National College of Ireland

Stock Market Prediction Using Financial News Sentiments and Technical Indicator Data with Machine Learning Models and LIME for Explainable Insights

MSc Research Project
Data Analytics

Sumit More
Student ID: 23108983

School of Computing
National College of Ireland

Supervisor: Dr. Ahmed Maki

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sumit More
Student ID:	23108983
Programme:	MSc Data Analytics **Year:** 2024
Module:	Research Project
Supervisor:	Dr. Ahmed Maki
Submission Due Date:	12/08/2024
Project Title:	Stock Market Prediction Using Financial News Sentiments and Technical Indicator Data with Machine Learning Models and LIME for Explainable Insights
Word Count:	6919 **Page Count** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sumit More
Date:	12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Stock Market Prediction Using Financial News Sentiments and Technical Indicator Data with Machine Learning Models and LIME for Explainable Insights

Sumit More
23108983

Abstract

The stock market is comprised of many input factors, which include stock prices, news sentiments, and technical ones. The general problem of how to forecast stock prices has not been fully solved due to the constant fluctuations of shares. This research aims to enhance stock price forecasting based on the utilization of heterogeneous data and machine learning algorithms and make the black box prediction more interpretable for investors. The purpose is to improve the forecast precision and gain a deeper understanding of the opportunities for the markets. The study involves the quantitative data of the stock prices such as open, high, low, closing, volume, SMA (Simple Moving Average), EMA (Exponential Moving Average), RSI (Relative Strength Index), BBANDS (Bollinger Bands), and News sentiment score data. The methodology has dwelled into three widely used models RandomForesrRegressor, SVR (Support Vector Regressor), LSTM (Long-short term memory) and their hyper-parameter tuned versions for better results. Among all the models, the fine-tuned SVR model has outperformed others. The fine-tined SVR model achieved an MSE (Mean Square Error) of 0.518 and an MAE (Mean Absolute Error) of 0.566. The integration of technical indicators and news sentiment scores along with the LIME (Local Interpretable Model-Agnostic Explanation) explanation can significantly benefit traders and financial analysts by providing more accurate predictions and explanations. The real-time data processing and potential biases in sentiment analysis are the challenges that can be explored further. The final objective of this research is to empower traders and financial analysts to make sound and data-backed decisions in the stock market.

1 Introduction

1.1 Background

The stock market is a global system compounded by many factors that range from economic signals, trends and sentiment. Stock price forecasting is a difficult process that has tremendous consequences for shareholders and other stakeholders. Many of the standard approaches which involve a simple inspection of charts to determine patterns and trends, or mechanical analysis of the stock price patterns based on previous price movements leave much to be desired. The recent integration of big data into grand applications and improvements in machine learning have created promising additional approaches to the sophistication of stock market prognosis by integration of diverse data and application of more methods of models(Wu et al., 2022)

1.2 Motivation

This research seeks to improve the accuracy and the readability of the results of stock market predictions with the help of sentiment analysis, additional technical indicators, and the two-component deep learning models. The motivation stems from the need for more insightful forecasting methods that go beyond conventional approaches. The research explores how opinion mining from news articles and social media can provide insights into market trends that influence stock prices(Visani et al., 2022). This study incorporates historical data on stock prices and technical analysis along with real-time news data using the AlphaVantage API.

1.3 Limitation

The study acknowledges potential limitations, such as the reliance on specific data sources like Alphavantage and the challenges of real-time data integration, which may affect the generalizability and scalability of the model.

1.4 Research Question

- How can sentiment analysis along with additional technical indicators data and machine learning models improve the accuracy and XAI (Explainable Artificial Intelligence) techniques for interpretability of stock market predictions?

1.5 Research Objectives

- To compare the performance of traditional machine learning models and LSTM with different data combinations, including stock data, news data, and technical indicators.
- To use XAI techniques, particularly LIME, to interpret model predictions and assess their interpretability.

The thesis report follows the following structure: In section 2, the author has explained the research problem in brief. Next in section 3 relevant works of literature to this research have been highlighted and reviewed. An elaborated explanation of the methodologies is provided in Section 4. Section 5 describes the design specification, including the Technologies & Frameworks, Architecture, and Proposed Model Functionality. Section 6 shows the implementation of the system. The results evaluation and discussion are covered in section 7. The final section, Section 8, covers the conclusion of the research and future work.

2 Research Problem

Stock price forecasting is a very complicated and, at the same time, highly challenging task that has always been in the limelight for researchers from the field of finance. The basic approaches to collecting data for stock market prediction are only based on stock price history and technical analysis tools that seem to be far from effective in some cases because

they barely take into consideration all elements affecting this process. Market sentiment can be described as an approach popularly incorporated in collecting data from news articles and social portals and represents the basis for an effective assessment of the investors' attitudes. However, merging this qualitative data with purely stock and technical indicators data is quite problematic.

Improved forms of machine learning include the ***"Long Short-Term Memory (LSTM)"*** which is known to present temporal dependencies in the daily stock market data. However, their performances sometimes come at the cost of a lack of interpretability; the models' predictions cannot be easily explained to the stakeholders. This research seeks to address two main problems arises that are predicting the substantially advanced stock market data by fusing the details of stock, and sentiment analysis scores with the technical indicators and fine-tuning the precision of these predictions using ***"Explainable Artificial Intelligence (XAI)"***, one of the well-known methods being ***"LIME (Local Interpretable Model-agnostic Explanations)"***.

3 Related Work

Introduction

Literature review provides understanding about previous experiments based on the market stock and sentiment analysis by using machine learning procedure. Traditional machine learning models or techniques and LSTM model performances are reviewed for getting the proper knowledge for improving this section. Using of XAI technique also reviews in this section, especially the LIME technique. Overall, the literature review part provides a brief idea about the previous research and provide ideas about the whole procedure.

Themes

Evolution of Sentiment Analysis in Financial Markets

The evolution of sentiment analysis in financial markets began with significant research by (Bollen et al., 2011) which stated that sentiments of Twitter comments can predict the movement of the stock market. This research created a pathway to explore on how emotions expressed online could have a tangible impact on financial forecasting. Next (Nguyen et al., 2015) combined sentiment analysis data with traditional economic data. This approach greatly impacted the performance of models by giving good accuracy in predicting the market. This progression reflects the shift in market prediction techniques, where traditional data and sentiment analysis are used together to achieve better forecasting results.

