ResearchPaper_RatishMoondra-4.docx

by Ratish Moondra

AN NLP AND DRL APPROACH TO SENTIMENT-BASED STOCK MARKET ANALYSIS
FOR INVESTMENT DECISION

RATISH MOONDRA

Research Proposal

FEBRUARY 2024

Abstract

The stock market, critical for both global and national economies, has experienced significant growth and volatility across the world. It exerts significance influence on economic and decision-making ability of various stakeholder e.g. individual investors, financial analysts, traders, and market practitioners. Consequently, stakeholders are quite keen on achieving accurate prediction.

Despite great advancements in AI and ML, predicting future stock prices remains challenging but advantageous task at the same time and an area of active research. This study aims to integrate Deep Reinforcement Learning (DRL) and Sentiment Analysis (SA) to create a robust investment decision system. By utilizing NLP techniques and establishing corelation with historical price data, the framework aims to extract textual sentiments and improve investment decisions. This research contributes to enhancing Stock Market Prediction (SMP) methodologies and providing insights for navigating through the modern financial markets. From modelling perspective, the proposed research methodology will infuse Sentiment Analysis (to extract sentiment score i.e. positive, negative or neutral), combine it with Deep neural network (i.e. Long Short-Term Memory) & Unsupervised learning technique (Reinforcement Learning) to understand the hidden data patterns & corelations; thereby maximizing the reward function to help investor makes a right decision and improve profitability and reduce possible losses.

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CR	Cumulative Return
DRL	Deep Reinforcement Learning
FA	Fundamental Analysis
ISI	Investment Sentiment Index
LSTM	Long Short-Term Memory
MER	Maximum Error Rate
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SMP	Stock Market Prediction
SVM	Support Vector Machine
TA	Technical Analysis
VADER	Valence Aware Dictionary and sEntiment Reasoner

1. Background

Stock markets are an important pillar for Global and National businesses & economy. They have experienced an unprecedented growth increasingly lucrative but fairly complex and high-risk space due to its volatile nature or agility. Anticipating future stock prices, termed Stock Market Prediction, remains an area of immense interest for academics, researchers, and economists alike. Recent advancements in the field of Sentiment Analysis, Artificial Intelligence, and Machine Learning have introduced novel perspectives into decision-making processes for investors. Researcher have tried to use a range of traditional algo such as Random Forests, Timeseries, Support vector machine algorithms to latest deep neural network-based algorithms. Most recently (Zou et al., 2024) used Neural Networks such as cascaded LSTM framework at deciphering patterns from historical data and focussed on increasing Cumulative Return (CR), MER (maximum earning rate) and Average profitability thereby improving range from 5% to 52% depending on a combination of metric and respective stock index.

However, the challenge persists in how predictive/deep learning models react or adapt to unexpected/unforeseen market events influenced by different factors such as macro or micro economic tendencies, sovereign GDP/economy, market subtleties, companies' growth prospects, consumer prices, central bank policies, and industry-specific nuances. Sahu (Sahu et al., 2023) described a plethora of algorithms and their families being applied in this space and keep the researchers most current. He also underlines the fact that DRL agents, combining price prediction and trading signal, have been deployed for automated trading systems. The fusion of diverse data points such as Technical Analysis (TA) (i.e. a methodology to predict stock movement using historical prices (open, close, high, low), volume), and Sentiment Analysis (SA) for extracting the mood of the market, presents a promising avenue for employing greater influence on investment decisions.

This research endeavours to amalgamate DRL (Jang & Seong, 2023) and SA to formulate a robust decision-making system. DRL, processing multidimensional data spaces to generate actions without supervision, addresses challenges posed by incomplete information or various external economic factors. The proposed contributions include:

- Leveraging NLP for preprocessing texts such as tweets to extract sentiments
- Historical price data extraction for various stocks
- Utilization of LSTM with Reinforcement Learning for pattern/corelation in stock data.

2. Related Works

Recent years have seen the integration of NLP and DRL techniques in stock market analysis and has gathered significant attention. To delve into the topic, I embarked on a thorough literature journey, tapping into various academic and scientific platform such as Google Scholar, Scopus, Springer, ScienceDirect, IEEE Xplore to unearth articles from reputable research journals and publications. This section provides a comprehensive review of relevant literature, outlining various methodologies, algorithms, and approaches employed for predicting market movements in a categorical manner.

