Application of an LSTM Network to CEPEA Indicators to Estimate the Price of Arabica Coffee

Deivison Oliveira Costa
Bachelor's Degree in Computer Engineering
Instituto Federal de Minas Gerais
Bambuí, Brazil
0034436@academico.ifmg.edu.br

Artur Francisco Pereira Carvalho
Bachelor's Degree in Computer Engineering
Instituto Federal de Minas Gerais
Bambuí, Brazil
0034430@academico.ifmg.edu.br

***Abstract*—This paper presents a method for estimating the price of Arabica coffee using Long Short-Term Memory (LSTM) networks, a type of Artificial Neural Networks (ANNs). The approach leverages CEPEA indicators as input features for training and validating the LSTM model, which exploits temporal dependencies in coffee price data to predict future trends. Experimental results demonstrate the effectiveness of the proposed approach, with the trained LSTM network achieving a coefficient of determination of 0.9679 and a Mean Absolute Percentage Error (MAPE) of 4.7576 when applied to the separated test set. These findings support informed decision-making in the coffee market.**

***Index Terms*—Arabica coffee, Long Short-Term Memory, Artificial Neural Networks, CEPEA indicators, Price estimation.**

I. Introduction

The price estimation of Arabica coffee is a crucial task for stakeholders in the coffee industry, as it enables them to make informed decisions regarding production, trading, and investment strategies. Accurate price forecasting can contribute to minimizing financial risks and optimizing market operations. In recent years, ANNs have emerged as powerful tools for predicting time series data, and specifically, the LSTM networks have shown promising results due to their ability to capture temporal dependencies in the data.

This paper presents an innovative approach that leverages LSTM artificial neural networks for estimating the price of Arabica coffee. The focus of the study is to incorporate CEPEA indicators, provided by the Centro de Estudos Avançados em Economia Aplicada da Escola Superior de Agricultura "Luiz de Queiroz" da Universidade de São Paulo (CEPEA ESALQ/USP) official website, as input resources to train and validate the LSTM model. These indicators provide valuable information on various factors that affect the coffee market, such as production volumes, international trade, consumption patterns and weather conditions.

The application of LSTM networks in the domain of financial forecasting has gained significant attention in recent years. LSTM networks, a variant of recurrent neural networks (RNNs), are designed to handle long-term dependencies by utilizing memory cells. These cells can selectively store, read, and write information, allowing the network to capture patterns over extended periods. Such characteristics make LSTM networks particularly suitable for modeling and predicting complex time series data.

Several studies have explored the use of ANNs, including LSTM networks, for financial time series prediction. In the field of stock market forecasting, authors have successfully employed LSTM networks to estimate stock prices, predict trading volumes, and detect market trends [1] [2]. Moreover, LSTM networks have been applied in the energy sector for forecasting electricity prices, load demand, and renewable energy generation [3] [4]. However, the application of LSTM networks in the coffee market domain, particularly in the context of Arabica coffee price estimation, remains relatively unexplored.

The novelty of this study lies in its use of CEPEA indicators as input features for training and validating the LSTM network. CEPEA indicators encompass a wide range of influential factors that drive the Arabica coffee market, enabling the network to learn and capture the complex relationships between these indicators and the coffee price. By exploiting the temporal dependencies in the coffee price data, the LSTM network can effectively estimate future price trends, empowering stakeholders with valuable insights.

II. Problem Characterization and Related Work

A. Problem

The price estimation of Arabica coffee poses a significant challenge for stakeholders in the coffee industry. Accurate forecasting of coffee prices is essential for decision-making processes, including production planning, trading strategies, and investment decisions. By obtaining reliable price estimates, stakeholders can minimize financial risks and optimize their operations in the coffee market.

B. Similar Works

The application of ANNs, particularly LSTM networks, has gained prominence in predicting time series data. LSTM networks have proven effective in capturing temporal dependencies, making them well-suited for financial forecasting tasks.

