Exploring Advanced Techniques for Stock Market Prediction with Big Data-A Focus on LSTM Methodology

M.MOHANA SUNDHARAM
Assistant Professor
Dept Of Computer Science & Engineering
RVS College Of Engineering & Technology,
Coimbatore
mohantvsm@gmail.com

M.RANJITHA
712820104030
Dept Of Computer Science & Engineering
RVS College Of Engineering & Technology,
Coimbatore
ranjitha270803@gmail.com

P.BHARATHWAJ
712820104007
Dept Of Computer Science & Engineering
RVS College Of Engineering & Technology,
Coimbatore
bharathsparrow1938@gmail.com

C.P.SHIBHATH
712820104038
Dept Of Computer Science & Engineering
RVS College Of Engineering & Technology,
Coimbatore
shibhathcp@gmail.com

S.YOGESHWARAN
712820104047
Dept Of Computer Science & Engineering
RVS College Of Engineering & Technology,
Coimbatore
yogeshjuggler14@gmail.com

Abstract- Big data enables the identification of patterns, correlations, and anomalies in stock market data that may not be apparent through traditional analysis methods. This can help predict future price movements, detect market trends early, and identify potential investment opportunities or risks. Our primary objective is centered around enhancing the efficiency of stock market data prediction, which plays a critical role in formulating effective trading strategies for investors. The successful forecasting of future stock prices is integral to maximizing investor profitability. Big data encompasses vast volumes of both structured and unstructured data, offering immense potential in stock prediction and analysis. A multitude of tools are utilized for stock market prediction, given the dynamic and complex nature of financial markets. Among these techniques, Long Short-Term Memory (LSTM) stands out as a promisingmethod for predictive analysis using time series data. Our proposal involves utilizing LSTM to predict future stock prices based on closing price data, with the aim of optimizing investment returns for investors.

Keywords- Big data, Long Short-Term Memory, Recurrent Neural Network.

I.INTRODUCTION:

Stock market prediction has remained a crucial and challenging field of study. The emergence of social media and technological advancements, particularly big data, has sparked widespread interest and enthusiasm among researchers in this domain.

Meanwhile, traders, major financial institutions, and corporations have leveraged big data to generate trade analytics, especially in high-frequency trading scenarios. Numerous tools are employed for stock market prediction due to the dynamic and intricate nature of the market. Accurately forecasting future prices holds the potential to enhance investors' profit margins through strategic stock investments. Many existing systems face challenges in adapting to market fluctuations, rendering them less effective during volatile periods or unexpected events. Traditional and machine learning models may struggle with feature selection and adapting to new relevant indicators. Deep learning, particularly LSTM, offers a promising avenue for gaining insights into the unpredictable dynamics of the stock market.

II.PREMILINARY KNOWLEDGE:

In this paper, to develop an advanced forecasting system for Apple stock prices using Long Short-Term Memory (LSTM) networks. The project seeks to harness the capabilities of deep learning to enhance the accuracy and adaptability of stock price predictions, providing valuable insights for investors and stakeholders.

A. LONG SHORT-TERM MEMORY:

Developing an efficient sequence learning model involves crafting a multi-layer LSTM architecture with carefully chosen hyperparameters. This architecture encompasses input layers, LSTM layers, and output layers, each configured to optimize performance. Determining the appropriate number of LSTM layers, neurons within each layer, and activation functions is crucial in achieving desired outcomes. Additionally, incorporating dropout layers helps mitigate overfitting, ensuring the model's generalizability and reliability.

