




ISSN: 2617-6548

URL: www.ijirss.com

Aligning social signals with market outcomes: A reinforcement learning approach to noisy sentiment data in stock forecasting

 Yangjun Lu

City University of Hong Kong, Hong Kong, China.

(Email: yjluedu@163.com)

Abstract

Stock price prediction relies heavily on market sentiment, but it is still difficult to identify truly meaningful signals from social media because of the volume of noise and flimsy engagement measurements. In this paper, sentiment signal extraction is formulated as a delayed reward recommendation issue using a unique reinforcement learning framework. The method uses actual market input instead of static labels by simulating the task as an agent that learns to suggest tweets based on their potential to increase predicted accuracy. Due to the scarce and delayed nature of market signals, the agent may differentiate between those that are useful and those that are misleading or irrelevant by using a reward function that aligns with 48-hour post-publication stock movements. When combined with a Long Short-Term Memory (LSTM) network for price forecasting, the suggested approach shows that, in contrast to employing sentiment variables directly, integrating the RL-based suggestion enhances prediction performance. Tests conducted on Apple stock data from 2014 to 2016 demonstrate that using agent-selected tweets improves the system's R-squared scores and reduces prediction errors. The findings indicate that dynamically adjusting comment selection in unstable financial settings can be achieved using reinforcement learning. This study combines market-based incentive design, financial text mining, and reinforcement learning to provide a viable method for improving the reliability of sentiment-driven stock predictions. To further validate this framework's robustness, future research may expand it to include more equities and explore integrating it with other market signals.

Keywords: Long short-term memory neural network, Reinforcement learning, Signal extraction, Social data, Stock prediction.

DOI: 10.53894/ijirss.v8i5.9176

Funding: This study received no specific financial support.

History: Received: 1 July 2025 / Revised: 31 July 2025 / Accepted: 4 August 2025 / Published: 7 August 2025

Copyright: © 2025 by the author. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.

Transparency: The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Publisher: Innovative Research Publishing

1. Introduction

The machine learning (ML) community has shown interest in intelligent trading systems in conventional stock trading. For example, a comprehensive assessment of machine learning studies [1] found that supervised learning methods are often

used in systems that consider financial trading as a market prediction issue. Even if supervised learning-based prediction has proven successful, a formulation based on sequential decision-making systems may be more effective due to the addition of usual trading operations expenses [2].

In this regard, a number of Reinforcement Learning (RL) studies [2-5] have shown advantages in resolving financial trading challenges by using a decision-making process that determines the optimal course of action to optimize long-term success on a particular asset. Researchers usually use feature representations that is only related to asset price time series while creating RL systems [2-4]. According to some methods of machine learning in the finance industry [1, 6], integrating textual data from external financial news with characteristics extracted from price time series yields positive outcomes in the supervised learning-based market forecasting assignment.

Recent effective exploration of natural language processing approaches to extract characteristics for a sequential decision-making process has been noted by some RL writers [5, 7, 8]. There are still many unanswered problems, especially regarding market sentiment momentum and the instability of RL approaches, despite the fruitful efforts along this promising new path. In the first, the market's dominant sentiment for a particular asset is extracted and captured [6] and in the second, the instability and difficulty of generalizing RL approaches [9] are discussed. These issues are exacerbated by the financial market environment's stochasticity [10].

Numerous studies integrating sentiment analysis and long short-term memory neural networks have been conducted in recent years to anticipate market prices. However, social data is noisy, which makes it difficult to extract useful information from it. This study uses reinforcement learning to identify truly important social data. To predict stock prices, we formulate the signal extraction problem as a reinforcement learning task, where the agent's goal is to suggest tweets that are the most relevant and instructive.

