**Stock Price Prediction using LSTM and Sentiment analysis.**

**ABSTRACT**

Stock prices in general say about the price of the stock in the market. Prediction of the stock price has been a big task for the people who do stock marketing as their primary job for survival, not only for the people who do marketing as their primary task but also for the people who do it as side work [4]. Many people doing stock market prediction as their primary job can predict the stock price based on some factors namely, Fundamental analysis (impact and correlation of stock prices of other companies, past performances, records, profits, and debts of the company), Technical analysis (Continuation Pattern, Reversal Pattern, Leading Indicator and Lagging indicator), Sentimental analysis (Client emotions such as positive, negative, neutral) [1]. The main question now arises is whether a machine can predict the stock price without human interaction. As stock price prediction is difficult for a machine, we can overcome this problem using some machine learning algorithms like RNN, and LSTM (Long Short Term Model) and using sentimental analysis same as the in-person analysis [3]. And here we are using the main thing that every stock price prediction should be considering which is heuristic analysis.

**Key Words:** Stock parameters, Stock Analysis, Stock price prediction, Recurrent neural network (LSTM), Analysis (Fundamental, Technical, Sentimental).

1. **INTRODUCTION:**

Stock price prediction is finding out the future movement of the stocks. The accurate prediction of stock movement results in high profits for an investor, at the same time wrong prediction gives inevitable losses for the investor. Stock price prediction is complex yet a challenging task due to the volatile and non-linear nature of the stock market. So, before investing in the stock market every investor needs to predict the stock price or say the stock market, to avoid incurring losses in stocks and increase the chances of making profits from the invested stocks [1].

Sentimental analysis is the mechanism in which the given input will be characterized into three forms namely, positive, negative, and moderate. Each of the three forms of characterization has a different value to represent it, which goes in the format of 1(positive) for the positive kind of connotation sentence, -1(negative) for the negative kind of the connotation and the zero for the moderate connotation of the sentence [2]. Sentimental analysis is often termed opinion mining where every unit of the sentence is categorized as one of the three forms of the type. Another way to utilize this idea is to use the know-how we benefit from analyzing human reviews of an agency on social media and discover a manner to attach it to the agency's inventory marketplace tendencies to make higher predictions of the agency's inventory expenses [1].

LSTM is the Long short-term memory that is a model used in machine learning and comes under the recurrent neural network (RNN). Here it is used in the processing of the output, using the recurrence or feeding the output again and again into the LSTM several times to get an accurate and precise prediction of the stock price. The output of every stage is inserted as input again for processing the output.

1. **SYSTEM STUDY:**

**2.1 RELATED WORK**

The initial focus of our literature survey was to explore generic online learning algorithms and see if they could be adapted to our use case i.e., working on real-time stock price data.

Stock prices change because of supply and demand. Stock prices never vary in isolation: the movement of one tends to have an avalanche effect on several other stocks as well. This aspect of stock price movement can be used as an important tool to predict the prices of many stocks at once [2].

Prediction of the future movement of stock prices has always been a challenging task for researchers. Investors are highly interested in the research area of stock price prediction. Time series forecasting analyses past data and projects estimates of future data values. This project attempts to model a predictor by a recurrence relation derived from past values [5]. A comparative study of LSTM and Deep Neural Network for Stock Market Forecasting has been conducted by Efficient Market Hypothesis (EMH) and states that LSTM is better compared with Machine Learning Approaches.

Impact of Financial Ratios, Technical, and Sentiments analyses on Stock Price Prediction -- tells increasing in efficiency and accuracy of the stock price prediction model.

**2.2 EXISTING SYSTEM:**

Traditional approaches to stock market analysis and stock price prediction include fundamental analysis, which looks at a stock's past performance and the general credibility of the company itself, and statistical analysis, which is solely concerned with number crunching and identifying patterns in stock price variation. The latter is commonly achieved with the help of Genetic Algorithms (GA) or Artificial Neural Networks (ANNs).

Alter to the above analysis an alternative approach to stock market analysis is introduced i.e., to reduce the dimensionality of the input data and apply feature selection algorithms to shortlist a core set of features (such as GDP, oil price, inflation rate, etc.) that have the greatest impact on stock prices or currency exchange rates across markets.

The latter other methods introduced like Historical Data Analysis, Multi-Source Multiple Instance Learning, Support Vector Machines (SVM), and Independent Component Analysis (ICA) are used for Stock Market Prediction.