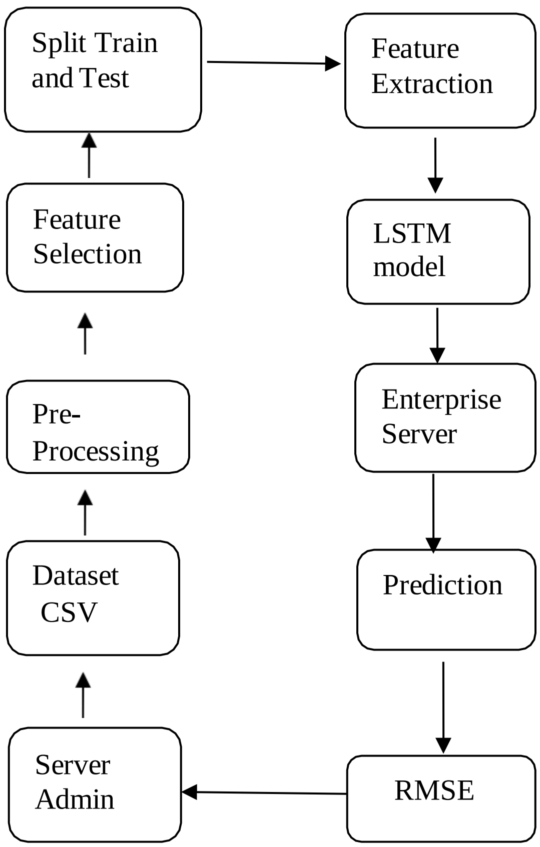


Figure:1. LSTM Architecture

B. CANDLESTICK:

Candlestick charts serve as a widely utilized tool in technical analysis for visualizing price movements within financial markets. Each candlestick comprises a body and wicks, with the body indicating the price range between opening and closing prices, while the wicks represent the highest and lowest prices during the specified time period. Various candlestick patterns can offer insights into potential trend reversals or continuations. Moreover, candlestick patterns are commonly employed to identify critical support and resistance levels. Observing a reversal pattern at such levels may suggest a forthcoming price reversal. Traders frequently integrate candlestick patterns with other technical indicators for further confirmation.

C.BULLISH CANDLESTICK:

A bullish candlestick signifies upward momentum in the market, characterized by a closing price higher than the opening price, indicating increased buying pressure. This candlestick, often depicted in green or with a filled body, reflects optimism among traders. Bullish candlestick patterns, formed by one or more candlesticks on a price chart, are widely utilized in technical analysis to anticipate potential upward price movements. Each candlestick comprises open, close, high, and low prices for a specified time period, with the body representing the price range between the open and close, while the wicks indicate the extreme price levels reached.

While bullish candlestick patterns offer valuable insights for stock market prediction, it's essential to complement them with other analytical techniques and indicators. Traders frequently integrate these patterns into a comprehensive technical analysis approach to enhance decision-making in a effective way.

D.BEARISH CANDLESTICK:

Bearish candlestick patterns are instrumental in technical analysis, aiding traders and investors in anticipating potential downward price movements in the stock market. These patterns, formed by one or more candlesticks on a price chart, serve as indicators of potential bearish (downward) momentum. An example is the bearish engulfing pattern, characterized by a small bullish candle followed by a larger bearish candle that completely engulfs the previous one, signaling a shift from bullish to bearish sentiment. Given the dynamic nature of markets, continuous monitoring is essential to adapt to changing conditions effectively.

While bearish candlestick patterns offer valuable insights, they should be used alongside other indicators and methods for making informed trading decisions. It's important to consider market conditions and the evolving nature of financial markets when interpreting these patterns, which suggest potential reversals of uptrends or the initiation of downtrends in stock prices. Traders rely on these patterns to anticipate downward price movements and adjust their strategies accordingly.

III.A. DATA PREPARATION AND EXPLORATION:

The historical stock price dataset was sourced from the Kaggle webpage, comprising stock prices spanning from January 1, 2003, to February 5, 2024, and formatted as comma-separated values (.csv). The dataset encompasses various types of price information for individual stocks. Upon acquiring the dataset, we proceed to analyze and delineate the distinct characteristics and behaviors exhibited by the stock prices.

