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**

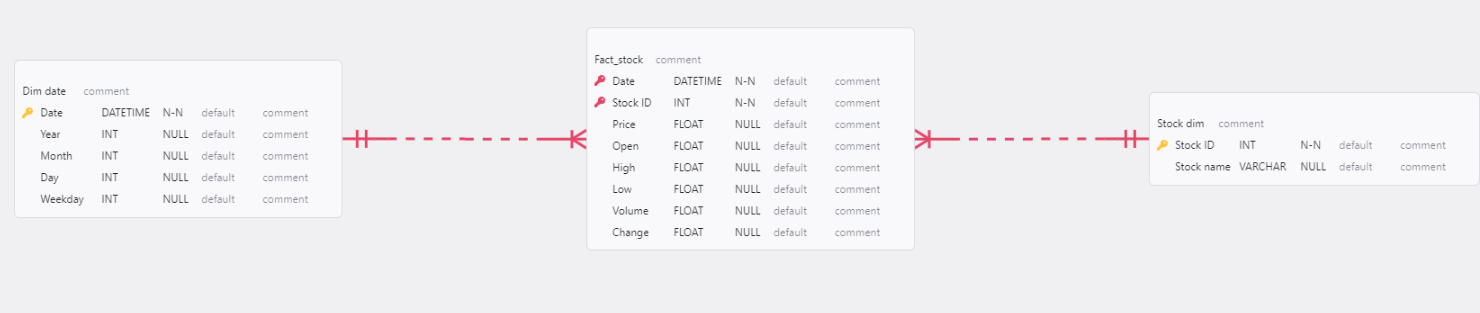
**Star Schema Overview** The star schema is a type of data warehouse schema that is optimized for querying large datasets. It consists of a central fact table that references multiple dimension tables. This structure is chosen for its simplicity and efficiency in handling complex queries and aggregations,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

2.Statistic

3. Stationary

|  |  |
| --- | --- |
| 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 |

As you see before differ and after differ,

4. Seasonal, Trend, Residual

5. Tính tự tương quan

6. Rolling mean …

## 2.4 ARIMA Model

- Model specification and parameter selection  
- Model fitting and evaluation

## 2.5 LSTM Model

- Architecture of LSTM networks  
- Hyperparameter tuning  
- Training and validation

## 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

- Forecast accuracy and performance metrics

## 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

# 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)