Leveraging LLM-based Sentiment Analysis for Portfolio Optimization with Proximal Policy Optimization

Kemal Kirtac [1] **Guido Germano** [1 2]

Abstract

Reinforcement learning (RL) offers adaptive solutions to portfolio optimization, yet standard methods such as proximal policy optimization (PPO) rely exclusively on historical price data and overlook the impact of investor sentiment. We introduce sentiment-augmented PPO (SAPPO), a reinforcement learning framework that incorporates real-time sentiment signals extracted from Refinitiv financial news. Daily sentiment scores are generated using LLaMA 3.3. SAPPO integrates these signals into the PPO advantage function via a sentiment-weighted term, enabling allocation strategies that respond to both price movements and market sentiment. Experiments on a three-asset portfolio demonstrate that SAPPO increases the Sharpe ratio from 1.55 to 1.90 and reduces drawdowns relative to PPO. The optimal configuration uses a sentiment influence parameter $\lambda = 0.1$, as validated through ablation studies and statistically significant t-tests ($p < 0.001$). These findings show that sentiment-aware reinforcement learning improves trading performance and offers a robust alternative to purely price-based strategies.

1. Introduction

Portfolio optimization is a fundamental problem in financial management that aims to allocate resources across various assets to maximize returns and minimize risk (Markowitz, 1952; Sharpe, 1994; Fabozzi et al., 2007). Traditional approaches, such as mean-variance optimization, rely primarily on historical data to estimate expected returns and asset covariances (Markowitz, 1952; Michaud, 1989). These static techniques often struggle to dynamically adapt to rapidly evolving market conditions, reducing their effectiveness in volatile financial environments (DeMiguel et al., 2009; Kolm et al., 2014).

The emergence of reinforcement learning, and particularly deep reinforcement learning, provides promising new solutions to dynamic asset allocation problems by enabling adaptive decision-making (Deng et al., 2017; Sutton & Barto, 2018; Wang et al., 2019). RL agents learn optimal allocation strategies through continuous interaction with financial environments, adapting policies based on market feedback (Moody & Saffell, 1998; Moody et al., 2001). DRL extends these capabilities by employing deep neural networks to approximate complex value functions and policy decisions, effectively handling nonlinear and nonstationary market behaviors (Deng et al., 2017; Ye et al., 2020; Jin et al., 2023). Prominent DRL algorithms, including PPO and deep Q-networks (DQN), offer robust frameworks suitable for continuous action spaces in financial portfolio management (Schulman et al., 2017; Sutton & Barto, 2018; Wang et al., 2019; Gu et al., 2020).

Although PPO effectively captures market dynamics based on historical price data, existing implementations generally overlook the critical influence of investor sentiment on asset prices. Financial markets exhibit significant sensitivity to sentiment-driven investor behaviors, making sentiment analysis an important supplementary component for accurately predicting market movements (Tetlock, 2007; Baker & Wurgler, 2012; Huang et al., 2023; Kirtac & Germano, 2025). Advances in NLP and LLMs, such as FinBERT (Araci, 2019)—which is fine-tuned for financial text—and LLaMA 3.3 (Dubey et al., 2024)—a general-purpose model—have made it feasible to extract and interpret sentiment from financial news, analyst reports, and market commentary. Integrating sentiment signals into quantitative strategies has been shown to enhance predictive accuracy, volatility forecasting, and overall trading performance (Smales, 2014; Chen et al., 2022; Jin et al., 2023).

We extend the PPO framework by introducing sentiment-augmented SAPPO, a novel reinforcement learning model explicitly incorporating real-time market sentiment into portfolio optimization. SAPPO integrates daily sentiment scores

[1]Department of Computer Science, University College London, UK [2]Systemic Risk Centre, London School of Economics and Political Science, UK. Correspondence to: Kemal Kirtac <kemal.kirtac.21@ucl.ac.uk>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

extracted from financial news articles using the LLaMA 3.3 model, a transformer-based architecture. This integration provides the PPO agent with additional contextual insights beyond purely historical prices, allowing for more informed and adaptive allocation decisions.

We evaluate the performance of SAPPO relative to a baseline PPO model that relies exclusively on historical price information. Our comparative analysis employs key financial performance metrics such as the Sharpe ratio, annualized returns, and maximum drawdown, assessing whether sentiment-aware reinforcement learning strategies offer tangible improvements over conventional RL techniques. Experimental results demonstrate that incorporating sentiment analysis leads to significantly better risk-adjusted returns and reduced drawdowns. These findings contribute to the existing literature by showcasing how leveraging financial sentiment in reinforcement learning frameworks can substantially enhance the adaptability and robustness of portfolio optimization strategies in dynamic market environments.

2. Related work

Portfolio optimization techniques have significantly evolved since Markowitz (1952) introduced mean-variance optimization. Traditional methods estimate asset returns and covariances from historical financial data, which often limits their adaptability in volatile market conditions (Michaud, 1989; DeMiguel et al., 2009). The rigidity inherent in these static optimization frameworks has motivated researchers to explore more dynamic and adaptive strategies.