Numerous studies have successfully utilized ANNs, including LSTM networks, in the domain of financial time series prediction. For instance, Moghar and Hamiche [1] employed

an LSTM recurrent neural network based on deep learning to predict stock prices. The study demonstrated the effectiveness of LSTM networks in stock market forecasting.

Similarly, Fischer and Krauss [2] explored the use of LSTM networks for financial market predictions. Their research focused on modeling the relationships between historical financial data and future market trends. The results highlighted the capabilities of LSTM networks in capturing complex patterns and making accurate predictions.

In addition to stock market forecasting, LSTM networks have found application in the energy sector. Liu et al. [3] utilized LSTM neural networks for short-term load forecasting. The study demonstrated the suitability of LSTM networks for predicting electricity demand, a crucial task for energy market participants.

Furthermore, Wang et al. [4] proposed a novel hybrid LSTM-based model for electricity price forecasting. By leveraging the memory cells of LSTM networks, the model achieved accurate predictions of electricity prices, aiding market participants in decision-making processes.

However, the specific application of LSTM networks in estimating the price of Arabica coffee remains relatively unexplored. While LSTM networks have shown success in various domains, their potential in the coffee market domain, especially when combined with CEPEA indicators, has not been extensively investigated.

The novelty of this study lies in its integration of CEPEA indicators as input features for training and validating the LSTM network. CEPEA indicators encompass a broad range of factors influencing the Arabica coffee market, including production volumes, international trade, consumption patterns, and climatic conditions. By incorporating these indicators into the LSTM model, the network can learn the intricate relationships between these factors and the price of Arabica coffee, enhancing the accuracy of price estimation.

III. Theoretical foundation

A. LSTM Algorithm

The LSTM algorithm has emerged as a powerful tool in the field of artificial intelligence and machine learning. With the increasing complexity and abundance of sequential data, such as time series, natural language, and audio streams, the need for models capable of capturing long-term dependencies became evident. The LSTM algorithm, which was introduced by Hochreiter and Schmidhuber in 1997 [5], addresses the vanishing gradient problem and enables effective learning of long-range dependencies in sequential data.

LSTM has gained significant attention due to its ability to capture and retain information over extended periods, making it particularly suited for processing sequential data. Traditional RNNs often struggle with long-range dependencies, as the gradients tend to either explode or vanish during backpropagation. The LSTM architecture overcomes this limitation by incorporating memory cells and gating mechanisms that allow it to selectively retain and discard information based on its relevance.

The LSTM algorithm has found applications in various domains, including natural language processing, speech recognition, time series forecasting, and even financial market analysis. Its ability to handle sequential data efficiently has made it a preferred choice for tasks involving temporal dependencies and long-term patterns.

B. History of LSTM Algorithm

The history of the LSTM algorithm dates back to 1997 when Sepp Hochreiter and Jürgen Schmidhuber published their groundbreaking paper titled ”Long Short-Term Memory” [5]. At the time, traditional RNNs were widely used for processing sequential data. However, RNNs faced a significant challenge known as the vanishing gradient problem, where the gradients propagated through the network during training either vanished or exploded, leading to difficulties in capturing long-term dependencies.

Hochreiter and Schmidhuber [5] recognized the limitations of traditional RNNs and proposed the LSTM architecture as a solution to the vanishing gradient problem. The LSTM algorithm introduced the concept of memory cells, which allowed the network to store and access information over extended periods. The key insight was the addition of gating mechanisms that regulated the flow of information within the LSTM units.

The original LSTM architecture comprised the input gate, forget gate, output gate, and a memory cell. These components worked together to control the flow of information and selectively retain or forget relevant information at each time step. The input gate determined how much new information should be stored in the memory cell, while the forget gate decided which information to discard. The output gate controlled the amount of information to be output from the cell. By incorporating these gating mechanisms, LSTM units could effectively capture and retain essential information over varying time intervals.

The LSTM algorithm quickly gained recognition for its ability to learn and exploit long-term dependencies in sequential data. Researchers and practitioners across various fields began adopting LSTM for tasks such as speech recognition, machine translation, handwriting recognition, and more. Its success in handling sequential data paved the way for advancements in natural language processing and time series analysis.