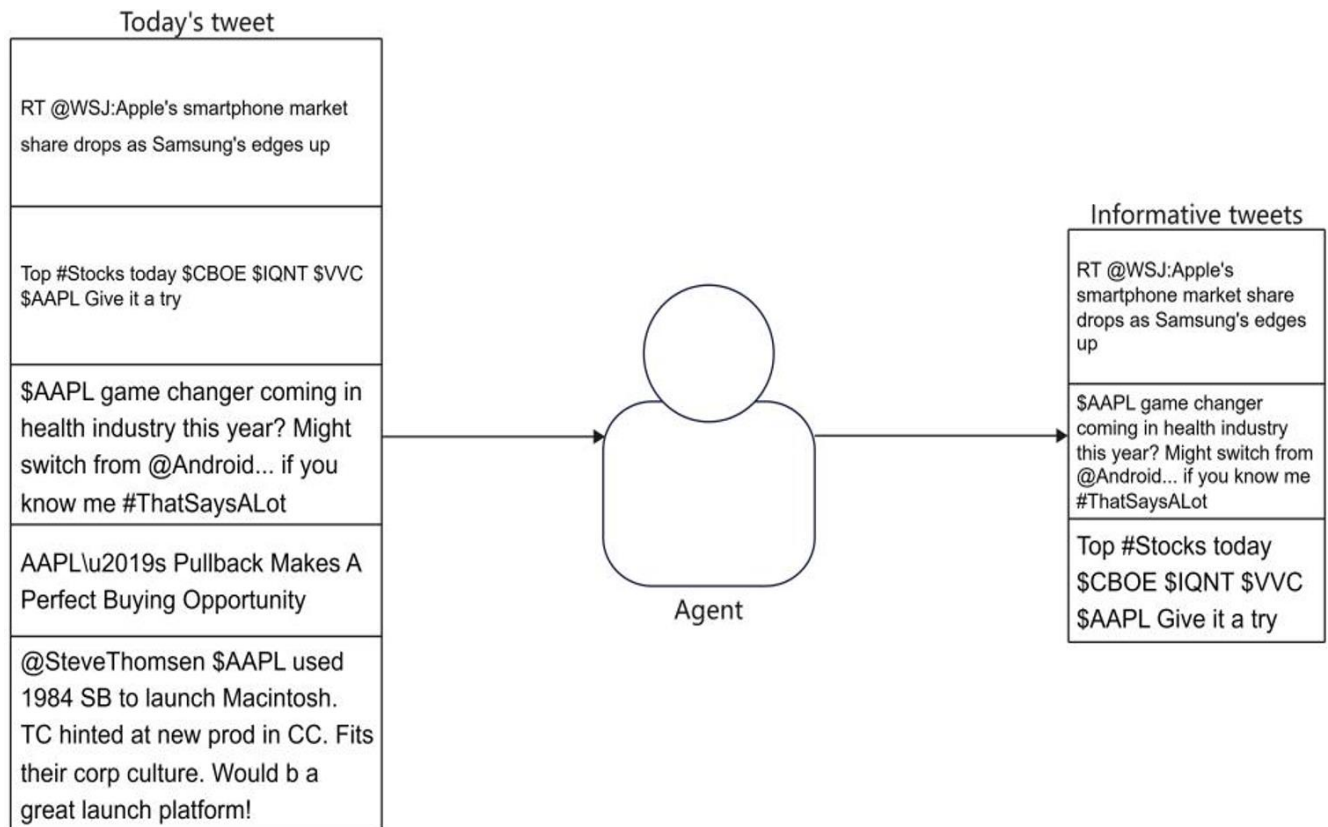


Figure 1.
An ideal agent that recommends useful tweets.

We start by introducing the two main restrictions that form the reinforcement system. The first restriction is that a comment's actual worth may become apparent after some time. The incentive for tweet recommendations is not instantaneous and is not provided immediately. The second limitation is that, although there may be many comments, not all of them are worthwhile. The agent must have the ability to filter out noise and select comments with high value.

The agent's objective in the task, which we classified as a reinforcement learning issue, is to determine whether to suggest a tweet based on its context. Stock data and tweet characteristics are included in the state representation. The activity is organized as both a recommendation and a non-recommendation. The immediate input from the market is intended to be the reward. The tweet suggestion scenario for each day may be seen as a standalone episode: the agent sees a tweet, decides what to recommend, and then gets paid by the market according to how "correct" or helpful that choice was in retrospect. By using this framework, the model is able to learn from real market results and identify patterns that may be hard to explicitly define.

