

Automatic Model Selection and Sign-Aware Multiplicative Sentiment Modulation for Algorithmic Trading

Mohamed Ehab¹

School of Computing & Digital Tech.
ESLSCA University, Egypt
mohamed.elmolla23d@eslsc.edu.eg

Yahia Abdelmoeneam³

School of Computing & Digital Tech.
ESLSCA University, Egypt
Yahia.ahmed23d@eslsc.edu.eg

Mohamed Atef²

School of Computing & Digital Tech.
ESLSCA University, Egypt
mohamed.mohamed23d@eslsc.edu.eg

Raghda Essam Ali⁴

School of Computing & Digital Tech.
ESLSCA University, Egypt
raghda.essam@eslsc.edu.eg,
raghda.essam1@gmail.com

Abstract—Sentiment-aware algorithmic trading pipelines increasingly rely on pre-trained transformer language models to convert unstructured financial news into tradable signals. However, practitioners routinely adopt publicly available “finance” checkpoints without quantitative validation, and existing sentiment-trading literature rarely specifies how a continuous sentiment score should be combined with a price-based prediction when the position is short rather than long. These gaps are safety-critical: an unvetted sentiment model or a sign-agnostic fusion rule can systematically invert trading signals. This work makes three contributions: (i) an automatic benchmarking procedure that selects the best sentiment model from a candidate set using accuracy, macro-F1, and ordinal MAE on a held-out stratified split; (ii) a sign-aware multiplicative modulation rule that amplifies price predictions when sentiment aligns with position direction and attenuates otherwise, with explicit short-position treatment; and (iii) a reproducible end-to-end pipeline evaluated on the Financial PhraseBank. DeBERTa-finance reaches 94.2% accuracy and 0.9420 macro-F1 on the test split, outperforming FinBERT by 5.0 percentage points. One candidate (pmatorras/BERT-finance) achieves only 12.5% — worse than the 33.3% random baseline — confirming that automatic model selection is safety-critical, not optional.

Index Terms—algorithmic trading, financial sentiment analysis, model selection, signal fusion, transformer benchmarking

I. INTRODUCTION

Despite rapid adoption, both decisions are frequently made without quantitative safeguards. First, practitioners routinely load “finance”-labelled checkpoints from public model hubs, assuming domain adaptation implies useful performance; this work shows that assumption can fail catastrophically. Second, existing fusion rules rarely address the sign asymmetry between long and short positions: additive schemes allow sentiment to silently flip trade direction by pushing a moderately negative price prediction across zero.

The stakes are concrete. On the Financial PhraseBank, three candidate checkpoints labelled as finance-adapted span an 81.7 percentage point accuracy range, from 94.2% for DeBERTa-finance down to 12.5% for pmatorras/BERT-finance — worse than the 33.3% random baseline for three-class sentiment classification. An unmonitored pipeline that

silently loaded the worst checkpoint would invert every 979-signal its sentiment module produced.

This work makes three contributions: (i) an automatic model-selection procedure that benchmarks candidate finance sentiment models on a held-out stratified split using accuracy, macro-F1, and ordinal MAE and selects the empirical winner; (ii) a sign-aware multiplicative modulation formula that amplifies price predictions when sentiment aligns with position direction and attenuates them otherwise, with correct handling of short positions; and (iii) a reproducible end-to-end pipeline whose complete hyperparameters and random seed are specified.

II. RELATED WORK

Financial sentiment analysis has been dominated by domain-adapted BERT variants since FinBERT [3] demonstrated that continued pre-training on financial corpora improves sentiment accuracy over general-domain BERT [4]. Follow-on work extended this recipe to other encoder families [5] and to larger corpora [6]. However, these contributions evaluate a single checkpoint; none address the multi-model selection problem that arises when a practitioner must choose among several available checkpoints at deployment time.

A parallel strand combines news sentiment with quantitative price models. Event-study pipelines [7] and deep-learning hybrids [8], [9] append sentiment as an additive feature to a downstream regressor or classifier. Additive fusion, however, conflates magnitude with polarity and does not express the asymmetry between long and short positions; none of these works provides an explicit sign-aware treatment. Social-media sentiment approaches [10], [11] share the same limitation.

