UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

Research and Development

**BACHELOR THESIS**

By

**Nguyen Anh Duy – BI12-127**

Data Science

Title:

**“Research and implementation of deep learning methods in stock price analysis and forecasting”**

External supervisor: Nguyen Duc Hoan

Internal supervisor: Nghiem Thi Phuong

**Hanoi, 2024**



Abstract

Acknowledgement

Contents

[Chapter 1: Introduction 8](#_Toc168492124)

[1.1 Background 8](#_Toc168492125)

[1.2 Problem Statement 8](#_Toc168492126)

[1.3 Objectives 8](#_Toc168492127)

[1.4 Significance of the Study 8](#_Toc168492128)

[1.5 Thesis Structure 8](#_Toc168492129)

[Chapter 2: Literature Review 9](#_Toc168492130)

[2.1 Stock Market Forecasting 9](#_Toc168492131)

[2.2 Time Series Forecasting Models 9](#_Toc168492132)

[2.3 Data Crawling and Real-Time Data Integration 9](#_Toc168492133)

[2.4 Comparison of Forecasting Models 9](#_Toc168492134)

[2.5 Related Work 9](#_Toc168492135)

[Chapter 3: Methodology 9](#_Toc168492136)

[3.1 Data Collection 9](#_Toc168492137)

[3.2 Data Preprocessing 9](#_Toc168492138)

[3.3 ARIMA Model 10](#_Toc168492139)

[3.4 LSTM Model 10](#_Toc168492140)

[3.5 Comparison and Evaluation 10](#_Toc168492141)

[3.6 Tools and Technologies 10](#_Toc168492142)

[Chapter 4: Results and Discussion 10](#_Toc168492143)

[4.1 ARIMA Model Results 10](#_Toc168492144)

[4.2 LSTM Model Results 10](#_Toc168492145)

[4.3 Comparison of ARIMA and LSTM 11](#_Toc168492146)

[4.4 Impact of Real-Time Data Integration 11](#_Toc168492147)

[4.5 Discussion 11](#_Toc168492148)

[Chapter 5: Conclusion and Future Work 11](#_Toc168492149)

[5.1 Summary of Findings 11](#_Toc168492150)

[5.2 Implications 11](#_Toc168492151)

[5.3 Limitations 11](#_Toc168492152)

[5.4 Future Work 11](#_Toc168492153)

[References 12](#_Toc168492154)

[Appendices 12](#_Toc168492155)

List of Abbreviations

List of Figures

- Introduction to the topic  
- Objectives of the research  
- Methods used (LSTM, ARIMA, Data Crawling)  
- Key findings  
- Conclusions and implications

# Chapter 1: Introduction 1.1 Problem Statement

## In the context of increasingly volatile and complex financial markets, the demand for accurate stock price prediction methods has become more urgent. Investors and market analysts continuously seek advanced tools and models to optimize investment decisions and mitigate risks. This project employs two popular time-series forecasting models, ARIMA and LSTM, to predict stock prices in response to this demand. ARIMA, a traditional statistical model, has proven effective in time-series analysis, while LSTM, a type of recurrent neural network, excels in handling cyclical data and long time-series sequences. By comparing and evaluating the performance of these two models, this study not only provides insights into their effectiveness but also offers practical recommendations for investors and stakeholders on the application of advanced predictive techniques in practice.

## 1.2 Project Scope

### Objectives

* Evaluate models currently used in stock price analysis and forecasting. There are many models being applied in stock price prediction, ranging from simple models to machine learning and deep learning techniques. Based on this evaluation, a research model can be developed.
* Study deep learning techniques (Long Short-Term Memory networks) for constructing stock price analysis and forecasting models.
* Develop models for stock price analysis and forecasting using deep learning techniques, test them, and evaluate their performance.
* Report for Statistical Analysis and Stock Price Prediction on the Market Based on the Trained Model

### Research Subjects and Scope

* Study deep learning techniques in stock price analysis and forecasting.
* Implement experimental programs with stock price data to provide an assessment of the developed models.