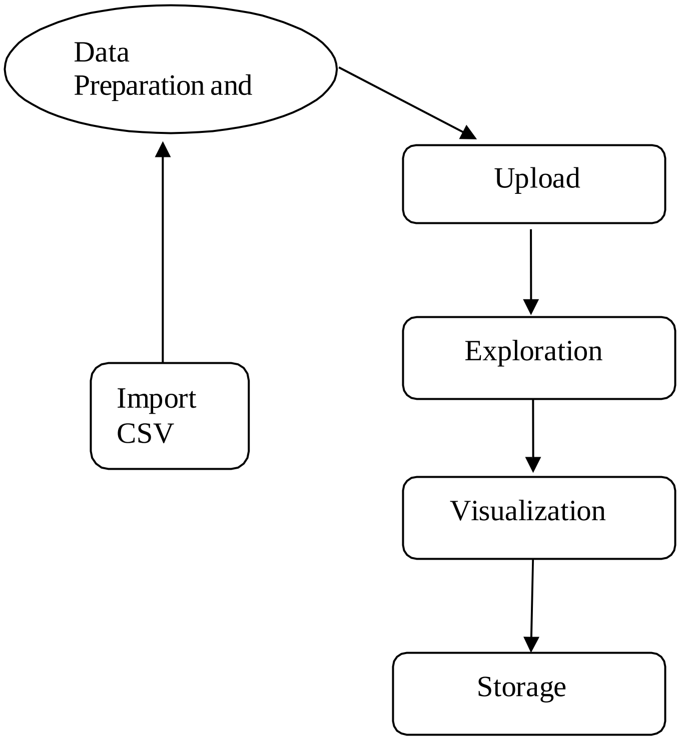


Figure:2. Data Preparation and Exploration

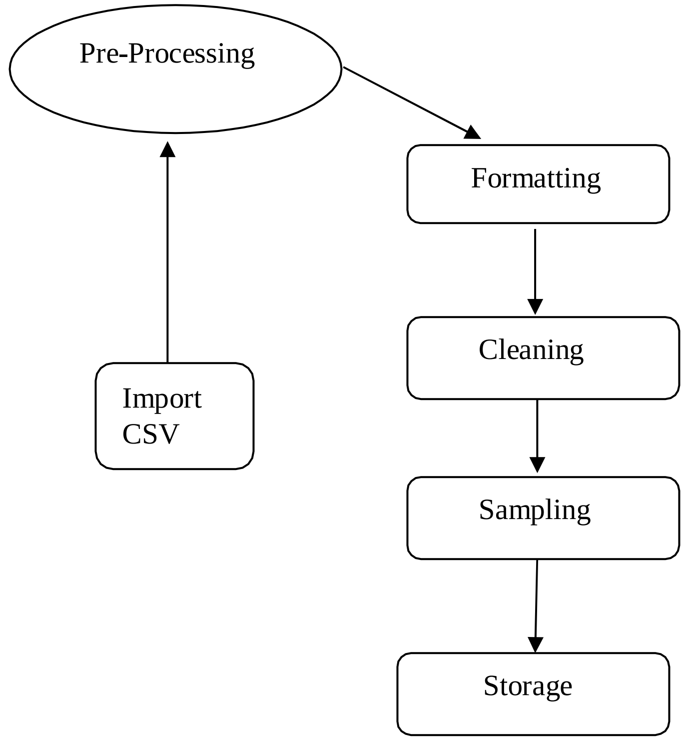


Figure:3. Data Pre-Processing

B.DATA PRE-PROCESSING:

The process of data Pre-processing for stock prediction through big data encompasses various steps:

1. Data Formatting: The dataset utilized, sourced from Kaggle, may contain attributes with unclear names or irrelevant features that do not contribute to the efficacy of the proposed work. It's crucial to address these issues to ensure the dataset's suitability for analysis.
2. Data Cleaning: This phase focuses on rectifying or removing missing entries within the dataset. Rows with incomplete columns are typically eliminated, and redundant entries are also discarded to streamline the dataset. This step is vital for maintaining data integrity and accuracy in subsequent analyses.
3. Data Sampling: Sampling techniques are applied to the dataset to optimize the algorithm's performance. Working with a sample dataset can enhance computational efficiency, although it's important to acknowledge that processing time may increase with larger datasets. Sampling helps streamline the analysis process while still providing representative insights into the dataset's characteristics.