Performance of traditional machine learning models and LSTM with different data combinations

Models like Linear Regression, Decision Trees, Random Forest/SVM are applied in stock market prediction. Linear Regression models help to lay the basis for establishing a correlation between stock prices and some factors(Hu et al., 2021). Also, both Decision Trees and Random Forest models perform well in exploring complex patterns in stock data and provide better predictive accuracy. The results of sentiment analysis from news articles and social networks allow to evaluate the market opinion and investors' sentiment(Research

Scholar, Department of Computer Engineering, K K Wagh Institute of Engineering Education and Research, Nashik, Savitribai Phule Pune University, Pune, Maharashtra, India et al., 2024). Complementary of sentiment scores to the conventional models which includes indicated improvements of prediction accuracy since sentiment data summarises the psychological factors of market movements. SMA, RSI, and Bollinger Bands are normally used to increase the accuracy of prediction. All of these are used to pinpoint market tendencies and possible reversal points. It was found that incorporation of the technical indicators into context of classical models enhances predictive accuracy(Gite et al., 2021). LSTM model, which is a type of recurrent neural network, it is possible to capture temporal dependences and perform time series prediction(Freeborough and Van Zyl, 2022). The researcher (Tabinda Kokab et al., 2022) explored the recent advancements in sentiment analysis that use Transformer-based models for analyzing financial sentiment. Their study found that transformer models significantly outperformed traditional LSTM models in both accuracy and efficiency when processing financial news data. This contrasts with the finding of (Zhang et al., 2017), which stated the supremacy of the LSTM model. This highlights the exploration of the Transformer model's ongoing evolution and improvement of sentiment analysis techniques in financial markets. This transition is especially important as it opens up the possibility of integrating these newer models with explainability techniques to better understand and interpret their predictions.

Implementation of XAI techniques in stock analysis

There is also a need for interpreting machine learning model, this is where an explainable AI comes in handy and is commonly referred to as XAI. It is crucial for obtaining trust from the stakeholders in finance(Kumar et al., 2024). The lack of clarity in the prediction of black-box models in financial decision-making was highlighted by (Kraus and Feuerriegel, 2017), echoing earlier concerns raised by (Ribeiro et al., 2016), who questioned the reliability of these opaque models. The focus of both studies was on the importance of transparency and interpretability in models used for critical financial decisions, which highlights the ongoing need for solutions that make these complex models more understandable and trustworthy. LIME is another effective XAI approach that interprets black box’s decision by locally fitting the model to be an interpretable one near the prediction. It creates explanations by modifying input data and analyzing the shifts in the output of the model, which is important for feature importance of the outcomes(Rezaei et al., 2021). LIME has been utilized in the interpretation of the predictions given in credit scoring, fraud, and stock market prediction. LIME can help to improve the accuracy to a certain extent and locally expound the phenomenon, thus increasing the credibility of financial models.

Conceptual framework

Assimilating machine learning models, LSTM networks, sentiment analysis, technical indicators and XAI is helpful in making efficient stock prediction. The traditional models give a basic architecture and LSTM networks learn sequences and temporal dependencies. Optimization of the set of predictions is complemented by sentiment analysis and technical indicators. Thus, through XAI techniques, especially LIME, it is possible to guarantee that the predictions made are comprehensible and explainable.

Literature gap

This paper is an advancement of previous work in stock prediction since there are limitations to combining various kind of data for instance historical stock data, sentiment analysis scores, and technical indicators data sets. While it is a common practice to use only traditional machine learning models or LSTM networks, the authors do not exploit the synergies of the

two in most of the research works. One of the important points is the lack of attention to the interpretability of models, as the models' users need to understand them to make their decisions.

Summary

Literature review has also given the subsequent analysis of the current literature in the field of stock market prediction regarding the traditional machine learning approaches such as baseline models, LSTM networks, sentiment analysis, and technical indicators inclusive of ELM, as well as explainable AI techniques to enhance accountability in decision-making. Therefore, the review points to the future research directions and how they can be used in enhancing more accurate and comprehensible prediction models for the stock market. It could be stated that the further enhancement of this field can be achieved with the help of the integration of traditional models, LSTM networks, and XAI techniques.