2.1. Researches/Papers related to Deep Reinforcement Learning

(Zou et al., 2024) proposed a novel automated stock trading system that leverages cascaded Long Short-Term Memory (LSTM) networks within a Deep Reinforcement Learning (DRL) framework. By combining the sequential learning capabilities of LSTM networks with the decision-making power of reinforcement learning, the authors achieved impressive range from 5% to 52% depending on the metric (CR, MER and average profitability per trade) and stock index. Similarly, (Aken et al., 2023) explored the application of DRL in stock trading models, emphasizing the adaptability and learning capacity of reinforcement learning agents. (Awad et al., 2023) and (Yousefi, 2022) extended this research by employing DRL techniques for stock market prediction, exhibiting its potential in forecasting market trends and optimizing trading decisions. (Jang & Seong, 2023) attempted to extend the application of DRL to portfolio optimization, signifying the efficacy of combining reinforcement learning with modern portfolio theory and not just the price prediction. In other part of finance such as cryptocurrency price prediction, (Kang et al., 2022) applied 1-dimentional CNN and GRU architectures (1DCNN-GRU) and their experiments showed that their methodology outstripped existing models with a lowest RMSE of 43.933, 3.511 and 0.00128 on the Bitcoin, Ethereum and Ripple dataset respectively. Moreover, (Lawi et al., 2022) implemented LSTM and GRU architecture on grouped time-series data to accurately forecast stock prices, their experiments demonstrated highest accuracy 97.37% (MAPE) and 96.60% (RMSPE) for validation results of the testing data using GRU Model-1.

2.2. Researches/Papers related to Sentiment Analysis

The relationship between sentiment analysis and stock market prediction has been extensively explored in the literature. (Koukaras et al., 2022) utilized microblogging sentiment analysis coupled with ML to predict stock market trends based on social media sentiments. By analysing the sentiment expressed in online conversations and social media posts, the authors demonstrated that the best results were obtained when tweets were analysed using VADER and SVM. The results were 76.3% and 67% for F-score and Area Under Curve (AUC) respectively.

2.3. Researches/Papers related to Hybrid or multiple modelling

(Tao, 2023) integrated Linear Regression, LSTM, and Random Forest Regression models for predicting BMW stock prices, leveraging the complementary strengths of different algorithms to enhance predictive performance; He concluded MLR (multiple linear regression) produced best RMSE of 1.238 (lowest among MLR, LSTM and Random Forest). (Lin et al., 2023) proposed a dynamic ensemble model for stock prediction based on Deep Reinforcement Learning (DRL), demonstrating that MSE touched 0.011 and 0.005, Sharpe ratio (SR) touched 2.20 and 1.53, and CR touched 1.38 and 1.21 in SSE 50 and NASDAQ 100 datasets.

2.4. Researches/Papers related to ML in Quantitative Finance

(Sahu et al., 2023) provided a comprehensive overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance, offering insights into recent advancements and challenges in the field. By reviewing the current state-of-the-art methodologies and applications, the authors shed light on the evolving landscape of predictive modelling in financial markets, emphasizing the growing importance of data-driven approaches in decision-making processes.

In summary, the literature review highlights the diverse range of methodologies and techniques employed in sentiment-based stock market analysis, highlighting the growing significance of NLP and DRL approaches in stock market arena. From deep reinforcement learning-based trading systems to sentiment analysis-driven predictive models, researchers continue to explore innovative strategies for extracting valuable insights from textual data and leveraging them to inform investment decisions in dynamic financial markets.

3. Research Questions

Most important question to understand is if fusion of NLP and DRL effectively predict market moves, thereby generating an informed action/decision.

This research aims to address following questions:

1. How can SA be leveraged to extract valuable insights from diverse text sources?

2. How can DRL algorithms be integrated with sentiment analysis to develop decision-making framework for stock trading?

3. To what extent does the fusion of NLP and DRL improve the accuracy of predicting market moves?

4. Aim and Objectives

The primary aim is to develop a novel model for predicting market moves based on sentiment analysis using NLP and DRL techniques.