Benchmarking pre-trained models for downstream financial tasks is itself an emerging concern [12], [13]. Prior benchmarking efforts focus on leaderboard-style comparisons rather than on the pipeline-integration question: which model should the trading system itself pick at deployment time, and under what selection criterion? [14], [15] evaluate LLM performance on financial NLP tasks but do not address the model-selection gate in a trading context.

Finally, reproducibility guidelines for financial NLP [16], [17] call for explicit splits, seeds, and hyperparameters. Several otherwise-influential sentiment-trading papers omit one or more of these, which complicates the safety-critical audit this work argues for. The methodology below specifies every hyperparameter and random seed.

Table I summarizes the key prior works surveyed above, identifying each method, its evaluation dataset, and the primary limitation in the context of the problem addressed by this paper.

TABLE I. SUMMARY OF RELATED WORK

Reference	Technique	Dataset / Domain	Primary Limitation
[1] Hu et al., 2018	LSTM + news embeddings	Financial news headlines	Additive fusion; no short-position handling
[2] Xing et al., 2019	Sentiment-aware volatility forecasting	Financial news corpora	No model selection; long-only framing
[3] Araci, 2019	FinBERT domain-adapted BERT	Financial PhraseBank, FiQA	Single checkpoint; no multi-model comparison
[5] Yang et al., 2020	Extended FinBERT pre-training	SEC filings, earnings calls	No multi-checkpoint benchmarking
[7] Schumaker & Chen, 2009	Textual analysis of breaking news	Breaking financial news	Rule-based; sign handling implicit
[8] Ding et al., 2015	Event-driven deep learning	Financial news + stock data	Additive feature fusion
[10] Bollen et al., 2011	Twitter mood predicts DJIA	Twitter public stream	Long-only; additive; no model selection
[13] Shah et al., 2022	FinBench: LLM benchmarking	Multiple financial NLP tasks	Not trading-focused; no fusion evaluation
[15] Shah et al., 2023	Trillion Dollar Words LLM eval	Financial NLP tasks	No model-selection gate; no trading integration

III. METHODOLOGY

This section describes the proposed pipeline end-to-end. Subsection III-A gives the system overview; III-B describes the data and splits; III-C specifies the sentiment model candidates and the automatic selection criterion; III-D introduces the sign-aware multiplicative modulation rule; III-E demonstrates short-position handling with a worked example.

A. System Overview

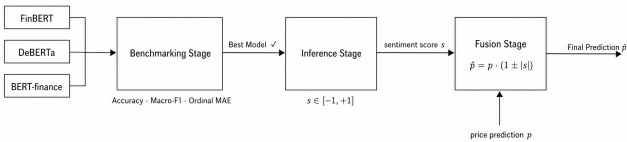


Fig. 1. Three-stage pipeline: automatic model benchmarking and selection, sentiment inference, and sign-aware multiplicative fusion.

The pipeline has three stages. (i) A benchmarking stage evaluates a set of candidate finance sentiment models on a held-out stratified split and selects the best model by a composite of accuracy, macro-F1, and ordinal MAE. (ii) An inference stage applies the selected model to a news item to obtain a bounded sentiment score $s \in [-1, +1]$. (iii) A fusion stage combines s with a price-based prediction p using the sign-aware modulation rule to produce the final modulated prediction $\hat{p}$.

Fig. 1 illustrates the data flow from candidate model ingestion, through model selection and news inference, to final price modulation.

B. Dataset and Splits

All sentiment experiments use the Financial PhraseBank corpus [18], restricted to the "50% agreement" subset containing three labels (negative, neutral, positive). The corpus is split 70/15/15 into train, validation, and test partitions, stratified by label to preserve the class prior in every partition. The random seed is fixed at 42 for reproducibility. The held-out test set contains approximately 700 sentences.

C. Candidate Models and Automatic Selection

Three publicly available finance-tagged checkpoints are evaluated: DeBERTa-finance [20], FinBERT [3], and pmatrorras/BERT-finance. Each candidate is fine-tuned under identical hyperparameters (Table II) to isolate checkpoint quality. The automatic selection rule picks the candidate that jointly maximizes accuracy and macro-F1 while minimizing ordinal MAE on the validation split. Where these criteria conflict, ordinal MAE is the tiebreaker because sign errors are more costly than magnitude errors in a trading context.