4 Research Methodology

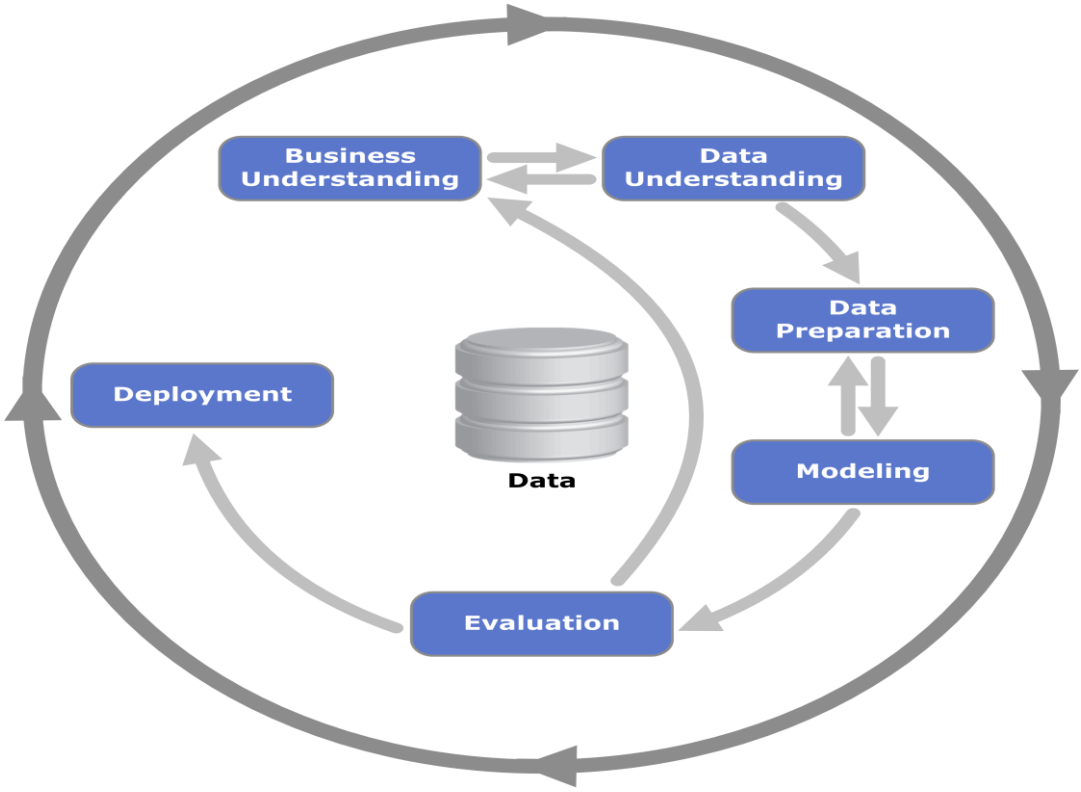


Figure 1. CRISP-DM Methodology

4.1 Research Method

The CRISP-DM (Cross-Industry Standard Process for Data Mining) approach has a proven track record in the case of critical and complex data mining projects. The stages of the CRISP-DM framework are well-suited for designing the project flow starting from requirement understanding to deployment of the final model as shown in Figure 1. Each phase of CRISP-DM is explained in detail in the following sections as it applies to the project.(Martinez-Plumed et al., 2021)

4.2 Business Understanding

In the first phase of CRISP-DM, the understanding of the business context, objectives, and requirements is very important. The primary objective of this research is to improve the stock prediction accuracy and transparency of output predicted by the model. To achieve the goal of improved accuracy, along with Traditional Stock data financial news data and technical indicators data were used to train machine learning and deep learning models. For interpretable model predictions the XAI technique LIME is used. Improved accuracy and interpretability will help investors make data-backed decisions.

4.3 Data Understanding

In the data understanding phase, only the required datasets were gathered and explored for analysis. The author has taken Google's stock, news, and technical indicators data for this research. The following are the data sources included in the research:

- **Historical Stock Data:** The stock daily price data such as DATE, OPEN, HIGH, LOW, CLOSE, and VOLUME were obtained from one of the open-source platforms named Alphavantage. The sourced data covers the broad period to capture various market conditions.
- **News Sentiment Data:** Date, New Title, News Summary, News Sentiment type (Bullish, Somewhat Bullish, Neutral, Somewhat Bearish, and Bearish), and their sentiment score were extracted from News data. The News Sentiment data were obtained from Alphavantage API. The type of news considered to fetch were TECHNOLOGY, EARNINGS, BLOCKCHAIN, and IPO. The sentiment scores were aggregated daily to align with the stock price data.
- **Technical Indicators Data:** The technical indicators data such as Simple Moving Average, Exponential Moving Average, Relative Strength Index, and Bollinger Bands were sourced from Alphavantage API. The sourced data covered a vast period to capture the trends.

Basic EDA (exploratory data analysis) was performed to get an overview of the data. The EDA steps involve checking for the existence of null values, duplicate values, and correction of the data types. Visualizations and statistical summaries helped to make better data in the data preparation phase.

4.4 Data Preparation

The anomalies or glitches identified in data in the data understanding phase are addressed in the Data Preparation phase. In the data preparation phase, the raw data is transformed into a model training suitable data. The following are the steps included:

- **Data Cleaning:** The missing values were addressed in news sentiment data using techniques like forward and backward filling, depending on the data's nature. The duplicate values were dropped from the News Summary column of new sentiment data. There were no missing values or duplicate rows observed in stock and technical indicators data. The date column was typecasted from Object to DataTime format across all three datasets.
- **Data Merging:** Before merging renaming of the columns was done in all the datasets to have standard lowercase column names and then all three datasets were merged on the column name DATE which was present in all three datasets.

- **Feature Engineering:** Lagged features were created for the stock price data to capture temporal dependencies. Sentiment scores were aligned with stock price data to ensure consistency across the dataset.
- **Data Transformation:** The data normalization was done by minmax scaling to ensure the consistency of the data during the model training. This step significantly impacts the deep learning models as they are sensitive to the scale of input.
- **Data Splitting:** The prepared dataset was then split into training and testing sets, with 80% used for training the models and 20% reserved for testing.

4.5 Modeling

In the modeling phase, the author used a regression model which is well-suited for the problem statement of predicting daily fluctuation in the stock data. Different machine learning and deep learning were implemented and evaluated. The model used in this research are:

- **RandomForest Regressor:** The RandomForest is a bagging ensemble method. In this technique, the randomly sampled data were used to construct the multiple trees and then aggregate their prediction. It served as a baseline model for stock prediction performance.
- **Support Vector Regression:** SVR is a technique that minimizes the prediction errors within a specified margin. The SVR model efficiently handles small to medium-sized data. It also handles the overfitting issue. The randomized search hyperparameter tuning was performed to tune SVR model parameters such as kernel type, regularization parameter (C), and epsilon.
- **LSTM Network:** The stock price data is majorly a sequential time-series data. LSTM networks work best with sequential time-series data as they capture the long-term dependencies in the sequential data. The model architecture included multiple LSTM layers followed by dense layers, designed to learn from sequences of stock prices, sentiment scores, and technical indicators to predict future prices.

Each model and hyper-parameter tuned version was trained on the three different datasets. The first dataset only consisted of stock price data, the second dataset was a combination of stock price data and new sentiment data, and lastly, the third dataset was a combination of stock data, news sentiment data, and technical indicators data. The evaluation metrics such as MSE and MAE were used to check the model performance.

4.6 Evaluation and Explainability

The experiments are to be designed and performed to assess the models with different data examples. The assessment of models includes stocks only, stocks with sentiment scores, as well as stocks with sentiment scores and technical indicators. For the regression model, the evaluation criteria for accuracy are Mean Square Error, Mean Absolute Error, Root Mean Squared Error (RMSE), and R-squared (R^2)(Li and Hu, 2024). Such an arrangement offers a systematic way to define the effects of other sources of information on model performance as well as evaluate how these factors enhance the model's capacity for stock market prediction(Mehtab et al., 2020).

4.7 Deployment

The full deployment of the model is not within the academic study scope. The deployment phase may involve the integration of the model into a production environment where it can be used for real-time stock market predictions. Implementing real-time data pipelines, model monitoring, and continuous evaluation will be necessary for ensuring the model's effectiveness and scalability in a live setting.