Over the years, researchers have made further refinements and variations to the LSTM architecture. Modifications, such as peephole connections, bidirectional LSTM, and stacked LSTM, have been introduced to enhance the model’s capabilities and adaptability to different types of sequential data.

Today, the LSTM algorithm stands as one of the most influential contributions to the field of deep learning. Its impact extends beyond academia, with widespread adoption in industry applications ranging from text generation and sentiment analysis to autonomous driving and financial market predictions.

C. Functioning of LSTM Algorithm

The LSTM architecture consists of LSTM units or cells, which are interconnected to form a recurrent neural network. Each LSTM unit contains several key components: the input gate, forget gate, output gate, memory cell, and hidden state.

The input gate regulates the flow of new information into the memory cell. It determines how much of the incoming information should be stored and updated in the memory cell. The input gate takes into account the current input, x_t, the previous hidden state, h_{t-1}, and their corresponding weight matrices, W_{xi} and W_{hi}, along with the bias term, b_i. These values are passed through a sigmoid activation function, represented as σ, to produce the input gate activation, i_t (1):

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{1}$$

The forget gate determines which information should be discarded from the memory cell. It considers the same inputs as the input gate and calculates the forget gate activation, f_t, using a sigmoid activation function (2):

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{2}$$

The output gate controls the flow of information from the memory cell to the hidden state. Similar to the input and forget gates, it takes into account the current input, x_t, the previous hidden state, h_{t-1}, and their corresponding weight matrices, W_{xo} and W_{ho}, along with the bias term, b_o. The output gate activation, o_t, is computed using a sigmoid activation function (3):

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{3}$$

The memory cell, c_t, is the heart of the LSTM unit. It stores and updates information over time by combining the input gate, the forget gate, and the current input, x_t, through a series of mathematical operations. These operations involve element-wise multiplication, addition, and applying the hyperbolic tangent activation function, $\tanh$, to capture the relevant information (4):

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{4}$$

Finally, the hidden state, h_t, is calculated by multiplying the updated memory cell, c_t, with the output gate activation, o_t, after applying the hyperbolic tangent activation function (5):

$$h_t = o_t \tanh(c_t) \tag{5}$$

The hidden state, h_t, represents the output of the LSTM unit, capturing essential information relevant to subsequent tasks or predictions, as per the LSTM model in Fig. 1.

By using these interconnected LSTM units, the LSTM algorithm can effectively learn and capture long-term dependencies in sequential data, making it well-suited for tasks involving temporal dynamics and extended context.

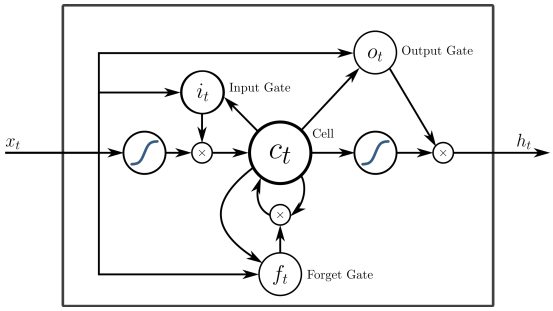

Fig. 1. LSTM model

D. LSTM Algorithm for Time Series Price Prediction

The LSTM algorithm has emerged as a powerful tool for time series price prediction. Its ability to capture long-term dependencies and handle sequential data makes it particularly well-suited for forecasting stock prices, commodity prices, and other financial indicators.

Traditional approaches to time series forecasting often rely on statistical methods or simple models that assume linear relationships or stationary data. However, financial markets exhibit complex dynamics and non-linear patterns, making them challenging to predict accurately. LSTM models address these challenges by capturing temporal dependencies and patterns in the data, enabling more accurate and robust predictions.

Researchers and practitioners have extensively explored the application of LSTM in financial time series prediction, including stock prices. Moghar and Hamiche [1] conducted a study to evaluate the effectiveness of LSTM for stock price prediction. They compared LSTM models with traditional forecasting methods, such as autoregressive integrated moving average (ARIMA) and support vector regression (SVR).