The following is a summary of this study's primary contributions: In order to assess the informativeness of social media posts, we proposed a novel reward function based on actual market outcomes. Our work connects social media processing, financial signal analysis, and reinforcement learning, offering a framework for reliable comment recommendation that aligns with economic impact.

2. Related Works

Given its learning principles that map observable conditions to possible transactions aiming for large payouts, RL approaches seem to be a logical answer for developing effective trading strategies. By presenting the Recurrent RL (RRL) architecture with a variant of the policy gradient approach for the direct approximation of parametric policies from gradient ascent over previous actions and rewards, the first trading RL paper, published in 1997, rose to prominence [4]. Because of how impactful this groundbreaking work was, scholars are continually expanding on its fundamental concepts and framework [2, 3].

RRL [10] belongs to the class of policy-based reinforcement learning, which has issues with high variance and convergence to local optima. On the other hand, value-based techniques like Deep Q-Learning (DQN) are also prominent in this field [11] have bias but show less volatility. The actor-critic approaches later appear as a hybrid effort to remedy the aforementioned shortcomings. In the video gaming industry, for example, the Asynchronous Advantage Actor-Critic (A3C) algorithm has shown superior performance compared to earlier state-of-the-art DQN techniques [11]. This comparison was quickly repeated in an active trading issue study [6], which again came to similar results about the A3C architecture's potential.

Using supervised learning to extract market information from textual news for market prediction has long been a promising approach [12, 13]. Furthermore, a substantial amount of research uses price time series and textual news information for market forecasting, according to surveys on machine learning approaches.

[1, 6]. Only lately, however, have a few RL works looked into this strategy. Feuerriegel and Prendinger [8] supplemented their Q-learning algorithm with a sentiment dictionary, sometimes known as a lexicon, to extract word-level sentiment from textual news. In order to generate sentiment-charged reward features that serve as stand-ins for investors' sentiment, other studies [7] used the inverse RL approach. However, our method integrated these emotion reward characteristics into a supervised market predictor rather than employing them to train a trading strategy. In order to enhance the RL market's state representation, Yunan et al. [5] recently used deep learning-based techniques to produce word embeddings and market forecasts. The aim of reinforcement learning in our study is to extract useful information from social data, and we use the Proximal Policy Optimization (PPO) method, which is based on Schulman's Trust Region Policy Optimization [14].

Reinforcement learning has been used increasingly in recommendation issues, particularly in cases where input is non-deterministic or delayed, in the context of recommendation and feedback alignment. An RL-based trading agent that integrates Twitter sentiment has been developed by researchers in the financial and social spheres [15], highlighting the potential of external signals in directing policy optimization. To enable more precise long-term credit assignment, this study treats the daily tweet set as an episode and models the market's post hoc reaction as a reward.

Few studies have included market responses as a delayed feedback mechanism for natural language prioritization, despite the fact that previous market-aware modeling studies have mostly concentrated on predictive modeling of price fluctuations [16]. In the context of market-based incentive design and external validation, connecting textual qualities with downstream economic indicators has been proven to be beneficial in recent work in financial text analysis [17]. These methods usually depend on static labeling or direct monitoring. Our system, on the other hand, uses market returns as a reward signal to determine how informative tweets are, offering a grounded and dynamic assessment criterion that does not require human annotation.

Recent studies on comment suggestion [12, 17] have mostly focused on engagement-based heuristics and supervised learning algorithms in the context of tweet recommendation and signal extraction. While some [18] employ likes, retweets, or user votes as a stand-in for quality, such systems use attributes like comment length, sentiment polarity, or user information to estimate helpfulness. These signals, however, are often noisy and influenced by popularity. The problem of matching tweet value with real-world consequences, particularly in high-noise areas like financial discourse, is not well addressed by more recent work [19], which examined robust signal extraction using semi-supervised techniques and pre-trained language models.

