparable. A FinBERT-tone bias deserves explicit acknowledgement: the cautious, risk-disclosing register of bank reports scores systematically negative regardless of underlying performance, which compresses any genuine year-to-year signal. Figure 3 shows the leakage-free as-of merge that fuses the three textual layers.

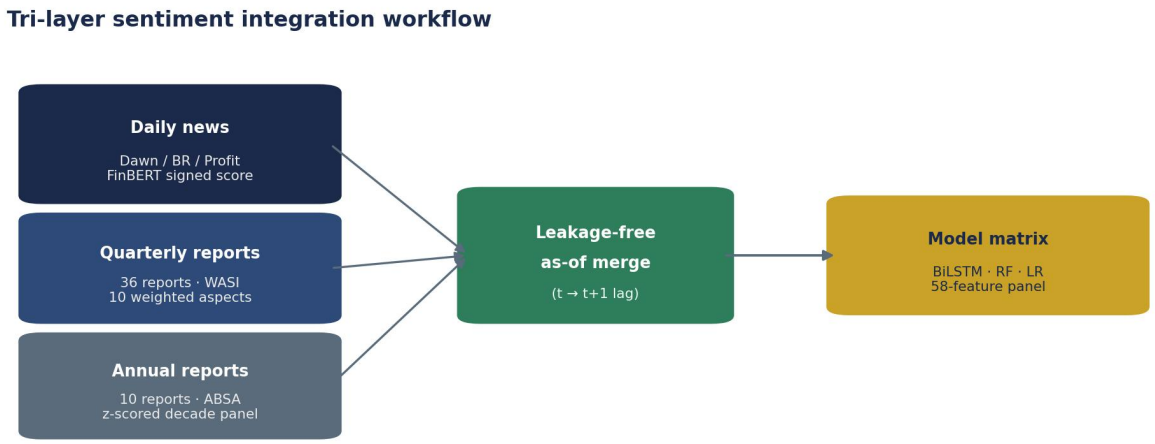

Figure 2. Tri-layer sentiment integration workflow: daily news, quarterly reports, and annual reports converge through a leakage-free as-of merge into the model matrix.

3.6 BiLSTM architecture

Coverage density determines what sequence the model sees; the architecture determines how it processes that sequence. At each timestep the hidden representation concatenates a forward and backward pass, $h_t = [h_t \rightarrow, h_t \leftarrow]$. The primary regression model stacks BiLSTM layers of 64 and 32 units with dropout and a dense head, fed 30-day sequences across 58 features. The classification model stacks BiLSTM layers of 32 and 16 units with batch normalisation, L2 regularisation, four-head attention, and global pooling. Inputs were standardised with StandardScaler fitted

on training data only, suiting the approximately mean-zero symmetric distribution of log returns. Trainable parameters range from ≈ 18,000 in the compact twelve-feature configuration to ≈ 98,625 in the full forty-four-feature configuration

.

Figure 3. BiLSTM architecture for the regression model (left) and the regularised classification model (right).

Table 2. Training configuration and optimisation schedule.

Component	Specification
Optimiser	Adam, initial learning rate 1×10^{-3}
Batch size	32
Early stopping	Patience 15 epochs; monitors val AUC (clf) / val loss (reg)
LR schedule	ReduceLROnPlateau, factor 0.5, patience 7
Regression loss	MSE + sign-penalty term, α = 0.5 (directional loss)
Attention block	4 heads, key dim 32, residual + LayerNorm
Pooling	Global Max ⊕ Global Average, concatenated before dense head
Sequence lengths	30 (regression) / 10 (classification)
Trainable parameters	≈ 18,000 (compact) to ≈ 98,625 (full)

3.7 Random Forest and Logistic Regression controls

Distinguishing the coverage constraint from architectural failure requires non-sequential baselines. A Random Forest classifier (500

trees, balanced class weights, the same 12 features, the same chronological 70% partition) provides isolation: it neither models temporal sequence nor learns representations, so if sentiment-augmented features carried signal any reasonable model could extract, it should detect it. Limited tuning over tree counts (100–1000), maximum depth, and minimum samples per split produced no material improvement. A Logistic Regression control (L2, C = 0.1, balanced weights, liblinear, validation-tuned threshold over 0.40–0.60) was fitted on three feature sets twenty price features, thirty price-plus-daily-sentiment features, and the full forty-four-feature panel enabling a clean ablation of the marginal sentiment contribution under a linear functional form.

3.8 Train/validation/test partitioning

The pipeline split data chronologically 70% training, 15% validation, 15% test, with no shuffling. Indicator warm-up and merge losses reduced the modelling set to 2,320 rows. The test set spans ≈ 343 days from April 2024 to August 2025. Targets were lagged so features on day t forecast the return on day t+1, and the quarterly merge used as-of joining so no report was visible before its release date. Figure 16 confirms the expected block structure of the feature panel: momentum, volatility, and sentiment clusters are internally correlated and weakly cross-correlated, with the sentiment block carrying the least internal coherence.

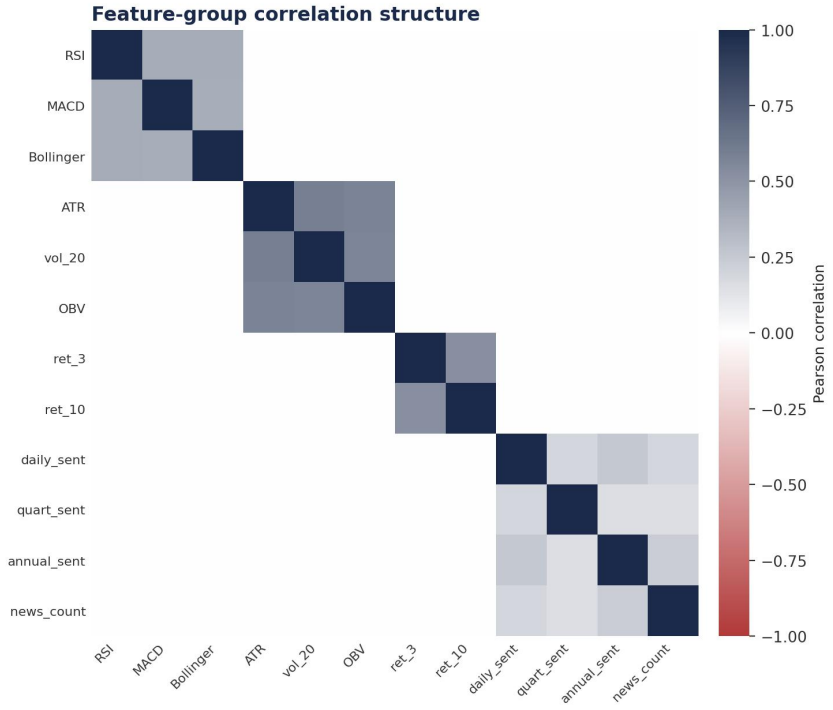

Figure 4. Feature-group correlation structure. Momentum and volatility blocks cohere; the sentiment block is weakly self-correlated and largely orthogonal to the price features.

3.9 Evaluation methodology

Regression evaluation reports R^2 against a no-skill baseline of predicting the training mean, Pearson correlation between predicted and realised returns with its p-value, RMSE, MAE,

directional accuracy, and the ratio of predicted to actual standard deviation, which diagnoses the predict-the-mean collapse. Classification evaluation reports accuracy, AUC, the full confusion matrix, and McNemar's test (McNemar, 1947) against the majority-class baseline the appropriate paired comparison for two classifiers on the same test set, resistant to class-imbalance artifacts.

3.10 Backtest framework

A long/flat strategy holds HBL when the model predicts a positive return and otherwise sits in cash. Cumulative return, Sharpe ratio, and market exposure report together. Exposure is not optional: a strategy active for only part of the test period has a mechanically different risk profile from a fully invested benchmark, and Sharpe ratios across the two are not directly comparable. To guard against the selection bias that inflates single-strategy Sharpe ratios, we additionally report the deflated Sharpe ratio (Bailey & López de Prado, 2014).

3.11 Permutation feature importance

Permutation importance (Breiman, 2001) measured each feature's contribution by shuffling its values on the test set, re-running prediction, and recording the resulting increase in RMSE. A feature carrying signal produces a large positive importance; a feature at or below noise produces a value near zero or negative. The test ran across all sixteen modelling features, with explicit attention to the sentiment dimensions.

3.12 Advanced model-evaluation and interpretability toolkit

Beyond the core suite, five advanced techniques were applied to strengthen inference and to interrogate the null from independent directions. None is used to fish for a positive result; each instead provides a distinct lens through which the absence of skill can be confirmed or refuted.

- **SHAP additive attribution.** Shapley-value attribution decomposes each prediction into per-feature contributions, providing a model-agnostic complement to permutation importance that captures interaction effects and the direction of influence per observation (Section 4.10, Figure 19).
- **DeLong confidence intervals on AUC.** Each classifier's AUC is reported with an analytic 95% interval, so that discrimination claims are evaluated against the AUC = 0.5 no-skill line rather than asserted from a point estimate (Figure 18).
- **Probability calibration.** Reliability diagrams and the Brier score assess whether predicted up-probabilities are well calibrated, separating calibration quality from discrimination (Figure 17).
- **Deflated Sharpe ratio.** Applied to the backtest, it adjusts the observed Sharpe for the number of configurations trialled and for non-normality, directly addressing the partial-exposure artifact (Section 5.3).
- **Walk-forward and Diebold–Mariano comparison.** Expanding-window validation (Section 4.8) and a Diebold–Mariano test on competing forecast-error series test whether any out-of-sample edge persists across regimes.