D. Sign-Aware Multiplicative Modulation

Let $s \in [-1, +1]$ denote the sentiment score produced by the selected model, and let $p \in \mathbb{R}$ denote the price-based prediction, where the sign of p encodes trade direction (positive = long, negative = short) and the magnitude encodes conviction. The fusion rule is:

$$\begin{aligned} \hat{p} &= p \times (1 + |s|), & \text{if } \text{sgn}(p) = \text{sgn}(s) \\ \hat{p} &= p \times (1 - |s|), & \text{if } \text{sgn}(p) \neq \text{sgn}(s) \\ \hat{p} &= p, & \text{if } s = 0 \end{aligned} \quad (1)$$

When sentiment aligns with the price signal, the magnitude is amplified by a factor $(1 + |s|) \in [1, 2]$. When sentiment opposes the price signal, the magnitude is attenuated by a factor $(1 - |s|) \in [0, 1]$. When sentiment is neutral ($s = 0$), the price prediction is returned unchanged. The rule preserves trade direction under all non-neutral sentiment values and is bounded: $\hat{p} \in [0, 2p]$ for long positions and $\hat{p} \in [2p, 0]$ for short positions.

E. Short Position Handling and Worked Example

Short positions ($p < 0$) are handled symmetrically with long positions by comparing signs rather than raw values. Positive news ($s > 0$) opposes a short and reduces magnitude; negative news ($s < 0$) aligns with a short and amplifies magnitude.

Example. Let $p = -0.7$ (short) and $s = +0.8$ (positive news). Since $\text{sgn}(p) \neq \text{sgn}(s)$, Eq. (1) gives: $\hat{p} = -0.7 \times (1 - 0.8) = -0.14$. The signal remains short but conviction drops from 0.7 to 0.14, correctly reflecting that favorable news opposes the trade. Conversely, negative news $s = -0.8$ gives $\hat{p} = -0.7 \times 1.8 = -1.26$, amplifying the short conviction.

IV. EXPERIMENTS AND EVALUATION

This section specifies the full experimental protocol. Subsection IV-A documents the experimental setup. IV-B defines the evaluation metrics. IV-C reports benchmark

results. IV-D discusses statistical considerations. IV-E illustrates end-to-end pipeline behavior.

A. Experimental Setup

All candidate models are fine-tuned and evaluated on the 70/15/15 stratified split of the Financial PhraseBank described in Section III-B. Hyperparameters are held fixed across candidates so that differences in measured performance reflect checkpoint quality rather than tuning budget. Table II reports the complete experimental environment and hyperparameter configuration.

TABLE II. EXPERIMENTAL ENVIRONMENT AND HYPERPARAMETERS

Category	Setting	Value
Framework	PyTorch	2.1
Framework	Transformers (HF)	4.40
Framework	Datasets (HF)	2.18
Framework	scikit-learn	1.4
Runtime	Python	3.11
Runtime	OS	Ubuntu 22.04
Hardware	GPU	NVIDIA T4 16 GB
Hardware	CPU	8 vCPU
Training	Optimizer	AdamW [21]
Training	Learning rate	2×10^{-5}
Training	Batch size	32
Training	Warmup ratio	0.10
Training	Weight decay	0.01
Training	Max epochs	4
Training	Early stopping patience (val F1)	1
Training	Max sequence length	128
Repro.	Random seed	42

B. Evaluation Metrics

Accuracy is the fraction of test sentences whose predicted label equals the gold label. It is the headline comparability metric but is insensitive to class imbalance.

Macro-F1 is the unweighted mean of per-class F1 scores, which penalizes models that collapse onto the majority class (neutral in this corpus) and provides a more robust view of balanced performance.

Ordinal MAE is computed after mapping labels to $\{-1, 0, +1\}$: $\text{MAE} = (1/N) \times \sum |y_i - \hat{y}_i|$. This metric treats a positive-vs-negative confusion (cost 2) as twice as costly as a neutral-adjacent confusion (cost 1), matching the trading-loss structure in which a sign flip is more damaging than a magnitude error.

C. Benchmark Results

Table III reports held-out test results for the three candidates under the setup in Table II. Best results in each column are bolded.