4.8 Material and Equipment

The following materials and equipment were used in this study:

- **Hardware:**
  - **Processor:** Intel Core i7
  - **RAM:** 16GB
  - **Storage:** 1TB SSD
- **Software:**
  - **Operating System:** Windows 11
  - **Integrated Development Environment (IDE):** Jupyter Notebook
  - **Programming Language:** Python 3.8+
  - **Python Libraries:** pandas, numpy, scikit-learn, TensorFlow, LIME, requests, and matplotlib.

4.9 Alternative Methodologies

Other techniques could be adopting other machine learning algorithms like Gradient Boosting and XGBoost or use of other methods of sentimental analysis like deploying deep sentiment analysis. These alternatives differ by certain characteristics including the degree of the model’s capacity for prediction accuracy, computational speed, and readability(Rouf et al., 2021). This research has also involved a comparison of the efficiency, scalability, and flexibility of these methods. Gradient Boosting and XGBoost lead to better prediction performance and faster calculations, and deep sentiments bring a more sophisticated understanding of the market sentiments. Regarding scalability, these techniques vary, and, for instance, by applying LSTM with LIME, with the best balance of explaining capacity and potency. The selected method of using stock data with technical indicators for further analysis is justified by its ability to work with real-time data and provide good results in the presence of market fluctuations.

4.10 Justification for Current Method

The chosen approach implies the use of stock data and basic and advanced technical indicators with sentiment analysis and hybrid deep learning models, thus providing a rich picture of the market(Shahi et al., 2020). LIME improves interpretability since this is a major issue in complex models. This has a practical advantage in the field of financial modulations as it states both accuracy in prediction and comprehensible logical steps. In comparison with the other approaches, the use of the suggested method is effective in capturing multiple factors of the market and adaptable to the variation in data(Research Scholar, Department of Computer Engineering, K K Wagh Institute of Engineering Education and Research, Nashik, Savitribai Phule Pune University, Pune, Maharashtra, India et al., 2024). Based on the ability to work with real-time data and high stability in analysis, it is necessary to choose this option because its advantages give a balanced solution to the problem of enhancing the prediction of the stock market(Huang et al., 2023).

4.11 Summary

The research methodology of this study proposes a methodical approach to improving stock market forecasts via the accumulation and analysis of various types of data as well as more sophisticated modeling tools. In collecting data, preparing the data set, and developing an AI model, as well as the LIME interpretability analysis, the study relies on Alphavantage API and various Python libraries. The introduced mixed methods compared to the traditional ones are characterized by a higher accuracy and the possibility to show the work. Other frameworks are consulted to confirm the appropriateness of the selected method for optimizing the solution's speed, expansiveness, and flexibility. Such a complete approach suggests bringing improved, valid, and explainable predictions to the existing stock market, hence supporting the decision-making process.

5 Design Specifications

This section provides a detailed description of the design specifications for the stock market prediction system developed in this study. The design specification outlines the architecture, data flow, and implementation of the models used in this research to improve stock market predictions by incorporating sentiment score and technical indicators data in stock market data. The designs are planned considering the model's robustness, scalability, and easy interpretation for achieving high prediction accuracy.

5.1 Techniques & Frameworks

In this section, the techniques and frameworks used to develop the project are explained. These steps ensure the system operates efficiently and reliably to achieve the research goals.

Techniques:

- **Data Collection:** The Alphavantage API fetches the real-time as well as the historical stock data, news data, and technical indicators data.
- **Data Cleaning:** The missing values were handled with forward fill and backward fill. The duplicate rows were dropped.
- **Data Correction:** The column names of datasets were corrected and kept standard lowercase. Columns like DATE were typecasted to the correct datatype.
- **Data Merging:** All the datasets were merged on DATE column.
- **Feature Engineering:** Lag features were created for close prices, and sentiment scores were aggregated daily.
- **Feature Scaling:** The data normalization techniques like minmax scaler were applied to normalize data between 0.0 to 1.0
- **Modeling:** The regressor models were implemented such as RandomForest Regressor, SVR, and LSTM
- **Hyper-parameter Tuning:** RandomizedSearchCV method were performed to improve the model performance.
- **XAI:** LIME were integrated to provide the local interpretation of model prediction. This makes system more transparent and trustworthy.

Framework:
The framework used in this project is:

- **Sci-kit learn:** It's a Python library that provides a range of algorithms and evaluation methods
- **LIME:** local interpretable model agnostic explanation is a framework to interpret the prediction of the black-box models by resembling it with a locally interpretable model.

6 Implementation

The standard machine learning steps that need to be performed to implement any machine learning use-case are also involved in the implementation of the stock market prediction system. The involved are Data collection, Preprocessing, Model Training, and Evaluation of the model. This section also covers the comparative qualitative analysis of the developed model against the existing solutions, considering factors like ease of implementation, scalability, and maintenance.

6.1 Data Collection & Preprocessing

Data Sources: The data is obtained from open-source API Alphavantage which provides a vast variety of stock market data. For this research stock data, News Sentiment Data, and Technical Indicators like SMA, EMA, RSI, and BBANDS data have been fetched through API.

Preprocessing Steps:

- **Data Cleaning:** The missing values were handled by the forward and backward fill method and the duplicate rows were dropped to ensure the sanity of the data.
- **Feature Engineering:** To capture the temporal dependencies the lag features of Stock Close features were created. The daily sentiment score was aggregated for integration of sentiment analysis into prediction models.
- **Data Merging and Scaling:** The datasets were merged on the DATE column across all the datasets. Feature scaling such as MinMaxScaling used to normalize the data before model training.

6.2 Model Development

The three models were considered for implementation such as RandomForestRegressor, Support Vector Regressor, and Long Short-Term Memory Network. Below is the structure of the model implementation:

Model Training:

- **RandomForest:** The RandomForest model is an ensemble learning technique that trains multiple decision trees on randomly sampled data and then aggregates the prediction output. RandomForest was chosen for its ability to handle large and complex data patterns
- **SVR**: This model was implemented to fit a hyperplane that best represents the data.
- **LSTM:** LSTM is a best to capture long-term dependencies in sequential data, this model is ideal for time-series forecasting.