C.FEATURE EXTRACTION:

Feature extraction is one of the most important parts in stock prediction process. Better market features always contribute to better predictions. The Candlestick model is used for extracting price movements in financial markets. It involves the interpretation of candlestick charts, which display the opening, closing, high, and low prices for a specific time period.

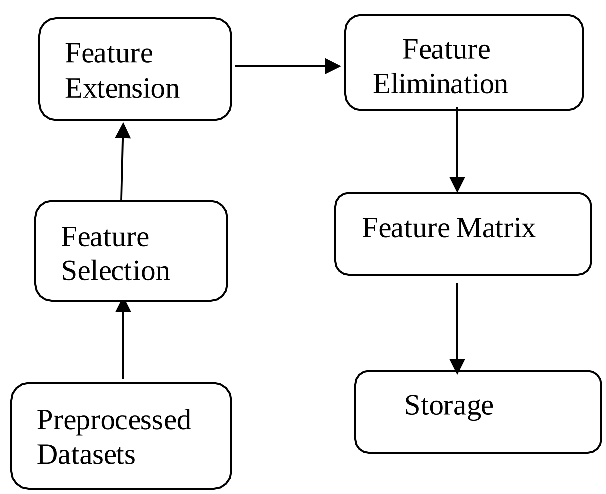


Figure.4. Feature Extraction

D.EVALUATION MEASURES:

1. Mean Absolute Percentage Error:

The Mean Absolute Percentage Error (MAPE) is frequently utilized to evaluate the effectiveness of prediction techniques. It serves as a measure of prediction accuracy within the domain of forecasting methods in machine learning, often expressed as a percentage.

2. Mean Absolute Error:

The Mean Absolute Error (MAE) represents the disparity between two values, calculated as the average deviation between predicted and actual values. It serves as a common metric for assessing prediction accuracy in regression analysis within the realm of machine learning. The mathematical expression for MAE is depicted in Equation.

3. Relative Root Mean Square Error

The Root Mean Square Error (RMSE) quantifies the dispersion of prediction errors in regression tasks. These errors, also known as residuals, signify the discrepancy between actual values and those predicted by a model, illustrating their distribution around the model. RMSE serves as a metric to gauge the concentration of data points around the optimal fitting model.

4. Mean Squared Error

The Mean Squared Error (MSE) evaluates the effectiveness of predictors, with lower values indicating higher quality. It is always nonnegative, with values closer to zero considered superior. MSE represents the second moment of the error, encompassing both the variance of the prediction model, which indicates the spread of predictions across data samples, and its bias, which reflects the proximity of the average predicted value to the observed value.