The study by Moghar and Hamiche [1] demonstrated that LSTM models outperformed traditional methods in terms of prediction accuracy. By leveraging the LSTM algorithm's ability to capture long-term dependencies, the models were able to identify and exploit complex patterns in the historical price data. The LSTM models exhibited superior performance in predicting stock prices, providing valuable insights for investors and traders.

The key advantage of LSTM in time series price prediction is its ability to capture both short-term fluctuations and long-term trends. The memory cells and gating mechanisms of LSTM enable the model to retain relevant information over extended periods, allowing it to learn and adapt to evolving market conditions. This makes LSTM models particularly effective for predicting price movements and identifying market trends.

The application of LSTM in time series price prediction extends beyond financial markets. It has been successfully employed in various domains, including energy markets, retail sales forecasting, and demand prediction. The flexibility and accuracy of LSTM models make them a valuable tool for decision-making in industries that rely on accurate and timely predictions of time-dependent variables.

The ability of LSTM models to learn and adapt to complex market dynamics has made them a valuable asset for investors, traders, and decision-makers in various industries.

E. Arabica Coffee Price Prediction

Several research papers have focused on predicting coffee prices using different techniques. Xu and Zhang [6] proposed a neural network-based model to forecast commodity prices, including coffee. Although his study covered several commodities, he did not specifically address the complexities of estimating the price of arabica coffee. Deina et al. [7] introduced a methodology based on extreme learning machines for predicting coffee prices, but their model did not incorporate CEPEA indicators and did not apply to the Brazilian market and its respective currency. Huy et al. [8] employed econometric techniques combined with neural networks, but their study was limited to a single neural network architecture.

The research methodology presented in this paper combines CEPEA indicators, which cover a long history of daily arabica coffee prices (1996-Current), with RNNs to estimate coffee prices. CEPEA's indicators capture the complexity and dynamics of the Brazilian coffee market, providing valuable information on price changes. Implementing an ANN with LSTM architecture, we took advantage of its ability to capture nonlinear relationships and adapt to changing patterns, modeling and predicting arabica coffee prices in the Brazilian market with some precision.

IV. Methodology

A. Objective

The main objective of this study is to estimate the price of Arabica coffee using the LSTM algorithm. The use of LSTM for this task is justified by its ability to capture long-term dependencies and patterns in sequential data. Hochreiter and Schmidhuber [5] demonstrated that LSTM can learn to bridge minimal time lags in excess of 1000 discrete time steps by enforcing constant error flow through constant error carousels within special units.

B. Model Development Environment

For the development of this work, we used Jupyter Notebook, a web application that allows the creation and sharing of documents that contain live code, equations, visualizations and narrative text. As a programming language we used Python version 3.10.10, which is a high-level, interpreted, interactive and object-oriented scripting language. The Python libraries used were: Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, Keras and Tensorflow.

C. Obtaining the Data Set

The dataset employed on this study was obtained from the CEPEA ESALQ/USP official website, which provides a daily history of arabica coffee prices, since setember of 1996 until the current date, the version that was used from the dataset contains data up to June 21, 2023. The dataset was downloaded in XLSX format, adapted to CSV format and imported into the Jupyter Notebook environment.

D. Data Preprocessing

As a preprocessing step, was removed the column of price in Reais (R$) and the column of date, because the LSTM network does not accept categorical data and because the objective of this study is a prediction of price in Dollars.

E. Data Standardization and Division of Training/Validation and Test Sets

The dataset was standardized placing this column of price in Dollars (US$) as being of type *float32* and being placed on a scale between 0 and 1. The dataset was divided into training and test set in the proportion of 70:30 respectively. A summary was also used, represented by Tab. I, plotting a histogram, as shown in Fig. 2, checking outliers with boxplot in Fig. 4 and a graph of arabica coffee price variation over time in Fig. 3 for a better idea of how the data is distributed in the dataset.