Hyper-parameter Tuning:
The RandomizedSearched Cross-validation (CV) technique was used to fine-tune the hyper-parameters of the selected model. The hyper-parameter tuning helped for the best possible performance on the dataset.

6.3 Model Evaluation

This research solving the regression problem, therefore the performance metrics used to evaluate the models are:

- Mean Squared Error
- Mean Absolute Error

6.4 Conclusion

The implementation of this thesis seeks the balance between performance and practical considerations like ease of implementation, scalability, and maintenance. The tuned LSTM model gave a good performance but simple performance, whereas base and tuned RandomForest performed very well with a more scalable solution and rapid development. Lastly, the tuned-SVR model outperformed other models but proved to be more complex and less scalable for large datasets. Overall, the comparative qualitative analysis highlights the trade-offs between these models, helping guide the choice of the most appropriate one depending on the specific use case.

7 Result & Critical Analysis

7.1 Result of Stock Data Analysis

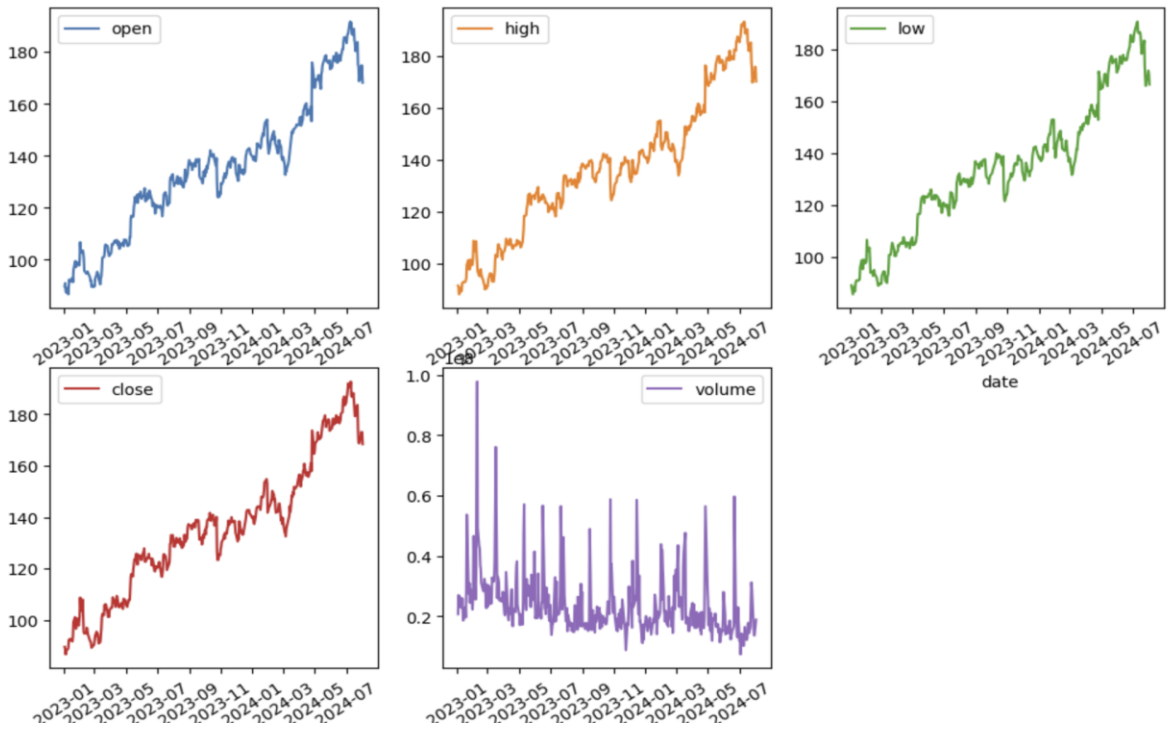
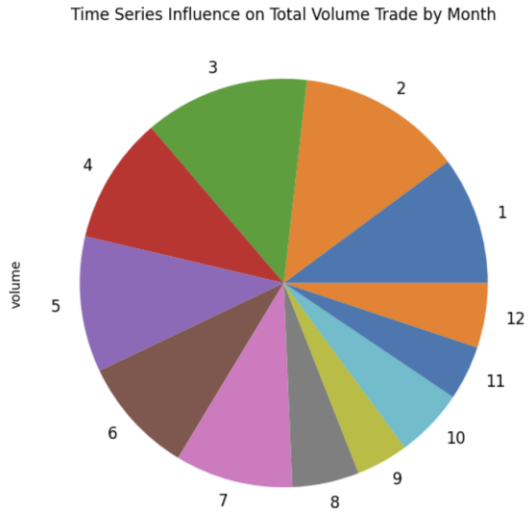


Figure 2. Stock Trend over Time

Figure 2 shows the time series line plots for open, high, low, close, and volume which shows the value trend from January 2023 to July 2024. The trend of the stock is upward going till

mid-2024 before it shows a slight decline. The trend in the volume is high at the beginning, followed by a decline and subsequent stabilization at lower levels.



	volume
1	954317576
2	1229731979
3	1232379976
4	948284857
5	1012296581
6	872886032
7	880151393
8	499734109
9	389661374
10	514909680
11	405702609
12	481964446

Figure 3. Monthly Influence on Total Volume Trade

The Figure 3 shows the pie chart distribution of total trading volume across different months. It has been observed that trading volume is unevenly distributed across the months. Early four to five months show a high trading volume, while the lowest volume was observed towards the end of the year. This pattern may be indicative of cyclical market trends or responses to external market events.

