SENTIMENTAL ANALYSIS FOR STOCK MARKET PRICE PREDICTION SYSTEM

Batch Number: CSD-G04

Roll Number	Student Name
20211CSD0043	Darshan M
20211CSD0072	Umme Kulsum
20211CSD0125	Gaurav H
20211CSD0129	Srivatsa K S

Under the Supervision of,

Dr. Manjunath K V

Associate Professor

School of Computer Science and Engineering Presidency University

Name of the Program: Computer Science And Data science
Name of the HoD: Dr. Saira Banu Atham
Name of the Program Project Coordinator: Dr. Manjula H M
Name of the School Project Coordinators: Dr. Sampath A K

Problem Statement

Organization: Ministry of Commerce and Industries Category (Hardware / Software Both) : Software Problem Description:

The stock market is highly influenced by public sentiment, which is often reflected in news articles, financial blogs, and social media platforms. Investors often make decisions based on emotional responses to market news, leading to price fluctuations.

Github Link

This Github link provided have public access permission.

GITHUB_LINK

Literature Survey

The increasing complexity and volatility of financial markets have highlighted the need for more intelligent, data-driven forecasting techniques. Traditional statistical and rule-based models often fail to incorporate unstructured, sentiment-rich data such as news articles and social media content. Hence, recent literature has shifted toward Sentiment Analysis-based approaches for stock price prediction using Natural Language Processing (NLP) and Deep Learning (DL) methods.

1. **Significance of Sentiment in Financial Markets**

Several studies have demonstrated a strong correlation between market sentiment and stock price movements. Investors' emotions, reflected in news headlines, tweets, and reports, often lead to market reactions even before traditional technical indicators show signs. Therefore, sentiment extraction and interpretation play a pivotal role in improving market forecasts and trading strategies

2. Traditional Machine Learning Techniques

Earlier sentiment classification models employed **Support Vector Machines (SVMs)** and **Random Forests**, which offered decent performance in basic classification tasks. However, they lacked the ability to process sequential dependencies and contextual meaning inherent in natural language. Their inability to scale or generalize across varying financial contexts posed significant limitations.

Literature Survey

Singh et al. – Feature-Specific Sentiment Analysis on E-commerce Data
Applied sentiment filtering on product reviews using MongoDB. Though not financial, the system highlighted the effectiveness of cleaning noisy data and filtering irrelevant sentiment for better accuracy.

Lu & Chen – Opinion Analysis on Microblogs
Used Support Vector Machine (SVM) to classify microblog content with 90%+ precision. Their modular system (data collection, corpus processing, sentiment detection) is a foundational structure still used in sentiment systems today.

Batool et al. (2018) – Knowledge Enhancement for Tweet Sentiment Classification
This research enhanced sentiment classification accuracy for Twitter data by implementing synonym binding and knowledge enhancement modules. The findings are particularly significant for financial sentiment analysis, where enhancing basic sentiment models can lead to more accurate stock market predictions.

Literature Survey

Bhardwaj et al. (2015) – Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty
This study explores the role of sentiment analysis in forecasting stock market trends in India, particularly focusing on Sensex and Nifty. It highlights the application of supervised learning methods like Naïve Bayes and SVM and lexicon-based approaches to classify financial sentiment, using real-time data scraped with Python**.**

Kumar et al. (2017) – Financial News Sentiment Analysis for Stock Price Prediction
This paper explores sentiment analysis techniques applied to financial news articles to predict stock price fluctuations in India. The researchers utilized machine learning models such as Random Forest and Decision Trees to assess the impact of news sentiment on market performance.

Ravi et al. (2020) – Hybrid Approach for Sentiment Analysis of Financial News and Social Media
This study proposes a hybrid approach combining machine learning models and lexicon-based techniques to analyze financial news and social media posts. The goal was to improve the prediction accuracy of stock market indices like Sensex and Nifty, demonstrating how a combined approach leads to better forecasting results.

Objectives

- **To Predict Stock Market Trends Using Sentiment Data:**
  Build a system that utilizes sentiment extracted from social media, news articles, and other sources to predict stock price movements for Sensex and Nifty indices, aiming for real-time predictions of market behavior based on public sentiment.

- **To Integrate Sentiment Classification with Stock Data:**
  Combine sentiment data with historical stock market data to develop a predictive model that can accurately forecast future price trends. This system would take into account both sentiment polarity and stock performance over time.