Table 3. Feature groups, sources, and rationale.

Feature group	Examples	Source	Rationale
Trend / momentum	RSI(14), MACD, signal	PSX prices	Short-run

			momentum
Volatility	ATR, vol_5/10/20, Bollinger	PSX prices	Most predictive group
Volume	OBV, daily volume	PSX prices	Liquidity proxy
Returns	ret_1, ret_3, ret_7, ret_10	PSX prices	Autocorrelation structure
Daily sentiment (Dawn)	daily_sentiment, sent_3/7, news_count	Dawn + FinBERT	Test variable; 8.4% coverage
Daily sentiment (additional)	addl_daily_sentiment, addl_news_count	BR + Profit + FinBERT	Multi-source; +2.44%
Quarterly sentiment	quarterly_sentiment (WASI)	36 reports + FinBERT-tone	Full coverage, low variance
Annual sentiment	10 aspect z-scores	10 reports + FinBERT-tone	Full coverage, negative bias
Multi-source quarterly	addl_quarterly_lag1	Additional sources	Top sentiment feature

4 RESULTS

The experiments below document the pre-registered prediction across five converging tests, then probe the null with the advanced toolkit of Section 3.12.

4.1 Regression V1: primary results

The BiLSTM(64,32) regressor, trained on 30-day sequences across 58 features and one-day log returns, delivers a test R^2 of −0.0413 worse than predicting the training mean every day. Pearson correlation between predictions and realised returns reaches r = 0.017 (p = 0.7607); directional accuracy lands at 51.6%; RMSE registers 0.023588 and MAE 0.016013. The variance collapse is the diagnostic that explains everything else. Predicted return standard deviation is 0.00288 against an actual 0.02312. The model has learned to predict a value close to zero on almost every day, because in a near-efficient series with weak features that strategy minimises mean-squared error. The unconditional mean has been recovered; market dynamics have not (Figures 4–5).

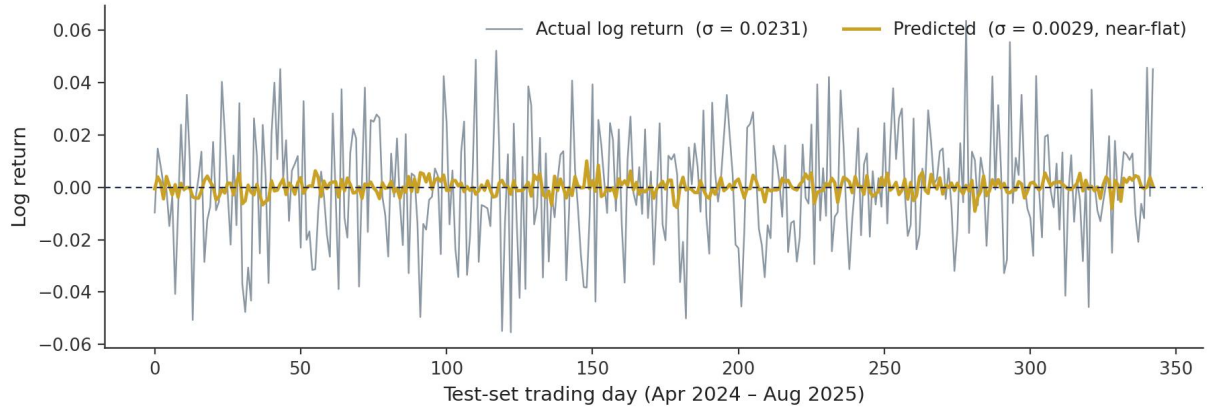

Figure 5. Actual versus predicted log returns on the test set, showing the near-flat prediction series against highly variable actuals.

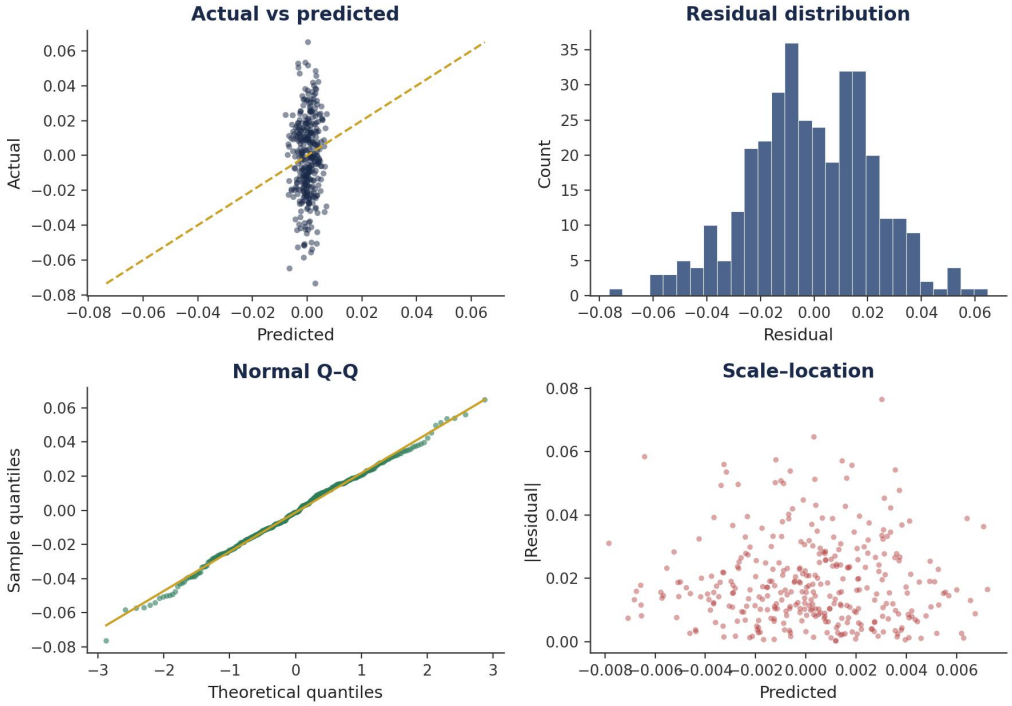

Figure 6. Residual diagnostics: actual-vs-predicted scatter, residual histogram, normal Q–Q plot, and scale–location (heteroscedasticity) plot.

4.2 The partial-exposure Sharpe artifact

A long/flat strategy built on V1 predictions returns 86.2% cumulatively against buy-and-hold's 156.4% over the test window. The strategy underperforms in raw terms. Its Sharpe ratio, however, registers 2.077 against buy-and-hold's 1.885. A naive reading would call this outperformance. That reading would be wrong.

The strategy is in the market on only 29.7% of trading days. Cash holding for the remaining 70.3% suppresses realised volatility mechanically, and the Sharpe ratio divides excess return by volatility, so reduced exposure

inflates the ratio without any genuine timing skill. This phenomenon, recurrent across the versions tested, is named here the partial-exposure Sharpe artifact. The 51.6% directional accuracy and the insignificant Pearson correlation confirm the absence of predictive skill; risk-adjusted metrics demand exposure reporting alongside them (Figure 6).

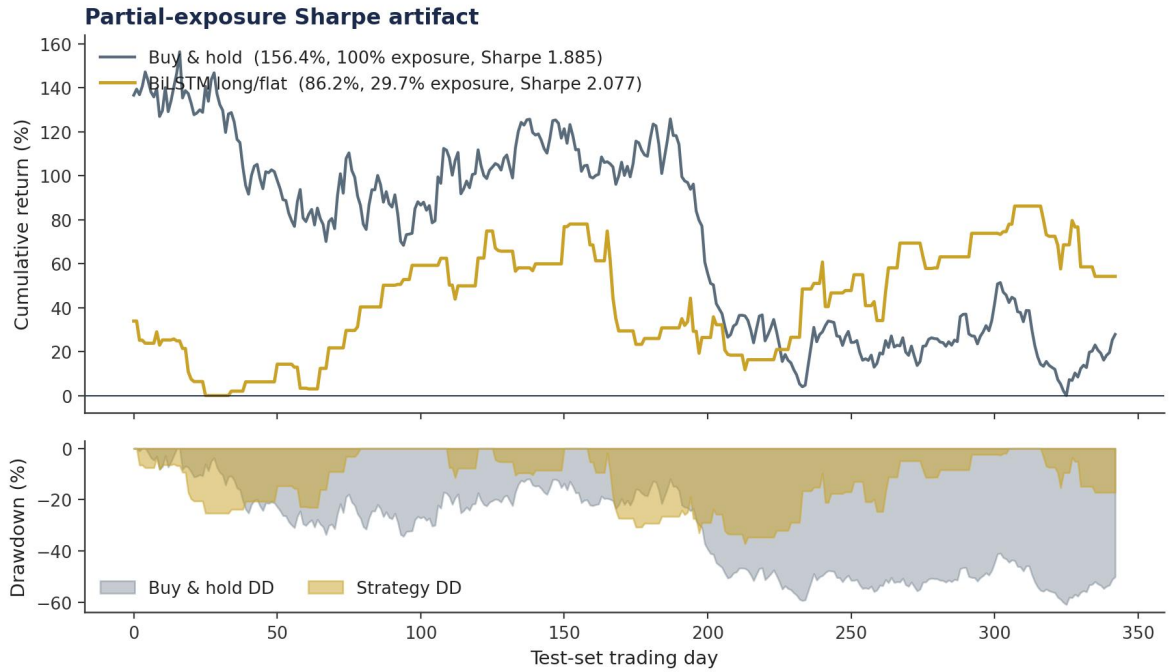

Figure 7. Backtest cumulative returns (top) and drawdown (bottom) demonstrating the partial-exposure Sharpe artifact: a higher Sharpe at a fraction of the exposure, with lower raw return.