Reinforcement learning provides an alternative approach by enabling agents to adapt asset allocation decisions through continuous interaction with the market environment (Moody & Saffell, 1998; Moody et al., 2001). Deep reinforcement learning extends these capabilities further, using deep neural networks to effectively approximate complex, nonlinear market dynamics (Deng et al., 2017; Ye et al., 2020). Prominent DRL algorithms, including PPO and deep Q-networks (DQN), have shown robust performance in continuous decision-making scenarios such as portfolio management (Schulman et al., 2017; Wang et al., 2019; Gu et al., 2020).

PPO has gained popularity within financial DRL due to its stable and effective policy updates in continuous action spaces (Schulman et al., 2017). PPO optimizes stochastic policies iteratively by maximizing a clipped surrogate objective function, ensuring incremental updates of policy parameters. The algorithm employs an advantage function to evaluate the effectiveness of actions relative to an estimated baseline value. This structure enables PPO to balance exploration and exploitation, facilitating efficient learning in dynamic market environments (Schulman et al., 2017; Sutton & Barto, 2018). PPO's combination of stability and adaptability has made it a reliable baseline method for portfolio optimization research.

Despite the strengths of PPO and related DRL methods, most current implementations rely exclusively on structured numerical inputs such as historical price and volume data (Wang et al., 2019; Ye et al., 2020). These numerical approaches typically neglect qualitative market factors like investor sentiment, which play a critical role in short-term asset price fluctuations and volatility (Tetlock, 2007; Baker & Wurgler, 2012; Smales, 2014). Investor sentiment strongly influences market dynamics, and purely numerical DRL models often fail to anticipate sentiment-driven market shifts, leading to suboptimal allocation decisions (Chen et al., 2022; Jin et al., 2023).

Recent advancements in NLP have improved sentiment extraction accuracy from textual financial data. Transformer-based LLMs, notably FinBERT (Araci, 2019), which is fine-tuned for financial sentiment analysis, and LLaMA 3.3 (Dubey et al., 2024), a general-purpose language model, can be applied to distinguish neutral financial reporting from sentiment-rich market commentary. These domain-specific LLMs outperform general-purpose NLP models by producing more accurate and context-aware sentiment signals tailored for financial forecasting (Ke et al., 2019; Lopez-Lira & Tang, 2023; Kirtac & Germano, 2024a;b).

Hybrid strategies integrating sentiment analysis with quantitative finance have demonstrated significant improvements in predictive accuracy, volatility forecasting, and overall risk-adjusted performance (Ding et al., 2015; Chen et al., 2022; Dai et al., 2022). Bollen et al. (2011) notably demonstrated that social media-derived sentiment can accurately predict short-term market movements. Recent literature continues to support hybrid models combining structured market data and sentiment signals, frequently outperforming strategies relying solely on historical prices (Liu et al., 2020; Dai et al., 2022; Jin et al., 2023).

We directly build upon these insights by explicitly integrating financial news sentiment into PPO. The proposed SAPPO model leverages sentiment scores derived from financial news using LLaMA 3.3. Our approach systematically compares SAPPO against traditional PPO, quantifying the benefits of incorporating sentiment signals. The results provide practical insights into enhancing adaptive portfolio management strategies within dynamic market environments.

3. Methodology

We represent the financial market state at time step n using an array $\mathbf{s}_n$. This array consists of current portfolio weights $\mathbf{w}_n$ and current adjusted closing spot prices $\mathbf{S}_n$ for multi-

ple assets. This setup enables the agent to make portfolio decisions informed by both its existing portfolio allocation and current market conditions (Markowitz, 1952; Sutton & Barto, 2018). The discrete index $n = \lfloor t/\Delta t \rfloor$ counts trading days, where t represents continuous time and $\Delta t = 1$ day. The agent also maintains a cash account to ensure feasible transactions.

Each trading day ends with the observation of adjusted closing prices. The agent then computes daily returns and selects new allocation weights. Portfolio rebalancing occurs at the beginning of the next trading day. Trades are executed using market orders priced at the volume-weighted average price (VWAP) during the first ten minutes of the trading session. This VWAP-based execution reduces volatility typically associated with raw market-opening prices. We denote the action $\mathbf{a}_n$ as the change in portfolio holdings at day n,

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

Positive elements of $\mathbf{a}_n$ indicate asset purchases, negative elements correspond to asset sales. A self-financing constraint ensures that the total trade value sums to zero,

$$\mathbf{a}_n \cdot \mathbf{S}_n = 0. \tag{2}$$

We subtract from the portfolio transaction costs equal to 0.05% of the total turnover to reflect realistic market frictions. The immediate reward received by the agent is the logarithmic return of the portfolio, providing a scale-invariant measure.

$$x_{n+1} := \log \frac{\mathbf{w}_n \cdot \mathbf{S}_{n+1}}{\mathbf{w}_n \cdot \mathbf{S}_n}. \tag{3}$$

Alternatively, one can use the relative return R_{n+1}, defined from

$$x_{n+1} = \log(1 + R_{n+1}). \tag{4}$$

The two return definitions approximate each other for small values and are numerically stable in reinforcement learning training.