Objectives:

- To Analyse existing SA techniques and DRL algorithms applicable to financial markets.
- To Propose an integrated framework that combines NLP for SA and DRL for decision-making in stock trading.
- To Capture and preprocess textual data from social media for sentiment analysis.
- To Evaluate effectiveness (accuracy) of the model in predicting market moves and capturing sentiment-driven fluctuations in stock prices.
- To Interpret patterns and correlations discovered by the model in the stock market dataset.

5. Significance of the Study

One of the key challenges in stock market has been to accurately predict the stock prices because it can cause significant losses. A lot of modelling work has been carried out in the past; however, a wider adoption is yet to be seen. This study will be integrating NLP and DRL techniques, and contributes to explore methods for extracting insights from texts and combine them with historical price data for better price prediction.

Another aspect to look at it is the development of a novel predictive model that combines SA with DRL introduces a new approach to stock market forecasting by maximising the associated reward function.

The insights gained from this research will have practical implications for investors, financial analysts, and market practitioners. Accurately predict market moves based on sentiment analysis (& its impact on price movement) can affect investment decisions, improve trading strategies, and mitigate risks specially in volatile market conditions.

The fusion of NLP and DRL techniques represents a significant technological innovation in the field of financial analysis. By harnessing the power of ML and NLP, this research pushes the boundaries of computational finance and opens new avenues for exploring the complex dynamics of financial markets.

6. Scope of the Study

6.1. Scope

The study will be based on data available in public domain (say price data via yahoo finance or tweets form Stock Twits, X i.e. twitter). No external survey or questionnaire activity will be conducted. The study will use EDA and effectiveness of model developed for prediction. The study will also analyse the performance of the model in relation to sentiment captured and impact on the trading decision. The study is also focused in exploring the DRL (LSTM+RL) with Sentiments Analysis and how this combined model performs which help the investor improve decision making and reduce possible losses.

6.2. Limitation

It's important to acknowledge the limitations of the study, which may include constraints on data availability, computational resources, and the simplifying assumptions made in the model. These limitations will be carefully considered and discussed in the research findings.

While the primary focus of the study is on predicting market moves based on sentiment analysis, there are opportunities for future research to explore additional aspects such as alternative data sources, refining the RL model, and extending the application of these techniques to other financial markets or asset classes.

In summary, the scope of this study encompasses the development and evaluation of a predictive model for sentiment-based stock market analysis using NLP and DRL techniques, with a focus on understanding the relationship between sentiment and market direction.

7. Research Methodology

7.1. Data Collection

For this study, data will be collected from various sources, including price data (from Yahoo Finance, as tabularized in Table 1: Price Data for Stock Market Prediction) and social media platforms (e.g., X aka Twitter, Stock Twits as tabularized in Table 2: Tweets or Media Article Data Description). The data will cover a specified timeframe relevant to the study's objectives and will be collected using automated web scraping tools and APIs.

Table 1: Price Data for Stock Market Prediction

Column	Type	Description
Date	Date	The date of the trading day.
Open	Numeric	The opening price of the stock on that trading day.
High	Numeric	The highest price reached by the stock during the trading day.
Low	Numeric	The lowest price reached by the stock during the trading day.
Close	Numeric	The closing price of the stock on that trading day.
Adjusted Close	Numeric	The adjusted closing price, which factors in any corporate actions, such as dividends or stock splits, that occurred before the next trading day.
Volume	Numeric	The total number of shares traded on that trading day.

Table 2: Tweets or Media Article Data Description

Column	Type	Description
Message	Text	The content of the tweet
Context	Text	Stock that message focuses
Date	Date	Date of Tweet

7.2. Data Preprocessing

Pre-processing Steps for Table 1: Price Data for Stock Market Prediction

The price data for stock market prediction undergoes rigorous preprocessing to ensure its suitability for analysis. The following steps are undertaken:

- Handling Missing Values: Any missing values within the dataset are addressed using appropriate strategies such as imputation or removal, ensuring data completeness.
- Normalization: Numeric features such as 'Open,' 'High,' 'Low,' 'Close,' 'Adjusted Close,' and 'Volume' are normalized to maintain consistency in scale across the dataset, facilitating model convergence.
- Feature Scaling: Feature scaling such as Min-Max scaler will be applied thereby preventing any single feature from controlling the analysis.
- Outlier Detection and Treatment: Outliers will be identified and treated appropriately, to mitigate their impact on model performance.