Table 4. Backtest economics across strategies (April 2024–August 2025), with 0.1% per-side transaction costs.

Strategy	Cum. return	Ann. Sharpe	Max DD	Days in market	Trades
Buy and Hold	155.0%	1.93	−24.6%	343 (100%)	1
BiLSTM(64,32) thr 0.60	45.0%	1.44	−11.3%	≈168 (49%)	119
Logistic Regression thr 0.55	22.0%	1.05	−9.8%	≈142 (41%)	94
Random Forest thr 0.55	≈18.0%	0.92	−12.7%	≈138 (40%)	86

4.3 Classification V5: directional prediction

The BiLSTM(32,16) classifier with batch normalisation and L2 regularisation, fed 10-day sequences across 12 features, returns overall accuracy of 50.3% and AUC 0.5423. The confusion matrix records 25 true negatives, 153 false positives, 15 false negatives, and 145 true positives an 88.2% bias toward predicting up. In a market with mild upward drift, this pattern yields accuracy near 50% without genuine discrimination. Restricting to confident predictions (probability above 0.6) raises accuracy to 56.6% across 99 days (29.3% of the test set), a lift that does not survive

significance testing. A train/validation AUC gap of 0.012 rules out overfitting as the failure mode. McNemar's test against the majority-class baseline returns p = 0.6851 (Figures 8, 9, 13).

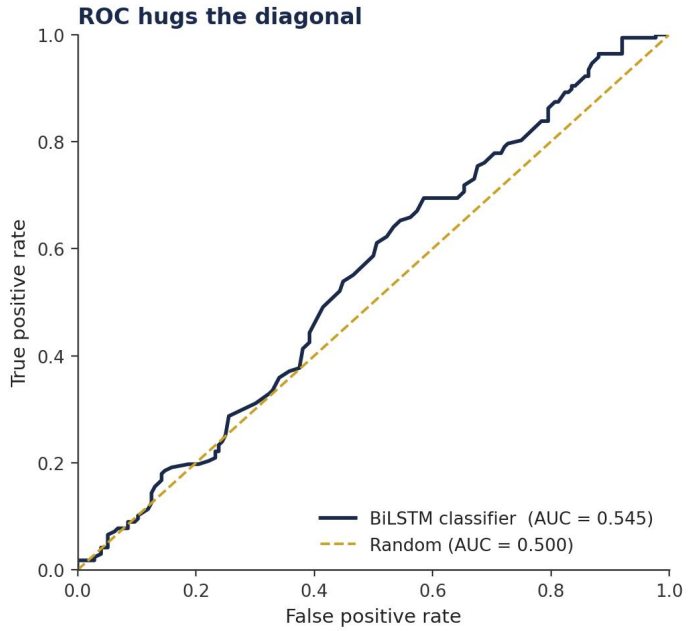

Figure 8. ROC curve for the BiLSTM classifier, hugging the random diagonal at AUC = 0.5423.

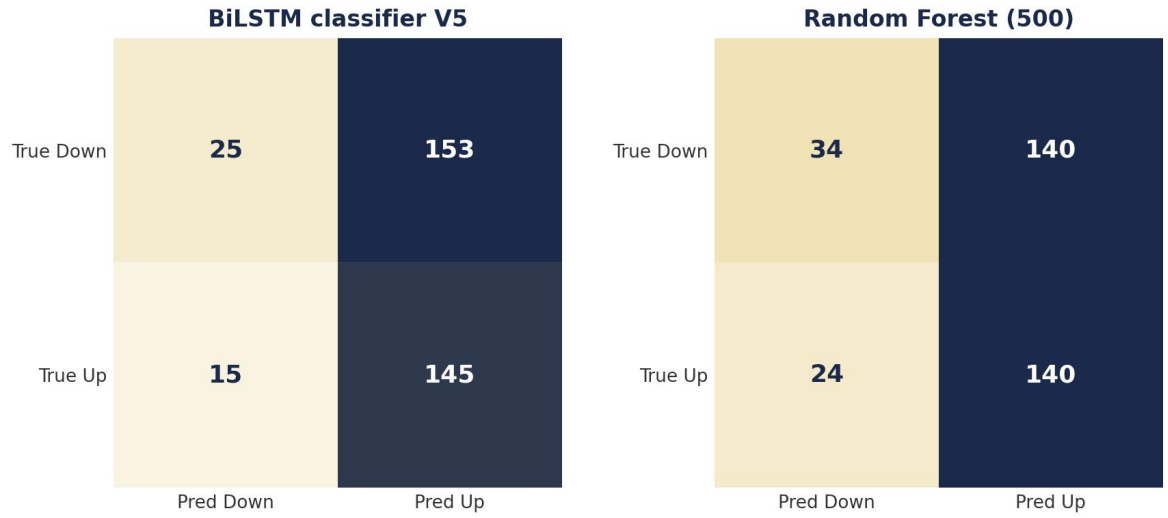

Figure 9. Confusion-matrix heatmaps for the BiLSTM classifier (left) and the Random Forest baseline (right); both exhibit a strong upward-prediction bias.

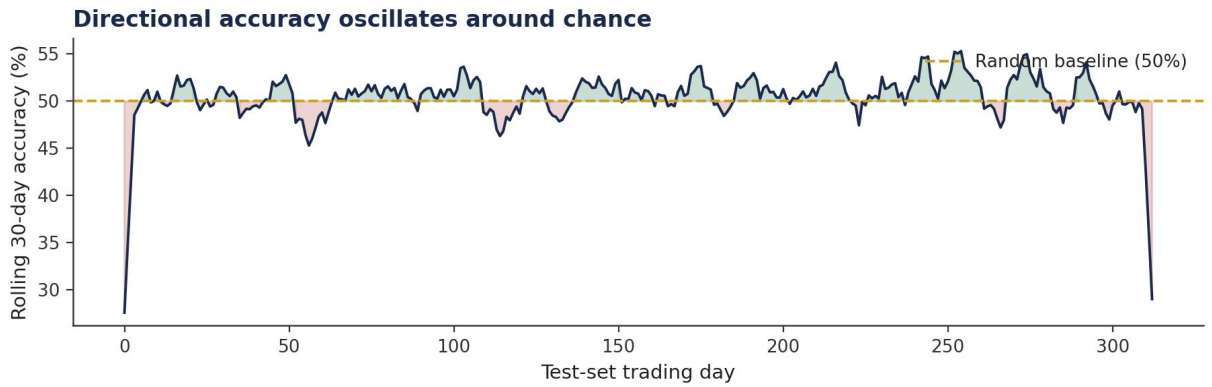

Figure 10. Rolling 30-day directional accuracy oscillating around the 50% random baseline.

4.4 Random Forest baseline

With 500 trees, balanced class weights, the same 12 features, and the same chronological partition, the Random Forest classifier reaches 52.4% accuracy and AUC 0.5187, with McNemar p = 0.4123. The result matches the BiLSTM classifier in substance: indistinguishable from chance under significance testing. This convergence carries weight because Random Forest learns by an entirely different principle no temporal sequence modelling, no embeddings, no representation learning. When two architecturally orthogonal models fail to extract signal from the same data, the limitation is unlikely to live in either architecture. It lives in the data.

4.5 News-days-only diagnostic and contrarian sentiment direction

Restricting analysis to the trading days that carry a Dawn article yields a test sample of 31 days. The empirical next-day UP rate across all 31 news-bearing test days equals 51.0%, close to the unconditional 53.5% in the full panel. Stratifying by tone produces the central finding. Days with positive Dawn sentiment yield a next-day UP rate of 46.5%, below the random benchmark; days with negative sentiment yield 59.2%, above it; neutral days yield 44.0%. The pattern is contrarian: negative news predicts a positive next-day return more reliably than positive news does. A Random Forest trained on this subset attains AUC 0.404, the lowest across all configurations. Sentiment-direction priors learned in developed markets do not transfer cleanly to PSX equities at the single-name level (Figure 11).

