Sentiment-augmented reinforcement learning for portfolio optimization with large language models

Abstract

Reinforcement learning (RL) has become a promising approach for portfolio optimization, but conventional models rely almost exclusively on structured price and return data. They typically neglect unstructured signals such as investor sentiment, which can anticipate short-term market movements. We introduce sentiment-augmented PPO (SAPPO), which incorporates daily asset-level sentiment extracted from Refinitiv financial news using large transformer-based models. Sentiment is embedded both in the observation space and in a modified advantage function, producing sentiment-guided updates. On a portfolio of ten U.S. equities, SAPPO improves Sharpe ratio from 1.67 to 2.07 and annualized return from 57% to 83%, with volatility reduced by 3 percentage points. Robustness checks show the improvement persists across sentiment models, hyperparameters, and market regimes. These results demonstrate that multimodal RL can leverage financial text to yield more adaptive trading strategies under uncertainty.

1 Introduction

Portfolio optimization aims to allocate capital across assets to maximize returns for a given level of risk [Markowitz, 1952]. Classical frameworks such as mean–variance optimization and factor models assume stable distributions of returns and correlations. These assumptions often fail in practice, where structural breaks, news shocks, and investor psychology drive volatility [DeMiguel et al., 2009]. Deep reinforcement learning (DRL) offers an adaptive alternative by framing allocation as sequential decision-making [Moody et al., 2001, Deng et al., 2017]. Proximal Policy Optimization (PPO) [Schulman et al., 2017] has been especially effective in financial settings due to its balance of sample efficiency and stability [Ye et al., 2020, Wang et al., 2019].

Yet existing DRL agents rely almost exclusively on numerical features. They miss unstructured signals—particularly textual sentiment—that influence prices before they are fully reflected in returns. Empirical evidence shows that media tone and investor sentiment affect returns, volatility, and liquidity [Tetlock, 2007, Baker and Wurgler, 2012, Smales, 2014]. Advances in large language models (LLMs), such as BERT [Devlin et al., 2019], FinBERT [Araci, 2019], OPT [Zhang et al., 2022], and LLaMA 3.3 [Dubey et al., 2024], allow scalable sentiment extraction from vast amounts of financial text. Most prior work treats sentiment as an external regressor for predictive models; few integrate it directly into an RL agent's policy learning process.

We propose sentiment-augmented PPO (SAPPO), which extends PPO in two ways: (i) sentiment vectors are concatenated with market states to inform the agent's perception of the environment, and (ii) sentiment modifies the advantage function, biasing updates toward allocations consistent with market tone. This dual integration allows SAPPO to be both sentiment-aware and sentiment-guided. Empirical evaluation on U.S. equities shows SAPPO consistently outperforms PPO and index benchmarks. Beyond finance, SAPPO exemplifies how multimodal RL can operationalize unstructured signals to improve sequential decision-making.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

2 Methodology

2.1 Data and preprocessing

We align trading days with RL steps. Daily adjusted closing prices for 10 liquid U.S. stocks (covering technology, retail, financials, and industrials) were collected for 2012–2019. The first seven years serve as training, with 2019 reserved for testing. Orders are placed at the close and executed the following day at a 10-minute VWAP, with transaction costs set at 0.01% per turnover.

For sentiment, we use Refinitiv financial news, extracting article headlines and short text for each firm. To reduce redundancy, near-duplicate articles are pruned using cosine similarity over TF-IDF vectors within a 5-day window. Each article is scored with LLaMA 3.3 fine-tuned for financial sentiment classification, producing values in $[-1, 1]$. Daily asset-level sentiment vectors $\mathbf{m}_n$ are computed as a time-decayed average of available headlines, normalized across assets.

2.2 State, action, and policy

The baseline market state $\mathbf{s}_n$ comprises previous portfolio weights $\mathbf{w}_{n-1}$ and current price relatives $\mathbf{S}_n$. SAPPO extends this with sentiment $\mathbf{m}_n$, yielding $\mathbf{s}'_n = (\mathbf{w}_{n-1}, \mathbf{S}_n, \mathbf{m}_n)$.

Actions $\mathbf{a}_n$ represent target portfolio weights, sampled from a Gaussian parameterized by the policy network. Allocations are projected to the simplex to enforce long-only, fully invested portfolios. The policy π_θ and critic V_ϕ are 2-layer MLPs with 64 ReLU units, trained using Stable-Baselines3 [Raffin et al., 2021]. Hyperparameters follow standard PPO defaults: learning rate 3×10^{-4}, γ=0.99, batch size 64, and clip parameter 0.2.

2.3 Sentiment-weighted advantage

The key innovation is the sentiment-modified advantage:

$$A'(\mathbf{s}_n, \mathbf{a}_n) = A(\mathbf{s}_n, \mathbf{a}_n) + \lambda\, \mathbf{w}_n \cdot \mathbf{m}_n, \tag{1}$$

where A is the standard PPO advantage, $\mathbf{w}_n$ the resulting weights after projection, and λ regulates sentiment strength. Intuitively, this term rewards allocations aligned with positive sentiment and penalizes exposure to negative sentiment. We tune λ via grid search on training data; λ=0.1 yields the best balance of return and stability.