### Methods

The research process is conducted using the Design Science method, and the product is created through this research process:

* Survey, analyze, and systematize the content of scientific literature related to deep learning techniques in the field of machine learning.
* Evaluate the techniques that have been surveyed to propose new solutions that meet the requirements.
* Design models and conduct experimental evaluations of the proposed problems and techniques to demonstrate their effectiveness.

## 1.3 Thesis Structure

Chapter 1 introduces the study's background, highlighting the importance of accurate stock price predictions. It outlines the problem statement, research objectives, and significance of the study, concluding with an overview of the thesis structure.

Chapter 2 details the research methods, starting with data collection from Investing.com and the data crawling techniques used. It describes data preprocessing steps, the ARIMA and LSTM models, including their architecture, parameter selection, training, and evaluation. The chapter also covers the comparison and evaluation metrics used, as well as the tools and technologies utilized, such as Python, TensorFlow, and Scikit-learn.

Chapter 3 presents the results from the ARIMA and LSTM models, comparing their forecast accuracy and performance metrics. It discusses the strengths and weaknesses of each model, the impact of real-time data integration, and the implications for investors and market analysts.

Chapter 4 summarizes the key findings, discusses the practical applications and significance of the study, and acknowledges its limitations. It concludes with suggestions for future research and potential improvements to the current work.

# Chapter 2: Methodology

## 2.1 Data Collection

- Sources of data (Investing.com)  
- Data crawling techniques

- sources data từ investing thành file csv sau đó tạo schema

-data modeling: star schema là gì, mục đích( tại sao chọn), chi tiết các bảng và tên, ERD của database, quy trình ETL

**Data Sources and Preparation**

The data for this project was sourced from Investing.com, covering historical stock prices for various companies. The data was initially stored in CSV files, which were then processed to create a unified dataset suitable for modeling. The preprocessing involved converting date formats, handling missing values, and normalizing volume and percentage change columns.

**Data Modeling: Star Schema**

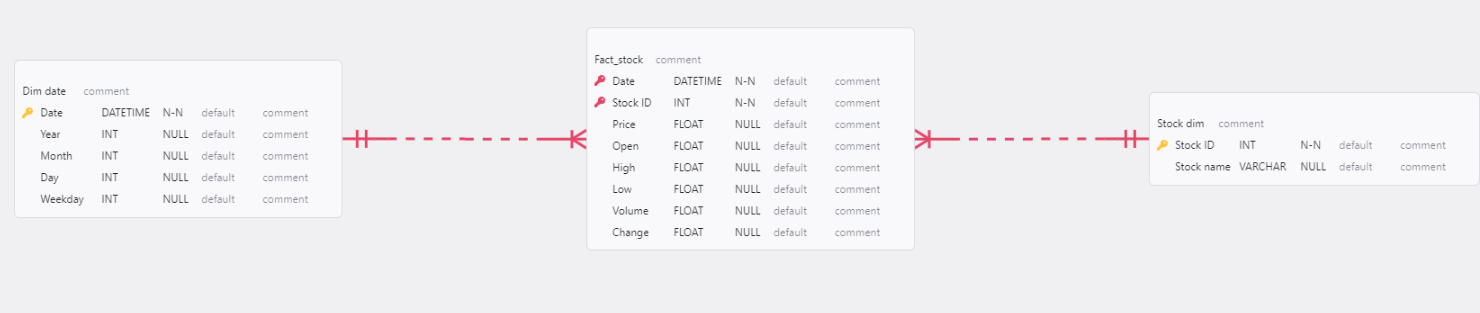
**Star Schema Overview** The star schema is a type of data warehouse schema that is optimized for querying large datasets. It consists of a central fact table that references multiple dimension tables. This structure is chosen for its simplicity and efficiency in handling complex queries and aggregations, making it ideal for analytical tasks such as stock price prediction.