- **To Improve Sentiment Analysis Accuracy Using Machine Learning Models**:
  Use advanced machine learning algorithms such as Naïve Bayes and SVM to classify the sentiment of financial news and social media content, improving the accuracy of stock price predictions based on sentimen

Existing Methods Drawbacks

1 Traditional Machine Learning Models:

- **Support Vector Machines (SVM)**: SVM is used for classifying the sentiment in financial text and predicting stock price movements. However, SVMs are better suited for structured data and are less efficient when working with unstructured textual data.
- **Random Forests (RF)**: RF models can analyze structured financial data like historical stock prices and economic indicators but struggle with textual data. They can capture some non-linear relationships but are limited when extracting insights from unstructured data

2 Text Vectorization Techniques:

- **Term Frequency-Inverse Document Frequency (TF-IDF)**: This method converts text into numerical features that can be processed by machine learning models. However, TF-IDF doesn't capture the context of words well, and it can miss out on deeper sentiment or contextual meaning, especially in financial news where word context is crucial.

3 Deep Learning Models:

- **Convolutional Neural Networks (CNNs)**: Although CNNs are commonly used in image recognition tasks, they have also been applied to text analysis, including sentiment analysis. However, CNNs are not optimal for processing sequential text data as they don’t capture long-range dependencies well.
- **Recurrent Neural Networks (RNNs)**: RNNs, including variants like **Long Short-Term Memory (LSTM)** networks, are designed to handle sequential data, making them effective for text data like financial news or social media posts. LSTMs are particularly good at capturing long-term dependencies and trends in text data.

Existing Methods Drawbacks

Challenges of Existing Methods:

- **Contextual Ambiguity**: Financial texts are often filled with context-dependent phrases, jargon, and even sarcasm, making it difficult for traditional models to capture the true sentiment.

- **Misinformation**: The stock market is highly sensitive to rumours and misinformation. Distinguishing between valid financial news and fake news remains a challenge.

- **Interpretability:** Many deep learning models, especially those based on LSTMs and transformers, work as black boxes, making it difficult to understand how the model arrives at its decisions. This limits their adoption in industries where decision transparency is crucial.

- **Data Quality:** Financial sentiment data is often noisy, inconsistent, and unbalanced. (e.g., more positive news than negative news), which complicates model training and reduces performance).

Proposed Methods

The proposed model (Fig 1.1) is designed as follows:

(1) **Embedding Layer**: Transforms financial text into dense numerical vectors based on NLP methods.

(2) **LSTM Layer:** Identifies sequential dependencies between sentiment patterns to forecast stock market variation.

(3) **Attention Mechanism:** Identified the most influential sentiment indicators responsible for directing stock price variation.

(4) **Dense Layers**: Utilizes nonlinear transformations to improve the accuracy of predictions.

(5) **Dropout Layers:** Prevents overfitting by randomly disabling neurons during training.

(6) **Output Layer:** Forecasts stock price direction based on sentiment insights aggregated**.**