Pre-processing Steps for Table 2: Tweets or Media Article Data Description

As detailed in Figure 1: Sentiment Analysis next page, The tweets or media article data undergoes preprocessing to enhance its suitability for analysis within the context of stock market prediction. The following steps are undertaken:

- Text Cleaning: The text content of tweets undergoes cleaning to remove any special characters, HTML tags, URLs, or non-alphanumeric characters, ensuring data integrity.
- Text Normalization: The text will be converted to lowercase to standardize the text format, followed by tokenization to segment it into individual words or phrases.
- Stop words: Commonly occurring words with little semantic meaning, known as stop words, will be removed to improve the quality of the text data.
- Stemming or Lemmatization: Stemming or lemmatization techniques will be applied to normalize words to their base form, reducing the dimensionality of the text data and aiding in subsequent analysis.
- Entity Recognition (Optional): Entity recognition techniques may be employed to identify specific entities such as company names mentioned in the context column.

Through these preprocessing steps, both sets of data are prepared meticulously, ensuring they meet the stringent requirements for analysis within the scope of stock market prediction in the context of this thesis.

7.3. Sentiment Analysis

NLP techniques will be employed for sentiment analysis of the textual data. Sentiment lexicons and machine learning algorithms, such as Support Vector Machines (SVM) or Recurrent Neural

Networks (RNNs), will be utilized to classify the sentiment expressed in the text as positive, negative, or neutral. Additionally, sentiment scoring methods may be applied to quantify the intensity of sentiment.

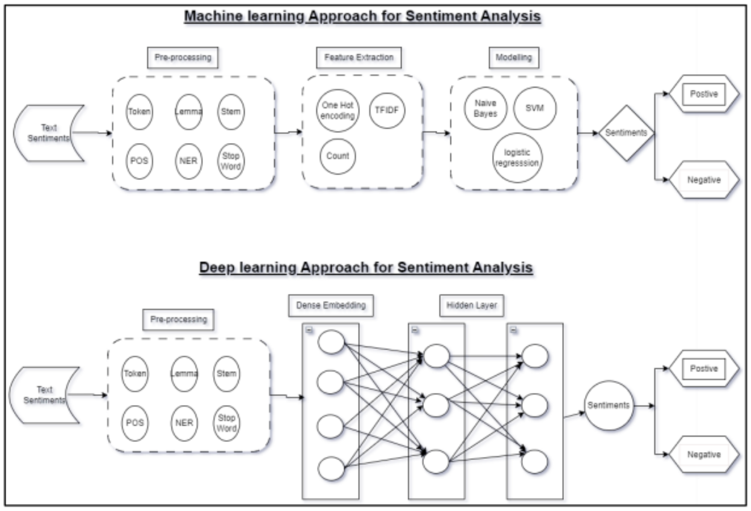

Figure 1: Sentiment Analysis

7.4. Reinforcement Learning Framework

A Reinforcement Learning (RL) framework (brief architecture is displayed in

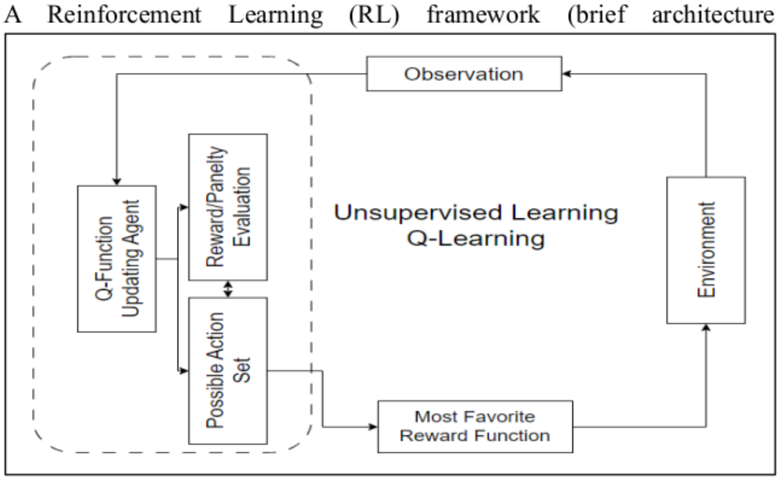

Figure 2: Unsupervised Learning (Reinforcement Learning)) will be developed to model the decision-making process in stock trading based on the sentiment analysis results. RL algorithms combined with Deep Neural network (LSTM) in Figure 3: LSTM Network (in conjunction with

RL architecture in Figure 2) will be employed to learn optimal trading strategies by interacting with the market environment and maximizing a defined reward function.