And Stock Price Prediction using Linear Regression based on Sentiment Analysis is introduced in the domain of deep learning and neural networks.

**2.3 PROPOSED SYSTEM:**

In this project, we are performing both fundamental analysis and technical analysis on time series data and trying to predict the future price of the stock. And increasing the stock price predicting efficiency with help of sentiment analysis [1]. That is our project aim. The stock parameters, analysis, and all data sets are going to be very large and time series and there is a need to extract the feature from finance market raw data/historical data, which is required for data analysis, then divide it as testing and training data, training the algorithm to predict the price and the final step it to visualize the data. So, we are using Long Short-Term Memory (LSTM), a Recurrent Neural Network (RNN) type [5].

LSTM makes small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things [5]. LSTM networks are ideal for exploring how variation in one stock's price can affect the prices of several other stocks over a long time.

**3. METHODOLOGY:**

**3.1 LSTM ALGORITHM**

Dataset: In this project, two main datasets have been used –

* Yahoo finance stock data from 01-01-2004 to 07-05-2021. The data includes open, closed, high, and low, values of a given day.
* Publicly available Twitter data. This incorporates the timestamp and tweet text for every tweet of a particular period. Since predictions are being made on daily basis, tweets are split by day using their timestamps.

Prediction: The training and prediction are done using Long Short-term memory (LSTM) architecture. LSTM architecture is a part of recurrent neural networks and is generally used in the field of deep learning. LSTM has feedback connections, which makes it very useful to process entire sequences of data. Before feeding data to LSTM [4], data should be processed and normalized. The other dataset that is used for sentiment analysis is obtained from Twitter; this data should also be processed before using it to do sentiment analysis.

Data Processing:

First, we use pandas to obtain data from yahoo finance. As our target value is the Close value, a target data frame is created with only a close column. The data is then normalized and converted so that all the values lie between 0 and 1. Data is then divided into two parts i.e., training data (70 percent) and testing data (30 percent) [2].

Second, the data from Twitter is obtained using Tweepy. Tweepy is a library that is used for accessing the Twitter API. After obtaining the tweets from the API they are cleaned so that all links and special characters are removed. After cleaning they are then divided according to their polarity that is positive polarity means the tweet is positive and negative polarity means the tweet is negative.

Stock and Date Selection: For this project, Apple, Google, and Microsoft's stock values have been used. Their stock values are publicly available. The values used are from 01- 01-2004 to 07-05-2021.

LSTM networks are ideal for exploring how variation in one stock's price can affect the prices of several other stocks over a long period.

LSTM, as depicted in the below Figure, has layers and dropout regularization.

Diagram

Description automatically generated

3.1 LSTM LAYERS

[4]LSTMs are a special subset of RNNs that can capture context-specific temporal dependencies for long periods. Each LSTM neuron is a memory cell that can store other information i.e., it maintains its cell state. While neurons in normal RNNs merely take in their previous hidden state and the current input to output a new hidden state, an LSTM neuron also takes in its old cell state and outputs its new cell state. An LSTM memory cell, as depicted in the following Figure, has the following three components, or gates:

Diagram

Description automatically generated

3.2 LSTM

**Forget gate:** the forget gate decides when specific portions of the cell state are to be replaced with more recent information. It outputs values close to 1 for parts of the cell state that should be retained, and zero for values that should be neglected.

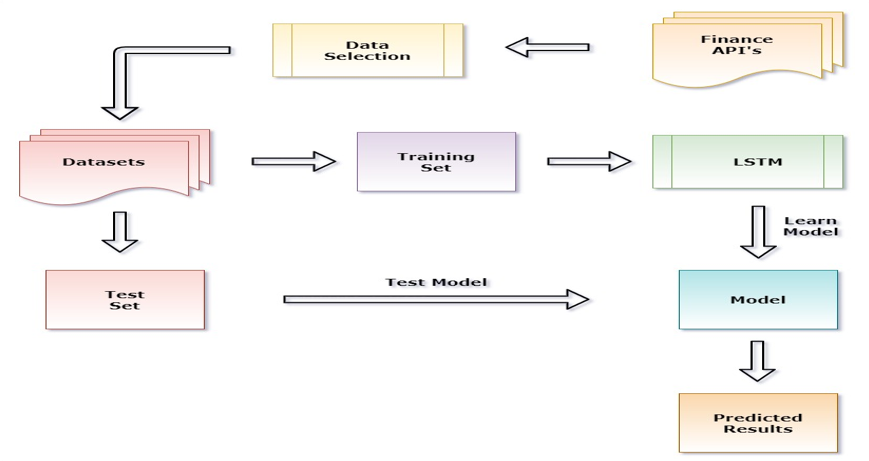
**Input gate:** based on the input (i.e., previous output o(t-1), input x(t), and previous cell state c(t-1)), this section of the network learns the conditions under which any information should be stored (or updated) in the cell state.