3. Approach

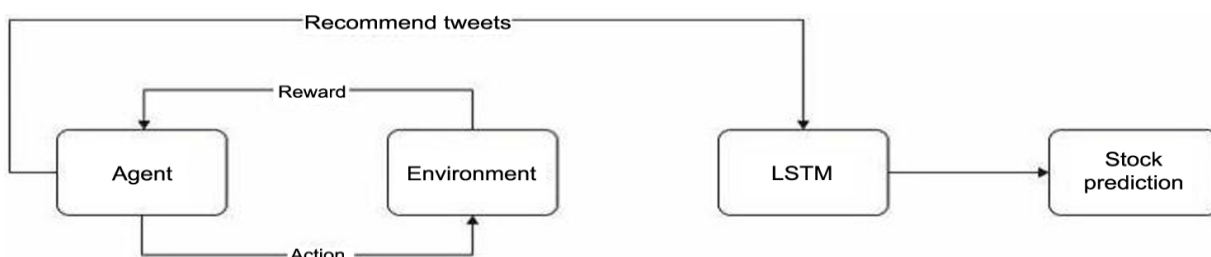


Figure 2.
Integrate the recommendation into the prediction.

3.1. Reinforcement Learning System and Reward Design

State, action, and reward are the three fundamental components that must be specified in order to create a reinforcement learning system. The current stock data, number of likes, number of commenters' followers, and embeddings of the comment text are all part of our system's state. Our system's activity is only set up to suggest or not recommend.

Since the incentive signal instructs the agent on what constitutes "valuable" tweets, its design is crucial. Instead of using flimsy measures like likes or retweets, we want the incentive in this case, to represent market validation of a tweet. Based on the correlation between tweets and the success or failure of the market movement forecast, the agent should develop its own concept of value. Based on the trade's 48-hour market performance, we can determine the payoff for suggesting a certain tweet. For instance, the agent's choice to suggest a tweet that expresses pessimism about the stock price should be rewarded positively if the token's price does, in fact, decrease considerably over the course of the next 48 hours. On the other hand, suggesting a tweet that was positive but the token price fell, would result in a negative reward. Here is how we really put the reward function into practice:

- +1 for recommending a tweet that correctly anticipated the outcome, or not recommending a tweet that incorrectly anticipated the outcome.
- +0.01 to recommend a tweet that correctly predicted the slight change in the token's price, or not to recommend a tweet that incorrectly predicted the slight change in the token's price.
- -1 for recommending a tweet that was misleading or wrong, or not recommending a tweet that correctly anticipated the outcome.

We determine the reward r_i for suggesting a tweet c_i based on how well the tweet matches the market trend observed within 48 hours after publication. This incentive motivates the algorithm to suggest tweets whose hypotheses or insights are then confirmed by actual market activity. Similar to reinforcement learning from human input, this method replaces objective market feedback with "human feedback."

It is difficult for RL algorithms to handle a 48-hour reward delay. When input is delayed, the issue of credit assignment becomes more complicated. However, after the market result is recovered, we can calculate the reward since each episode can be regarded as ending once the outcome is known.

Proximal Policy Optimization is used to train the reinforcement system to ensure its stability. Additionally, its computational efficiency is a factor in its selection.

3.2. Long-Short-Term-Memory Neural Network Design

The study's prediction model is based on an LSTM network, which builds on previous developments in stock price forecasting. The ability of LSTMs to capture and learn dependencies over long periods of time sets them apart as an evolution of conventional RNNs. Their complex architecture, which incorporates memory cells and gating mechanisms, allows the network to choose to store or ignore information across long data sequences. This capability, when combined with fully connected layers, makes them especially well-suited for time-series forecasting and sequence classification applications. The open price, high price, low price, close price, and volume of the stock from the previous L days are intended to be the inputs of the LSTM in this study. The feature of market sentiment is optional. The model's projection of the closing price is its output.

A hyperparameter search concerning the sequence time span L was conducted to optimize the model's performance. Additionally, an early stopping mechanism was implemented during training to reduce overfitting and ensure the model's robustness. This mechanism halts the training process if the loss value on the validation set does not decrease for 50 consecutive epochs.

4. Experiments

4.1. Dataset and Data Preprocessing

The gathered dataset includes two subsets that report on the public tweets of relevant stocks (Twitter dataset) or the stock price history over time (Finance dataset). Both of these categories pertain to Apple stocks during the period from January 2, 2014, to March 31, 2016.