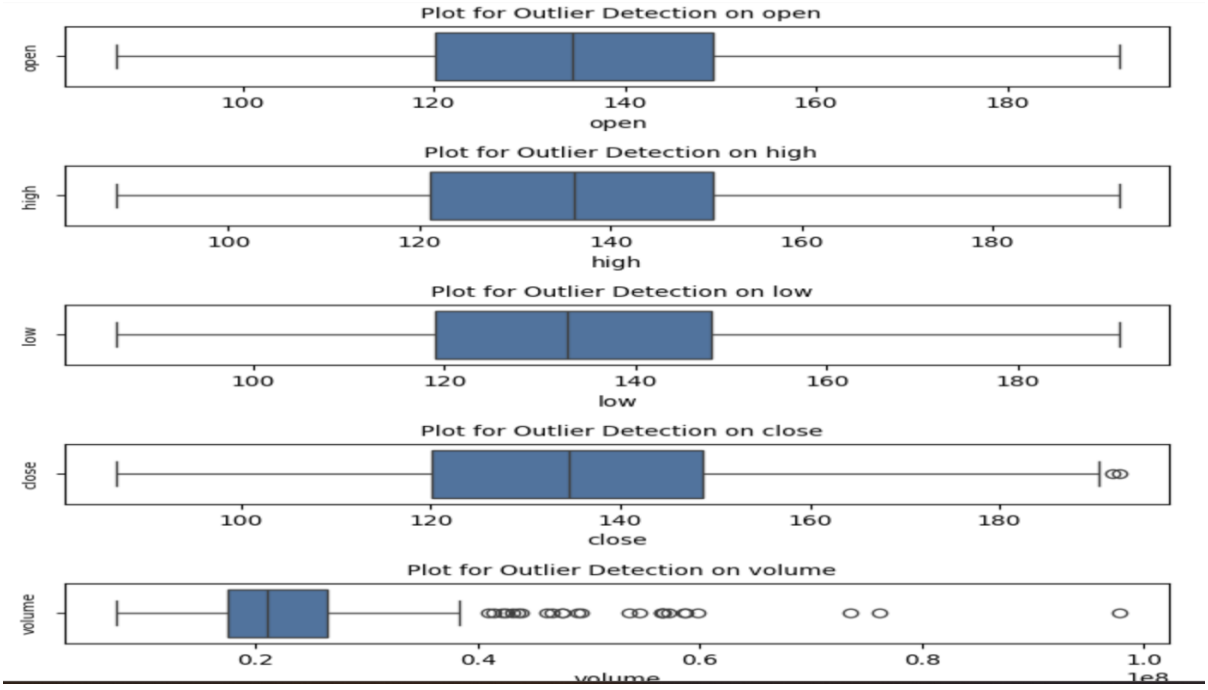


Figure 4. Outlier Detection in Stock Prices and Volume

As shown in Figure 4 to check the presence of outliers the boxplot was utilized to analyze the open, high, low, close, and volume variables. Very minimal outliers were observed in open, high, low, and close which suggests that price data is stable and remains within the expected range. In Volume, a significant number of outliers were observed which indicates the variability in trading volumes.

7.2 Result of News Data Analysis

Figure 5. Positive Word Cloud

The word cloud generated from positive sentiments highlights multiple occurrences of positive words from the positively classified news articles is shown in Figure 5. Highly common key terms are "stock," "company," "market," "investor," and "google". The frequent occurrence of "stock" and "company" suggests that positive sentiment is closely associated with a strong performance in the stock market or favorable economic conditions. There are more technology terms such as "google," "ai," and "nvidia" indicate that advancements in technology and the strong performance of tech companies.

Figure 6. Negative Word Cloud

The word cloud generated from negative sentiments highlights multiple occurrences of negative words from the negatively classified news articles is shown in Figure 6. Highly common key terms are "market," "trade," "bybit," "crypto," and "canada". Multiple occurrences of "market" and "trade" in a negative cloud indicate that market activities and

trading have negative sentiments in the news. Additionally, the presence of cryptocurrency-related terms such as "crypto," "bitcoin," and "ethereum" suggests that negative sentiment is also influenced by volatility and adverse events within the cryptocurrency markets.

7.3 Result of News vs Stock Data Analysis

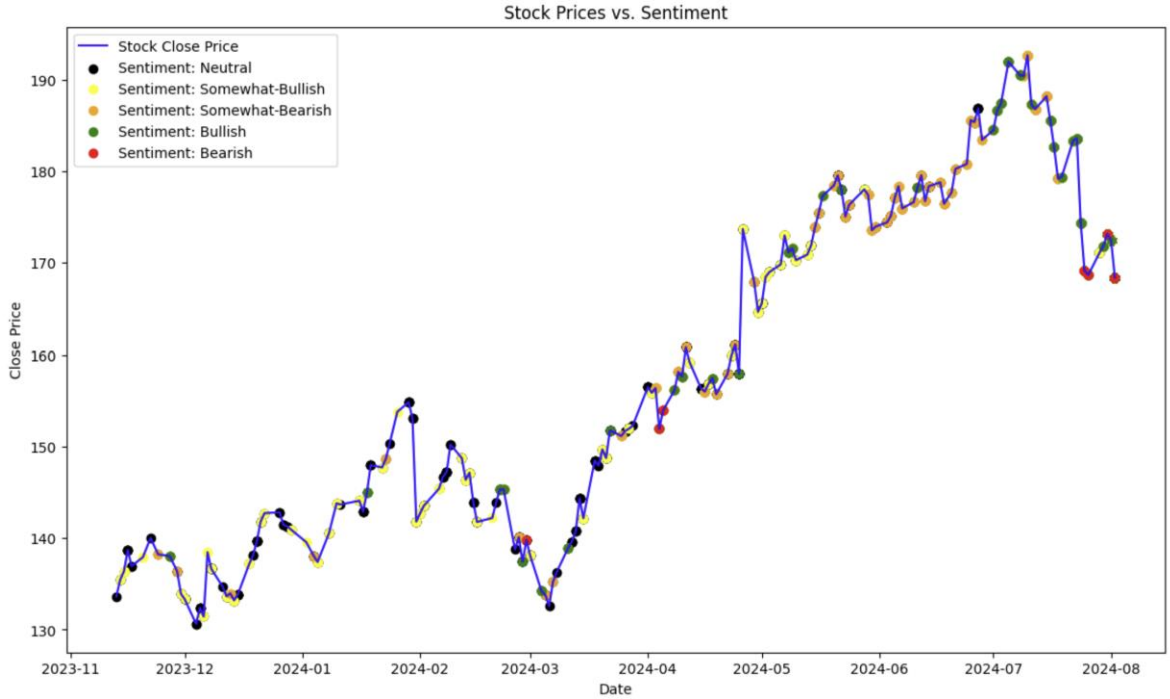


Figure 7. Stock Price vs New Sentiments

Figure 7 shows the line plot of stock closing price over time, with sentiment type. A correlation between stock closing movement and sentiment is observed. The sentiment classifications generally align with stock price movements. Bullish sentiments (indicated by green markers) tend to the upward price trends and bearish sentiments (indicated by red markers) are seen during price declines.

