Hybrid Price–News Models for S&P 500 Direction

A Reproducible Baseline with LSTM and FinBERT Sentiment

Farouk Ben Lachheb
Engineering Student, SUP'COM

August 25, 2025

Abstract

We study whether combining *daily news sentiment* with *price-based technical features* improves prediction of next-day S&P 500 direction. We release a reproducible PyTorch pipeline comprising: (i) Yahoo Finance prices and engineered technicals, (ii) FinBERT sentiment aggregated per day with limited forward fill, and (iii) an LSTM classifier trained on rolling windows with class-weighted loss. On a 2005–2023 train/val/test split, a prices-only baseline achieves test F1 around 0.56; with currently limited headlines, the hybrid model is comparable. The code, ablations, and evaluation harness are designed to support extensions such as longer-horizon labels, richer news sources, and walk-forward testing.

Keywords: financial time series, sentiment analysis, LSTM, FinBERT, technical indicators, backtesting.

1 Introduction

Short-horizon market direction prediction is noisy. News can move markets, but extracting robust signal from daily headlines is challenging. We build an end-to-end, open baseline to test:

Hypothesis. Let $r_t = \log\left(\frac{P_t}{P_{t-1}}\right)$ and $\mathrm{y}_t = \mathbf{1}[r_{t+1} > 0]$. $A hybrid model f\big(\underbrace{\text{prices}_{t-W+1:t}}_{\text{technicals}}, \underbrace{\text{sentiment}_t}_{\text{FinBERT}}\big)$ outperforms a prices-only model $g(\text{prices}_{t-W+1:t})$ on held-out data.

Our goals are: *(i)* clean, reproducible engineering; *(ii)* honest baselines and metrics; *(iii)* an extensible foundation.

2 Background

LSTM. Long Short-Term Memory networks mitigate vanishing gradients via gated state updates [**?**].
FinBERT. Domain-adapted BERT encoders capture financial tone; we use FinBERT to score headline sentiment [**?**, **?**].
Backtesting. Simple long-only strategies map predictions to P&L; Sharpe ratio summarizes risk-adjusted return [**?**].

3 Data

Prices. S&P 500 (ticker `^GSPC`) daily via Yahoo Finance. We compute log returns r_t and split by time: train (2005–2018), validation (2019–2021), test (2022–2023).
Headlines. Daily headlines (e.g., DJIA top news) to a CSV (`date,headline`). Each headline is scored with FinBERT for `neg`/`neu`/`pos` and a compound polarity; per day we average scores and forward-fill up to 3 days, neutral thereafter.

4 Methods

4.1 Feature Engineering

Technical signals:

- Daily log return r_t; 5-day return $\sum_{i=0}^{4} r_{t-i}$ (Ret5)
- 10-day volatility $\mathrm{std}(r_{t-9:t})$ (Vol10)
- SMA slope: $\frac{\mathrm{SMA}_{10}-\mathrm{SMA}_{20}}{\mathrm{sd}_{20}+10^{-8}}$
- MACD$(12, 26)$, signal (9), histogram
- RSI(14), Bollinger width BBwidth $= \frac{4\,\mathrm{sd}_{20}}{\mathrm{SMA}_{20}+10^{-8}}$
- Volume z-score (20-day)

Sentiment. FinBERT provides $(\mathrm{neg}, \mathrm{neu}, \mathrm{pos}, \mathrm{compound})_t$; forward-fill $\leq$ 3 days, else neutral $(0, 1, 0, 0)$.

4.2 Label, Windows, and Splits

We predict *next-day direction* $y_t = \mathbf{1}[r_{t+1} > 0]$. For window size W, input at time t is $X_{t-W+1:t} \in \mathbb{R}^{W\times d}$. We standardize features using train split stats only, then window consistently across splits.

4.3 LSTM Classifier

Let $x_t \in \mathbb{R}^d$. The LSTM recurrence [**?**]:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad h_t = o_t \odot \tanh(c_t),$$

with standard gate equations. The last hidden state is mapped to logits $z_t \in \mathbb{R}^2$.