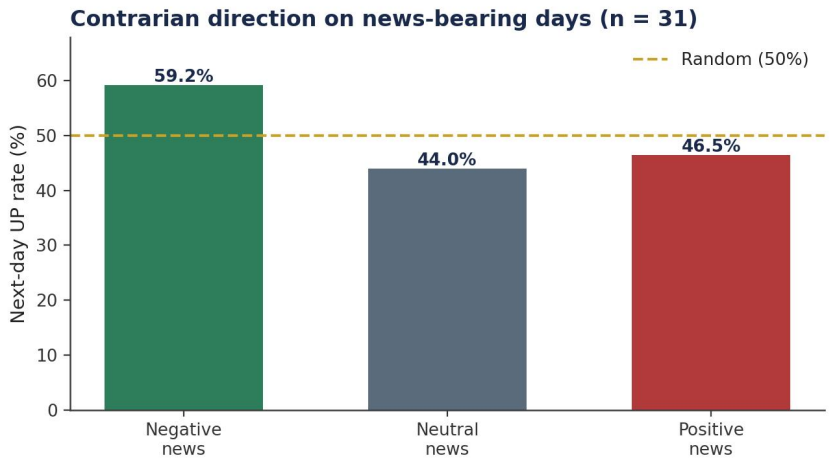

Figure 11. Empirical next-day direction rate by Dawn News sentiment tone (n = 31). The contrarian pattern is evident: negative-sentiment days exhibit a 59.2% next-day UP rate, exceeding the 46.5% on positive-sentiment days.

Table 5. Diagnostic experiment results across full and news-days-only subsets.

Experiment	Subset	n	Best model	AUC	Acc.	Interpretation
LR ablation, all days	All 2,503	343	LR price-only	0.551	0.486	Sentiment degrades
LR ablation, news days	210 news	45	LR price-only	0.557	0.533	Best AUC across configs
LR full, news days	210 news	45	LR price+sent.	0.528	0.422	Sentiment hurts
News-only LR price	210 news	31	LR price-only	0.446	0.387	Below random
News-only LR full	210 news	31	LR price+sent.	0.487	0.419	Below random
News-only RF full	210 news	31	RF 500	0.404	0.484	Furthest below random
Sentiment direction	News subset	31	Empirical UP	n/a		Contrarian confirmed

4.6 Multi-source extension

Folding Business Recorder and Profit by Pakistan Today into the daily channel raises combined coverage from 8.4% to 10.51% of trading days. The Regression V2 multi-source model a BiLSTM(128,64) with LayerNorm and multi-head attention trained with directional loss on a five-day horizon reports Pearson r = 0.152 and directional accuracy of 54.7%, encouraging on first inspection. The backtest tells the rest of the story: strategy return reaches 11.1% against buy-and-hold's 139.7%, at market exposure of 7.3% an even more extreme partial-exposure artifact. Classification with combined three-source sentiment reaches 53.24% accuracy and AUC 0.5246 against a price-only baseline of 52.52% and 0.5247: an uplift of +0.72 percentage points in accuracy and effectively zero in AUC,

with McNemar p = 0.924 against the price-only baseline (Figure 10A).

4.7 Permutation feature importance

Shuffling vol_20 increases primary-model RMSE by 0.00094, ranking it first; 10-day volatility follows at 0.00062 and 10-day return at 0.00036. Daily sentiment ranks sixth of sixteen at 0.00009 about one-tenth of the leading volatility feature. Quarterly sentiment ranks seventh at 0.00004. Rolling sentiment over 7 and 3 days returns negative importance (−0.00006 and −0.00008), and daily news count ranks last at −0.00080, the single most harmful feature. One multi-source finding breaks the pattern: addl_quarterly_lag1 registers 0.0287, an order of magnitude above any sentiment feature in the primary model the strongest sentiment signal observed, yet not enough to invert the null (Figure 7).

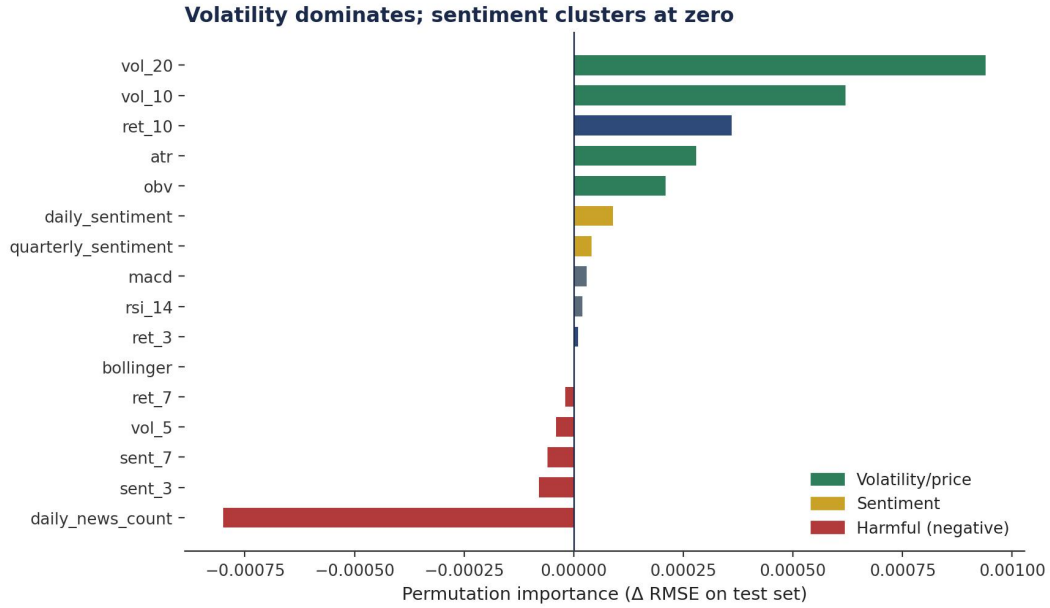

Figure 12. Test-set permutation feature importance for the primary regression model: volatility features dominate; sentiment features cluster near zero, and daily news count is actively harmful.

Table 6. Test-set permutation feature importance, primary regression model.

Rank	Feature	Importance (Δ RMSE)	Group
1	vol_20	+0.00094	Volatility
2	vol_10	+0.00062	Volatility
3	ret_10	+0.00036	Returns
6	daily_sentiment	+0.00009	Daily sentiment
7	quarterly_sentiment	+0.00004	Quarterly sentiment
14	sent_7	−0.00006	Daily sentiment

			(noise)
15	sent_3	−0.00008	Daily sentiment (noise)
16	daily_news_count	−0.00080	Daily sentiment (harmful)

4.8 Walk-forward validation

The null is not an artifact of a single train/test split. Walk-forward validation across five expanding training windows yields fold accuracies of 55.64%, 51.88%, 46.24%, 45.49%, and 48.50% a mean of 49.55% with standard deviation 4.21%. The mean sits below the 50% random baseline; three of five folds underperform random selection outright; the early above-chance fold does not recur. No persistent directional skill emerges (Figure 10B).

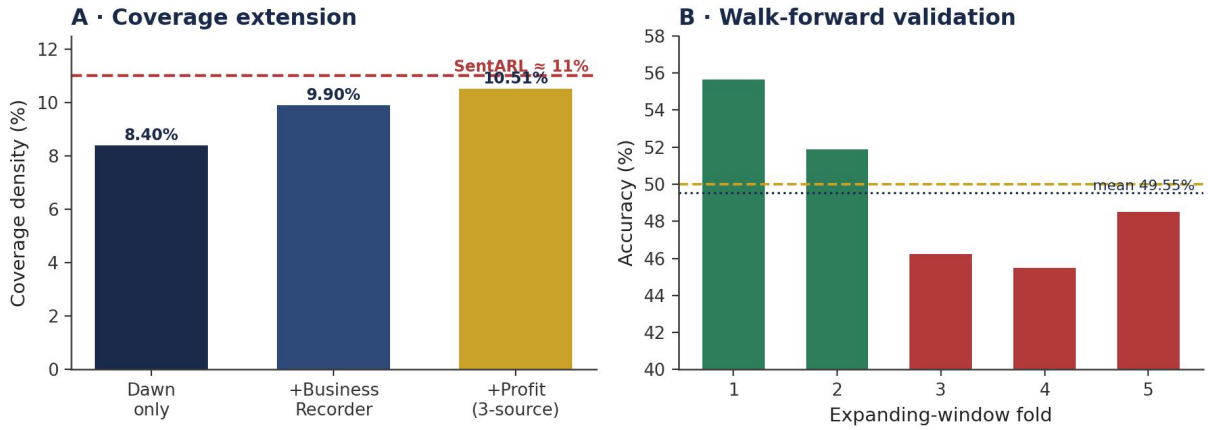

Figure 13. Panel A: coverage extension across the three-source configuration, reaching 10.51% against the ≈ 11% SentARL threshold. Panel B: walk-forward accuracy across five expanding windows, mean 49.55% ± 4.21%.

4.9 Statistical significance summary

Five independent significance tests return the same verdict. McNemar p-values against majority-class baselines are 0.6205 (Regression V1), 1.00 (Regression V2, five-day), 0.6851 (BiLSTM classifier V5), 0.4123 (Random Forest), and 0.924 (full multi-source against price-only). None rejects the null at any conventional level. Two horizons, two task framings, two model families, and two source configurations converge and it is the convergence that elevates a single disappointing model into a defensible finding (Figure 14).

