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

Can stock prices be predicted and thus profited from? The Efficient Market Hypothesis (EMH) posits that this is not feasible in efficient markets, where information is disseminated rapidly (Campanella et al., 2016; Urquhart and Hudson, 2013; Shonkwiler, 2013; Malkiel, 2003). If this hypothesis held entirely true, the field of investment finance would not exist. According to EMH, rational investors in efficient markets have already accounted for all available information in current stock prices, making future prices unpredictable and unprofitable (Manahov and Hudson, 2014; Malkiel, 2003). EMH asserts that future stock prices will be determined only by unforeseen events, which, once occurring, will be reflected in stock prices (Urquhart and Hudson, 2013; Malkiel, 2003). Thus, EMH advocates believe stock prices follow a Random Walk (RW) model, implying that sustained prediction and profit from stock price movements is impossible, with any gains being purely by chance (Urquhart and Hudson, 2013; Malkiel, 2003).

While EMH and the random walk model have been influential, turning from theory to doctrine in finance (Manahov and Hudson, 2014), the Adaptive Market Hypothesis (AMH) has emerged as an alternative (Lo, 2017). AMH contends that investors, as humans, exhibit both rational and irrational behaviors and adapt to changing environments (economy, technology, etc.) (Lo, 2017). This creates periods when markets are efficient, as EMH assumes, and periods when markets are predictable and do not follow a random walk, thus allowing for profitability (Lo, 2017). Therefore, AMH suggests that stock prices can be predicted under certain conditions, unlike EMH.

Stock prices are generally determined by supply and demand forces driven by traders' buy or sell decisions (Thomsett, 2015). Two main schools of thought in predicting stock prices are technical analysis and fundamental analysis (Thomsett, 2015; Rockefeller, 2011). Technical analysis relies on historical stock price and trading volume data (Rockefeller, 2011; Lorenzo, 2013), while fundamental analysis evaluates a company's potential to generate economic value (profitability, long-term growth, etc.) (Thomsett, 2015). Analysts, regardless of their preferred method, forecast stock prices to make trading decisions (buy, sell, hold). Historically, analysts have belonged exclusively to one camp, holding opposing views on the efficacy of fundamental versus technical analysis (Schwager and Turner, 1995). However, practitioners now often use a combined approach, integrating both technical and fundamental indicators in their analysis (Thomsett, 2015; Rockefeller, 2011). This combined approach leverages the strengths of both methods (Thomsett, 2015; Schwager and Turner, 1995).

Recent advancements in hardware and software have increased the role of computing in stock markets and trading (Lo, 2017; Tkac and Verner, 2016; Ibidapo et al., 2017). Machine learning methods for stock price forecasting have become popular and successful (Tkac and Verner, 2016; Ibidapo et al., 2017; Lo, 2017). These methods, including Artificial Neural Networks (ANN) and Support Vector Regression (SVR), have been widely used to learn from past stock movements and generate future forecasts (Cavalcante et al., 2016; Nassirtoussi et al., 2014; Atsalakis and Valavanis, 2009).

However, surveys indicate that machine learning research has mainly focused on technical indicators, often overlooking fundamental indicators (Ibidapo et al., 2017; Cavalcante et al., 2016; Atsalakis and Valavanis, 2009). This focus on technical indicators contrasts with the benefits of using both types together, as argued by finance practitioners and researchers. This raises several questions:

* What are the consequences of machine learning researchers' preference for technical over fundamental indicators?
* Is one type of analysis significantly better than the other, and can their effects be isolated?
* Does combining technical and fundamental analysis yield better performance than using them separately?

Another issue with forecasting models is treating the relationship between drivers and stock prices as static, whereas it is dynamic (Cavalcante et al., 2016; Cavalcante and Oliveria, 2015). Financial time series data, like stock prices, are non-stationary, and "Concept Drift" occurs over time, requiring models to adapt to changing relationships (Cavalcante and Oliveria, 2015). Changes in political, economic, and regulatory environments contribute to these shifts (Cavalcante et al., 2016). The stock market's dynamic nature, with states like trending, non-trending, chaotic, bullish, and bearish, impacts stock prices. Indicators like the Volatility Index (VIX) measure market sentiment and expectations of volatility (Achelis, 2000; Rockefeller, 2011). Clustering techniques have been applied to financial time series for forecasting (Aghabozorgi et al., 2015; D’Urso et al., 2013; Fu, 2011) and developing localized models (Tsinaslanidis and Kugiumtzis, 2014; Cherif et al., 2011; Wu and Lee, 2015). However, these approaches have mainly targeted stock price fluctuations rather than market state sensitivity, raising additional questions:

* How does adapting forecasting models for different market states affect performance?
* How can this approach be best implemented and evaluated?
* Can clustering methods capture market states and integrate them into stock price prediction?