**Output gate:** depending on the input and cell state, this portion decides what information is propagated forward (i.e., output o(t) and cell state c(t)) to the next node in the network.

Accuracy plays an important role in stock market prediction. Although many algorithms are available for this purpose, selecting the most accurate one continues to be the fundamental task in getting the best results.

To achieve this, we proposed a model that can analyze the past study & present data and, gives an idea in order of creating methodologies to select stocks. Our proposed model overcomes the lack of implementation of financial analysis methods.

The final look of the proposed system looks like, the following Figure.

****

2.3 PROPOSED SYSTEM

**ALGORITHM**

Step 1: Start.

Step 2: Collect historic data of companies' stocks from Finance API.

Step 3: Perform Data Preprocessing after getting the historic data from the market for a particular share.

Step 4: If required, do Step 3 for all companies' stocks.

Step 5: Import the dataset to the data structure and perform data exploration to read the open price, close price, volume, and date.

Step 6: Apply Fundamental and Technical analysis on pre-processed data. (we consider earning per share ratio, price to earning to growth ratio for fundamental and moving average [exponential average with open, close, high, low] for technical).

Step 7: Describe the data set in rows and columns (Open, High, Low, Close, Adj Close, Volume as columns and count, mean, std, min, 25%, 50%, 75%, and max as rows).

Step 8: Do a feature scaling on the pre-processed data so that the data values will vary from true and false.

Step 9: Create a data set with 90% of pre-processed data as train data and 10% as test data.

Step 10: Build the RNN (Recurrent neural network) for the Step 6 data set and Initialize the RNN by using a sequential repressor.

Step 11: Add the first LSTM layer and some Dropout regularization for removing unwanted values on Step 9 output as input data. Take input of x-train and y-train from Step 7 output. And do training.

Step 12: Add multiple LSTMs to have high efficiency in data. At last, add the output layer.

Step 13: Train the LSTM model with sentimental analysis. To get the best accuracy.

Step 14: Compiling the RNN (LSTM) by adding adman optimization and the loss as (RMSE) root mean square error.

Step 15: Calculate the Close Price of companies' stocks and update it to Adj Close.