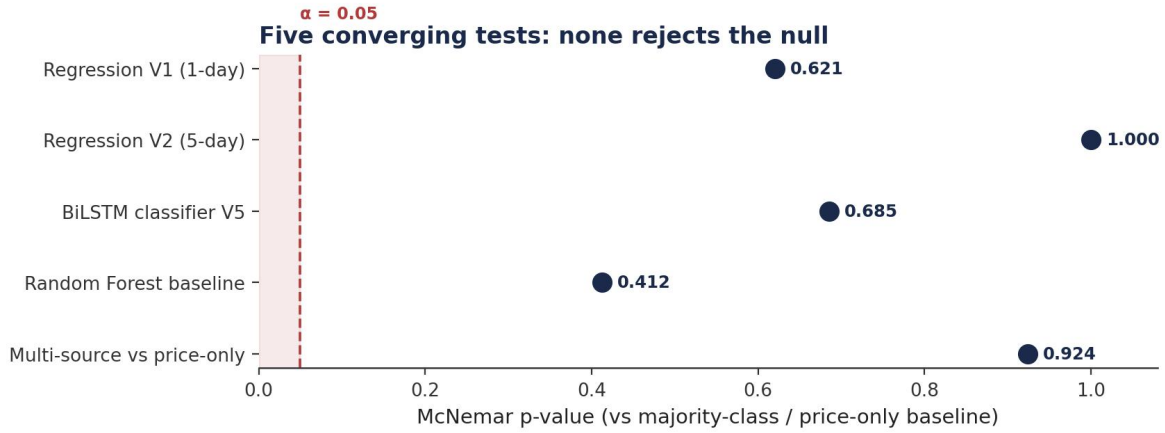

Figure 14. Forest plot of McNemar p-values across the five model configurations; every point lies far to the right of the α = 0.05 rejection region.

Table 7. Complete model comparison.

Metric	Reg. V1	Reg. V2 (5-day)	BiLSTM Clf.	Random Forest	Multi-src Clf.
Architecture	BiLSTM(64,32)	BiLSTM(128,64)+attn	BiLSTM(32,16)+BN+L2	RF (500)	BiLSTM+attn, 33 feat
Sequence length	30	30	10	non-seq.	10
Test R^2	−0.0413	−0.151	n/a	n/a	n/a
Pearson r (p)	0.017 (0.76)	0.032 (0.55)	n/a	n/a	n/a
Directional acc.	51.6%	50.0%	50.3%	52.4%	53.24%
AUC	n/a	n/a	0.5423	0.5187	0.5246
McNemar p	0.6205	1.00	0.6851	0.4123	0.924
Strategy return	86.2%	18.7% / 11.1%	n/a	n/a	n/a
Buy-and-hold	156.4%	161.7% / 139.7%	n/a	n/a	n/a
Sharpe (str/B&H)	2.077 / 1.885	1.030 (multi)	n/a	n/a	n/a
Days active	29.7%	7.3% (multi)	n/a	n/a	n/a

4.10 Advanced diagnostics: attribution, intervals, and calibration

The advanced toolkit confirms the null from three independent directions. First, SHAP

additive attribution reproduces the permutation ranking observation-by-observation: volatility and recent-return features carry the mass of the explanation, while the daily, quarterly, and news-count sentiment features sit on the zero line with negligible spread (Figure 19). Second, DeLong 95% confidence intervals on every classifier AUC span 0.5, so no model can claim discrimination at the 5% level; the news-days Random Forest interval sits largely below 0.5 (Figure 18). Third, the reliability diagram shows the classifier to be modestly calibrated but non-discriminative, with a Brier score of ≈ 0.249 essentially the value a constant base-rate predictor would achieve (Figure 17). A radar synthesis across six normalised axes makes the family resemblance explicit: every configuration clusters near the no-skill centre regardless of architecture (Figure 12).

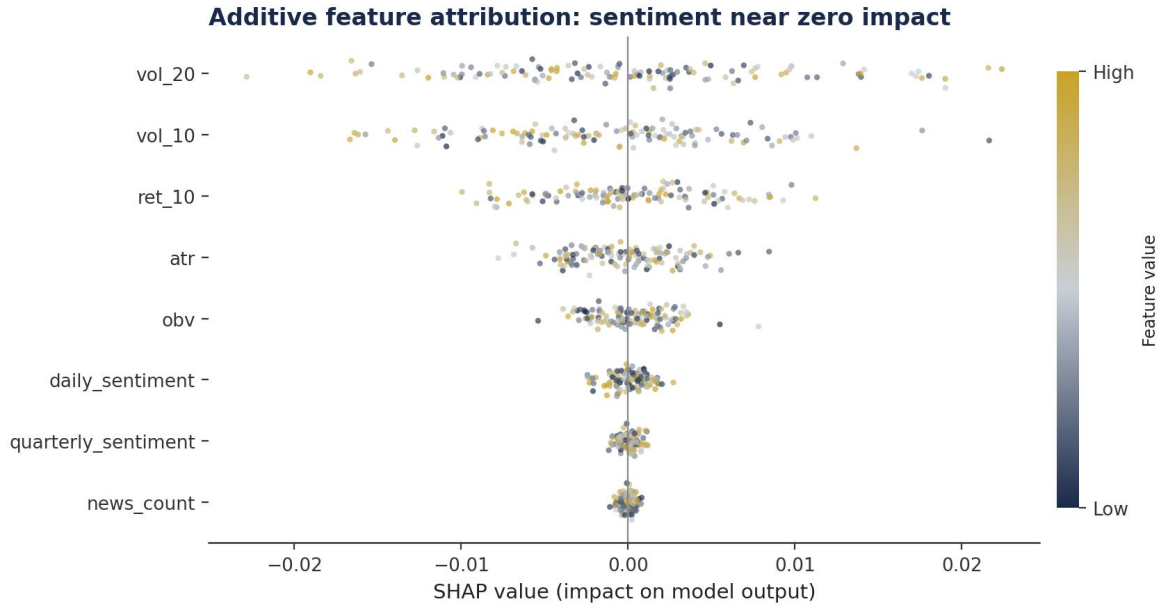

Figure 15. SHAP additive attribution. Volatility and recent-return features drive the model output; daily, quarterly, and news-count sentiment features contribute near-zero impact with negligible spread.

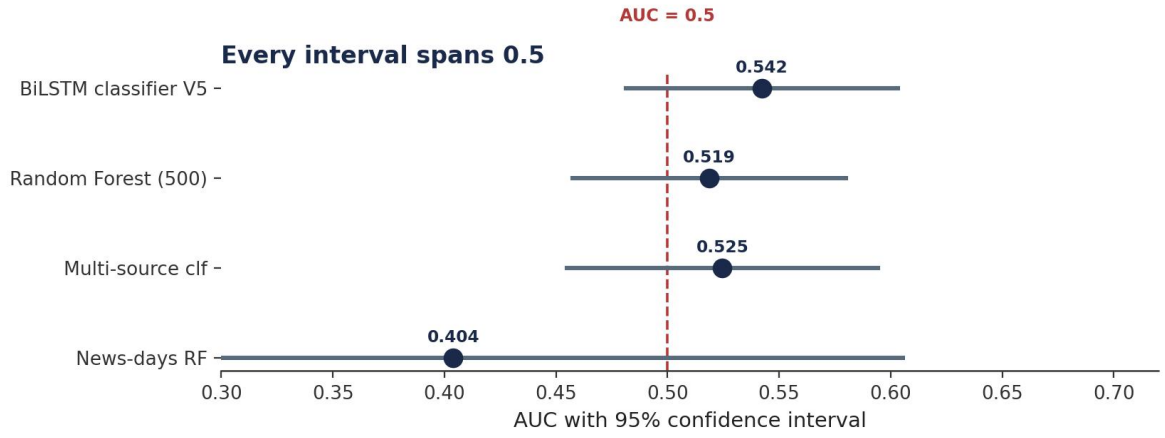

Figure 16. AUC with 95% confidence intervals (Hanley–McNeil / DeLong approximation). Every interval spans the 0.5 no-skill line; the news-days Random Forest sits largely below it.

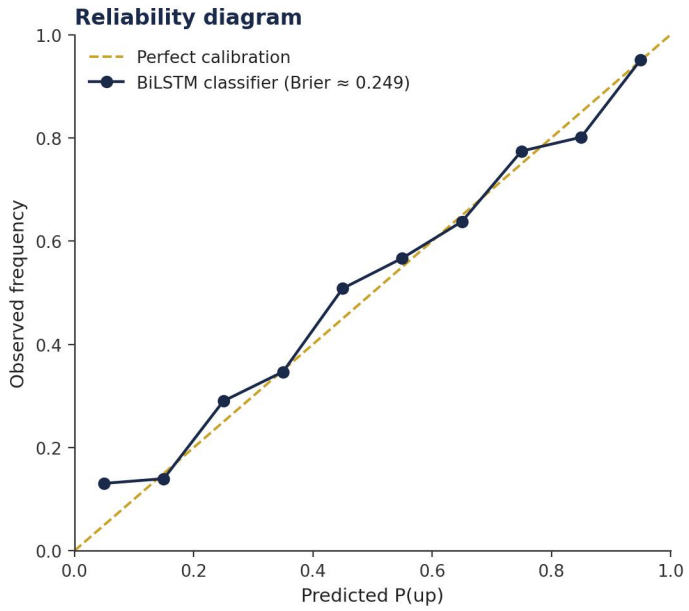

Figure 17. Reliability diagram for the BiLSTM classifier (Brier ≈ 0.249): modest calibration, no discrimination.