4.1.1. Finance Dataset

Yahoo Finance provided the stock information. Features pertaining to daily opening and closing prices, daily trading volume, the highest and lowest prices were extracted at each time (day) t . It is important to note that the time series under examination do not fully cover the 365 days of the calendar year since the stock market is recorded during the working days (Monday through Friday, omitting a few holidays throughout the year). The stock price value $S_t \in \mathbb{R}$ was then remapped into categorical classes y_t for each time t , showing either a positive or negative trend, as seen below.

$$y_t = \begin{cases} +1, & \text{if } \Delta = S_t - S_{t-2} > 0 \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

4.1.2. Twitter Dataset

The Tweets may be obtained in two ways. To begin with, Twitter offers an API for downloading Tweets. However, it is not an option for this paper due to the rate restriction and history limit. The second method involved scraping tweets directly from the Twitter website. This approach was used to collect daily Tweets for equities of interest from January 2014 to March 2016. Each sample in the raw tweet data collection was represented as a four-element vector that included the text, likes, followers, and publication date. All gathered tweets were released during the same period as the financial data

collection. It was important to synchronize tweets published on weekends (or other holidays) with those posted on Mondays (or the day after the holiday ended), since tweets could be produced every day.

In order for the NLP algorithm to effectively extract the pertinent information, tweets must be pre-processed to remove unusual symbols, URLs, emojis, and other content. Therefore, the cleaning procedure includes eliminating tweets that contain http or .com, eliminating @user, #hashtag, tabs, and excessive spaces, and eliminating tweets that are not in English. The sentiment of each cleaned tweet was then determined by feeding it into VADER, a lexicon- and rule-based sentiment analysis program that is especially sensitive to feelings expressed on social media. Additionally, the cleaned tweets are sent into SBERT, which may be used to calculate tweet embeddings. The sample in the Twitter dataset should be a vector with a length of 390 after the preprocessing step.

4.2. Experimental Settings

Several tests were conducted by altering various variables that could influence the model's performance to thoroughly examine the proposed technique and assess its effects. The primary consideration was whether the Long Short-Term Memory model performs better with the suggested tweets. A consistent prediction time frame was maintained across all configurations to ensure experimental uniformity and facilitate straightforward comparison of trained model outputs.

4.2.1. Sequence Time Span

The model may be able to grasp longer-term relationships in the data and obtain additional contextual information from extended sequences. Longer sequences may have drawbacks, too, particularly in fields like the stock market, where it can be difficult to identify significant patterns and trends due to noisy and volatile data. Conversely, shorter sequences could provide the model with more rapid and targeted input, which would aid in capturing more current patterns and trends. This study compares outcomes with various L, with L = 1, 2, 5, 10, in order to determine the optimal L.

4.2.2. Discount Factor

The degree of shortsightedness in the reinforcement learning system is determined by the discount factor. This study has compared several discount factors, such as $\gamma = 0.1, 0.5$, and 0.8 , in order to increase the ultimate forecast accuracy.

4.2.3. Recommendation Contribution

Two experiments using distinct feature sets, one with the suggested tweets and the other without the market sentiment feature, were conducted to evaluate the effectiveness of our reinforcement learning system.

4.3. Evaluation Metrics

Determining how well the model fits the data in regression analysis requires assessing the model's performance. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) are the four key metrics used in this study to evaluate the model.

