UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

**UNDERGRADUATE SCHOOL**



**BACHELOR THESIS**

By

Dang Van Tuan - BA11-096

Data Science

Academic years: 2020 - 2024

Title:

**Big Data Stock Forecaster - A Real-time LSTM-based Stock Price Prediction Platform**

**External Supervisor:** Nguyen Huy Hoang

FPT Software, 17 Duy Tan, Hanoi

**Internal Supervisor:** Dr. Nghiem Thi Phuong

University of Science and Technology of Hanoi

Hanoi, July 2024

Big Data Stock Analysis

**Title Page**

* Title of the Thesis
* Your Name
* Affiliations
* Date

**Abstract**

* A brief summary of the thesis, including the research question, methodology, main findings, and conclusions.

**Acknowledgements**

* Optional section to thank those who assisted in the research.

**Table of Contents**

* Lists sections of the thesis and page numbers for easy navigation.

**List of Figures and Tables**

* Optional but recommended if your thesis includes multiple figures and tables.

**Abbreviations**

* A list of abbreviations used in the thesis.

**Chapter 1: Introduction**

* **Background:** Provide context for your research.
* **Research Question:** Define the specific question or hypothesis your thesis addresses.
* **Significance of the Study:** Explain the importance and potential impact of your research.
* **Scope and Limitations:** Outline the scope of your study and acknowledge any limitations.

**Chapter 2: Literature Review**

* **Theoretical Framework:** Discuss the theoretical aspects of big data in stock market analysis.
* **Previous Research:** Summarize previous studies related to your research topic.
* **Gap in Literature:** Identify the gap your research aims to fill.

**Chapter 3: Methodology**

* **Data Collection:** Describe the sources of your data and how it was collected.
* **Data Analysis:** Explain the techniques and tools used for analyzing the data (e.g., machine learning models, statistical analysis).
* **Validity and Reliability:** Address how you ensured the validity and reliability of your analysis.

**Chapter 4: Results**

* **Data Description:** Provide a detailed description of the dataset(s) used.
* **Analysis:** Present the results of your data analysis, including tables, figures, and graphs to support your findings.
* **Interpretation:** Interpret the results in the context of your research question.

**Chapter 5: Discussion**

* **Findings:** Discuss the implications of your findings in relation to the research question and literature review.
* **Limitations:** Acknowledge the limitations of your study and their impact on the findings.
* **Future Research:** Suggest areas for future research based on your findings.

**Chapter 6: Conclusion**

* **Summary:** Briefly summarize the research conducted and the main findings.
* **Contributions:** Highlight the contributions of your research to the field.
* **Practical Implications:** Discuss any practical implications of your findings for the stock market analysis and decision-making.

**References**

* List all the sources cited in your thesis in a consistent format, following the citation style recommended by your institution.

**Appendices**

* Include any additional material that supports your thesis but is too detailed for the main body, such as data tables, code, questionnaires, etc.

**Trang Tiêu Đề**

* Tiêu đề báo cáo
* Tên tác giả (và đồng tác giả nếu có)
* Tên cơ quan, trường đại học
* Ngày

**Tóm Tắt (Abstract)**

* Mô tả ngắn gọn về mục tiêu nghiên cứu, phương pháp, kết quả chính, và ý nghĩa của nghiên cứu.

**Lời Cảm Ơn (Acknowledgements)**

* Bày tỏ lòng biết ơn đến những người đã hỗ trợ trong quá trình nghiên cứu.

**Mục Lục (Table of Contents)**

* Liệt kê các chương, mục phụ, hình ảnh, bảng biểu với số trang tương ứng.

**Danh Sách Hình Ảnh và Bảng Biểu (List of Figures and Tables)**

**Chương 1: Giới Thiệu (Introduction)**

* Bối cảnh và ngữ cảnh của nghiên cứu.
* Mô tả vấn đề và tầm quan trọng của việc dự đoán giá cổ phiếu.
* Mục tiêu nghiên cứu và câu hỏi nghiên cứu.
* Giới thiệu sơ lược về LSTM và ứng dụng trong dự đoán giá cổ phiếu.

**Chương 2: Tổng Quan Lý Thuyết (Literature Review)**

* Các công trình nghiên cứu liên quan.
* Giới thiệu về Big Data và vai trò của nó trong phân tích tài chính.
* Tổng quan về mạng LSTM và cách thức hoạt động.
* Các phương pháp dự đoán giá cổ phiếu trước đây.

**Chương 3: Phương Pháp Nghiên Cứu (Methodology)**

* Mô tả chi tiết về cách thu thập dữ liệu Big Data.
* Kiến trúc của hệ thống dự đoán: bao gồm cách thiết kế mạng LSTM, quá trình xử lý và chuẩn bị dữ liệu.
* Mô tả các tiêu chí đánh giá hiệu suất mô hình.

**Chương 4: Triển Khai và Kết Quả (Implementation and Results)**

* Mô tả quá trình triển khai hệ thống.
* Trình bày và phân tích kết quả dự đoán giá cổ phiếu.
* So sánh hiệu suất với các phương pháp khác.