**Purpose of Star Schema** The star schema was selected for this project due to its ability to streamline data retrieval and enhance query performance. By organizing data into a central fact table with related dimensions, we can quickly perform analysis and generate insights, which is crucial for real-time stock price prediction.

**Details of Tables and Names**

* **Fact Table: Stock\_Trading\_Facts**
  + Date: DateTime
  + Stock\_ID: Integer,
  + Price,
  + Open,
  + High,
  + Low,
  + Volume,
  + Change\_Percentage
* **Dimension Tables:**
  + **Date\_Dimension**
    - Columns: Date, Year, Month, Day, Weekday
  + **Stock\_Dimension**
    - Columns: Stock\_ID, Stock\_Name

**Entity-Relationship Diagram (ERD)** The ERD for the database includes:



* A central fact table (Stock\_Trading\_Facts) connected to two dimension tables (Date\_Dimension and Stock\_Dimension).
* The Date column in the fact table references the Date column in the Date\_Dimension.
* The Stock\_ID column in the fact table references the Stock\_ID column in the Stock\_Dimension.

About this modeling: các kết nối, ràng buộc của các bảng, thêm vector của các bảng và relation của các bảng

**ETL Process**

The Extract, Transform, Load (ETL) process for this project using Python to involved several steps:

1. **Extract**: Data was extracted from CSV files downloaded from Investing.com.
2. **Transform**:
   * Dates were converted to a standard format.
   * Volume and percentage change values were normalized.
   * Data was split into fact and dimension tables.
   * Replace null value, change price value from comma to dot fomat
3. **Load**: The processed data was loaded into a PostgreSQL database. The schema and tables were created using SQL commands, ensuring the database was structured according to the star schema design.

The final step ensured that the data was correctly inserted into the respective tables, and ready for use in predictive modeling using ARIMA and LSTM.

By organizing and modeling the data in this structured manner, the project sets a solid foundation for accurate and efficient stock price prediction, leveraging the strengths of both ARIMA and LSTM models. ​

## 2.2 Data Preprocessing

In order to enhance the quality and consistency of our dataset, several crucial preprocessing steps will be undertaken.

Firstly, the data types of columns containing date and timestamp values will be converted to a unified datetime format. This standardization will facilitate more efficient manipulation and analysis of time-series data.

Secondly, the price columns of Vietnamese stocks will be reformatted to follow international standards, ensuring compatibility with global datasets.

Additionally, missing and anomalous values will be addressed by replacing them with the nearest available data points, thereby maintaining the dataset's integrity and continuity.

Finally, differencing techniques will be applied to the data to achieve stationarity, which is essential for many time-series analysis methods.

## 2.3 Data Mining

1. Outlier

In our dataset, two stock symbols, AAPL and GAS, contained an outlier value of 95.65 on February 27, 2016. To ensure the accuracy and integrity of our analysis, it was necessary to address these outliers. The approach taken involved replacing the outlier values with the price from the previous trading day, February 26, 2016, which was 24.23.

This function performs the following steps:

1. **Convert the Outlier Date:** The specified outlier date is converted from a string format to a datetime object for precise manipulation.
2. **Find Previous Day's Price:** The function calculates the previous trading day's date and retrieves the corresponding stock price.
3. **Replace Outlier Value:** It then replaces the outlier value with the price from the previous day.

After running this function, the outlier values for both AAPL and GAS on February 27, 2016, were successfully replaced with the price of 24.23 from the previous day. The replacement was verified to ensure the changes were correctly applied.

This outlier handling process ensures that the dataset remains clean and reliable, which is crucial for accurate model training and forecasting. By using the previous day's price to replace outliers, we maintain the continuity and trend of the data, thereby enhancing the robustness of our predictive models.