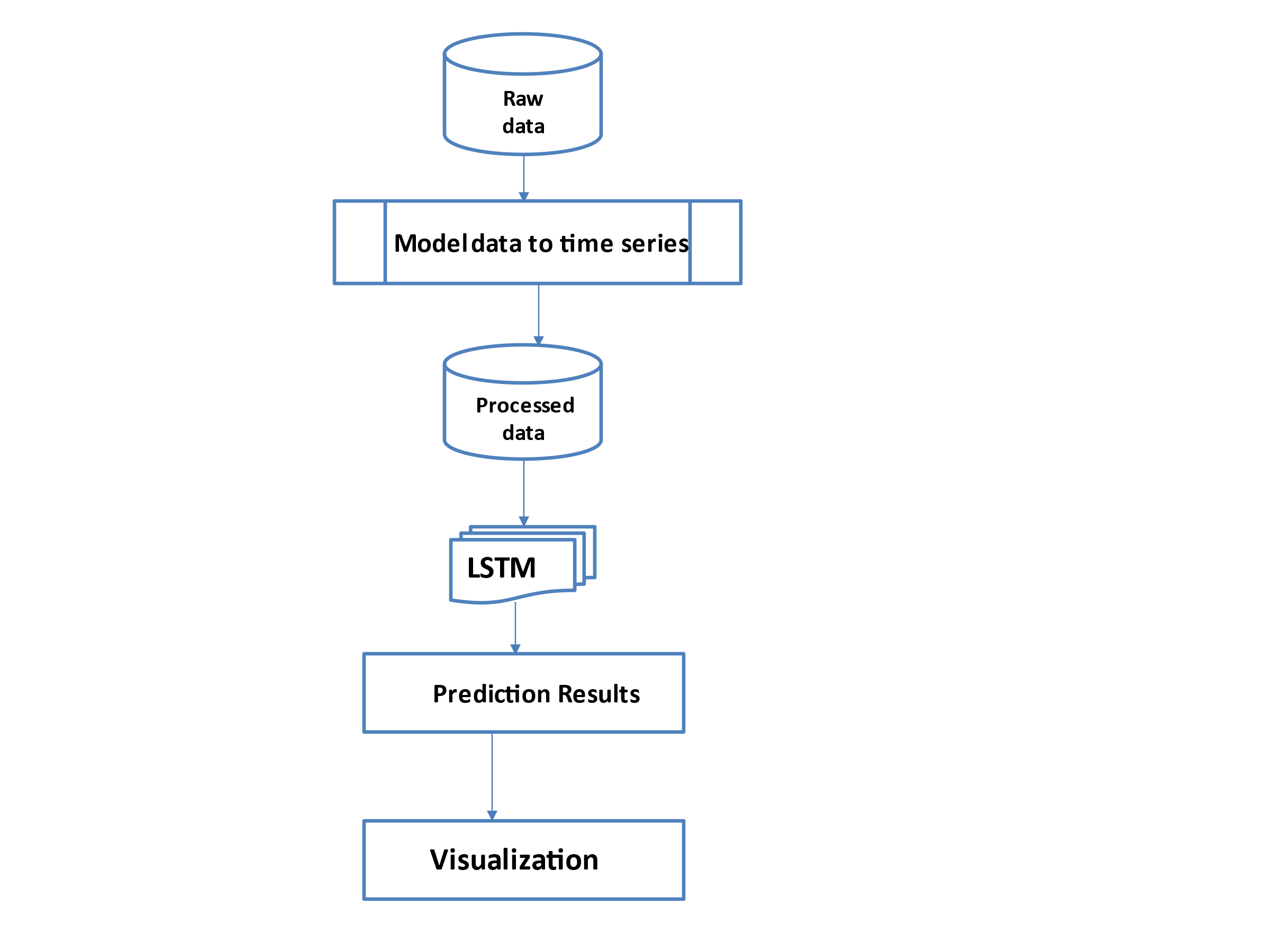
Step 16: Making the predictions and visualizing the results using plotting techniques.

Step 17: Stop.

**4. SYSTEM ARCHITECTURE:**

**Design diagram (Data Flow Diagram)**

Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and report generation.