5. Results Analysis

5.1. Sequence Time Span

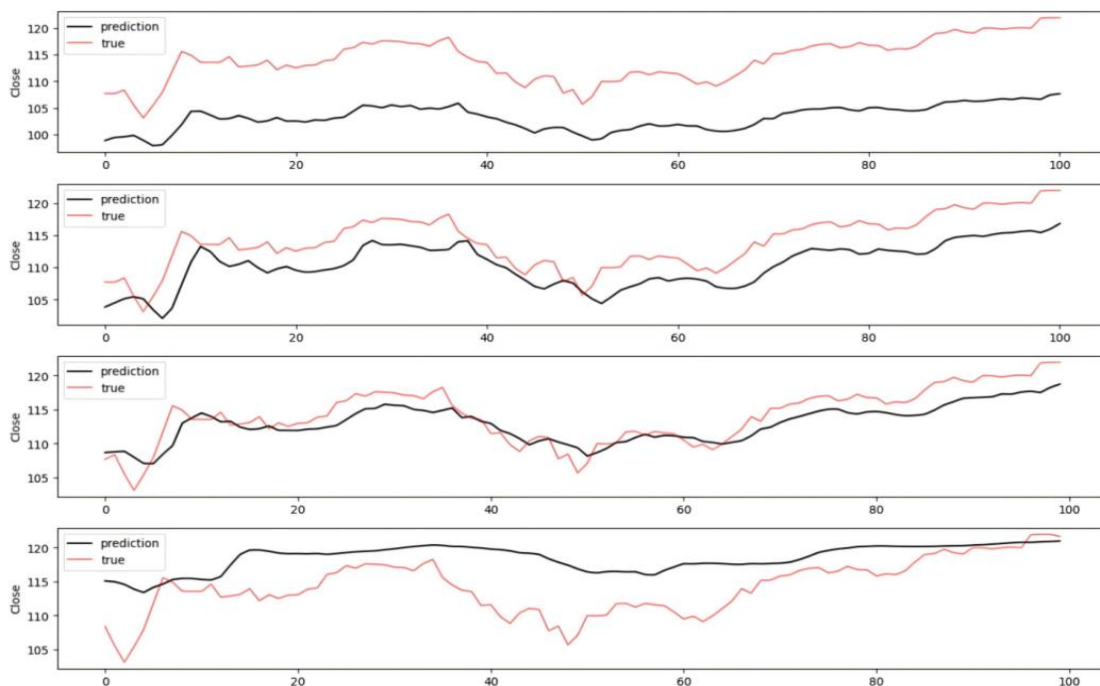


Figure 3.

Prediction curve with different sequence length L. Arranged from top to bottom are the cases for L = 1, 2, 5, 10.

Table 1.
Evaluation Metrics.

Model	MSE	MAE	RMSE	R2
L=1	119.63	10.76	10.94	-6.48
L=2	15.92	3.65	3.99	0.01
L=5	4.60	1.79	2.14	0.71
L=10	28.56	4.61	5.34	-0.79

With the lowest MSE and the highest R2, the model with $L = 5$ performs the best when compared to models with various time spans L . This is consistent with the widely held belief that recent trends are more predictive of future movements.

5.2. Discount Factor

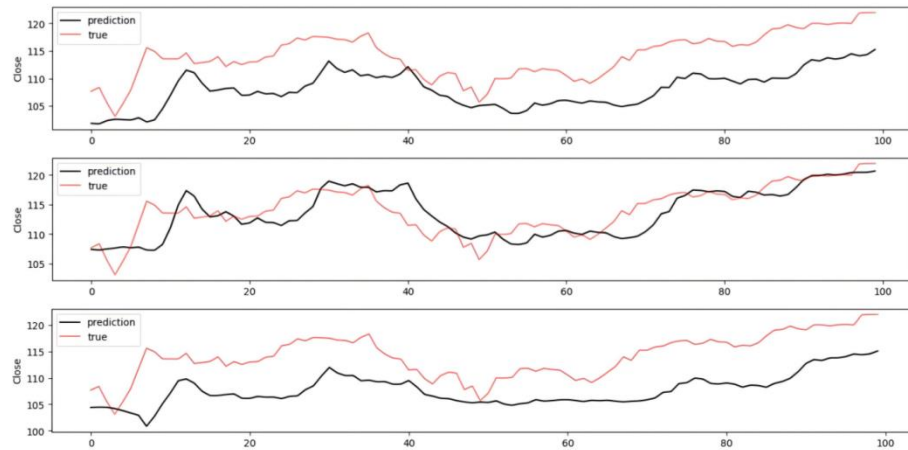


Figure 4.

Prediction curve with different discount factor γ . Arranged from top to bottom are the cases for $\gamma = 0.1, 0.5, 0.8$.

Table 2.
Evaluation Metrics.

Model	MSE	MAE	RMSE	R2
$\gamma = 0.1$	41.90	5.99	6.47	-1.58
$\gamma = 0.5$	7.32	2.01	2.70	0.55
$\gamma = 0.8$	46.87	6.38	6.85	-1.89

Based on the assessment criteria, we may conclude that $\gamma = 0.5$ is the best discount factor γ among the options. When the reinforcement system strikes a balance between now and future rewards, it operates well, as indicated by a medium γ .