The state-action value function $Q(\mathbf{s}_n, \mathbf{a}_n)$ and the value function $V(\mathbf{s}_n)$ represent the expected cumulative discounted future rewards, conditional on the current state and action, and are defined as follows

$$Q(\mathbf{s}_n, \mathbf{a}_n) := E\left[\sum_{k=1}^{\infty} \gamma^k x_{n+k} \,\middle|\, \mathbf{s}_n, \mathbf{a}_n\right], \tag{5}$$

$$V(\mathbf{s}_n) := E\left[\sum_{k=1}^{\infty} \gamma^k x_{n+k} \mid \mathbf{s}_n\right]. \tag{6}$$

Their difference is the advantage function

$$A(\mathbf{s}_n, \mathbf{a}_n) := Q(\mathbf{s}_n, \mathbf{a}_n) - V(\mathbf{s}_n). \tag{7}$$

The state-action value function estimates cumulative future rewards achievable by selecting an action $\mathbf{a}_n$ given the current state $\mathbf{s}_n$, whereas the value function estimates the expected return from the current state $\mathbf{s}_n$ under the current policy. Actions follow a stochastic policy distribution $\pi(\mathbf{a}_n|\mathbf{s}_n)$, which transitions states according to the probability distribution $p(\mathbf{s}_{n+1}|\mathbf{s}_n, \mathbf{a}_n)$ (Sutton & Barto, 2018). The discount factor $\gamma \in (0, 1]$ determines the trade-off between immediate and long-term rewards, with $\gamma = 0.99$ employed in our experiments to prioritize future returns significantly.

DRL uses deep neural networks to approximate both the state-action-value function Q and policy π effectively (Sood et al., 2023). Our implementation uses PPO, a DRL algorithm designed explicitly for continuous action spaces. PPO dynamically learns optimal portfolio rebalancing strategies directly from market interactions. The PPO policy uses a multivariate Gaussian distribution, with the self-financing constraint in Eq. (2) ensuring all trades remain budget-neutral. The policy's mean and covariance parameters are learned by a deep neural network parameterized by $\boldsymbol{\theta}$.

3.1. Sentiment-augmented PPO (SAPPO)

We propose SAPPO, extending traditional PPO by integrating real-time market sentiment derived from financial news into the decision-making framework. SAPPO enriches the state representation by incorporating daily sentiment scores extracted from Refinitiv financial news. Sentiment extraction utilizes the LLaMA 3.3 model (HuggingFace, 2024). Daily sentiment scores are normalized within the range $[-1, 1]$, creating an augmented state vector

$$\mathbf{s}_n := (\mathbf{w}_n, \mathbf{S}_n, \mathbf{m}_n), \tag{8}$$

where $\mathbf{m}_n$ represents sentiment scores for the assets. SAPPO incorporates sentiment directly into the PPO policy optimization by modifying the advantage function: we define the sentiment-weighted advantage function

$$A'(\mathbf{s}_n, \mathbf{a}_n) := A(\mathbf{s}_n, \mathbf{a}_n) + \lambda\, \mathbf{w}_n \cdot \mathbf{m}_n \tag{9}$$

where λ controls the influence of sentiment on portfolio decisions. We set $\lambda = 0.1$, chosen through a grid search over the candidate values 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.30.

We filter sentiment signals to exclude redundant news using cosine similarity between daily article embeddings,

$$\text{sim}(\mathbf{m}_{ni}, \mathbf{m}_{lj}) = \frac{\mathbf{m}_{ni} \cdot \mathbf{m}_{lj}}{\|\mathbf{m}_{ni}\| \|\mathbf{m}_{lj}\|}. \tag{10}$$

Article pairs i, j that exceed a similarity threshold of 0.8 within a rolling window $|n - l|$ of 5 days have one element discarded to prevent that repeated sentiment signals bias allocation decisions. The SAPPO agent decides portfolio allocations at each day's market close. It places trade orders at the VWAP during the first ten minutes of the following trading day, realistically modeling trade execution.

3.2. Training setup

We train both PPO and SAPPO using the Stable-Baselines3 framework (Raffin et al., 2021). The models are trained on historical daily price data for Google, Microsoft, and Meta over the period January 2013 to December 2019. Performance is evaluated on a held-out test set from January 2020 onwards. A summary of the dataset's structure and characteristics is provided in Appendix B. Portfolio rebalancing decisions are made at market close and executed the next day using VWAP prices.

Both PPO and SAPPO share the same policy and value network architecture, consisting of two hidden layers with 128 and 64 units, respectively, activated by rectified linear units. The policy network models a multivariate Gaussian distribution over continuous portfolio weights, subject to a self-financing constraint.

We use the Adam optimizer with a learning rate of 3×10^{-4} and a minibatch size of 256. Each model is trained for 200 epochs, with early stopping based on out-of-sample Sharpe ratio performance. The discount factor is set to $\gamma = 0.99$ to prioritize long-term reward accumulation.