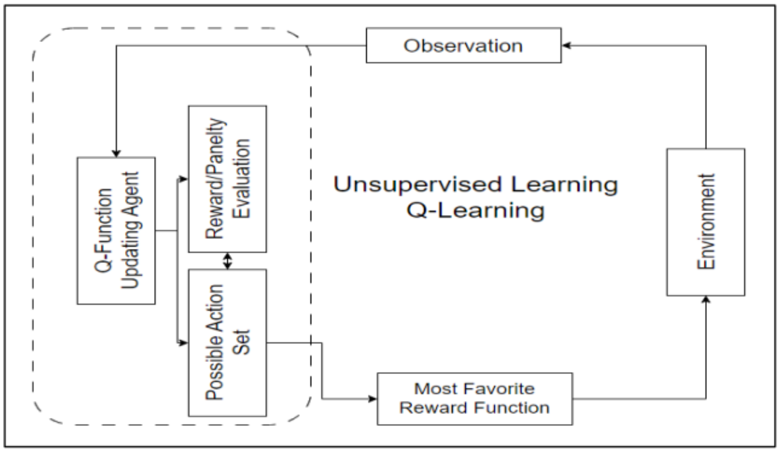

Figure 2: Unsupervised Learning (Reinforcement Learning)

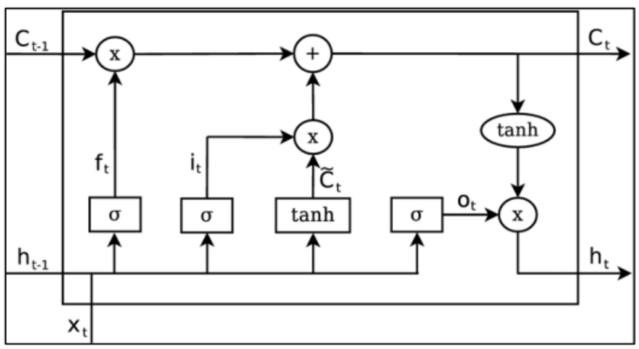

Figure 3: LSTM Network (in conjunction with RL architecture in Figure 2)

7.5. Model Evaluation

For evaluating the proposed NLP and DRL model combining historical market data and texts to generate the insight, it is important to capture the sentiment of the text correctly and corelate it with the historic price data. In order to check the effectiveness of such a model, several performance metrics such as Accuracy, Precision, Recall (Sensitivity), F1-Score can be considered.

Since the aim of this study is also to improve the trading decision and reduce the possible losses and hence some of the financial metrics may also be used to check the performance of the prediction. Some of the such business metrics are listed below with a brief explanation.

- Cumulative Returns: Cumulative returns measure the total returns generated by the investment strategy over a specified period.
- Sharpe Ratio: The Sharpe ratio measures the risk-adjusted return of an investment strategy.

These metrics shall provide a comprehensive evaluation of the model's predictive power, profitability, and risk-adjusted performance.

The findings of the model will be interpreted to identify patterns and correlations between sentiment and price moves. Visualizations, statistical analysis, and qualitative assessments will be used to gain insights into the impact of sentiment on stock prices and financial market dynamics. Various graph depicting superimposition of real price move v/s predicted price moves shall provide a better mean of deciphering the results and make a profitable strategy.

7.6. Limitations

Potential limitations of the research methodology include constraints on data availability, the inherent uncertainty and volatility of financial markets, and simplifying assumptions made in the model.

8. Expected Outcome

The proposed study aims to leverage NLP to extract the sentiment from the tweets relevant to the market movement. The application of LSTM with reinforcement learning will enable identification of the patterns and corelation within stock data enhancing the predictive capabilities of the model. The expected outcome of this study is to assist in developing an investment decision making system/signal for stock market by amalgamating state-of-the-art techniques from DRL and SA, thereby providing the individual investor, traders and market practitioners with more accurate and reliable tool for making informed decision in the dynamic market condition.