**Chương 5: Thảo Luận (Discussion)**

* Giải thích và thảo luận về kết quả.
* Đánh giá về ưu và nhược điểm của hệ thống.
* Khả năng ứng dụng và mở rộng của nghiên cứu.

**Chương 6: Kết Luận và Hướng Nghiên Cứu Tiếp Theo (Conclusion and Future Work)**

* Tóm tắt các phát hiện chính.
* Đề xuất về hướng nghiên cứu tiếp theo và cách cải tiến hệ thống.

**Tham Khảo (References)**

* Danh sách các tài liệu, bài báo, sách, và nguồn trực tuyến đã tham khảo.

**Phụ Lục (Appendices)**

* Bất kỳ dữ liệu bổ sung, mã nguồn, hoặc thông tin chi tiết khác liên quan đến nghiên cứu.

# Abstract

# Acknowledgement

I would like to express my gratitude to all those who gave me the possibility to complete this thesis in specific as well as academic support in general that I received during the time of my studying at the University of Science and Technology of Hanoi.

First and foremost, I want to express my sincere gratitude to Dr. Nghiem Thi Phuong, who served as my academic advisor at the university, for his ongoing encouragement and support. He has helped me by pointing out the gaps in my project so I may fill them in and strengthen my argument.

Secondly, I'd want to specifically thank Mr. Nguyen Huy Hoang, my supervisor at FPT Software, as well as the rest of the company's team for inspiring my ideas and ambitions. They have offered me many opportunities to study and practice.

Last but not least, I want to thank all of my friends and colleagues who have helped me with my work and motivated me to persevere through difficult periods when I've tried to manage work and school or encountered challenges.

This project would not have been completed successfully without the dedicated help of all mentors, family, and friends.

# List of Abbreviations

**API** Application Programming Interface.

**AI** Artificial intelligence.

**GCP** Google Cloud Platform

**CSS** Cascading Style Sheets

**CSV** Comma Separated Values.

**ELT** Extract, Load, Transform.

**ETL** Extract, Transform, Load.

**JSON** JavaScript Object Notation.

**HTTP** Hypertext Transfer Protocol

**MUI** Material UI

**QA** Quality Assurance

**REST** Representational State Transfer

**RDBMS** Relational Database Management System

**SQL** Structured Query Language

**SVG** Scalable Vector Graphics

**UI** User Interface

**XML** Extensible Markup Language

**XUL** XML User Interface Language

# List of Figures

[Figure 1 ELT Model](#_Toc164955512)

# List of Tables

**No table of figures entries found.**

Table of Contents

[Abstract 7](#_Toc168577815)

[Acknowledgement 8](#_Toc168577816)

[List of Abbreviations 9](#_Toc168577817)

[List of Figures 10](#_Toc168577818)

[List of Tables 11](#_Toc168577819)

[Chapter 1: Introduction 14](#_Toc168577820)

[1. Project Statement 14](#_Toc168577821)

[2. Project Scope 14](#_Toc168577822)

[3. Report Structure 14](#_Toc168577823)

[Chapter 2: Literature Review 16](#_Toc168577824)

[1. Overview of Big Data in Stock Market Analysis 16](#_Toc168577825)

[2. Previous Works on LSTM Models for Stock Prediction 16](#_Toc168577826)

[3. Importance of Real-Time Data Processing 16](#_Toc168577827)

[Chapter 3: Objectives 16](#_Toc168577828)

[1. Specific Goals of the Project 16](#_Toc168577829)

[2. Expected Contributions and Outcomes 16](#_Toc168577830)

[Chapter 4: Requirements Analysis 16](#_Toc168577831)

[1. Data Requirements 16](#_Toc168577832)

[2. Hardware and Software Requirements 16](#_Toc168577833)

[3. User Requirements 16](#_Toc168577834)

[Chapter 5: Methodology 17](#_Toc168577835)

[1. Data Collection Process (Real-Time Data Flow) 17](#_Toc168577836)

[2. LSTM Model Development 17](#_Toc168577837)

[3. Data Visualization Techniques 17](#_Toc168577838)

[Chapter 6: System Design & Implementation 17](#_Toc168577839)

[1. Architecture of the Big Data System 17](#_Toc168577840)

[2. Integration of LSTM with Real-Time Data 17](#_Toc168577841)

[3. Implementation of Data Visualization 17](#_Toc168577842)

[Chapter 7: Results and Discussions 17](#_Toc168577843)

[1. Evaluation of the LSTM Model 17](#_Toc168577844)

[2. Analysis of Real-Time Data Handling 17](#_Toc168577845)

[3. Discussion of Findings 17](#_Toc168577846)

[Chapter 8: Future Work 17](#_Toc168577847)

[1. Potential Enhancements 17](#_Toc168577848)

[2. Future Research Directions 17](#_Toc168577849)

[REFERENCES 18](#_Toc168577850)

[APPENDIX 19](#_Toc168577851)

# Chapter 1: Introduction

## Project Statement

In today’s fast-paced and ever-evolving financial markets, the ability of investors and analysts to navigate complexities and harness the power of emerging technologies is paramount for maintaining a competitive edge. The use of big data and real-time analytics platforms has become crucial for firms aiming to thrive in this environment. Big data in this context refers to the vast volume of trading data, market indicators, and global economic factors generated every minute. Effectively managing and leveraging this information through real-time platforms enables market participants to respond instantly to price fluctuations, market trends, and economic changes.