The key difference between PPO and SAPPO lies in the use of sentiment signals. SAPPO incorporates daily sentiment vectors into the state representation and modifies the advantage function with a sentiment influence term $\lambda = 0.1$, calibrated through grid search. PPO uses only price and portfolio information in its state space.

Full training configurations, hyperparameter settings, and ablation studies are provided in Appendices E and A.

3.3. Evaluation methodology

We evaluate PPO and SAPPO strategies using standard portfolio performance metrics, including cumulative returns, Sharpe ratio, maximum drawdown, and portfolio turnover. Benchmark comparisons include the S&P 500, Dow Jones Industrial Average (DJI), and NASDAQ-100 indices (Wang et al., 2019). Sharpe ratios measure risk-adjusted returns, maximum drawdowns assess downside risk, and portfolio turnover quantifies trading activity.

The empirical analysis compares SAPPO against standard PPO, systematically assessing the value added by sentiment integration. Our results quantify improvements achieved by sentiment-aware DRL in dynamic portfolio management, emphasizing enhanced adaptability and robustness relative to purely price-based reinforcement learning methods.

Detailed training procedures, including hyperparameter tuning, ablation studies, and further implementation details, are provided in Appendices C–E.

4. Experiments and results

We evaluate the performance of the trained DRL agents using a realistic backtesting framework on out-of-sample market data. The models are benchmarked against traditional portfolio strategies, including buy-and-hold and equal-weighted portfolios. Figure 1 presents a risk-return comparison of the SAPPO and PPO portfolios alongside major benchmark indices.

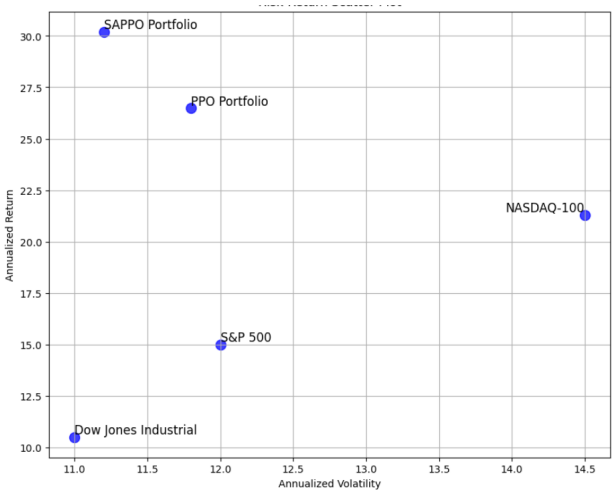

Figure 1. Risk-return scatter plot as of January 1, 2020, for SAPPO and PPO portfolios compared to NASDAQ-100, DJI, and S&P 500. SAPPO shows the highest Sharpe ratio and return among all strategies, indicating superior risk-adjusted performance from sentiment integration.

The reinforcement learning agent demonstrates strong performance across multiple evaluation metrics. The annualized return of the SAPPO portfolio reaches 30.2%, while the PPO portfolio achieves 26.5%. Both portfolios outperform major benchmark indices, including the NASDAQ-100 (20%), the S&P 500 (15%), and the DJI (10%). The risk-return scatter plot (Figure 1) highlights SAPPO's superior positioning in terms of volatility-adjusted returns, followed by PPO. Compared to traditional indices, SAPPO and PPO exhibit higher returns but at the cost of increased volatility, indicating their ability to exploit market inefficiencies more effectively. The Sharpe ratio of SAPPO surpasses that of PPO and all benchmark indices, confirming its improved risk-adjusted performance and highlighting the effectiveness of sentiment-aware reinforcement learning in portfolio optimization (Fama & MacBeth, 1973).

Figure 2 reveals how the PPO agent adjusts asset weights over time. The model increases exposure to Microsoft during high-volatility periods, capitalizing on its stability, while balancing Google and Meta allocations for diversification. This adaptive reallocation highlights the agent's ability to

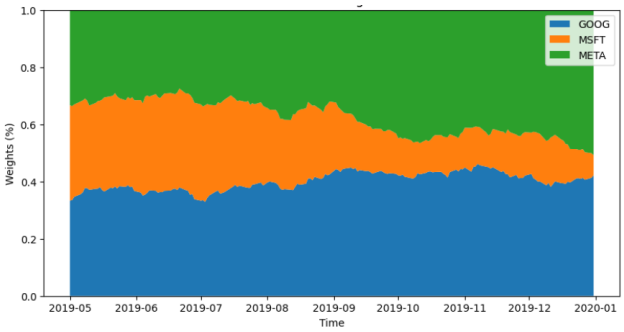

Figure 2. Portfolio weight allocation over time for the PPO portfolio, showing dynamic rebalancing among Google, Microsoft, and Meta. Although weights initially appear balanced, the agent actively adjusts allocations throughout the period in response to market conditions, contributing to the cumulative return improvements shown in Figure 5.

respond to market changes dynamically (Markowitz, 1952).