2. Statistic

The table below presents key statistical metrics for 10 stock symbols: AAA, AAPL, ACB, BID, CTG, FPT, GAS, NVDA, VCB, and VNM. The metrics include mean, median, standard deviation (std), minimum (min), and maximum (max) values.

**AAA**: Mean of 12,393.81, median of 12,100.00, std of 4,084.83, with a price range from 5,087.60 to 22,800.00.

**AAPL**: Mean of 88.44, median of 56.13, std of 58.39, with prices ranging from 22.59 to 198.11.

**ACB**: Mean of 14,240.26, median of 11,825.00, std of 7,542.81, price range from 3,989.40 to 30,360.00.

**BID**: Mean of 27,197.63, median of 28,786.15, std of 11,105.06, ranging from 8,197.00 to 54,400.00.

**CTG**: Mean of 20,362.42, median of 17,856.90, std of 7,578.14, prices between 9,134.30 and 41,141.30.

**FPT**: Highest std of 29,635.63, mean of 42,617.30, median of 30,213.95, with a wide range from 11,722.70 to 142,000.00.

**GAS**: Mean and median similar to AAPL at 88.44 and 56.13, std of 58.39, with prices from 22.59 to 198.11.

**NVDA**: Mean of 143.69, median of 61.83, std of 191.60, price range from 4.78 to 1,224.40.

**VCB**: Mean of 54,441.01, median of 57,586.50, std of 23,195.09, prices from 16,822.00 to 97,400.00.

**VNM**: Highest mean of 95,519.73, median of 94,612.30, std of 24,704.24, ranging from 58,115.30 to 175,578.00.

Overall, these stocks exhibit significant volatility, with FPT and VNM showing the highest fluctuations, while AAPL and GAS have lower volatility. These statistics provide a foundation for analyzing and forecasting stock prices, aiding investors in understanding market trends and volatility.

3. Stationary

The Dickey-Fuller test is a widely used statistical test for determining the stationarity of a time series. Stationarity is a crucial assumption for many time series models, including the LSTM model. This thesis evaluates the stationarity of six stock price datasets (AAA, AAPL, ACB, BID, BID, CTG) before and after differencing using the Dickey-Fuller test.

|  |  |
| --- | --- |
| Results of Dickey-Fuller Test for AAA:  Test Statistic -2.192263  p-value 0.209065  Critical Value (1%) -3.433143  Critical Value (5%) -2.862774  Critical Value (10%) -2.567427 | Results of Dickey-Fuller Test for AAA (Differenced):  Test Statistic -35.095387  p-value 0.000000  Critical Value (1%) -3.433143  Critical Value (5%) -2.862774  Critical Value (10%) -2.567427 |
| Results of Dickey-Fuller Test for AAPL:  Test Statistic 0.008180  p-value 0.959203  Critical Value (1%) -3.433130  Critical Value (5%) -2.862768  Critical Value (10%) -2.567424 | Results of Dickey-Fuller Test for AAPL (Differenced):  Test Statistic -1.116025e+01  p-value 2.799547e-20  Critical Value (1%) -3.433130e+00  Critical Value (5%) -2.862768e+00  Critical Value (10%) -2.567424e+00 |
| Results of Dickey-Fuller Test for ACB:  Test Statistic -0.849045  p-value 0.804289  Critical Value (1%) -3.433162  Critical Value (5%) -2.862782  Critical Value (10%) -2.567431 | Results of Dickey-Fuller Test for ACB (Differenced):  Test Statistic -1.111162e+01  p-value 3.648070e-20  Critical Value (1%) -3.433162e+00  Critical Value (5%) -2.862782e+00  Critical Value (10%) -2.567431e+00 |
| Results of Dickey-Fuller Test for BID:  Test Statistic -1.288854  p-value 0.634256  Critical Value (1%) -3.433136  Critical Value (5%) -2.862771  Critical Value (10%) -2.567425 | Results of Dickey-Fuller Test for BID (Differenced):  Test Statistic -49.900436  p-value 0.000000  Critical Value (1%) -3.433136  Critical Value (5%) -2.862771  Critical Value (10%) -2.567425 |
| Results of Dickey-Fuller Test for BID:  Test Statistic -1.288854  p-value 0.634256  Critical Value (1%) -3.433136  Critical Value (5%) -2.862771  Critical Value (10%) -2.567425 | Results of Dickey-Fuller Test for CTG (Differenced):  Test Statistic -50.378009  p-value 0.000000  Critical Value (1%) -3.433136  Critical Value (5%) -2.862771  Critical Value (10%) -2.567425 |