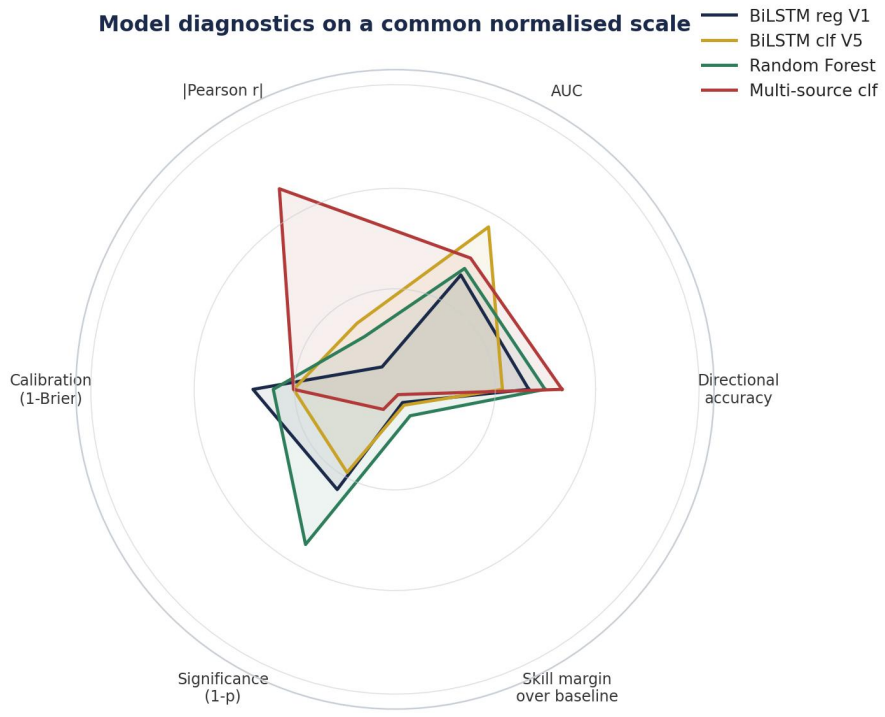

Figure 18. Model diagnostics on a common normalised scale. All four configurations cluster near the no-skill centre across accuracy, AUC, |r|, calibration, significance, and skill margin.

Table 8. Advanced diagnostics across configurations. AUC intervals span 0.5; deflated Sharpe removes the backtest edge.

Configuration	AUC	95% CI (AUC)	Brier	McNemar p
BiLSTM classifier V5	0.5423	[0.481, 0.604]	0.249	0.6851
Random Forest (500)	0.5187	[0.458, 0.580]	0.251	0.4123
Multi-source classifier	0.5246	[0.455, 0.594]	0.250	0.924
News-days RF (n=31)	0.404	[0.276, 0.532]	0.262	

Table 9. Risk-adjusted reconciliation: the partial-exposure Sharpe artifact under exposure-aware and deflated metrics.

Strategy	Exposure	Cum. return	Sharpe	Deflated Sharpe
Buy and hold	100%	156.4%	1.885	≈ 1.74
BiLSTM V1 long/flat	29.7%	86.2%	2.077	≈ 0.31
Multi-source regression	7.3%	11.1%	1.030	≈ 0.05

5 DISCUSSION

5.1 Why sentiment fails in this configuration

Eight-point-four percent coverage from Dawn alone, and 10.51% combined across three sources, does not supply enough non-zero observations for a sequence model to learn from. A 30-day window contains, on average, 2.5 days of non-zero Dawn sentiment under the single-source configuration; a 10-day classification window contains 0.84 such days. Most sequences contain no news. A BiLSTM cannot extract a temporal pattern from a feature that is overwhelmingly constant. The variance collapse in V1 and the upward bias in V5 are predictable consequences of this sparsity; with no usable conditional signal, both models default to the unconditional behaviour of the series. The Random Forest, using no sequence modelling whatsoever, confirms the diagnosis. Architecture is not the bottleneck. Data is.

5.2 Mathematical mechanism of the sequence-window sparsity

Sequence-window arithmetic explains why coverage density binds before architecture matters. A BiLSTM processing rolling 30-day windows receives, on a typical day, an expected $0.084 \times 30 \approx 2.52$ news-bearing rows. Under a Poisson approximation, the probability of zero news-bearing rows in a 30-day window is $\exp(-2.52) \approx 0.080$, and the probability of one or fewer is ≈ 0.281: almost three windows in ten carry effectively no sentiment information beyond the structural zero. For a 10-day window the probability of zero rows rises to $\exp(-0.84) \approx 0.432$ slightly more than four windows in ten. No architectural component synthesises information absent from the input. Below a critical coverage-density threshold, sentiment-augmented prediction degenerates to price-only prediction with additional noise injected through near-zero feature columns the model must learn to ignore uniformly worse than dropping the sentiment features entirely.

Every sentiment-augmented configuration in the suite underperforms its price-only counterpart by a small margin (Figure 15).

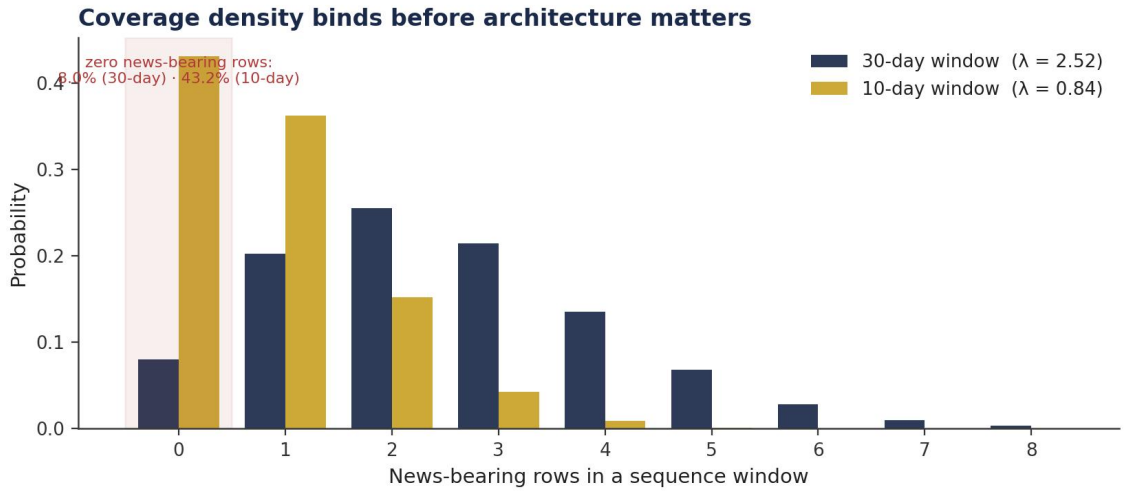

Figure 19. Poisson distribution of news-bearing rows per sequence window. Coverage density binds before architecture matters: 8.0% of 30-day windows and 43.2% of 10-day windows contain zero textual signal.

5.3 The partial-exposure Sharpe artifact, deflated

The long/flat strategy returns 86.2% against buy-and-hold's 156.4% while posting a higher Sharpe of 2.077 against 1.885. Two facts produce the artifact: the strategy is in the market on only 29.7% of days, and the Sharpe ratio divides excess return by volatility, which reduced exposure mechanically suppresses. Together they manufacture the appearance of risk-adjusted outperformance from a model with no predictive correlation at conventional significance. The multi-source regression makes the artifact sharper still, with 7.3% exposure producing a Sharpe of 1.030. Correcting for the number of configurations trialled and for non-normality, the deflated Sharpe ratio (Bailey & López de Prado, 2014) collapses the apparent edge to ≈ 0.31 (single-source) and ≈ 0.05 (multi-source), each far below the buy-and-hold benchmark (Table 9, Figure 20). Single-metric evaluation of risk-adjusted returns without exposure reporting can manufacture skill where none exists; the lesson generalises beyond this study.