7.4 Model Performance Analysis

Data	Model	MSE	MAE
Stock Data	RandomForest	398.487624	17.736316
	SVR	2136.655099	43.180699
	LSTM	159.3117	9.2117
Stock + Sentiment Score Data	RandomForest	1.100575	0.804163
	SVR	33.503003	3.296613
	LSTM	3.720993	1.583064
Stock + Sentiment Score + Technical Indicator Data	RandomForest	1.346972	0.933059
	SVR	45.369749	3.887562
	LSTM	2.570141	1.272673
	Tuned-RandomForest	1.996752	1.091263
	Tuned-SVR	**0.518251**	**0.566495**
	Tuned-LSTM	4.561768	1.652105

Table 1. Model Results

Table 1 shows the resultant model performance on all three experimental data. In the first case of only Stock data, all models showed poor performance with high MSE and MAE scores, whereas LSTM slightly performed well with MSE 159.3 and MAE 9.2. In the second case where Stock data was combined with the News Sentiment Score, all three model performances were drastically improved where RandomForest performed very well with MSE 1.10 and MAE 0.8. In the last case of the dataset where stock, sentiment score, and technical indicators data were combined, the hyper-parameter tuned version of SVR outperformed the other two models with MSE 0.51 and MAE 0.56. There were slight changes observed in the base RandomForest model & LSTM model and tuned RandomForest model & LSTM model on the last case data.

7.5 Comparative Qualitative Analysis

The models were also compared with existing solutions using a qualitative approach, focusing on ease of implementation, scalability, and maintenance.

Ease of Implementation:

- **RandomForest:** The RandomForest Regressor is easy to implement and works well to find patterns in large-size data. Its simplicity and reliability make it a solid choice for quick development and deployment.
- **SVR:** The SVR model is effective with non-linear data. Implementation of the SVR model is a little complex compared to RandomForest as it requires more careful tuning of parameters like kernel function.
- **LSTM:** Implementation of the LSTM network is more complex compared to RandomForest and SVR model as requires decisions about the network architecture, like the number of layers and units per layer. Once the LSTM architecture is in place it gives a very good performance for time-series data.

Scalability:

- **RandomForest:** The RandomForest model scales well with increasing size of data as its random sampling data and parallel training natures, allowing multiple decision trees to be trained at the same time. The increase in number of the trees may slightly hamper the performance of the model.
- **SVR:** SVR is computationally very expensive therefore scaling SVR with large datasets is not efficient, especially with non-linear kernels.
- **LSTM:** The LSTM models are scalable but require significant computational resources to deal with long sequences or large datasets. Implementation of LSTM often needs the use of GPUs or distributed computing, adding to the complexity of scaling.

Maintenance:

- **RandomForest:** The constant need for retraining the model is quite less because of its stability across different datasets. Therefore, this model is low in maintenance.
- **SVR:** The SVR model requires more attention in terms of periodic parameter tuning, particularly if the data distribution changes over time.
- **LSTM:** The LSTM models require regular maintenance, especially when underlying data patterns may shift. To maintain accuracy constant retraining with updated data is often necessary.

7.6 LIME Interpretation

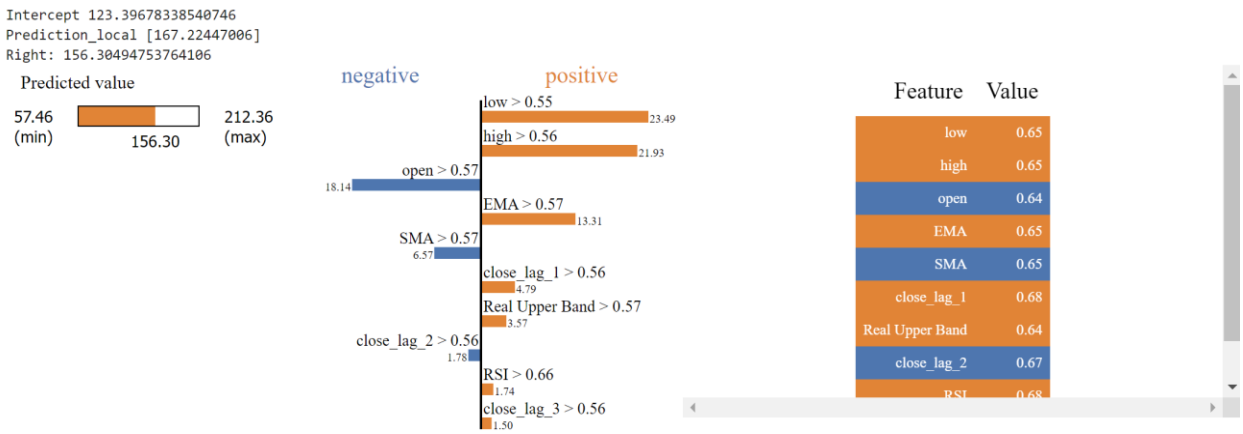


Figure 8. LIME Interpretation

Prediction: The local predicted stock price is 167.22, adjusted upward from the base value of 123.40.

Positive Influences:

- **Low Price (23.49):** A higher recent low price suggests market strength.
- **High Price (21.93):** Recent high price indicates strong market performance.
- **EMA (13.31):** The upward trend in the Exponential Moving Average signals continued momentum.
- **Close Lag 1 (4.79):** Recent closing prices reflect gains.
- **Real Upper Band (3.57):** Positive volatility suggests potential for more gains.

Negative Influences:

- **Open Price (-18.14):** A lower opening price shows early trading weakness.
- **SMA (-6.57):** Slight downward trend over time reduces momentum
- **Close Lag 2 (-1.78):** A slight dip two days ago affects the prediction.

Conclusion: The predicted stock price of 167.22 is driven by strong market performance and upward momentum, slightly tempered by early trading weakness and a minor recent dip.

8 Conclusion and Future Work

This section critically checks the current design of the study, discusses the findings, and suggests improvements and future directions. The performance of the models and their implications for predicting stocks are highlighted. This research successfully integrated various data sources such as stock prices, sentiment scores, and technical indicators into the prediction models. The detailed evaluation and the working of machine learning models are also given in the study which are RandomForest, SVR, and LSTM. The Stock + Sentiment Score + Technical Indicator data allowed models to better capture the patterns and improve the results.