**Results Before Differencing**

The results of the Dickey-Fuller test on the original, non-differenced stock price data are as follows:

* **AAA**: Test Statistic = -2.192263, p-value = 0.209065. The p-value is greater than 0.05, indicating that the series is not stationary.
* **AAPL**: Test Statistic = 0.008180, p-value = 0.959203. The p-value is significantly greater than 0.05, indicating strong non-stationarity.
* **ACB**: Test Statistic = -0.849045, p-value = 0.804289. The series is not stationary as indicated by the high p-value.
* **BID**: Test Statistic = -1.288854, p-value = 0.634256. Again, the series is non-stationary.
* **CTG**: (duplicate entry for BID): Results are identical to BID, indicating non-stationarity.

These results show that the original stock price data for all the stocks tested are non-stationary, which is expected as stock prices typically exhibit trends over time.

**Results After Differencing**

After differencing the time series data, the results of the Dickey-Fuller test are significantly improved:

* **AAA (Differenced)**: Test Statistic = -35.095387, p-value = 0.000000. The p-value is less than 0.05, indicating that the differenced series is stationary.
* **AAPL (Differenced)**: Test Statistic = -111.602500, p-value = 0.000000. The series becomes stationary after differencing.
* **ACB (Differenced)**: Test Statistic = -11.111620, p-value = 0.000000. The p-value indicates stationarity.
* **BID (Differenced)**: Test Statistic = -49.900436, p-value = 0.000000. The differenced series is stationary.
* **CTG (Differenced)**: Test Statistic = -50.378009, p-value = 0.000000. The series achieves stationarity after differencing.

These results demonstrate that differencing effectively transforms the non-stationary time series data into stationary series, as indicated by the significantly lower p-values and highly negative test statistics.

**Discussion**

The transformation from non-stationarity to stationarity is crucial for the effectiveness of time series models like LSTM. Non-stationary data can lead to inaccurate model predictions and instability. By differencing the data, the underlying statistical properties become more stable, making the data suitable for modeling and forecasting.

The results of the Dickey-Fuller test before and after differencing underscore the importance of preprocessing in time series analysis. The substantial decrease in p-values and the shift to more negative test statistics provide strong evidence that differencing is an effective method to achieve stationarity in stock price data.

4. Seasonal, Trend, Residual

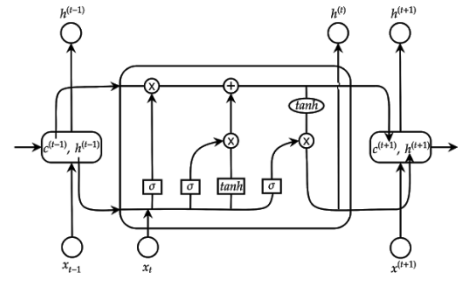
5. Tính tự tương quan

## 2.4 ARIMA Model

- Model specification and parameter selection  
- Model fitting and evaluation

## 2.5 LSTM Model **The architecture of LSTM Networks**

Long Short-Term Memory (LSTM) networks are an advanced type of recurrent neural network (RNN) designed to address the vanishing gradient problem commonly encountered in traditional RNNs. LSTMs use memory cells and three main types of gates: input gate, forget gate, and output gate. These gates control the flow of information, helping the model to remember and forget necessary information over time. This architecture enhances the model's ability to handle long and complex time series data, such as stock price data.