Proposed Methods

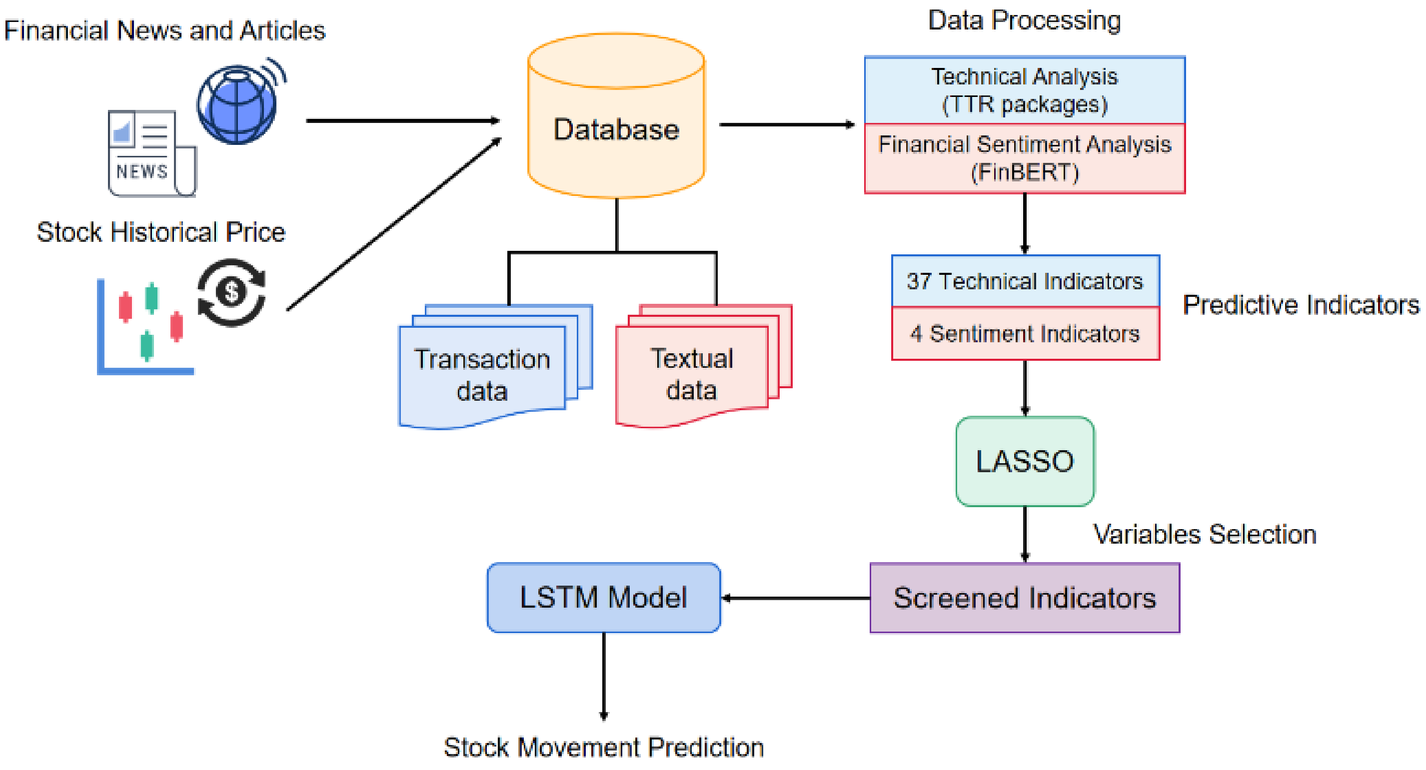


Figure 1.1 - Model Architecture

Software Componenets

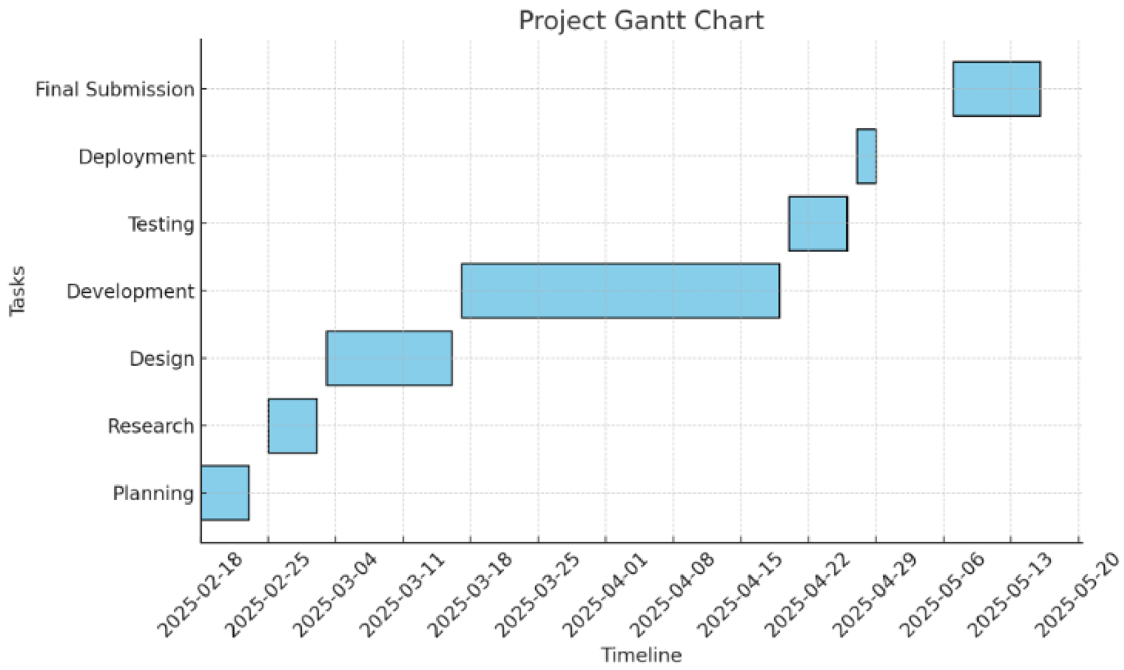
- **Python:** Core programming language used for implementing the project.
- **Google Colab:** Cloud platform for coding, execution, and GPU support.
- **FinBERT:** Pre-trained model for financial sentiment analysis.
- **PyTorch:** Deep learning library to run and fine-tune FinBERT.
- **Pandas:** Library for data loading, preprocessing, and analysis.
- **Matplotlib:** Used for plotting sentiment trends and visualizations.