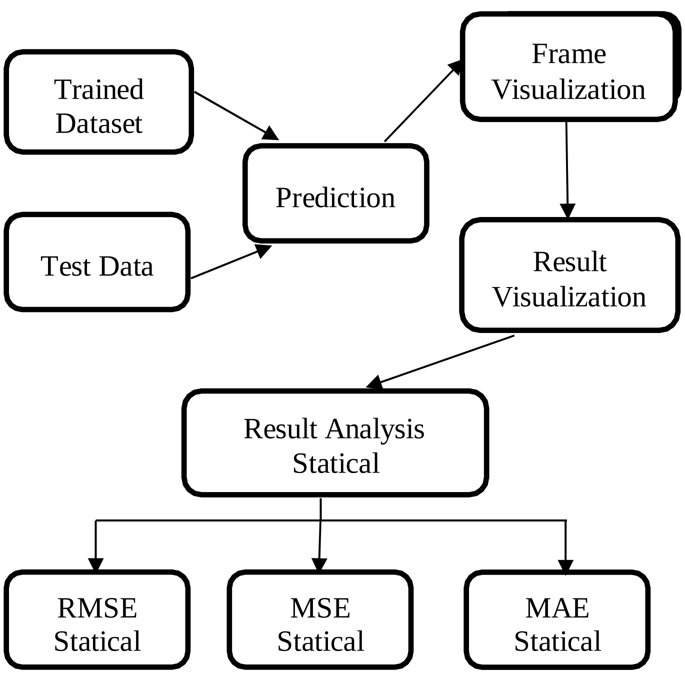


Figure:5. Evaluation Measures

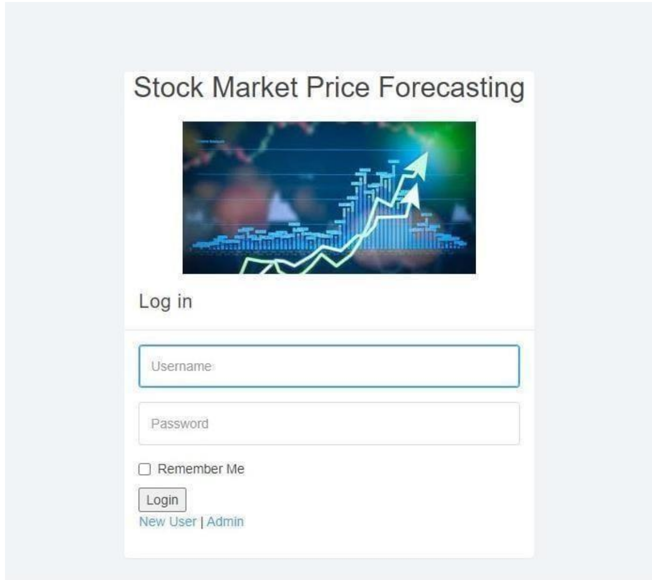
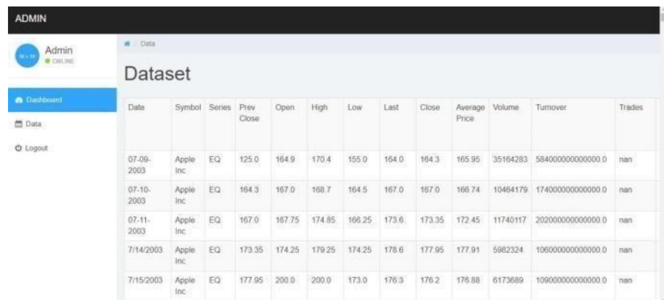
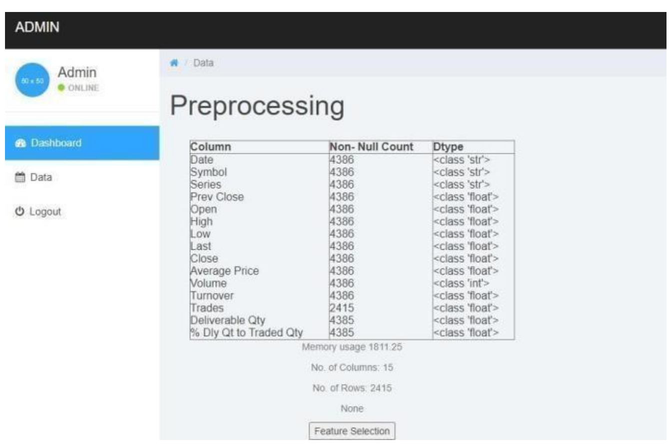


Figure:6



Date	Symbol	Series	Prev Close	Open	High	Low	Last	Close	Average Price	Volume	Turnover	Trades
07-09-2003	Apple Inc	EQ	125.0	164.9	170.4	155.0	164.0	164.3	165.95	35164283	584000000000000.0	nan
07-10-2003	Apple Inc	EQ	164.3	167.0	168.7	164.5	167.0	167.0	166.74	10464179	174000000000000.0	nan
07-11-2003	Apple Inc	EQ	167.0	167.75	174.85	166.25	173.6	173.35	172.45	11740117	202000000000000.0	nan
7/14/2003	Apple Inc	EQ	173.35	174.25	179.25	174.25	178.6	177.95	177.91	5982324	106000000000000.0	nan
7/15/2003	Apple Inc	EQ	177.95	200.0	200.0	173.0	176.3	176.2	176.88	6173689	109000000000000.0	nan