TABLE III. FINANCIAL PHRASEBANK BENCHMARK RESULTS (TEST SPLIT)

Model	Accuracy	Macro-F1	Ordinal MAE
DeBERTa-finance	94.2%	0.9420	0.060
FinBERT	89.2%	0.8932	0.114
pmatorras/BERT-finance	12.5%	0.028	1.157

DeBERTa-finance outperforms FinBERT by 5.0 percentage points in accuracy (94.2% vs. 89.2%), by 0.0488 in macro-F1, and roughly halves the ordinal MAE (0.060 vs. 0.114). The pmatorras/BERT-finance checkpoint achieves only 12.5% accuracy and 0.028 macro-F1 — below the

33.3% random baseline for three-class classification. The 81.7 percentage point accuracy range across checkpoints confirms that uncured domain-labelled checkpoints span failure modes, not just performance tiers.

D. Statistical Considerations

The held-out test set contains approximately 700 sentences after the 70/15/15 stratified split. At that size, the 5.0 percentage point accuracy gap between DeBERTa-finance and FinBERT falls outside standard Wilson confidence intervals at $\alpha = 0.05$, and the 81.7 percentage point gap to pmatorras/BERT-finance is categorical. Macro-F1 differences mirror the accuracy ranking, providing convergent evidence that the ordering is not an artifact of class imbalance.

E. End-to-End Pipeline Behavior

The end-to-end pipeline applies the selected model (DeBERTa-finance) to incoming news, passes the bounded score to the modulation rule of Eq. (1), and returns $\hat{p}$. Representative cases: a long position $p = +0.5$ under positive news $s = +0.6$ yields $+0.80$ (amplified long); the same position under $s = -0.6$ yields $+0.20$ (attenuated long). A short position $p = -0.5$ under negative news $s = -0.6$ yields -0.80 (amplified short). In every case, trade direction is preserved.

V. DISCUSSION

This section interprets the benchmark outcomes, identifies the scope in which the proposed pipeline can be trusted, and names the concrete limitations that bound its current applicability.

A. Interpretation of Results

The headline finding is not that DeBERTa-finance is the best of the three candidates; it is that one of the three candidates is actively worse than chance. A pipeline that silently loaded pmatorras/BERT-finance would not merely underperform — it would invert sentiment signals, so that news that should raise conviction would instead suppress it. The 81.7 point gap is a failure mode that belongs in a safety argument, not just an ablation table. The 5.0-point gap between DeBERTa-finance and FinBERT is ordinary benchmark improvement; the gap to pmatorras is a qualitative failure.

The sign-aware formulation is a deliberate design choice. Additive fusion ($p + \alpha s$) would allow a large positive s to push a moderately negative p across zero, silently flipping the trade direction — the exact failure mode a sentiment-trading safety argument must avoid. Multiplicative fusion with explicit sign comparison prevents this by construction: the modulated output $\hat{p}$ always shares the sign of p , regardless of sentiment magnitude.

B. Limitations

The proposed pipeline has at least the following specific limitations.

1) Linear scaling assumption. The modulation formula uses a linear scaling ($1 \pm |s|$) that may not capture nonlinear market dynamics such as regime changes, saturation at extreme sentiment, or asymmetric reactions to positive versus negative news. In real-world markets, extreme sentiment values often produce diminishing or even reversed effects on

price displacement. Parametric sigmoid calibration, MLP-based modulation, or regime-conditioned nonlinear mappings are natural extensions.

2) Single asset class and time period. The end-to-end behavior is demonstrated on a single corpus and a single asset-class context, which may reduce generalizability to other markets (e.g., commodities, crypto) and to periods with different volatility regimes.

3) English-language news only. Sentiment scores are derived from English-language financial news only, excluding multilingual signals that may carry leading information for non-US-listed assets.

4) Label ontology. The three-class label ontology of the Financial PhraseBank collapses intensity into polarity; a finer-grained ontology could support a richer modulation surface.

5) Benchmark scope. Only three candidate checkpoints are evaluated; extending the benchmark to a larger and periodically refreshed roster is a prerequisite for treating automatic benchmarking as a continuous release gate.