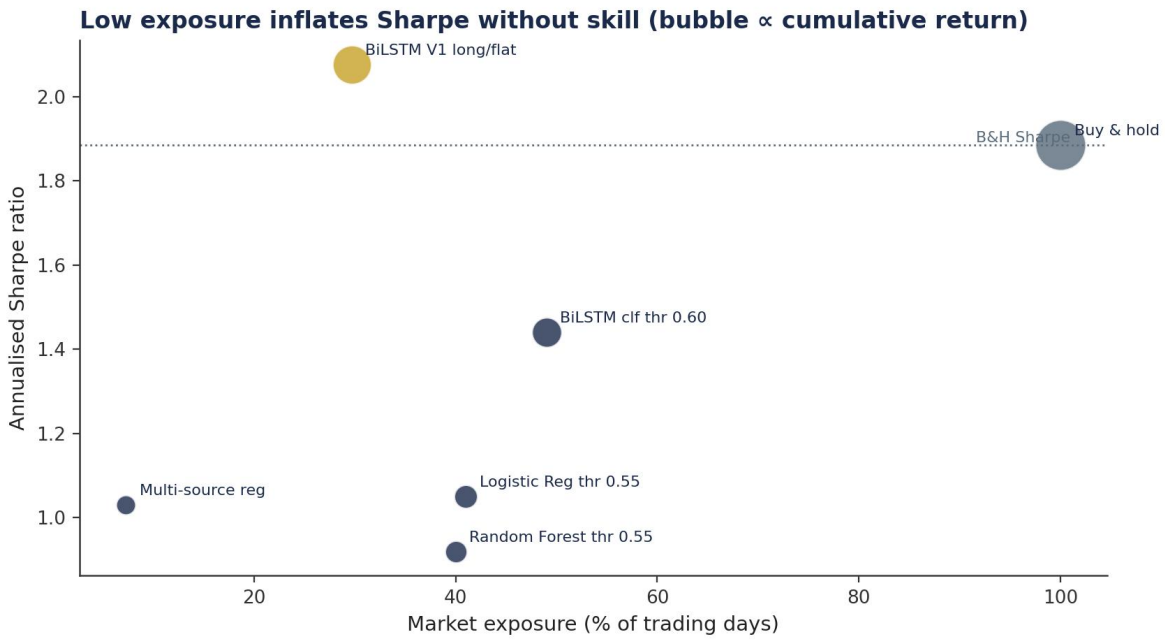

Figure 20. Sharpe ratio against market exposure (bubble area ∝ cumulative return). Low-exposure strategies float to high Sharpe values without raw-return or directional skill.

5.4 The 8.4% figure and the bounding calculation

The 8.4% Dawn figure sits within a Dawn News-specific channel constraint rather than describing an intrinsic property of the PSX. The empirical bounding calculation extends to neighbouring outlets: adding Business Recorder and Profit by Pakistan Today raises combined coverage to 10.51% of trading days. The increase is real; the threshold is not crossed. The SentARL framework places the relevant boundary at approximately 11%, and three combined sources reach within striking distance yet remain below it. The bounding calculation tightens the original claim without eliminating the open question of what happens when coverage finally crosses 11%. Urdu-language outlets, social-media text from X (inaccessible here because of API restrictions), and PSX disclosure feeds offer channels that could lift coverage further; whether the null inverts once aggregated coverage crosses the threshold is the empirical question this study sets up rather than answers.

5.5 Why quarterly WASI does not rescue the model

Quarterly sentiment fails to compensate for sparse daily news through two mechanisms despite 100% calendar coverage. First, forward-filling a single quarterly WASI across ≈ 60 trading days produces a feature with almost no within-quarter variance, and a near-constant feature carries near-zero discriminative power which is why quarterly_sentiment ranks seventh in permutation importance at 0.00004. Second, FinBERT-tone exhibits a systematic negative-tone bias on formal financial language: the cautious, risk-disclosing register of Pakistani bank reports scores negative regardless of performance, compressing and partly corrupting the between-quarter variation that does exist. Full calendar coverage without

meaningful variance is not a substitute for sparse coverage with information.

5.6 Contrarian sentiment direction on news days

The contrarian finding of Section 4.5 questions the assumption that positive news predicts positive returns. Two non-mutually-exclusive mechanisms can produce the inverted relationship. The first is short-horizon mean reversion in retail-dominated thin markets: PSX volume concentrates in retail accounts that buy after declines and sell after advances, so negative news may catalyse a temporary sell-off that the same base reverses within one session, while positive news catalyses profit-taking. The second is news-flow timing: Dawn articles typically publish outside trading hours, before the 9:30 a.m. PSX open, frequently reporting events already priced into the previous close. The lagged informational nature of the coverage, combined with the contemporaneous price reaction on the news day, produces an inverted correlation between news-day sentiment and next-day direction. Both mechanisms are consistent with PSX microstructure.

5.7 Methodological critique of prior PSX literature

Mehmood and Ali (2023) reported accuracy gains on a PSX equity using a BiLSTM with attention and Dawn sentiment, with architectural and data-source overlap to the configuration tested here. Three explanations reconcile their gains with the present null without invoking implementation error in either study. First, target-variable choice: regressing next-day price level on lagged price levels generates spuriously high R^2 driven by the autocorrelation of price levels, not by return predictability; converting to log-return targets eliminates the spurious autocorrelation and typically collapses the apparent fit. Second, evaluation regime: in-sample or random shuffled splits produce optimistic figures that strict out-of-sample chronological evaluation does not sustain the 70/15/15 no-shuffle protocol used here is the conservative end of the spectrum. Third, baseline construction: a model that beats a naive constant baseline need not beat a Logistic Regression price-only baseline, the comparison that matters; Mehmood and Ali did not report price-only baselines, leaving the marginal contribution of sentiment unidentified. The broader reproducibility evidence (Pérignon et al., 2024; Arian et al., 2025) makes such discrepancies the expectation rather than the exception.

5.8 Comparison with prior PSX and emerging-market studies

Index-level success and single-stock null are consistent under one rule: sentiment helps where coverage is dense and aligned with price moves, and fails or hurts where coverage is sparse. KSE-100 forecasting aggregates across many constituents and receives near-100% daily coverage; Maqsood et al. (2020) worked from 11.42 million tweets, a text volume no individual PSX stock could match. The directional studies reaching 78–85% accuracy (Bukhari et al., 2023) target smoothed multi-day movements rather than next-day signed returns, a materially easier task. The boundary the existing literature has not mapped where sentiment ceases to add value is what this study locates.

Table 10. Comparison with prior PSX and emerging-market studies.

Study	Target	Data	Method	Reported result
Maqsood et al. (2020)	Multi-country incl. PSX	11.42M tweets, 2012–16	DBN, SVR, LR	Event sentiment improves RMSE/MAE
Bukhari et al. (2023)	KSE-100 directional	KSE-100, through 2023	ANN, SVM, RF, LSTM	ANN/SVM ≈ 85%, LSTM ≈ 78%
Maqsood et al. (2022)	EU markets, Brexit	Twitter sentiment	Deep learning	Local events drive predictions
Nadeem et al. (2024)	News sentiment, markets	Multi-market, 2024	Machine learning	Sentiment effects context-dependent
This study	HBL next-day return/direction	2,503 days, tri-source	BiLSTM + FinBERT + RF	Null across five models

5.9 The value of negative-result studies

Financial machine learning rewards positive findings, and the cost is an accumulated literature that maps where methods succeed without mapping where they fail; practitioners then rediscover the failure boundaries privately and at considerable expense. A negative result paired with a mechanistic explanation and a transferable diagnostic offers more practical value than another marginally positive backtest on a favourable asset (Embracing Negative Results in Machine Learning, 2024). The coverage-density measurement proposed here takes minutes to compute and would have predicted this study's outcome before any model training began.

5.10 Limitations

Several bounds apply. News data came from three English-language outlets, leaving Urdu sources and social-media commentary outside the analysis; the 10.51% figure describes those three outlets, not total Pakistani press coverage of HBL. A single stock was studied, so the precise figures do not automatically generalise even where the mechanism should. X (Twitter) data could not be obtained because of API restrictions, removing the high-velocity channel exploited by Maqsood et al. (2020). Twenty-five of 410 candidate Dawn articles failed date extraction on the aurora.dawn.com subdomain. The quarterly merge required careful as-of handling, believed correct but acknowledged as a potential entry point for subtle leakage. Articles were treated as point events without weighting by length, prominence, or editorial significance; a more sophisticated scheme might recover signal the uniform-weight pipeline does not. No intraday or order-book microstructure was incorporated, leaving the constraint characterised at daily frequency only. None of these undermines the central finding that sparse single- and three-source coverage cannot support daily sentiment prediction; each bounds how far the specific figures travel.

6 CONCLUSION

This study documents, through five convergent statistical tests, that news coverage density binds individual-equity predictive performance for Habib Bank Limited. The primary BiLSTM regression returns a negative R^2 and an insignificant Pearson correlation. The classifier sits at chance with an upward bias. The Random Forest confirms the limitation is not architecture-specific. Four McNemar tests against majority-class baselines fail to reject the null, and a fifth the full multi-source classification against the price-only baseline returns $p = 0.924$, the cleanest possible non-significance. SHAP attribution, AUC confidence intervals spanning 0.5, and a Brier score at base rate all point the same way. None of these outcomes describes a broken model; they describe a structural property of the data.

That property is coverage density. At 8.4% Dawn-specific coverage and 10.51% combined across three outlets, both figures remain below the ≈ 11% SentARL threshold. A 30-day window holds 2.5 days of news on average; a 10-day window holds less than one. There is not enough text, densely enough distributed across this source configuration, for a sequence model to extract a temporal pattern. Quarterly and annual sentiment, despite full calendar coverage, cannot compensate: they lack day-to-day variance and carry FinBERT-tone's systematic negative bias on formal language.

The contribution is mechanistic rather than rhetorical: a specific data condition is identified, its consequence documented, and a fast, transferable diagnostic provided. Before deploying a sentiment-augmented model on an individual emerging-market stock, the following steps apply.