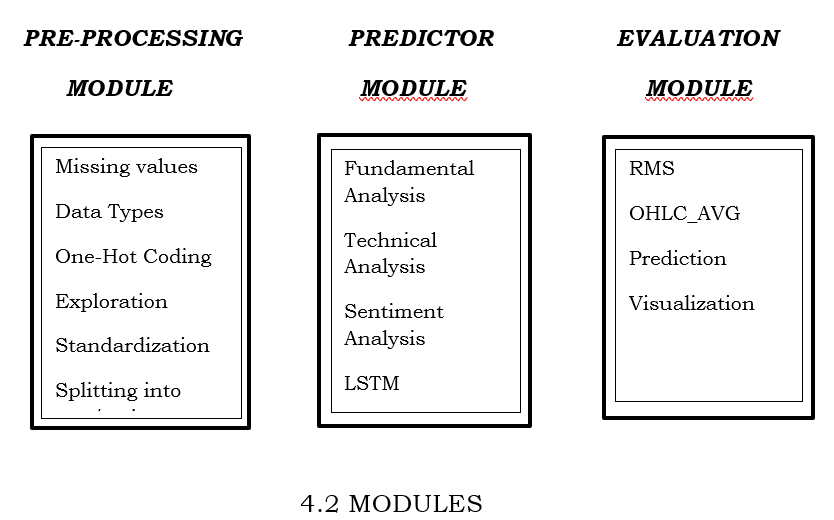


4.1 DATA FLOW DIAGRAM

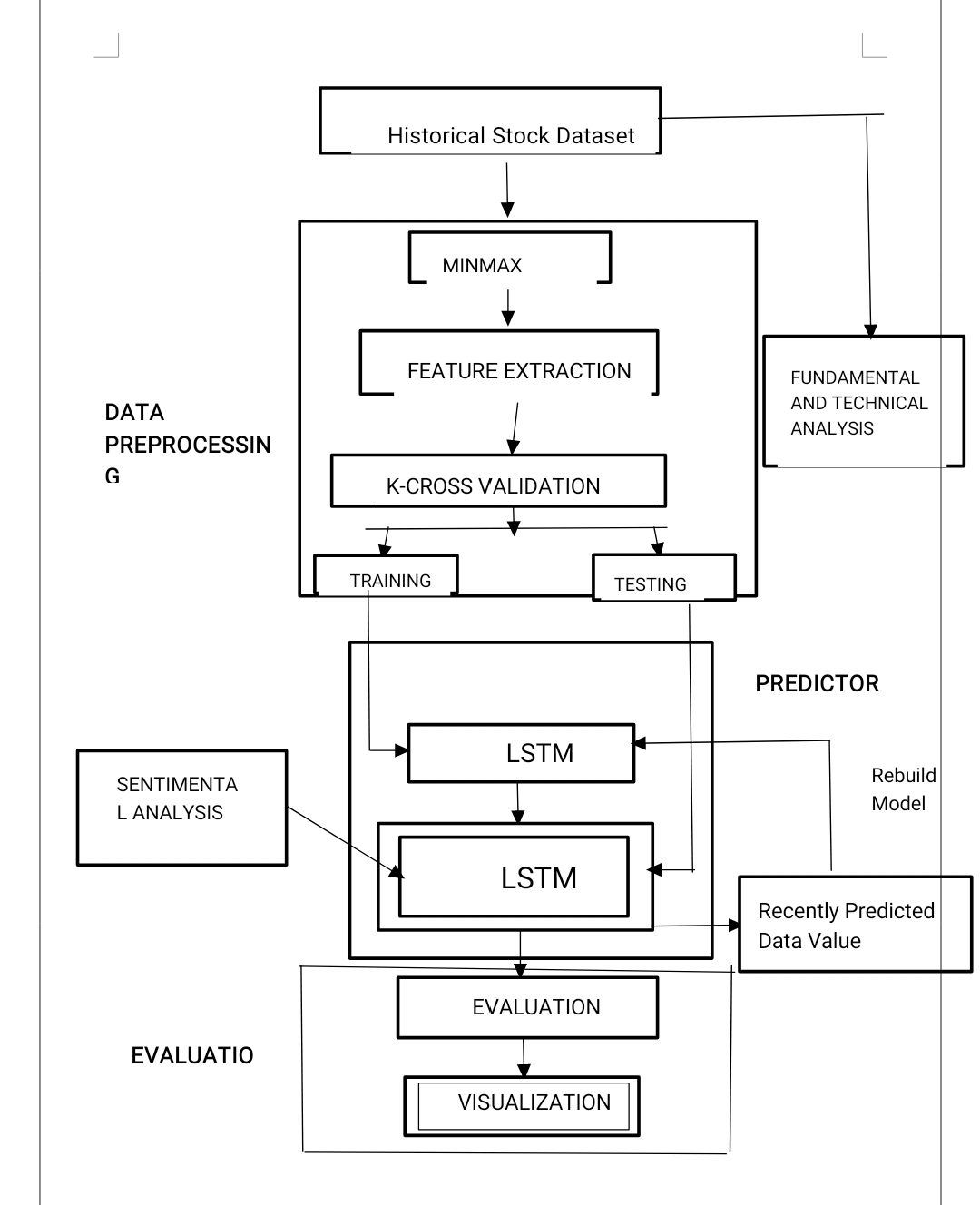
DFD graphically reprepresents functions, or processes, which capture, manipulate, store, and distribute data between a system and its environment and between components of a system.

1. **Graphical Representation**

Modules: There are 3 Modules present in the design.



**ORGANISATION:**

****

**Research And Experimental Work Done:**

This project focuses on the volatile nature of the stock market making it an area that needs an abundance of analysis with the old data predicated.

The prices of stocks are governed by the principles of demand and supply. Stock markets are closely linked with the world of economics — the rise and fall of share prices can be traced back to some Key Performance Indicators (KPIs). The five most used KPIs are the opening stock price ('Open'), end-of-day price ('Close'), intraday low price ('Low'), intra-day peak price ('High'), and total volume of stocks traded during the day ('Volume'). In the project, we will develop a stock price predictor model that uses previous stock prices and data will be treated as training sets for the predictor to predict the stock prices of a particular share. This predictor model develops a procedure.

The proposed model uses time series analysis to predict a share price for a required period. This model also provides features of shares which are Opening price, day High, day Low, previous day Close price, predicated Close price, Date of trading, and Total Trade Quantity (Volume).

**5. RESULTS:**

The output after the data set is trained is illustrated in the following Figure.