Hardware Requirements

- **CPU:** Intel i5 or higher (for general processing tasks).
- **RAM:** Minimum 8 GB (recommended 16 GB for smooth performance).
- **GPU:** NVIDIA Tesla T4 / RTX 3060 or higher (for model inference and faster processing).
- **Storage:** Minimum 256 GB SSD (for fast data access and storage).
- **Internet:** Stable connection (for using Google Collab and accessing online models).
- **Device:** Laptop or desktop with compatible GPU or cloud-based system (like Google Collab with GPU/TPU).

Timeline of the Project (Gantt Chart)

- **Planning Phase**: February 18, 2025 – February 23, 2025
- **Research Phase**: February 25, 2025 – March 2, 2025
- **Design Phase**: March 3, 2025 – March 16, 2025
- **Development Phase**: March 17, 2025 – April 19, 2025
- **Testing Phase**: April 20, 2025 – April 26, 2025
- **Final Submission**: May 7, 2025 – May 16, 2025



Output

The graph shows predicted stock prices over the next 60 days using your AI model (likely LSTM + FinBERT).

The Y-axis represents the predicted stock price, and the X-axis shows the day number (from 0 to 60).

The blue line with dots indicates significant fluctuations in predicted prices, showing both rises and sharp drops reflecting real market volatility.

The model captures short-term trends well, suggesting it's sensitive to sentiment-based signals, though the high variation may imply market uncertainty or data sensitivity