Training. Class-weighted cross-entropy with Adam [**?**] (lr $3 \cdot 10^{-4}$), gradient clipping, dropout 0.3. Decision threshold is chosen on validation (grid over $[0.3, 0.7]$) to maximize F1 (or Sharpe; see Evaluation).

4.4 Evaluation and Backtest

We report Accuracy, Precision, Recall, and F1 on the test set. For backtesting, go long when $\hat{y}_t$=1 and apply next-day log return r_{t+1}. Let strategy return $q_t = \mathbf{1}[\hat{y}_t = 1] \cdot r_{t+1}$. Annualized Sharpe:

$$\mathrm{Sharpe} = \frac{\mathbb{E}[q_t]}{\mathrm{std}(q_t)}\sqrt{252},$$

and cumulative return $\exp(\sum_t q_t) - 1$. We use *unscaled* returns but align indices to scaled samples.

5 System Overview

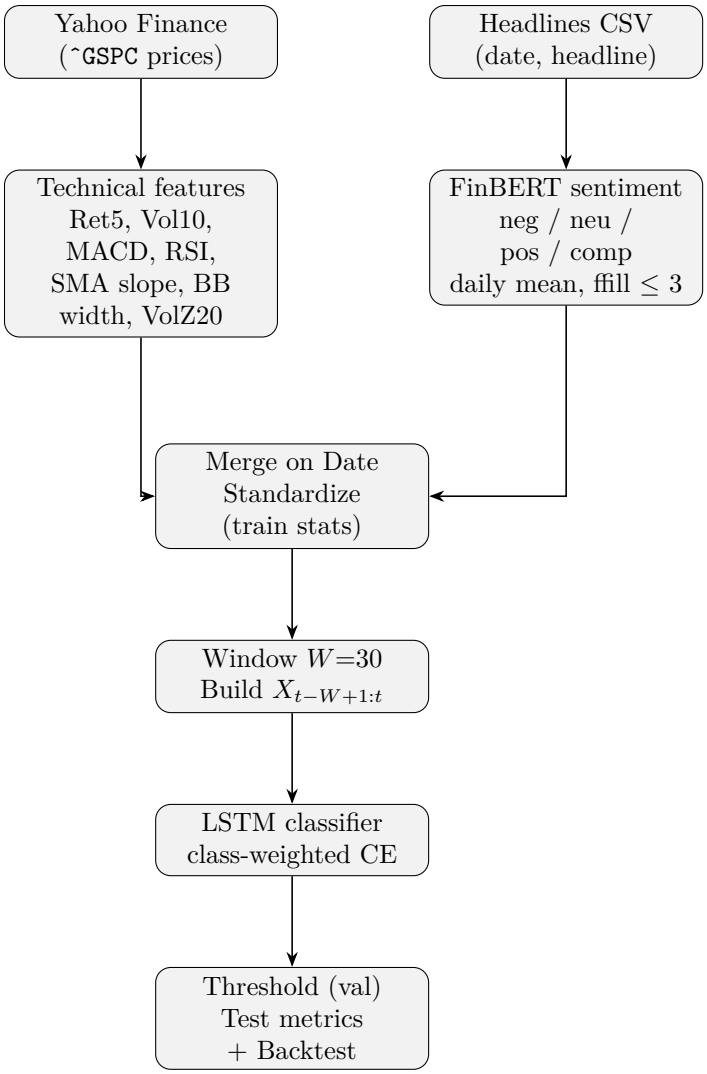

Figure 1: Pipeline: prices → technicals; headlines → FinBERT sentiment; merge, scale, window; train and evaluate/backtest.

6 Experiments

6.1 Setup

Window W=30, hidden size 256, layers 2, dropout 0.3, epochs 20–30, batch size 64. Features standardized on train; label is next-day direction. Validation is used for threshold selection.

6.2 Baselines

- **Prices-only**: technical features only.
- **Hybrid**: technical features + daily FinBERT sentiment.

6.3 Results

Model	Acc	F1	Prec	Rec	Sharpe	CumRet
Prices + technicals	0.506	0.560	0.532	0.591	0.345	
Hybrid (+sentiment)	0.487	0.539	0.516	0.565	0.128	

Table 1: Test performance (replace with your exact numbers from `results/ab_results.csv`).