9. Required Resources

9.1. Hardware Resources:

Processor:	Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz (or equivalent)
RAM:	8GB
Graphics:	Based on the requirements an external GPU can be used
Storage:	SSD preferred for faster data access
OS:	Windows 10, macOS, or Linux

9.2. Software and Tools:

Programming Language:	Python 3.7 as it has a support for most of the latest ml algo
Libraries and Framework:	TensorFlow, Pytorch, Keras
Data Visualization Tools:	Matplotlib, Tableau

9.3. Textual Data Processing Tools:

Preprocessing textual data: NLTK (Natural Language Toolkit), Spacy

Web Scrapping: Beautiful Soup

Text processing and analysis: spacy

9.4. IDE – interactive development environment

Jupyter Notebook IDE (based on requirements free GPU resources can be used)

VS-Code

10. Research Plan

Research project plan from Dec-2023 to May-2024 has been listed Figure 4: Research Project Plan. This will be further granularized as the program proceeds. It is using a scaling of 3 days so a plan duration of 6 means ~18 to ~19 days of effort.

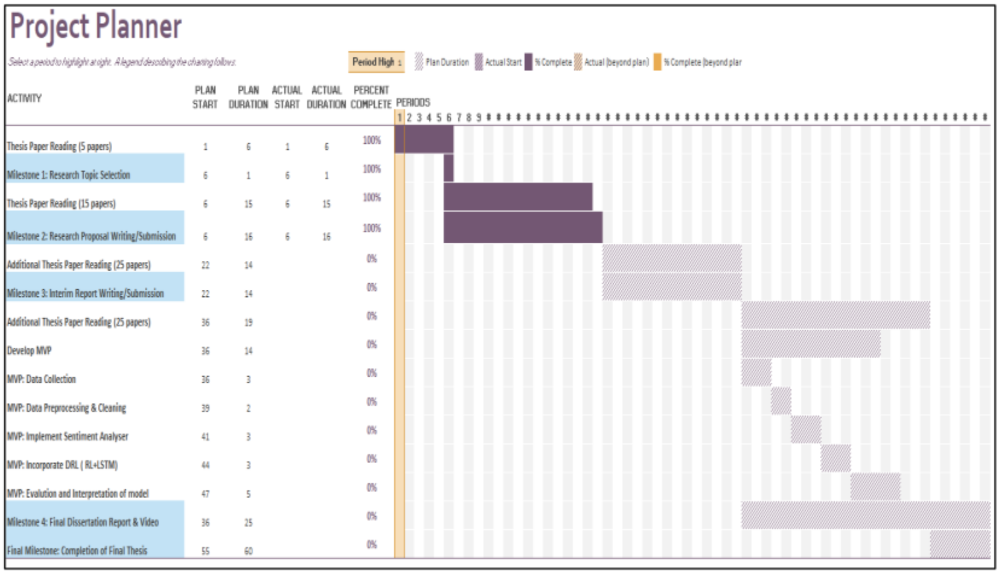

Figure 4: Research Project Plan

11. Risk and Contingency Plan

Table 3: Risk and Contingency Plan

S#	Possible Risk	Contingency Plan
1	Impact on the Research plan timelines due to some professional commitments or responsibilities.	It is possible and at times it is unavoidable however research plan proposed above provisions some buffer.
2	Quality of the thesis is not up to the mark.	Regular catchup with Thesis Supervisor can help get the timely feedback on the improvement of the quality of thesis.
3	Given the outcome of the thesis is quite challenging i.e. combine SA with LSTM+RL, it is possible that it does not produce the expected outcome.	Meeting milestone as listed in the Research plan on time, putting a system of checks and balances, regular catchup with Thesis Supervisor should keep the Research on track and achieve the expected outcome.
4	Computational Resource challenges – training a large set of data using LSTM and RL may require significant computational resource and time.	Optimizing the model by hyperparameter tuning, train at small dataset instead of large datasets, consider using the cloud platform could be some of those options that can help mitigate this risk