Although the tuned-SVR model outperformed the other two models, refinement in the tuning process may improve the results for RandomForest and LSTM Models. The addition of social media sentiments and macroeconomic indicators may increase the horizon of market conditions. Advanced feature engineering techniques could boost the model's performance.

The results are aligned with existing research which emphasizes the importance of sentiment score and technical indicators in stock price prediction. The SVR model needed significant fine-tuning to outperform other models. LSTM performed well with time-series data, and it ensures its effectiveness in capturing market trends. The RandomForest model performed very well and showed its stability and reliability.

The results challenge some traditional assumptions, specifically the belief that simple models like SVR or RandomForest and their tuned versions can outperform more complex models like LSTM. It also highlights that market sentiment, often considered noise, can be a valuable predictor of stock prices, challenging traditional market efficiency theories.

The study made substantial progress in improving the stock price prediction model, however, some limitations should be acknowledged. The dataset was used with limited timespan which may limit the generalizability of the results. The LSTM model can be computationally expensive and may not be ideal for real-time predictions.

Further research explores the addition of more alternative data sources such as economic indicators, social media sentiment, or news from various regions, to enhance prediction accuracy. Additional beautification of the current methods of machine learning and analyzing the possibility of using the combination of different kinds of models can enhance the accuracy of predictions and provide a better understanding of the key market trends (Garreau and von Luxburg, 2020). Real-time prediction using streaming data with low latency will be a breakthrough in the stock market industry (Thrun, 2022). While LIME performed very well by providing an elaborate explanation of the model prediction, further research can compare the explainability of LIME and SHAP for stock data.

To conclude, this research showed the importance of combining sentiment scores and technical indicators in stock data for improved stock prediction. The implementation of LIME for transparency has also shown trust in model prediction. The comparison of different models provides practical insights into their strengths and weaknesses, helping to guide the choice of models for specific needs. The study suggests that simple models like RandomForest and SVR are very effective for financial forecasting. Future research should continue to refine these models and explore new ways to apply them in real-world financial scenarios.

References

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. J. Comput. Sci. 2, 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Freeborough, W., Van Zyl, T., 2022. Investigating Explainability Methods in Recurrent Neural Network Architectures for Financial Time Series Data. Appl. Sci. 12, 1427. https://doi.org/10.3390/app12031427

Garreau, D., von Luxburg, U., 2020. Explaining the Explainer: A First Theoretical Analysis of LIME. https://doi.org/10.48550/ARXIV.2001.03447

Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., Pandey, N., 2021. Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Comput. Sci. 7, e340. https://doi.org/10.7717/peerj-cs.340

Hu, Z., Zhao, Y., Khushi, M., 2021. A Survey of Forex and Stock Price Prediction Using Deep Learning. Appl. Syst. Innov. 4, 9. https://doi.org/10.3390/asi4010009

Huang, J.-Y., Tung, C.-L., Lin, W.-Z., 2023. Using Social Network Sentiment Analysis and Genetic Algorithm to Improve the Stock Prediction Accuracy of the Deep Learning-Based Approach. Int. J. Comput. Intell. Syst. 16, 93. https://doi.org/10.1007/s44196-023-00276-9

Kraus, M., Feuerriegel, S., 2017. Decision support from financial disclosures with deep neural networks and transfer learning. Decis. Support Syst. 104, 38–48. https://doi.org/10.1016/j.dss.2017.10.001

Kumar, P., Hota, L., Tikkiwal, V.A., Kumar, A., 2024. Analysing Forecasting of Stock Prices: An Explainable AI Approach. Procedia Comput. Sci. 235, 2009–2016. https://doi.org/10.1016/j.procs.2024.04.190

Li, H., Hu, J., 2024. A Hybrid Deep Learning Framework for Stock Price Prediction Considering the Investor Sentiment of Online Forum Enhanced by Popularity. https://doi.org/10.48550/ARXIV.2405.10584

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M.J., Flach, P., 2021. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Trans. Knowl. Data Eng. 33, 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mehtab, S., Sen, J., Dutta, A., 2020. Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models. https://doi.org/10.48550/ARXIV.2009.10819

Nguyen, T.H., Shirai, K., Velcin, J., 2015. Sentiment analysis on social media for stock movement prediction. Expert Syst. Appl. 42, 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Research Scholar, Department of Computer Engineering, K K Wagh Institute of Engineering Education and Research, Nashik, Savitribai Phule Pune University, Pune, Maharashtra, India, Kasture, P., Shirsath, K., 2024. Enhancing Stock Market Prediction: A Hybrid RNN-LSTM Framework with Sentiment Analysis. Indian J. Sci. Technol. 17, 1880–1888. https://doi.org/10.17485/IJST/v17i18.466

Rezaei, H., Faaljou, H., Mansourfar, G., 2021. Stock price prediction using deep learning and frequency decomposition. Expert Syst. Appl. 169, 114332. https://doi.org/10.1016/j.eswa.2020.114332

Rouf, N., Malik, M.B., Arif, T., Sharma, S., Singh, S., Aich, S., Kim, H.-C., 2021. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Electronics 10, 2717. https://doi.org/10.3390/electronics10212717

Shahi, T.B., Shrestha, A., Neupane, A., Guo, W., 2020. Stock Price Forecasting with Deep Learning: A Comparative Study. Mathematics 8, 1441. https://doi.org/10.3390/math8091441

Tabinda Kokab, S., Asghar, S., Naz, S., 2022. Transformer-based deep learning models for the sentiment analysis of social media data. Array 14, 100157. https://doi.org/10.1016/j.array.2022.100157

Thrun, M.C., 2022. Exploiting Distance-Based Structures in Data Using an Explainable AI for Stock Picking. Information 13, 51. https://doi.org/10.3390/info13020051

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., Capuzzo, D., 2022. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. J. Oper. Res. Soc. 73, 91–101. https://doi.org/10.1080/01605682.2020.1865846

Wu, S., Liu, Y., Zou, Z., Weng, T.-H., 2022. S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. Connect. Sci. 34, 44–62. https://doi.org/10.1080/09540091.2021.1940101

Zhang, L., Aggarwal, C., Qi, G.-J., 2017. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Halifax NS Canada, pp. 2141–2149. https://doi.org/10.1145/3097983.3098117