5.3. Recommendation Contribution

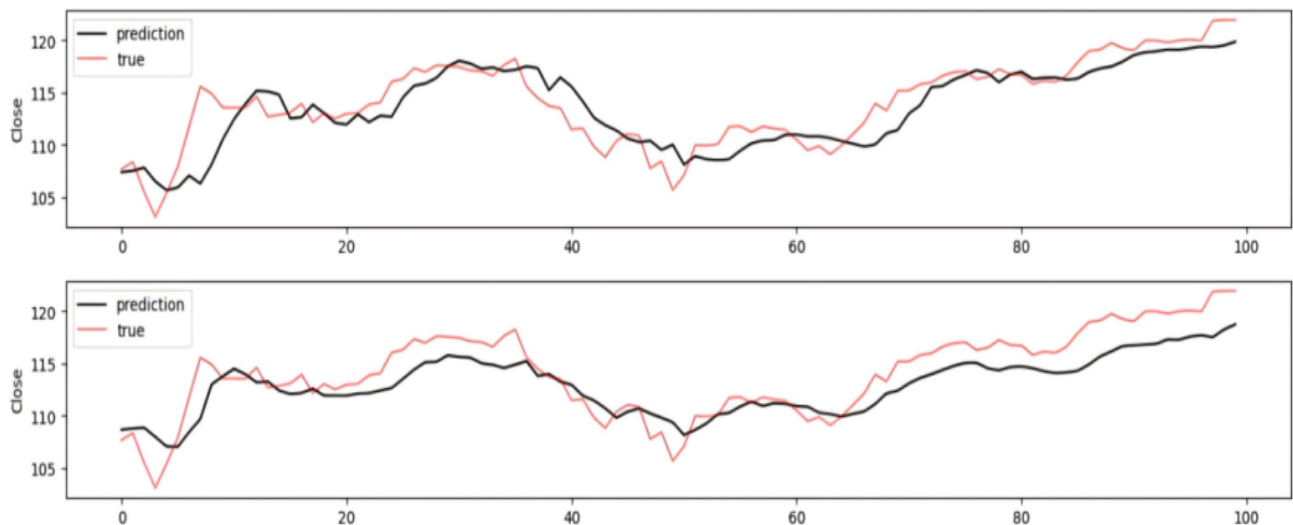


Figure 5.

Prediction curve without tweet recommendation. The upper curve corresponds to the LSTM with RL4TR. The lower curve corresponds to the LSTM.

Table 3.
Evaluation Metrics.

Model	MSE	MAE	RMSE	R2
LSTM with RL for Tweet Recommendation	4.43	1.56	2.10	0.73
LSTM	4.60	1.79	2.14	0.71

When it comes to the function of tweet recommendation, the results show that adding a reinforcement system to the models has advantages. It seems to improve forecast accuracy.

The incorporation of the reinforcement learning recommendation system in the examples provided results in a better R2, suggesting that it may offer more insights into market movements influenced by public mood. Risk managers may find this method especially useful in predicting how the market will respond to news and events in real time.

5.4. Comparison to Supervised Learning and Heuristics

Another option is supervised learning, which involves training a classifier and classifying comments as "useful" or "not useful" depending on the results. This approach is simpler and may perform well in a static setting. However, due to its three inherent advantages, we frame this as a reinforcement learning (RL) problem:

- Constant adaptation: RL picks up new information on its own.
- Policy optimization: The recommendation policy is directly optimized using RL.
- Dealing with delayed reward: RL deals directly with delayed feedback.

Even if supervised learning has its place, reinforcement learning could be more aligned with our long-term objectives, particularly in dynamic settings.

Heuristic techniques, such as giving experienced users or certain keywords priority, are straightforward yet fragile. They are not flexible and fail to recognize important but subtle cues. Beyond hard-coded logic, our RL-based method enables pattern recognition.

6. Conclusion

The idea of redefining the sentiment analysis issue as a reinforcement learning task is investigated in this work. According to the research, the performance of the prediction model may be significantly enhanced by the suggested reinforcement learning approach for determining the actual market mood.