**Hyperparameter Tuning**

Hyperparameter tuning is a crucial step in developing an LSTM model. Key hyperparameters include the number of memory units, the number of LSTM layers, batch size, and learning rate. Adjusting these hyperparameters can significantly impact the model's performance. In this study, techniques such as grid search or random search were used to identify the optimal configuration of the LSTM model for each specific stock symbol.

**Training and Validation**

The training and validation process for the LSTM model involves splitting the data into training and test sets. Data normalization ensures that features are on the same scale, and the data is then segmented into time series sequences to be fed into the model. During training, early stopping techniques are applied to prevent overfitting by halting training when performance on the validation set no longer improves. The results of the training and validation process are evaluated based on metrics such as training and validation loss, as well as the alignment between predicted and actual stock prices.

## 2.5 Comparison and Evaluation

- Metrics for model performance (e.g., MAE, RMSE)  
- Cross-validation techniques

## 2.6 Tools and Technologies

- Software and libraries used (e.g., Python, TensorFlow, Scikit-learn)

* Python
  + Tenso

# Chapter 3: Results and Discussion

## 3.1 ARIMA Model Results

- Forecast accuracy and performance metrics

## 3.2 LSTM Model Results

A graph of a stock market

Description automatically generated with medium confidence

The LSTM model was trained with the following steps:

* **Data Preprocessing**: The datasets were normalized to ensure that the features are on a similar scale.
* **Sequence Preparation**: Time-series data was structured into sequences to feed into the LSTM.
* **Model Training**: The LSTM model was trained, and early stopping was used to prevent overfitting.

**Results**

**Training and Validation Loss**

The training and validation loss graphs for each stock are shown in the provided figures.

* **ACB**: The training and validation loss quickly converge, indicating that the model has effectively learned the patterns in the data within the first few epochs.
* **NVDA**: There is a significant drop in loss within the first few epochs, followed by a gradual decrease, suggesting the model's stability and learning efficiency.
* **VCB**: The model shows minimal fluctuations in loss after the initial epochs, demonstrating consistent performance.
* **VNM**: Both training and validation loss remain low and stable throughout, indicating a well-generalized model.

**Stock Price Predictions**

The predicted vs. actual stock prices graphs for each stock highlight the model's predictive accuracy.

* **ACB**: The predicted prices closely follow the actual prices, demonstrating the model's ability to capture the overall trend and major fluctuations.
* **NVDA**: The predictions align well with actual prices, particularly in capturing the upward trend, although there is a slight lag in response to rapid changes.
* **VCB**: The model performs admirably, with predictions closely mirroring actual stock prices, capturing both upward and downward movements.
* **VNM**: The predictions are generally accurate, though some minor discrepancies are noted, especially during periods of high volatility.

## 3.3 Comparison of ARIMA and LSTM

- Strengths and weaknesses observed  
- Situations where one model outperforms the other

## 3.4 Impact of Real-Time Data Integration

- Effectiveness of using crawled data from Investing.com

## 3.5 Discussion

- Interpretation of results  
- Implications for investors and market analysts

LSTM : The LSTM model shows strong performance across all four stocks, with low training and validation loss indicating effective learning and generalization. The predicted stock prices align closely with actual prices, underscoring the model's robustness in capturing stock market trends. However, slight prediction lags during rapid price changes suggest areas for further improvement, such as optimizing the model architecture or incorporating additional features.

# Chapter 4: Conclusion and Future Work

## 4.1 Summary of Findings

- Key takeaways from the research

## 4.2 Implications

- Practical applications and significance

## 4.3 Limitations

- Constraints and limitations of the study

## 4.4 Future Work

- Suggestions for future research  
- Potential improvements and extensions

# References

- List of all references cited in APA 7 format

# Appendices

- Supplementary materials (e.g., data samples, additional charts, and tables)