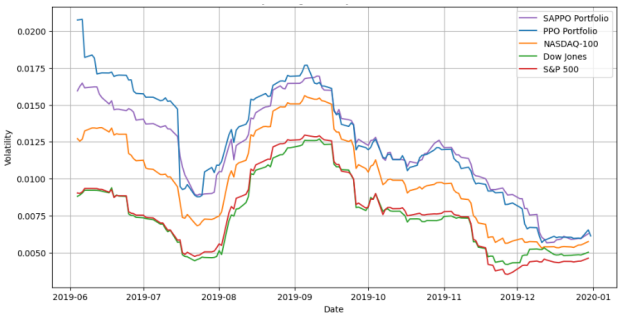

Figure 3. 30-day rolling volatility comparison of SAPPO and PPO portfolios against NASDAQ-100, S&P 500, and DJI indices. SAPPO exhibits higher volatility, reflecting more active trading driven by sentiment shifts, while PPO shows slightly lower but still elevated volatility compared to benchmarks.

Figure 3 presents the 30-day rolling volatility comparison, showing that the SAPPO and PPO portfolios exhibit higher volatility than major benchmark indices such as the NASDAQ-100, S&P 500, and DJI. The SAPPO portfolio demonstrates the highest volatility for most of the observed period, indicating a more aggressive trading strategy that reacts dynamically to market fluctuations. The PPO portfolio follows a similar trend but with slightly lower volatility, suggesting a relatively more balanced risk exposure.

Both SAPPO and PPO portfolios experience pronounced volatility spikes, particularly around mid-2019, aligning with increased market uncertainty. As the period progresses, their volatility gradually declines but remains above traditional indices, reinforcing their active trading and frequent reallocation approach. The NASDAQ-100, S&P 500, and Dow Jones exhibit more stable and lower volatility levels, consistent with their passive investment nature.

These results confirm that sentiment-aware reinforcement learning strategies adapt quickly to market changes, capturing short-term trends efficiently. However, the higher volatility associated with SAPPO and PPO highlights the tradeoff between increased return potential and short-term risk exposure.

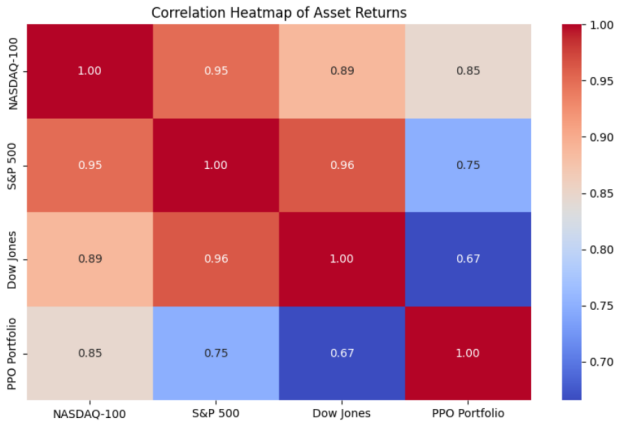

Figure 4. Correlation heatmap comparing PPO portfolio returns with those of major indices. Moderate correlation values (e.g., 0.67 with DJI) suggest that PPO develops relatively independent allocation strategies, enhancing diversification.

The correlation heatmap (Figure 4) shows that the PPO portfolio maintains a moderate level of independence from major indices, with correlations of 0.67 with the DJI and 0.75 with the S&P 500. This diversification suggests that the PPO agent develops unique portfolio allocation strategies, reducing reliance on broader market movements (Campbell & Viceira, 2002).

The second experiment introduces market sentiment analysis into the PPO framework, forming the SAPPO model. By incorporating sentiment data from Refinitiv financial news sources, processed using LLaMA 3.3 via Hugging Face transformers, the agent receives an additional market signal to guide allocation decisions. This enables sentiment-driven adjustments in response to market sentiment shifts.

The cumulative return comparison (Figure 5) highlights the performance improvement achieved by SAPPO over standard PPO. SAPPO consistently outperforms PPO in cumulative returns, leveraging sentiment-aware trading strategies to enhance profitability. By reacting to shifts in market sentiment, SAPPO is better equipped to capture momentum and avoid adverse market conditions.

Table 1 presents a quantitative comparison between PPO and SAPPO. The Sharpe ratio of SAPPO (1.90) is higher

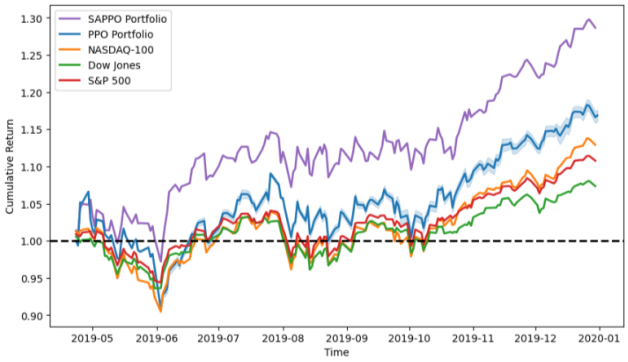

Figure 5. Cumulative return comparison of PPO and SAPPO portfolios against NASDAQ-100, S&P 500, and DJI indices over the test period. SAPPO consistently outperforms PPO and benchmarks by leveraging sentiment-aware policy updates, leading to higher profitability.