3 Results

3.1 Main performance

Table 1 reports out-of-sample (2019) results. SAPPO delivers an annualized return of 83% with Sharpe 2.07, compared to PPO's 57% and 1.67. Drawdowns are slightly deeper, but volatility is reduced, indicating better efficiency. SAPPO outperforms index benchmarks by large margins.

Table 1: Out-of-sample (2019) performance.

	Sharpe	Ann. Return	Max DD	Vol.	Turnover
PPO	1.67	57%	-8.4%	20.5%	3.5%
SAPPO	**2.07**	**83%**	-10.2%	**17.3%**	12.0%
NASDAQ-100	1.20	33%	-12.1%	16.1%	N/A
S&P 500	1.05	27%	-11.4%	13.0%	N/A
Dow Jones	0.95	21%	-9.8%	12.5%	N/A

3.2 Portfolio dynamics

Analysis of daily weights shows SAPPO reallocates more aggressively in response to sentiment shocks. For instance, during Apple's January 2019 earnings warning, SAPPO reduced exposure within two days, whereas PPO maintained overweight positions until losses materialized. This responsiveness

explains both SAPPO's higher turnover and its superior drawdown resilience. Although turnover is higher (12% vs. 3.5%), the incremental costs are small relative to excess returns.

3.3 Ablations and robustness

Ablation on λ confirms that moderate sentiment weighting improves outcomes, while excessive reliance ($\lambda \geq 0.25$) amplifies noise. When sentiment is included only as a state feature (not in the advantage), Sharpe rises modestly (1.78) but remains below SAPPO's 2.07, showing that the learning-signal modification is critical.

Replacing LLaMA 3.3 with FinBERT, BERT-base, or OPT-1.3B reduces performance but still beats PPO, demonstrating robustness. Stress tests with random sentiment perturbations (10% sign flips) yield Sharpe 1.91, above PPO's 1.67. SAPPO also generalizes to different regimes: on 2021 data, it achieves Sharpe 1.92 vs. PPO's 1.38, underscoring adaptability.

4 Discussion

SAPPO illustrates the power of integrating structured and unstructured signals within RL. Sentiment-weighted updates allow the agent to anticipate market moves rather than react solely to lagging price data. The approach generalizes beyond finance: any sequential decision problem with rich text signals (e.g., healthcare triage, supply-chain logistics) could benefit from sentiment-aware RL.

The higher turnover of SAPPO raises questions about practical execution. While our cost assumptions are modest, real markets impose liquidity and slippage constraints. Extensions could incorporate execution-aware RL modules or model sentiment uncertainty explicitly, weighting updates by confidence scores.

5 Limitations and Future Work

Our study has several limitations. First, we rely solely on Refinitiv news; social media, earnings calls, or analyst reports may provide complementary sentiment. Second, daily aggregation may obscure intraday dynamics where sentiment impacts prices more quickly. Third, while our trading simulation includes costs, it assumes frictionless execution at VWAP; live deployment would require order book modeling. Finally, LLM sentiment classifiers may misinterpret context, sarcasm, or domain-specific phrasing.

Future work will address these by incorporating multiple sentiment sources, exploring intraday horizons, and testing SAPPO in simulated order books with market impact. Broader applications include multi-asset portfolios and cross-market arbitrage. From a methodological perspective, confidence-calibrated sentiment and attention mechanisms could refine how unstructured signals enter the advantage function.

6 Conclusion

We introduced SAPPO, a sentiment-augmented PPO that embeds LLM-extracted news sentiment into both state and policy updates. On U.S. equities, SAPPO achieves substantial gains in return and Sharpe relative to PPO and benchmarks, with robustness across hyperparameters, sentiment models, and regimes. Crucially, ablations show that sentiment-guided updates—not just state augmentation—drive the improvements.

SAPPO demonstrates that multimodal RL can effectively harness unstructured information for financial decision-making. More broadly, it provides a blueprint for integrating textual signals into sequential control tasks, opening pathways for RL systems that adapt not only to numerical dynamics but also to the narratives shaping human behavior.

References

D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019. URL https://arxiv.org/abs/1908.10063.

M. Baker and J. Wurgler. Comovement and predictability relationships between bonds and the cross-section of stocks. *Review of Asset Pricing Studies*, 2(1):57–87, 2012. doi: 10.1093/rapstu/ras001.

V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009. doi: 10.1093/rfs/hhm075.

Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3): 653–664, 2017. doi: 10.1109/TNNLS.2016.2522401.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics, 2019.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi: 10.2307/2975974.

J. Moody, L. Wu, Y. Liao, and M. Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001. doi: 10.1109/72.935097.

A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Dormann, and N. Carnevale. Stable-baselines3: Reliable reinforcement learning implementations. GitHub repository, 2021. Available at `https://github.com/DLR-RM/stable-baselines3`.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

L. A. Smales. News sentiment and bank stock returns. *European Journal of Finance*, 20(11):925–938, 2014. doi: 10.1080/1351847X.2013.766625.