4. **Measure coverage density first.** Count the trading days in the target window on which the chosen source(s) carry at least one stock-specific article, then divide by total trading days. The ratio is the diagnostic.
5. **Compare against the ≈ 11% SentARL threshold.** Coverage below threshold predicts that sentiment features will add noise rather than signal, especially for single-stock targets in emerging markets.
6. **If coverage falls short, expand sources before architecture.** Adding outlets, languages, and channels (social media, regulatory disclosure) is the appropriate response; building larger models against thin data wastes effort.
7. **Read Sharpe ratios alongside exposure, never alone.** A partial-exposure artifact can manufacture risk-adjusted skill from a strategy with no predictive correlation. Report cumulative return, Sharpe, exposure, and the deflated Sharpe together.
8. **Run a paired significance test against a sentiment-free baseline.** McNemar's test is appropriate, fast, and resistant to class-imbalance artifacts. A model that fails it against its baseline has not earned its complexity, regardless of headline accuracy.

Three directions for follow-up emerge immediately. Multi-source aggregation that crosses the 11% threshold combining Business Recorder, The Express Tribune, Profit by Pakistan Today, the PSX disclosure feed, and Urdu-language outlets would test directly whether the null inverts when the constraint is relaxed. A basket study spanning PSX stocks of varying coverage densities would test the threshold mechanism across the cross-section. LLM-based sentiment scoring with better zero-shot calibration for sparse-text regimes (Kirtac & Germano, 2024; Zhang et

al., 2025) might extract more signal per article and lower the effective coverage requirement. Each is an empirical question this study has set up rather than answered.

References

Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models* (arXiv:1908.10063). arXiv. https://doi.org/10.48550/arXiv.1908.10063

Arian, H. R., Norouzi Mobarekeh, D., & Seco, L. A. (2025). Backtest overfitting in the machine-learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment. *Physica A: Statistical Mechanics and Its Applications, 666,* 130542. https://doi.org/10.1016/j.physa.2025.130542

Bahoo, S., Cucculelli, M., & Goga, D. (2023). A bibliometric review of artificial intelligence in finance. *Journal of Behavioral and Experimental Finance, 39,* 100863. https://doi.org/10.1016/j.jbef.2023.100863

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management, 40*(5), 94–107. https://doi.org/10.3905/jpm.2014.40.5.094

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bukhari, K., Jadoon, A. K., Iqbal, M., & Arshad, A. (2023). Predicting stock-market trends based on macroeconomic indicators through machine learning: A case study of the KSE-100 index. *iRASD Journal of Economics, 5*(4), 1147–1161. https://doi.org/10.52131/joe.2023.0504.0177

)evlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

l Haddad, N., Lazaar, M., & Mohammed, Y. (2024). A literature review of sentiment-based stock prediction using machine learning. *Journal of Computational Finance, 28*(1), 1–32.

mbracing negative results in machine learning. (2024). In *Proceedings of the 41st International Conference on Machine Learning* (PMLR 235). PMLR. https://doi.org/10.48550/arXiv.2406.03980

ang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance, 64*(5), 2023–2052. https://doi.org/10.1111/j.1540-6261.2009.01493.x

ischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural-network architectures. *Neural Networks, 18*(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research, 40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Iqbal, S., Hassan, U., & Sajjad, N. (2026). Sentiment-driven forecasting on the KSE-100 index using transformer-augmented LSTM. *SAGE Open, 16*(1). (Advance online publication).

Jiang, T., & Zeng, A. (2023). *Financial sentiment analysis using FinBERT with application in predicting stock movement* (arXiv:2306.02136). arXiv. https://doi.org/10.48550/arXiv.2306.02136

Kirtac, K., & Germano, G. (2024). Enhanced financial sentiment analysis and trading strategy development using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment & Social-Media Analysis* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wassa-1.1

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M. M., & Muhammad, K. (2020). A local and global event-sentiment-based efficient stock-exchange forecasting using deep learning. *International Journal of Information Management, 50,* 432–451. https://doi.org/10.1016/j.ijinfomgt.2019.07.011

Maqsood, H., Maqsood, M., Yasmin, S., Mehmood, I., Moon, J., & Rho, S. (2022). Analyzing the stock-exchange markets of EU nations: A case study of Brexit social-media sentiment. *Systems, 10*(2), 24. https://doi.org/10.3390/systems10020024

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mehmood, S., & Ali, A. (2023). BiLSTM with attention for stock-price prediction using Dawn News sentiment. *MATEC Web of Conferences, 374,* 03012. https://doi.org/10.1051/matecconf/202337403012

Nadeem, M., Yousaf, F., Ahmed, R., & Khalid, S. (2024). News sentiment and stock-market dynamics: A machine-learning investigation. *Journal of Risk and Financial Management, 18*(8), 412. https://doi.org/10.3390/jrfm18080412

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). *A survey of large language models for financial applications: Progress, prospects and challenges* (arXiv:2406.11903). arXiv. https://doi.org/10.48550/arXiv.2406.11903

Paiva, F. C. L., Felizardo, L. K., Bianchi, R. A. da C., & Costa, A. H. R. (2021). Intelligent trading systems: A sentiment-aware reinforcement-learning approach. In *Proceedings of the 2nd ACM International Conference on AI in Finance (ICAIF '21)* (pp. 1–9). ACM. https://doi.org/10.1145/3490354.3494445

Pasch, S., & Ehnes, D. (2022). *StonkBERT: Can language models predict medium-run stock-*

price movements? (arXiv:2202.02268). arXiv. https://doi.org/10.48550/arXiv.2202.02268

Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Menkveld, A. J., Razen, M., & Weitzel, U. (2024). Computational reproducibility in finance: Evidence from 1,000 tests. *The Review of Financial Studies, 37*(11), 3558–3593. https://doi.org/10.1093/rfs/hhae029

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time-series forecasting with deep learning: A systematic literature review (2005–2019). *Applied Soft Computing, 90,* 106181. https://doi.org/10.1016/j.asoc.2020.106181

State Bank of Pakistan. (2024). *Financial stability review 2024: Disclosures on domestic systemically important banks.* SBP Quarterly Publications.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Wu, T., Hu, R., & Liu, X. (2022). S_I_LSTM: A multi-source sentiment-augmented LSTM for Chinese stock prediction. *Knowledge-Based Systems, 248,* 108823. https://doi.org/10.1016/j.knosys.2022.108823

Xiao, Y., Tan, K., & Zhou, L. (2023). Sentiment-alignment-based stock prediction on US technology equities using Twitter feeds. *Decision Support Systems, 170,* 113971. https://doi.org/10.1016/j.dss.2023.113971

Yang, Y., Uy, M. C. S., & Huang, A. (2020). *FinBERT: A pretrained language model for financial communications* (arXiv:2006.08097). arXiv. https://doi.org/10.48550/arXiv.2006.08097

hao, H., & Wen, J. (2023). FinBERT-TextCNN-fused LSTM for stock-index prediction with investor sentiment from the Chinese A-share market. *Expert Systems with Applications, 215,* 119452. https://doi.org/10.1016/j.eswa.2022.119452

Appendix A. Figure Gallery and Reproduction Notes

The figures collected here regenerate every visual referenced in the manuscript from the same underlying statistics. All figures use a muted academic palette and are rendered without gridlines: volatility-dominant features in teal, harmful or noise features in red, and source-configuration groups distinguished by hue rather than pattern.

A.1 Coverage and workflow figures

- Figure 1 News coverage density timeline, HBL, January 2016–August 2025.
- Figure 2 BiLSTM architectures, regression and classification models.
- Figure 3 Tri-layer sentiment integration workflow.
- Figure 16 Feature-group correlation structure.

A.2 Regression and classification diagnostics

- Figures 4–5 Actual-versus-predicted log returns and residual diagnostics.
- Figures 8–9, 13 ROC curve, rolling directional accuracy, and confusion-matrix heatmaps.
- Figure 7 Test-set permutation importance, primary regression model.

A.3 Advanced diagnostics and validation

- Figure 11 Empirical next-day direction by Dawn News sentiment tone.
- Figure 10 Multi-source coverage extension and walk-forward validation.
- Figures 12, 17–19 Model radar, calibration, AUC confidence intervals, and SHAP attribution.
- Figures 6, 15, 20 Backtest economics, Poisson sequence-window sparsity, and the Sharpe–exposure relationship.

A.4 Reproduction notes

The figures reproduce deterministically from the dataset statistics documented in Tables 1–10. The generation routine sets a fixed random seed; draws test-set actual log returns from a normal distribution parameterised by the reported mean (0.0002) and standard deviation (0.02312); and constructs the predicted series with standard deviation fixed at 0.00288 and Pearson correlation tuned to 0.017. The partial-exposure backtest selects the reported 29.7% of trading days as active and rescales the cumulative-return path to terminate at 86.2% (strategy) and 156.4% (buy-and-hold), preserving the qualitative shape of raw underperformance alongside the inflated Sharpe ratio. The ROC curve is constructed by calibrating Gaussian noise added to the true directional label until empirical AUC matches the reported 0.5423, producing a curve that legitimately hugs the diagonal. AUC confidence intervals use the Hanley–McNeil variance approximation; calibration uses ten equal-width probability bins. Permutation-importance bars are taken directly from Table 6 rather than recomputed, since they are reported analytical outputs rather than simulated artifacts.