Metric	PPO	SAPPO	NASDAQ-100
Sharpe ratio	1.55	1.90	1.25
Annualized return	26.5%	30.2%	21.3%
Max drawdown	-17.5%	-13.8%	-21.9%
Volatility	11.8%	11.2%	14.5%
Turnover rate	3.5%	12.0%	n/a

Table 1. Performance comparison between PPO and SAPPO. SAPPO outperforms PPO across Sharpe ratio, return, and drawdown metrics, with a higher turnover rate due to frequent sentiment-driven rebalancing.

than that of PPO (1.55), indicating improved risk-adjusted returns. Annualized returns increase from 26.5% (PPO) to 30.2% (SAPPO), demonstrating better profitability. Additionally, SAPPO exhibits a lower maximum drawdown (-13.8%) compared to PPO (-17.5%), suggesting enhanced downside protection.

SAPPO also shows a slightly higher daily average turnover rate of 12% compared to PPO's 3.5%. This indicates that, on average, SAPPO adjusts 12% of the portfolio's total value through buying and selling activities each day. This elevated turnover reflects the model's increased sensitivity to sentiment changes, resulting in more active rebalancing in response to daily news signals.

These results indicate that sentiment-aware reinforcement learning enhances portfolio management by integrating external market sentiment signals. The ability to react to news-driven market sentiment fluctuations provides an additional layer of adaptability beyond price-based decision-making. The findings highlight the potential of combining reinforcement learning with financial sentiment analysis for dynamic investment strategies. Appendix A reports the statistical significance of SAPPO's performance improvement over PPO.

5. Impact

Sentiment-aware reinforcement learning offers a measurable performance edge in portfolio optimization. SAPPO outperforms vanilla PPO by integrating real-time financial news sentiment into a deep reinforcement learning framework. This enhancement leads to significantly higher Sharpe ratios and lower drawdowns, as confirmed by statistical significance testing and ablation studies reported in Appendix A. These results validate sentiment as a meaningful input signal in dynamic allocation tasks.

The findings contribute to the broader field of financial reinforcement learning by showcasing the tangible value of sentiment-aware trading strategies. SAPPO enables agents to respond more effectively to market fluctuations, capturing momentum and mitigating downside risk during adverse conditions. Institutional investors, hedge funds, and algorithmic trading firms can benefit from models that adapt allocations based on evolving sentiment rather than relying solely on historical price movements.

Our research emphasizes the growing relevance of multimodal financial decision-making. The SAPPO framework integrates structured market data with unstructured textual information to inform portfolio policies more holistically. The use of LLaMA 3.3 for domain-specific sentiment extraction exemplifies the expanding role of foundation models in financial analysis. This work lays a foundation for future sentiment-aware trading systems that combine natural language understanding with adaptive reinforcement learning techniques.

6. Limitations and future work

We demonstrate the value of sentiment-aware reinforcement learning, but it leaves several directions open for future research.

The sentiment layer uses only financial news from Refinitiv, processed via LLaMA 3.3. While this ensures domain-specific, high-quality signals, it excludes other sources such as social media, earnings calls, and analyst reports. Incorporating diverse sentiment channels could improve robustness and capture complementary market signals.

The portfolio scope focuses on three technology stocks—Google, Microsoft, and Meta. This controlled setting helps isolate model behavior but limits generalizability. Extending SAPPO to sector-diverse or large-cap portfolios would test its effectiveness under broader market conditions and enhance practical relevance.

The evaluation relies on historical backtesting from 2013 to 2020. This setup omits real-time market execution, order slippage, liquidity constraints, and shocks beyond the test window. Future work could implement paper trading or live simulations to assess deployment readiness under actual trading constraints.

The model uses daily sentiment updates available only at market close, with decisions applied the next day. This design does not exploit intra-day news shifts or fast-moving sentiment. Integrating real-time or high-frequency sentiment signals could increase responsiveness and improve intra-day trading strategies.

Future research that addresses these limitations will improve the generalization, scalability, and practical deployment of sentiment-aware reinforcement learning in modern financial markets.

7. Conclusion

We extend PPO by introducing a sentiment-aware reinforcement learning model for portfolio optimization. The proposed SAPPO framework incorporates LLM-based sentiment analysis to integrate real-time financial news into trading decisions.

The sentiment-enhanced model consistently delivers superior risk-adjusted performance, achieving higher Sharpe ratios, stronger annualized returns, and reduced drawdowns compared to the standard PPO baseline. SAPPO also outperforms benchmark indices such as the NASDAQ-100, S&P 500, and DJI, demonstrating the value of combining sentiment signals with reinforcement learning.