TABLE I
SUMMARY OF THE DATASET

count	mean	std	min	max	median
6672.0	**138.15**	**60.04**	**30.92**	**349.39**	**127.04**

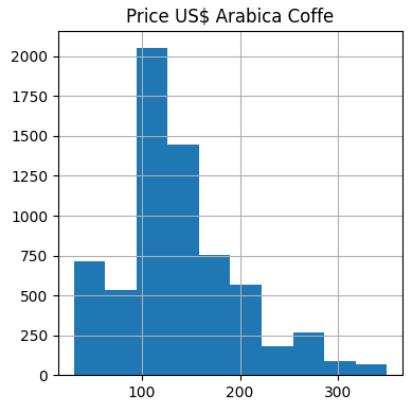

Fig. 2. Histogram of the dataset

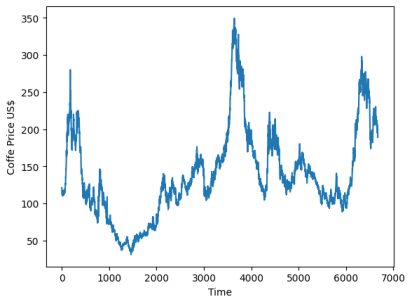

Fig. 3. Arabica coffee price variation over time

F. Treatment of Outliers

For this study, the removal of outliers was not used, because for this dataset, even the data considered outliers are important for the prediction of the price of arabica coffee.

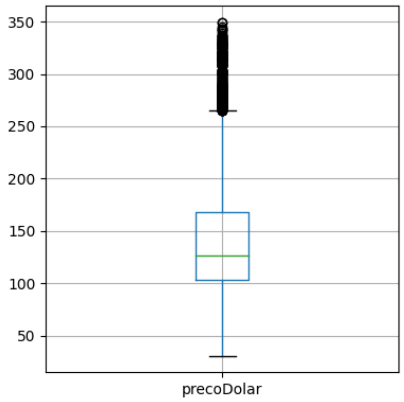

Fig. 4. Boxplot of the dataset

G. Treatment of Missing Data

This dataset has no missing count, so there is no need to use any technique to fill in missing data.

H. LSTM Model Training Specifications

The LSTM network was trained using the Adam optimizer and the Mean Squared Error (MSE) loss function. The network was trained for 1000 epochs, with a batch size of 16. Was used a learning rate of 0.0001, which is the default value of the Adam optimizer. The train set have 20% for valitation and 80% for training. The dataset was not shuffled, because the order of the data is important for the prediction of the price of arabica coffee.

I. LSTM Model Topology

The topology of the LSTM network used in this study have 5 LSTM layers, the last two layers are dense layers, to fit the shape of the output. The first layer have 16 neurons, the second and third layer have 32 neurons, the fourth layer have 16 layers and the last layer have 1 neuron. The activation function used in the LSTM layers is the hyperbolic tangent function (tanh) for all layers. This topology generated a total of 25274 parameters, where all of them are trainable.

V. Results and Discussion

The results of the LSTM network were satisfactory, after training the neural network, the network convergence curve, shown in Fig. 5, was plotted, and there were no overfitting problems or similar to the context.

Using the trained LSTM network, it was possible to perform the prediction of the test set that had been previously separated. After applying the network to this set, we were able to obtain a coefficient of determination of 0.9679 and a MAPE of 4.7576.

Finally, the LSTM forecast graph was plotted along with a histogram of that forecast. The forecast chart is shown in Fig. 6. This graph shows the price of arabica coffee in dollars over time, and the histogram shows the distribution of the forecast data. Also was used a histogram of the forecast data to compare with the histogram of the dataset, to check if the forecast data is similar to the dataset. This histogram is shown in Fig. 7.