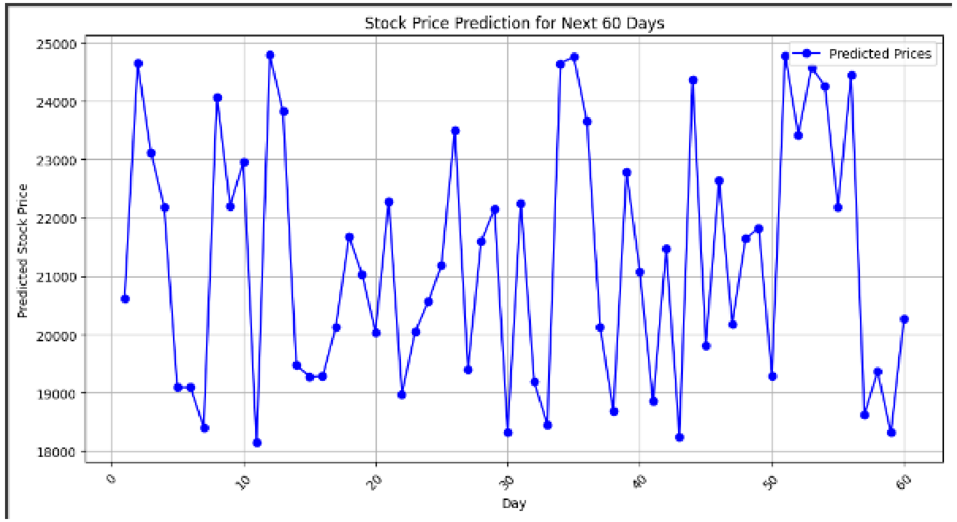
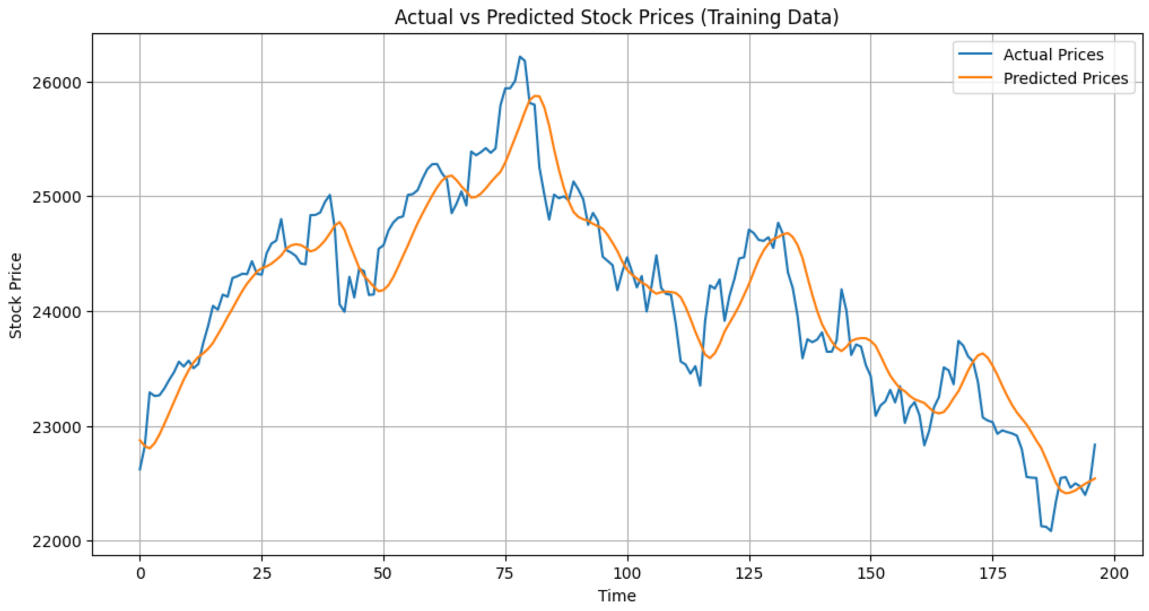


Fig 1.2 – Graphical Result

Output

1.Blue Line (Actual Prices): Represents the true stock prices over time in your training dataset.
2. Orange Line (Predicted Prices): Shows the prices predicted by your hybrid model (LSTM + FinBERT).
3. The lines follow a similar trend, indicating that your model has effectively captured the key patterns and fluctuations in stock price movement.
4. There are minor deviations in some areas, but the model generally tracks both upward and downward trends well, showing strong learning performance.
5. This visualization confirms that your model is well-fitted to training data, which aligns with the high R^2 score and accuracy shown earlier.



Output

The bar chart displays predicted stock prices over the 60-day period, with each bar representing one day's forecast. It shows that while prices fluctuate daily, most predictions fall within a similar range, indicating moderate variation. This visual helps quickly identify days with higher or lower expected prices. Overall, it provides an at-a-glance summary of the predicted trend without needing to read individual values.