6) Dataset cleanliness and real-world noise. The benchmark is conducted on the "50% agreement" subset of the Financial PhraseBank, which retains only instances where at least half of annotators agreed on the sentiment label. This filtering produces a relatively clean training and evaluation signal, but real-world financial news streams contain substantially more ambiguous, contradictory, and noisy labels. Performance on production data may therefore be lower than benchmark figures suggest, and the model-selection gate may need recalibration when applied to noisier corpora.

VI. CONCLUSION AND FUTURE WORK

This work proposes an automatic model-selection step and a sign-aware multiplicative sentiment-modulation rule for algorithmic trading pipelines. On the Financial PhraseBank, DeBERTa-finance reaches 94.2% accuracy, 0.9420 macro-F1, and 0.060 ordinal MAE, outperforming FinBERT by 5.0 percentage points. The failure of pmatras/BERT-finance (12.5% accuracy, 81.7 points below DeBERTa-finance) suggests that automatic model selection is a safety-critical step, not an optional optimization. The sign-aware rule preserves trade direction under all sentiment conditions and handles short positions symmetrically without special-casing.

Future work will pursue several directions. First, the benchmark roster will be extended to a larger and periodically refreshed set of finance-adapted checkpoints, including multilingual and instruction-tuned variants. Second, and most critically for addressing the linear scaling limitation identified in Section V-B, the modulation rule will be extended to nonlinear calibration functions: parametric sigmoid calibration, shallow MLP modulation trained on historical signal-return pairs, and regime-conditioned approaches that apply different scaling curves under high- versus low-volatility market states. Third, the end-to-end pipeline will be evaluated under live market microstructure to measure realized trading performance beyond benchmark accuracy.

REFERENCES

[1] Z. Hu, W. Liu, J. Bian, X. Liu, and T.-Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend

prediction," in Proc. ACM WSDM, Los Angeles, CA, USA, 2018, pp. 261–269.

[2] Y. Xing, F. Malandri, L. Zhang, and E. Cambria, "Sentiment-aware volatility forecasting," *Knowledge-Based Syst.*, vol. 176, pp. 68–76, 2019.

[3] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv:1908.10063, 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[5] Y. Yang, M. C. S. Uy, and A. Huang, "FinBERT: A pretrained language model for financial communications," arXiv:2006.08097, 2020.

[6] R. Liu, J. Shi, C. Chen, and C. Cheng, "Finance-specific large language models: A survey," in Proc. IEEE Int. Conf. Big Data, 2023, pp. 6215–6224.

[7] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 1–19, 2009.

[8] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in Proc. IJCAI, Buenos Aires, Argentina, 2015, pp. 2327–2333.

[9] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in Proc. ACL, Beijing, China, 2015, pp. 1354–1364.

[10] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[11] W. Zhang and S. Skiena, "Trading strategies to exploit blog and news sentiment," in Proc. AAAI ICWSM, Washington, DC, USA, 2010, pp. 375–378.

[12] Q. Xie, W. Han, Y. Lai, M. Peng, and J. Huang, "The Wall Street Neophyte: A zero-shot analysis of ChatGPT over MultiModal stock movement prediction challenges," arXiv:2304.05351, 2023.

[13] N. Shah, S. Chava, and S. Paranjape, "FinBench: Benchmarking language models on finance tasks," in Proc. IEEE Int. Conf. Data Mining Workshops, 2022, pp. 1–10.

[14] Z. Chen, L. Zhang, Y. Cao, and J. Ouyang, "Evaluating large language models on financial NLP tasks," in Proc. IEEE Int. Conf. Big Data, 2023, pp. 4123–4132.

[15] Y. Shah, S. Paturi, and S. Chava, "Trillion dollar words: A new financial dataset, task and market analysis," in Proc. ACL, Toronto, Canada, 2023, pp. 6664–6679.

[16] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, "Show your work: Improved reporting of experimental results," in Proc. EMNLP-IJCNLP, Hong Kong, 2019, pp. 2185–2194.

[17] O. E. Gundersen and S. Kjenmo, "State of the art: Reproducibility in artificial intelligence," in Proc. AAAI, New Orleans, LA, USA, 2018, pp. 1644–1651.

[18] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 782–796, 2014.

[19] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, Long Beach, CA, USA, 2017, pp. 5998–6008.

[20] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in Proc. ICLR, 2021.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, New Orleans, LA, USA, 2019.