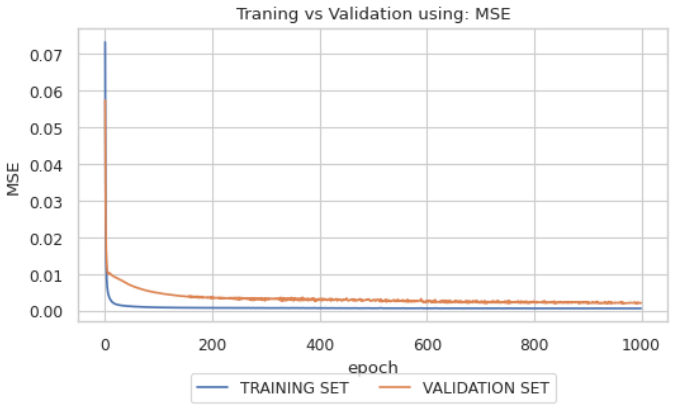

Fig. 5. Convergence curve of the LSTM network

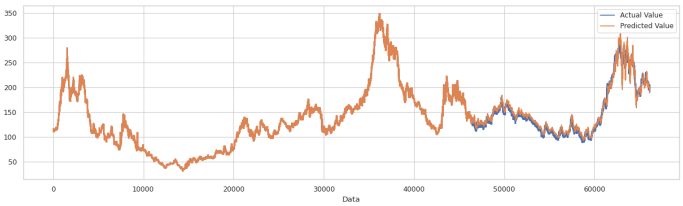

Fig. 6. Forecast chart of the LSTM network

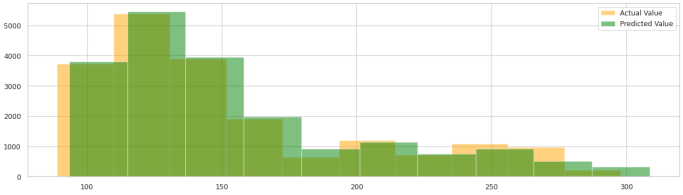

Fig. 7. Histogram of the forecast data

VI. Conclusion

Throughout the paper, we presented an innovative approach that utilizes LSTM artificial neural networks for estimating the price of Arabica coffee. The incorporation of CEPEA indicators as input resources in training and validating the LSTM model provided valuable information on various factors affecting the coffee market. The application of LSTM networks in financial forecasting, particularly in the domain of stock market and energy sector predictions, has shown promising results.

By applying the trained LSTM network to the test set, we achieved a coefficient of determination of 0.9679 and a MAPE of 4.7576. These results indicate a high level of accuracy in predicting the price of Arabica coffee using the LSTM model.

Furthermore, the forecast graph displayed the price of Arabica coffee over time (Fig. 6), and the histogram of the

forecast data was compared with the dataset to assess its similarity (Fig. 7). The histogram comparison indicated that the forecast data aligns well with the dataset.

Overall, the findings of this research highlight the effectiveness of LSTM networks in estimating the price of Arabica coffee, demonstrating their potential for minimizing financial risks and optimizing market operations in the coffee industry. Further research and exploration in this area can lead to enhanced forecasting techniques and better decision-making for stakeholders in the coffee industry.

REFERENCES

[1] A. Moghar and M. Hamiche, "Stock market prediction using LSTM recurrent neural network," in Procedia Computer Science, vol. 170, pp. 1168–1173, 2020.

[2] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," in European Journal of Operational Research, vol. 270, no. 2, pp. 654–669, 2018.

[3] C. Liu et al., "Short-term load forecasting using a long short-term memory network," in IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), pp. 1–6, 2017.

[4] Y. Wang et al., "A novel hybrid LSTM-based model for electricity price forecasting," in Energies, vol. 12, no. 19, p. 3779, 2019.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] X. Xu and Y. Zhang, "Commodity price forecasting via neural networks for coffee, corn, cotton, oats, soybeans, soybean oil, sugar, and wheat," in Proceedings of the Intelligent Systems in Accounting, Finance and Management, vol. 29, no. 3, pp. 169–181, 2022.

[7] C. Deina et al., "A methodology for coffee price forecasting based on extreme learning machines," in Proceedings of the Information Processing in Agriculture, vol. 9, no. 4, pp. 556–565, 2022.

[8] H.T. Huy et al., "Econometric combined with neural network for coffee price forecasting," in Proceedings of the Journal of Applied Economic Sciences, vol. 14, no. 2, 2019.