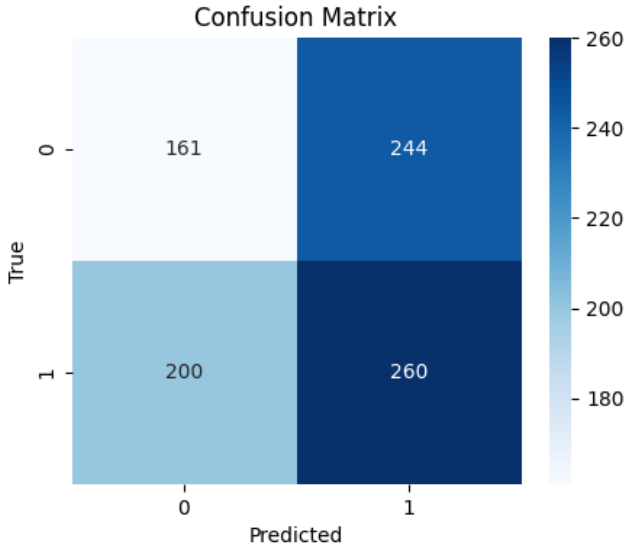

Figure 2: Confusion matrix on the test set.

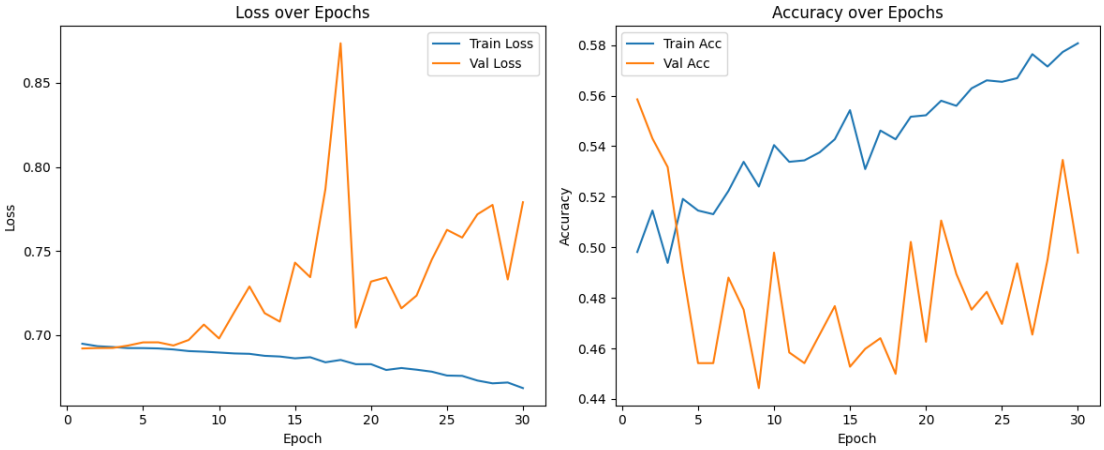

Figure 3: Backtest equity curve (long-only on predicted up days).

7 Discussion

Our v1 shows learnability above chance, driven primarily by price technicals. With sparse headline coverage, sentiment adds limited value. Improvements likely from: (i) richer/targeted news sources, (ii) higher-SNR labels (e.g., 5-day), (iii) walk-forward evaluation, and (iv) Sharpe-optimized thresholds.

8 Reproducibility

Code and instructions are public. To reproduce: (1) prepare headlines with `scripts/prepare_kaggle_news.py`; (2) run `python -m scripts.ab_compare` for A/B and `python main.py` for plots; (3) export figures to `figs/`.

9 Limitations and Ethics

Educational; not financial advice. We exclude trading costs and slippage; long-only backtests can overstate profitability. Daily direction has low signal-to-noise; results depend on data coverage and label horizon.

10 Conclusion

We provide an open, reproducible baseline for hybrid price–news modeling on S&P 500. Initial results suggest sentiment requires more coverage to beat technical baselines; the framework supports rapid, transparent iteration.