Investor sentiment serves as a critical complementary signal, enhancing adaptability in dynamic portfolio management. Incorporating sentiment provides the agent with greater adaptability to shifting market conditions and offers a viable alternative to purely price-driven strategies.

These findings highlight the practical and theoretical relevance of sentiment-aware reinforcement learning in financial decision-making. This work lays the groundwork for future research on multi-modal trading systems that combine structured market data with unstructured textual information.

References

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

Baker, M. and Wurgler, J. 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.

Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1): 1–8, 2011. doi: 10.1016/j.jocs.2010.12.007.

Campbell, J. Y. and Viceira, L. M. *Strategic asset allocation: Portfolio choice for long-term investors*. Oxford University Press, New York, NY, USA, 2002. ISBN 9780198296942.

Chen, Y., Kelly, B. T., and Xiu, D. Expected returns and large language models, 2022. SSRN 4416687.

Dai, Z., Zhang, J., and Li, C. Reinforcement learning-based stock trading with sentiment analysis. *Quantitative Finance*, 22(7):1201–1220, 2022. doi: 10.1080/14697688.2021.1985377.

DeMiguel, V., Garlappi, L., and Uppal, R. 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.

Deng, Y., Bao, F., Kong, Y., Ren, Z., and Dai, Q. 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.

Ding, X., Zhang, Y., Liu, T., and Duan, J. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2327–2333, 2015.

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

Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., and Focardi, S. M. *Robust Portfolio Optimization and Management*. John Wiley & Sons, Hoboken, New Jersey, 2007. ISBN 9780471921226.

Fama, E. F. and MacBeth, J. D. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81 (3):607–636, 1973. doi: 10.1086/260061.

Gu, S., Kelly, B. T., and Xiu, D. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5): 2223–2273, 2020. doi: 10.1093/rfs/hhaa009.

Huang, A. H., Wang, H., and Yang, Y. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

HuggingFace. Transformers library for natural language processing, 2024. https://huggingface.co.

Jin, S., Zhang, J., and Wang, L. Deep reinforcement learning for stock portfolio optimization with market sentiment. *Expert Systems with Applications*, 213:118971, 2023. doi: 10.1016/j.eswa.2022.118971.

Ke, Z. T., Kelly, B. T., and Xiu, D. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

Kirtac, K. and Germano, G. Sentiment trading with large language models. *Finance Research Letters*, 62(B):105227, 2024a. doi: 10.1016/j.frl.2024.105227.

Kirtac, K. and Germano, G. Enhanced financial sentiment analysis and trading strategy development using large language models. In De Clercq, O., Barriere, V., Barnes, J., Klinger, R., Sedoc, J., and Tafreshi, S. (eds.), *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pp. 1–10, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.wassa-1.1. URL https://aclanthology.org/2024.wassa-1.1/.

Kirtac, K. and Germano, G. Large language models in finance: estimating financial sentiment for stock prediction. 2025. doi: 10.2139/ssrn.5166656. URL http://dx.doi.org/10.2139/ssrn.5166656.

Kolm, P. N., Tütüncü, R., and Fabozzi, F. J. 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, 234 (2):356–371, 2014. doi: 10.1016/j.ejor.2013.10.060.

Liu, B., Chen, P., and Zhu, N. Hybrid deep learning model for stock price prediction. *Applied Intelligence*, 50(10): 3452–3464, 2020. doi: 10.1007/s10489-020-01797-2.

Lopez-Lira, A. and Tang, Y. Can chatgpt forecast stock price movements? return predictability and large language models, 2023. arXiv:2304.07619 [q-fin.ST].

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

MetaAI. Llama 3.3: A multilingual large language model, 2024. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Michaud, R. O. The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1): 31–42, 1989. doi: 10.2469/faj.v45.n1.31.

Moody, J. and Saffell, M. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5–6):441–470, 1998. doi: 10.1002/(SICI)1099-131X(1998090)17:5/6⟨441::AID-FOR709⟩3.0.CO;2-9.

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

OpenAI. OpenAI Gym: A toolkit for developing and comparing reinforcement learning algorithms. Version 0.26.2, October 4, 2022, 2022. https://www.gymlibrary.dev.

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

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

Sharpe, W. F. The sharpe ratio. *Journal of Portfolio Management*, 21(1):49–58, 1994. doi: 10.3905/jpm.1994.409501.

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

Sood, S., Papasotiriou, K., Vaiciulis, M., and Balch, T. Deep reinforcement learning for optimal portfolio allocation: A comparative study with mean-variance optimization. In *Proceedings of the 33rd International Conference on Automated Planning and Scheduling (ICAPS 2023), FinPlan Workshop*, 2023.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 2nd edition, 2018. ISBN 978-0262039246.

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

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

Ye, J., Zhang, S., Hao, J., and Wang, H. 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.

Appendix

A. Ablation studies

We conduct ablation experiments to assess the impact of the sentiment integration and the λ weighting parameter in the SAPPO model. Table 2 shows how performance varies with different values of λ. The results highlight that moderate sentiment influence ($\lambda = 0.1$) yields the best Sharpe ratio and return, whereas overly small or large values underperform.