Can stock prices be predicted and thus profited from? The Efficient Market Hypothesis (EMH) posits that this is not feasible in efficient markets, where information is disseminated rapidly (Campanella et al., 2016; Urquhart and Hudson, 2013; Shonkwiler, 2013; Malkiel, 2003). If this hypothesis held entirely true, the field of investment finance would not exist. According to EMH, rational investors in efficient markets have already accounted for all available information in current stock prices, making future prices unpredictable and unprofitable (Manahov and Hudson, 2014; Malkiel, 2003). EMH asserts that future stock prices will be determined only by unforeseen events, which, once occurring, will be reflected in stock prices (Urquhart and Hudson, 2013; Malkiel, 2003). Thus, EMH advocates believe stock prices follow a Random Walk (RW) model, implying that sustained prediction and profit from stock price movements is impossible, with any gains being purely by chance (Urquhart and Hudson, 2013; Malkiel, 2003).

While EMH and the random walk model have been influential, turning from theory to doctrine in finance (Manahov and Hudson, 2014), the Adaptive Market Hypothesis (AMH) has emerged as an alternative (Lo, 2017). AMH contends that investors, as humans, exhibit both rational and irrational behaviors and adapt to changing environments (economy, technology, etc.) (Lo, 2017). This creates periods when markets are efficient, as EMH assumes, and periods when markets are predictable and do not follow a random walk, thus allowing for profitability (Lo, 2017). Therefore, AMH suggests that stock prices can be predicted under certain conditions, unlike EMH.

Stock prices are generally determined by supply and demand forces driven by traders' buy or sell decisions (Thomsett, 2015). Two main schools of thought in predicting stock prices are technical analysis and fundamental analysis (Thomsett, 2015; Rockefeller, 2011). Technical analysis relies on historical stock price and trading volume data (Rockefeller, 2011; Lorenzo, 2013), while fundamental analysis evaluates a company's potential to generate economic value (profitability, long-term growth, etc.) (Thomsett, 2015). Analysts, regardless of their preferred method, forecast stock prices to make trading decisions (buy, sell, hold). Historically, analysts have belonged exclusively to one camp, holding opposing views on the efficacy of fundamental versus technical analysis (Schwager and Turner, 1995). However, practitioners now often use a combined approach, integrating both technical and fundamental indicators in their analysis (Thomsett, 2015; Rockefeller, 2011). This combined approach leverages the strengths of both methods (Thomsett, 2015; Schwager and Turner, 1995).

Recent advancements in hardware and software have increased the role of computing in stock markets and trading (Lo, 2017; Tkac and Verner, 2016; Ibidapo et al., 2017). Machine learning methods for stock price forecasting have become popular and successful (Tkac and Verner, 2016; Ibidapo et al., 2017; Lo, 2017). These methods, including Artificial Neural Networks (ANN) and Support Vector Regression (SVR), have been widely used to learn from past stock movements and generate future forecasts (Cavalcante et al., 2016; Nassirtoussi et al., 2014; Atsalakis and Valavanis, 2009).

However, surveys indicate that machine learning research has mainly focused on technical indicators, often overlooking fundamental indicators (Ibidapo et al., 2017; Cavalcante et al., 2016; Atsalakis and Valavanis, 2009). This focus on technical indicators contrasts with the benefits of using both types together, as argued by finance practitioners and researchers. This raises several questions:

* What are the consequences of machine learning researchers' preference for technical over fundamental indicators?
* Is one type of analysis significantly better than the other, and can their effects be isolated?
* Does combining technical and fundamental analysis yield better performance than using them separately?

Another issue with forecasting models is treating the relationship between drivers and stock prices as static, whereas it is dynamic (Cavalcante et al., 2016; Cavalcante and Oliveria, 2015). Financial time series data, like stock prices, are non-stationary, and "Concept Drift" occurs over time, requiring models to adapt to changing relationships (Cavalcante and Oliveria, 2015). Changes in political, economic, and regulatory environments contribute to these shifts (Cavalcante et al., 2016). The stock market's dynamic nature, with states like trending, non-trending, chaotic, bullish, and bearish, impacts stock prices. Indicators like the Volatility Index (VIX) measure market sentiment and expectations of volatility (Achelis, 2000; Rockefeller, 2011). Clustering techniques have been applied to financial time series for forecasting (Aghabozorgi et al., 2015; D’Urso et al., 2013; Fu, 2011) and developing localized models (Tsinaslanidis and Kugiumtzis, 2014; Cherif et al., 2011; Wu and Lee, 2015). However, these approaches have mainly targeted stock price fluctuations rather than market state sensitivity, raising additional questions:

* How does adapting forecasting models for different market states affect performance?
* How can this approach be best implemented and evaluated?
* Can clustering methods capture market states and integrate them into stock price prediction?

## 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

This thesis is organized into four main chapters, each detailing crucial aspects of the research on stock price prediction using ARIMA and LSTM models. 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

## 2.2 Data Preprocessing

- Cleaning and preparing data for analysis  
- Handling missing values and outliers

## 2.3 ARIMA Model

- Model specification and parameter selection  
- Model fitting and evaluation

## 2.4 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)