Figure:7



Column	Non- Null Count	Dtype
Date	4386	<class 'str'>
Symbol	4386	<class 'str'>
Series	4386	<class 'str'>
Prev Close	4386	<class 'float'>
Open	4386	<class 'float'>
High	4386	<class 'float'>
Low	4386	<class 'float'>
Last	4386	<class 'float'>
Close	4386	<class 'float'>
Average Price	4386	<class 'float'>
Volume	4386	<class 'int'>
Turnover	4386	<class 'float'>
Trades	2415	<class 'float'>
Deliverable Qty	4385	<class 'float'>
% Dly Qt to Traded Qty	4385	<class 'float'>

Memory usage 1811.25

No. of Columns: 15

No. of Rows: 2415

None

Feature Selection

Figure:8

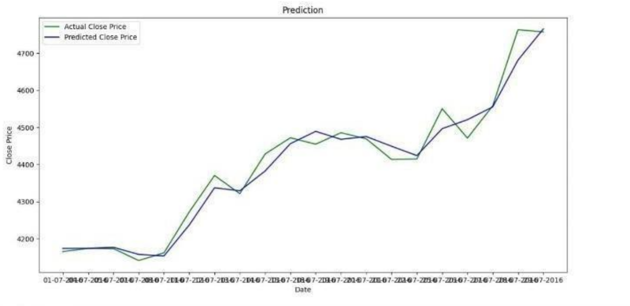


Figure:9

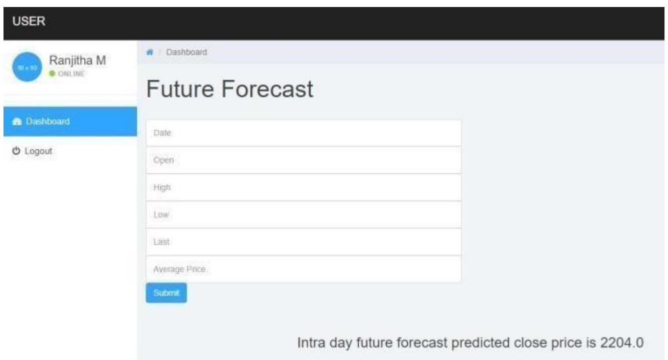


Figure:10

IV. CONCLUSION:

This study utilizes big data to conduct effective stock market analysis and prediction. Given the inherent uncertainty of the stock market, accurately forecasting stock values is crucial to avoid significant financial losses. The project aims to utilize Long Short-Term Memory (LSTM) neural networks for precise and dynamic stock price forecasting. By integrating advanced deep learning techniques, including candlestick patterns and external factors, along with comprehensive feature engineering, the system aims to enhance adaptability to changing market conditions. Furthermore, the project emphasizes transparency and interpretability in its framework. Through rigorous validation and continuous improvement efforts, the project strives to improve predictive accuracy within the ever-evolving landscape of financial markets.

.

V.REFERENCES

1. XIAOSONG ZHAO ,YONG LIU, AND QIANGFU ZHAO, “Cost Harmonization Light GBM-Based Stock Market Prediction” at Graduate School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima 965-0006, Japan Corresponding author: Xiaosong Zhao.

2. Banic L. et al.," Using Big data and sentiment analysis in product evaluation", Presented at MIPRO 2013, Croatia, May 2013.

3. Bing L. et al, " Public sentiment analysis in twitter data for prediction of a company's stock price movement.", Presented at IEEE 11th International Conference on e- Business Engineering,2014.

4. Burton G. Malkiel, 'The Efficient Market Hypothesis and Its Critics", The Journal of Economic Perspectives, Vol. 17, No. I (Winter, 2003), pp. 59-82.

5. Chodhury. R, K. Garg, "A hybrid machine learning system for stock market forecasting", Proceeding of World Academy of Science, Engineering and Technology, vol.29 (2008) ISSN 1307-6884.

6. Laszlo Gerencser, Balazs Torma and Zsanett Orlovits, “Fundamental Modelling of Financial Markets”, “ERCIM News,vol 78, pp 52, Jul 2009.