Figure,

Graphical user interface, text, application, Excel

Description automatically generated

5.1 DATA

The stock parameters as output after LSTM, FA, TA, and SA procedure as – Open, Close, Predicted Close, Volume of Bank of American Corporation (BAC) in the below Figure.

Graphical user interface, application, table

Description automatically generated

5.2 DATA

The final project output as Visualization is depicted in the below Figures.

For Bank of American Corporation (BAC): -

Graphical user interface, chart

Description automatically generated

5.3 GRAPH

For Netflix (NQ=F): -.

Graphical user interface, chart, line chart, scatter chart

Description automatically generated

Figure: 5.4 GRAPH

For National Stock Exchange of India of Tata Global Beverages Limited (TATA GLOBAL): -

Chart

Description automatically generated

5.5 GRAPH

The efficiency improvement can be seen in the below Figure. As rmse decreases, efficiency increases.Table

Description automatically generated

RMS VALUES

**Result Analysis: -**

Based on the methodology – LSTM with FA & TA, we noticed the efficiency and graph are showing major differences between the graph plotted by us and the graph in "OFFICIAL FINANCE API". As we can see below figure.

For National Stock Exchange of India of Tata Global Beverages Limited (TATA GLOBAL): -

Chart

Description automatically generated

5.6 RESULT ANALYSIS

For National Stock Exchange of India of Tata Global Beverages Limited (TATA GLOBAL) in "QUANDL": -A screenshot of a computer

Description automatically generated

5.7 RESULT ANALYSIS

And, with the methodology that we proposed in this project (LSTM with Sentiment Analysis) – we are getting good efficiency and graph plotted by us and in "OFFICIAL FINANCE API". As we can see in the below figure.

Graphical user interface, application

Description automatically generated

5.8 RESULT ANALYSIS

Since this application aims to overcome the draining efficiency faced in existing systems and we have shown this project by proposing a new system with methodology - LSTM with Sentiment Analysis).

**6. CONCLUSION:**

Stock markets are hard to monitor and require plenty of context when interpreting the movement and predicting prices. In this project, we proposed a Predicator. To prove an efficient and accurate approach among existing system(s) to predict the market price of a stock [3].

With the help of a predictor, we measured the accuracy of the following approach(es) LSTM algorithm, LSTM with fundamental & technical analysis, and LSTM with sentiment analysis. And proven among the three approaches (es), that the most suitable analysis for predicting the market price of a stock is LSTM with sentiment analysis [with having the lowest rmse value compared to the other two] [2].

This project will be a great asset for brokers and investors for investing money in the stock market application (like Web, Mobile application) since it is trained on a huge collection of historical data and built with the following approach - LSTM with sentiment analysis to predict stock price(s) of companies [5].

**Future Enhancement**

1. Future scope of this project will involve adding more parameters and factors like the financial ratios, multiple instances, etc. The more the parameters are considered more will be the accuracy.

2. We will try to collect historical data/raw data of more years to have more data points.

3. We will enhance the time series into minute-to-minute data set.

4. This project will be able to the extension of this stock prediction system would be to augment it with a news feed analysis from social media platforms such as Twitter, where emotions are gauged from the articles. This sentiment analysis can be linked with the LSTM to better train weights and further improve accuracy.

5. We will increase layers and units/cells within the LSTM architecture to improve the model accuracy.

6. This project viewing output upgradation like plotting includes stock parameters, company assets, and summary and shares detailed information is considered as a future enhancement.

**ACKNOWLEDGMENT**

To thank the people who helped us to complete this paper we are thankful. We take this opportunity to express our sincere gratitude towards our mentor for her guidance and support and we would like to thank the college for allowing us to work on this paper. Lastly, we would like to thank each, and every person who helped us in the completion of the paper especially my friends and Peers for their wide support.

**REFERENCES**

1. Bharathi S. and Geetha A. (2017). Analysis for effective stock market prediction." International Journal of Intelligent Engineering and Systems, 146–154.
2. Bollen, J. and Mao., H. (2018). "Twitter mood as a stock market predictor." IEEE Computer, 44(6), 91–94.
3. Joshi, K., N., B. H., and Rao, J. "Stock trend prediction using news sentiment analysis.
4. Kordonis, J., Symeonidis, S., and Arampatzis, A. "Stock price forecasting via sentiment analysis on Twitter.
5. M., H., E.A, G., Menon, V. K., and K.P., S. (2018). "Stock price forecasting via sentiment analysis on Twitter." Procedia Computer Science, 132(6), 1351–1362.