References

1. Zou, J., Lou, J., Wang, B., & Liu, S. (2024). A novel Deep Reinforcement Learning based automated stock trading system using cascaded LSTM networks. Expert Systems with Applications, 242. https://doi.org/10.1016/j.eswa.2023.122801
2. Tao, S. (2023). Predicting BMW Stock Price Based on Linear Regression, LSTM, and Random Forest Regression. In BCP Business & Management EMFRM (Vol. 2022).
3. Aken, J., Liang, D., Lin, Z., & Wang, C. (2023). The Application of Deep Reinforcement Learning in Stock Trading Models. Advances in Economics, Management and Political Sciences, 39(1), 215–223. https://doi.org/10.54254/2754-1169/39/20231973
4. Jang, J., & Seong, N. Y. (2023). Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory. Expert Systems with Applications, 218. https://doi.org/10.1016/j.eswa.2023.119556
5. Awad, A. L., Elkaffas, S. M., & Fakhr, M. W. (2023). Stock Market Prediction Using Deep Reinforcement Learning. Applied System Innovation, 6(6). https://doi.org/10.3390/asi6060106
6. Lin, W., Xie, L., & Xu, H. (2023). Deep-Reinforcement-Learning-Based Dynamic Ensemble Model for Stock Prediction. Electronics (Switzerland), 12(21). https://doi.org/10.3390/electronics12214483
7. Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges. In Applied Sciences (Switzerland) (Vol. 13, Issue 3). MDPI. https://doi.org/10.3390/app13031956
8. Yousefi, N. (2022). Deep Reinforcement Learning for Tehran Stock Trading. Indonesian Journal of Data and Science (IJODAS), 3(3), 97–106.
9. Zhang, J., & Lei, Y. (2022). Deep Reinforcement Learning for Stock Prediction. Scientific Programming, 2022. https://doi.org/10.1155/2022/5812546
10. Kang, C. Y., Lee, C. P., & Lim, K. M. (2022). Cryptocurrency Price Prediction with Convolutional Neural Network and Stacked Gated Recurrent Unit. Data, 7(11). https://doi.org/10.3390/data7110149
11. Lawi, A., Mesra, H., & Amir, S. (2022). Implementation of Long Short-Term Memory and Gated Recurrent Units on grouped time-series data to predict stock prices accurately. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00597-0

12. Koukaras, P., Nousi, C., & Tjortjis, C. (2022). Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. Telecom, 3(2), 358–378. https://doi.org/10.3390/telecom3020019
13. Fiorini, P. M., & Fiorini, P. G. (2021). A Simple Reinforcement Learning Algorithm for Stock Trading. Proceedings of the 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2021, 2, 824–830. https://doi.org/10.1109/IDAACS53288.2021.9660900
14. Canese, L., Cardarilli, G. C., di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., & Spanò, S. (2021). Multi-agent reinforcement learning: A review of challenges and applications. In Applied Sciences (Switzerland) (Vol. 11, Issue 11). MDPI AG. https://doi.org/10.3390/app11114948
15. Xu, R., Yang, H., & Hambly, B. (2021). Recent Advances in Reinforcement Learning in Finance. https://www.researchgate.net/publication/356598345
16. Lussange, J., Lazarevich, I., Bourgeois-Gironde, S., Palminteri, S., & Gutkin, B. (2021). Modelling Stock Markets by Multi-agent Reinforcement Learning. Computational Economics, 57(1), 113–147. https://doi.org/10.1007/s10614-020-10038-w

ResearchPaper_RatishMoondra-4.docx

ORIGINALITY REPORT

8%	5%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	2%
2	Submitted to Liverpool John Moores University Student Paper	1%
3	discovery.researcher.life Internet Source	1%
4	journalofbigdata.springeropen.com Internet Source	1%
5	fastercapital.com Internet Source	1%
6	www.mdpi.com Internet Source	1%
7	Wenjing Lin, Liang Xie, Haijiao Xu. "Deep-Reinforcement-Learning-Based Dynamic Ensemble Model for Stock Prediction", Electronics, 2023 Publication	1%

doaj.org

8	Internet Source		1 %
9	pure.au.dk	Internet Source	1 %
10	ebin.pub	Internet Source	1 %

Exclude quotes Off
Exclude bibliography On
Exclude matches < 1%