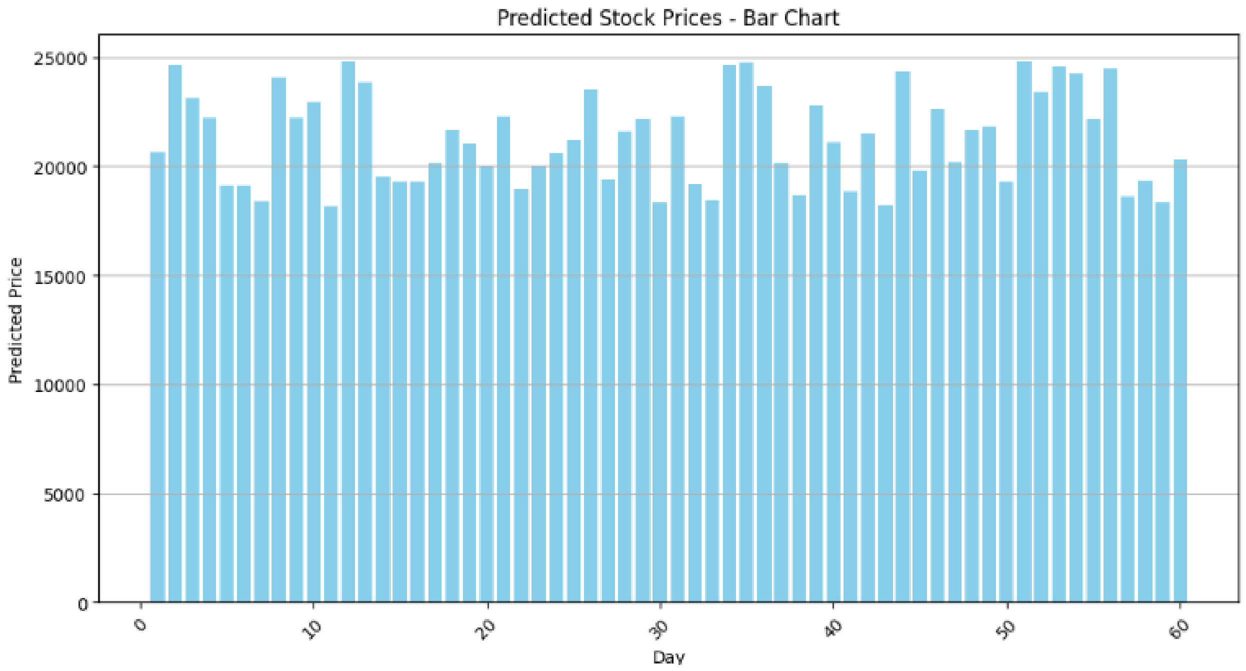


Fig 1.4 – Graphical Result

	Day	Predicted Price
0	1	20621.780832
1	2	24655.000145
2	3	23123.957593
3	4	22190.609389
4	5	19092.130483
5	6	19091.961642
6	7	18406.585285
7	8	24063.233020
8	9	22207.805082
9	10	22956.508045
10	11	18144.091460
11	12	24789.368965
12	13	23827.098486
13	14	19486.373775
14	15	19272.774770
15	16	19283.831569
16	17	20129.695701
17	18	21673.295021
18	19	21023.615130
19	20	20038.603981
20	21	22282.970263
21	22	18976.457025
22	23	20045.012540
23	24	20564.532903
24	25	21192.489890
25	26	23496.231730
26	27	19397.716475
27	28	21599.641069
28	29	22146.901982
29	30	18325.152889
30	31	22252.813963
31	32	19193.668866
32	33	18455.361151
33	34	24642.198761
34	35	24759.424232
35	36	23658.781437
36	37	20132.296384
37	38	18683.704798
38	39	22789.631186
39	40	21081.067456
40	41	18854.267644
41	42	21466.238371
42	43	18240.719648
43	44	24365.242815
44	45	19811.459871
45	46	22637.655990
46	47	20181.977533
47	48	21640.476148
48	49	21826.971955
49	50	19293.981189
50	51	24787.092394
51	52	23425.929764
52	53	24576.492591
53	54	24263.791453
54	55	22185.299852
55	56	24453.119645
56	57	18619.447514
57	58	19371.880037
58	59	18316.591022
59	60	20277.312315

Fig 1.5– Stock Predicted Price for next 60 Days

References

- Lin, F.; Cohen, W.W. Semi-Supervised Classification of Network Data Using Very Few Labels. In Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 9–11 August 2010. Available online:
- https://ieeexplore.ieee.org/abstract/document/5562771/
- (accessed on 25 July 2024).
- QuantConnect. (2021). Leveraging sentiment analysis for algorithmic trading. Quantitative Finance Journal, 9(5), 99-113. https://doi.org/10.2139/ssrn.3664972.
- Mehtab, Sidra & Sen, Jaydip & Dutta, Abhishek. (2020). Stock Price Prediction Using Machine Learning and LSTM- Based Deep Learning Models. 10.13140/RG.2.2.23846.34880.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007
- Can Yang, Junjie Zhai, Guihua Tao, "Deep Learning for Price Movement Prediction Using Convolutional Neural Network and Long Short-Term Memory", Mathematical Problems in Engineering, vol. 2020, Article ID 2746845, 13 pages, 2020. https://doi.org/10.1155/2020/2746845.
- Dutta, A.; Kumar, S.; Basu, M. A Gated Recurrent Unit Approach to Bitcoin Price Prediction. J. Risk Financial Manag. 2020, 13, 23. https://doi.org/10.3390/jrfm13020023.


Thank
You!