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

Y. Wang, X. Jin, H. Guo, and H. Xu. Deep reinforcement learning for portfolio optimization. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2190–2193, New York, NY, USA, 2019. ACM. doi: 10.1145/3357384.3358070.

J. Ye, S. Zhang, J. Hao, and H. Wang. Reinforcement-learning-based portfolio management with augmented asset movement prediction states. *Expert Systems with Applications*, 159:113594, 2020. doi: 10.1016/j.eswa.2020.113594.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, K. Dewan, A. Diwan, J. Dodge, K. Ethayarajh, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Appendix

A Reinforcement learning framework

This section formalizes the reinforcement learning setup for portfolio optimization. I define states, actions, rewards, and transitions to support both price-driven and sentiment-aware strategies.

State definition. The market state at day n is

$$\mathbf{s}_n = (\mathbf{w}_n, \mathbf{S}_n),$$

where $\mathbf{w}_n \in \mathbb{R}^d$ is the agent's portfolio allocation and $\mathbf{S}_n \in \mathbb{R}^d$ are adjusted closing prices for d assets.

Actions and rebalancing. The action $\mathbf{a}_n$ updates allocations via

$$\mathbf{w}_n = \mathbf{w}_{n-1} + \mathbf{a}_n,$$

subject to a self-financing constraint

$$\mathbf{a}_n \cdot \mathbf{S}_n = 0.$$

Reward signal. Daily reward is the log return

$$x_{n+1} = \log \frac{\mathbf{w}_n \cdot \mathbf{S}_{n+1}}{\mathbf{w}_n \cdot \mathbf{S}_n}.$$

Transaction costs of 0.01% per unit turnover are applied.

Policy and value function. The stochastic policy is parameterized as $\pi_\theta(\mathbf{a}_n \mid \mathbf{s}_n)$. PPO with clipping is used for stable training. SAPPO modifies the advantage function:

$$A'(\mathbf{s}_n, \mathbf{a}_n) = A(\mathbf{s}_n, \mathbf{a}_n) + \lambda\, \mathbf{w}_n \cdot \mathbf{m}_n,$$

where $\mathbf{m}_n$ is the daily sentiment vector.

—

B Theoretical rationale for sentiment-guided advantage

Financial markets are non-stationary due to shocks and evolving expectations. Sentiment acts as side information that captures latent regime shifts earlier than prices. By integrating sentiment into the advantage function, SAPPO effectively conditions on partially observable states and relaxes the strict Markov assumption. This modification functions as behaviorally-informed regularization and improves adaptability under volatility.

—

C Ablation studies

Varying sentiment weight λ. Performance improves as λ increases up to 0.10, but degrades beyond that due to overreaction to noise. Table 2 reports results.

Configuration	Sharpe	Annual Return	Max Drawdown
PPO baseline ($\lambda = 0$)	1.67	57.0%	-8.4%
$\lambda = 0.05$	1.87	72.6%	-7.2%
$\lambda = 0.10$	**2.07**	**83.0%**	**-10.2%**
$\lambda = 0.20$	1.80	68.4%	-12.3%

Table 2: Effect of sentiment weighting parameter λ.

Alternative sentiment models. FinBERT, BERT-base, and OPT-1.3B improve over PPO, but LLaMA 3.3 achieves the best performance (Sharpe 2.07).

Robustness. Noise injection and out-of-sample tests (2020) confirm SAPPO's resilience across regimes. Full robustness checks appear in the Supplementary Material.

—

D Dataset summary

The dataset integrates price and sentiment signals for 10 large-cap U.S. equities (AAPL, NVDA, CMG, TGT, CPRT, LDOS, SBUX, JPM, BAC, ADP).

- Market data: Yahoo Finance daily adjusted closes (2012–2020).

- Sentiment: Refinitiv news headlines, classified by LLaMA 3.3 and baselines.
- Training period: 2012–2018.
- Test period: 2019 (main) and 2020 (robustness).
- Execution: VWAP over first 10 minutes, 0.01% transaction costs.

—

E Implementation details

Agents are implemented in `PyTorch 2.0` with `Stable-Baselines3`. The environment is OpenAI Gym-compatible and enforces long-only, self-financing portfolio constraints.

Portfolio rebalancing occurs daily. Sentiment signals are appended to the observation vector and used in the modified advantage function. All runs use NVIDIA V100 GPUs with CUDA 11.7.

—

F Training and architecture

Architecture. Actor and critic are two-layer MLPs (128–64 units, ReLU). Actor outputs a Gaussian distribution projected onto the simplex. Critic outputs scalar $V(s_t)$.

Training. PPO with Adam optimizer (3×10^{-4}), rollout length 256, 10 epochs per batch, clip range 0.2, discount factor 0.99, GAE $\lambda = 0.95$.

SAPPO uses $\lambda_{\text{sent}} = 0.1$. Random seeds fixed for reproducibility.

—

Note on Supplementary Material

Extended robustness checks (qualitative examples, scalability to 50 assets, statistical tests, Sharpe ratio derivation, and baseline comparisons) are available in the Supplementary Material.