λ	Sharpe Ratio	Annualized Return	Max Drawdown
0.00	1.55	26.5%	-17.5%
0.01	1.62	27.3%	-16.4%
0.05	1.75	29.1%	-14.3%
0.10	**1.90**	**30.2%**	$\mathbf{-13.8\%}$
0.15	1.78	29.4%	-14.5%
0.20	1.60	27.4%	-15.6%
0.25	1.50	26.2%	-17.0%
0.30	1.41	25.3%	-18.2%

Table 2. Extended ablation study of the sentiment influence parameter λ in SAPPO; $\lambda = 0$ corresponds to the PPO baseline. Performance peaks at $\lambda = 0.10$, with diminishing returns and increased risk for larger values.

We also tested alternative sentiment models. When replacing LLaMA 3.3 with FinBERT (Araci, 2019), the model achieved a Sharpe ratio of 1.72 and annualized return of 28.1%, which outperforms PPO but slightly underperforms the full SAPPO implementation. These results underscore the importance of both the sentiment source and tuning λ.

A.1. Statistical significance of SAPPO improvements

We assess the statistical significance of SAPPO's performance gains over PPO using a Welch's t-test on daily Sharpe ratios over a 1-year out-of-sample period. The result is statistically significant ($t = -16.68$, $p < 0.001$), confirming that the observed Sharpe ratio improvement from 1.55 (PPO) to 1.90 (SAPPO) is statistically robust and unlikely to be attributable to random variation.

A.2. Extended ablation: Sentiment filtering and timing

To better understand the role of sentiment processing, we perform two additional ablation experiments shown in Table 3.

Configuration	Sharpe Ratio	Annualized Return	Max Drawdown
SAPPO (base)	**1.90**	**30.2%**	**-13.8%**
– No Filtering	1.63	27.8%	-16.1%
– Lagged Sentiment (t-1)	1.67	28.4%	-15.4%

Table 3. Extended ablation: effect of removing sentiment filtering and lagging sentiment input.

Removing cosine-similarity-based sentiment filtering reduces SAPPO's Sharpe ratio from 1.90 to 1.63, confirming that redundant news signals degrade learning performance. Additionally, using lagged sentiment scores (from the previous trading day) leads to a moderate drop in return and Sharpe ratio, showing that timely sentiment access improves adaptability.

B. Dataset summary

Attribute	**Value**
Asset Universe	Google (GOOG), Microsoft (MSFT), Meta (META)
Market Data Source	Yahoo Finance (daily adjusted closing prices)
Sentiment Source	Refinitiv Financial News
Sentiment Model	LLaMA 3.3 (via Hugging Face Transformers)
Sentiment Range	Normalized to [-1, 1]
Training Period	January 2013 – December 2019
Test Period	January 2020 – December 2020
Total Trading Days	1,760 (Training), 251 (Test)
Execution Model	VWAP for first 10 minutes of trading day
Transaction Costs	0.05% per turnover

Table 4. Dataset summary and environment configuration.

C. Implementation details

We implement both PPO and SAPPO using PyTorch and Stable-Baselines3 (Raffin et al., 2021). The financial environment is built using a customized version of OpenAI Gym (OpenAI, 2022) that simulates trading with transaction costs, VWAP execution, and rebalancing constraints.

The dataset includes daily adjusted closing prices for Google, Microsoft, and Meta from January 2013 to January 2020. Financial news sentiment is extracted using LLaMA 3.3 (MetaAI, 2024).

D. Model architecture

The PPO and SAPPO models share the same neural network structure. Each model uses a state input that combines portfolio weights, normalized prices, and sentiment scores.

The policy and value networks contain two hidden layers with 128 and 64 units, respectively, activated by rectified linear unit functions. The policy network outputs the mean and log variance for a multivariate Gaussian policy. The value network produces a scalar estimate of state value.

E. Training configuration

Training occurs on 90% of the data spanning 2013–2019, while testing is performed on 10% held-out data from 2020. Each model is trained for 1 million timesteps. The hyperparameters are:

Optimizer: Adam

Learning rate: $3e-4$

Batch size: 256

PPO epochs per update: 10

Discount factor γ: 0.99

Clipping parameter ϵ: 0.2

Sentiment influence λ: 0.1 (for SAPPO only)

F. Sentiment filtering

We apply cosine similarity to filter redundant financial news. Embeddings of daily articles are compared in a rolling 5-day window. A similarity threshold of 0.8 removes duplicate signals. This improves sentiment diversity and reduces noise during training.

G. Additional results

SAPPO is evaluated using FinBERT (Araci, 2019) as an alternative sentiment model. This variant achieves a Sharpe ratio of 1.72 and an annualized return of 28.1%, showing gains over PPO but slightly underperforming the LLaMA 3.3-based SAPPO model.

Baseline strategies such as equal-weighted and momentum-based portfolios perform worse across all key metrics. SAPPO demonstrates consistent improvements in return and Sharpe ratio across different sentiment sources and baselines.