The study's exclusive emphasis on Apple stocks is one of its limitations.

In order to assess the sentiment impact on other equities, future research should broaden its focus to include additional stocks, especially those less affected by market sentiment. A useful further layer of analysis might also be added by contrasting these results with those of other prediction (and sentiment analysis) models or methodologies. However, this research emphasizes how crucial it is to select educational tweets that accurately represent the market's mood signal.

References

- [1] H. Bruno, Miranda, V. A. Sobreiro, and H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction," *Expert Systems with Applications*, vol. 124, pp. 226-251, 2019. <https://doi.org/10.1016/j.eswa.2019.01.012>
- [2] D. Yue, B. Feng, K. Youyong, R. Zhiqian, and D. Qionghai, "Deep direct reinforcement learning for financial signal representation and trading," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 653-664, 2017. <https://doi.org/10.1109/TNNLS.2016.2522401>
- [3] A. M. Amine and C.-G. Lee, "Continuous control with stacked deep dynamic recurrent reinforcement learning for portfolio optimization," *Expert Systems with Applications*, vol. 140, p. 112891, 2020. <https://doi.org/10.1016/j.eswa.2019.112891>
- [4] J. Moody and L. Wu, "Optimization of trading systems and portfolios," in *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFER) (pp. 300-307)*. New York: IEEE, 1997. <https://doi.org/10.1109/CIFER.1997.618952>
- [5] Y. Yunan et al., "Reinforcement-learning based portfolio management with augmented asset movement prediction states," in *Proceedings of the AAAI Conference on Artificial Intelligence 34, 01 (apr 2020)*, 1112-1119, 2020. <https://doi.org/10.1609/aaai.v34i01.5462>
- [6] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653-7670, 2014. <https://doi.org/10.1016/j.eswa.2014.06.009>
- [7] S. Y. Yang, Y. Yu, and S. Almahdi, "An investor sentiment reward-based trading system using Gaussian inverse reinforcement learning algorithm," *Expert Systems with Applications*, vol. 114, pp. 388-401, 2018. <https://doi.org/10.1016/j.eswa.2018.07.056>
- [8] S. Feuerriegel and H. Prendinger, "News-based trading strategies," *Decision Support Systems*, vol. 90, pp. 65-74, 2016. <https://doi.org/10.1016/j.dss.2016.06.020>
- [9] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," 2018. <https://doi.org/10.1609/aaai.v32i1.11694>
- [10] R. S. Tsay, *Analysis of financial time series*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2010. <https://doi.org/10.1002/9780470644560>
- [11] M. Volodymyr et al., "Asynchronous methods for deep reinforcement learning," in *Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1928-1937)*. New York, NY: Proceedings of Machine Learning Research, 2016. <https://proceedings.mlr.press/v48/mniha16.html>
- [12] L. Chuhan, M. Shulin, and S. Maosong, "Comment selection for argument summarization in online debates," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 1021-1032.

- [13] 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 *Proceedings of the Eleventh ACM International Conference on Web Search and data Mining*, 2018, pp. 261-269.
- [14] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015, Vol. 37, pp. 1889–1897)*. Lille, France: *Proceedings of Machine Learning Research (PMLR)*, 2015.
- [15] P. Lima and C. Francisco, "Intelligent trading systems: A sentiment-aware reinforcement learning approach," in *Proceedings of the Second ACM International Conference on AI in Finance*, 2021.
- [16] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (Vol. 32, No. 1, pp. 7582–7589)*. Palo Alto, CA: AAAI Press, 2018. <https://doi.org/10.1609/aaai.v32i1.12255>
- [17] Y. Jianlong, Q. Libo, T. Zhiheng, and Z. L. Yue, "Modeling financial news impact with graph attention networks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4443–4453.
- [18] S. Aditi, S. Manish, and T. Chenhao, "Engagement-aware comment ranking in online news," in *Proceedings of the Web Conference*, 2021, pp. 2812–2823.
- [19] Y.-N. Chen, C.-S. Wu, and Y.-A. Chen, "Co-attentive ranking model for multi-turn response selection in retrieval-based chatbots," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 3308–3317)*, 2020.