In the competitive realm of financial markets, the ability to rapidly adapt to market conditions and predict future trends with accuracy is crucial. This project introduces a real-time LSTM-based stock price prediction platform, designed to leverage the untapped potential of big data analytics in forecasting. By integrating cutting-edge real-time data processing technologies, this platform aims to transform how traders and analysts predict market movements, offering a robust tool that enhances decision-making processes and improves financial outcomes.

## Project Scope

This thesis aims to develop a real-time, LSTM-based platform for stock price prediction, leveraging the power of big data. The primary objective is to create a model that outperforms traditional forecasting methods in terms of accuracy and timeliness. Specifically, the research questions we seek to answer are:

* What are the main challenges in deploying and maintaining a real-time stock price prediction system, and how can these challenges be overcome?
* What factors influence the performance of a real-time stock price prediction platform, and how can these factors be optimized?
* How can the availability and reliability of a real-time stock price prediction platform be ensured under high workload conditions?
* What strategies can be applied to optimize real-time data processing and reduce latency in a stock price prediction system?
* What is the effectiveness of a real-time platform in predicting sudden market events, and how can the system's responsiveness to these events be improved?

## Thesis Structure

**Introduction**: This section provides an overview of the thesis, its significance, and outlines the structure of the document.

**Literature Review**: This section frames the research within the context of existing literature, focusing on big data and LSTM models in stock forecasting. It highlights the gaps and how this research aims to fill them.

**Objectives**: A detailed explanation of what the project aims to achieve and its relevance in the current financial market landscape.

**Requirements Analysis**: An expanded view of the different types of requirements, offering a thorough preparatory analysis for the project.

**Methodology**: This section details the approach to managing real-time data and implementing LSTM, ensuring clarity on the technical processes and the rationale behind the chosen methods.

**System Design and Implementation**: A new chapter that provides in-depth details on the technical architecture and practical implementation of the system, supported by diagrams and technical specifications.

**Results and Discussion**: Focuses on what was achieved with the system, including a comprehensive analysis of its performance and comparison with traditional methods.

**Future Work**: Suggests potential future enhancements and research directions, providing a roadmap for continued improvement and innovation.

# Chapter 2: Literature Review

## Overview of Big Data in Stock Market Analysis

The integration of big data and real-time analytics has transformed the landscape of stock market analysis. This chapter provides a review of existing literature, focusing on the application of big data and Long Short-Term Memory (LSTM) models in stock price prediction. It aims to identify key concepts, current research trends, and existing gaps, setting the stage for the development of a real-time LSTM-based stock price prediction platform. This literature review will cover the definitions and characteristics of big data, the importance of real-time data processing, the role of LSTM models in forecasting, and the challenges faced in this domain.

## Big Data in Stock Market Analysis

### 2.1. What is Big Data

Big data refers to extremely large datasets that are too complex to be processed and analyzed using traditional data-processing techniques. These datasets are characterized by the "Four Vs": Volume, Velocity, Variety, and Veracity.

* **Volume**: The sheer amount of data generated which can be in terabytes or petabytes.
* **Velocity**: The speed at which data is generated and processed, often in real-time.
* **Variety**: The different types of data, including structured, semi-structured, and unstructured data.
* **Veracity**: The quality and accuracy of the data, which can affect the outcomes of data analysis.

### 2.2 Differences Between Big Data Projects and Traditional Data Projects

Big data projects differ from traditional data projects in several keyways:

* **Data Sources**: Big data projects often integrate diverse data sources, including social media, transaction records, sensor data, and more, whereas traditional projects typically rely on structured data from a limited number of sources.
* **Data Processing**: Big data projects use advanced processing techniques and tools like Hadoop, Spark, and NoSQL databases to handle large volumes and high velocities of data, whereas traditional projects use relational databases and simpler processing tools.
* **Analysis Techniques**: Big data analytics employs machine learning, artificial intelligence, and advanced statistical methods to uncover patterns and insights, whereas traditional data analysis might use basic statistical tools and predefined models.
* **Scalability**: Big data solutions are designed to scale efficiently with growing data volumes, while traditional data systems may struggle with scalability issues.

### 2.3. Impact of Big Data on Stock Market Analysis

Big data has a profound impact on stock market analysis in several ways:

* **Enhanced Predictive Accuracy**: By incorporating vast amounts of historical and real-time data, big data analytics can improve the accuracy of stock price predictions. This includes data from market transactions, economic indicators, and even social media sentiment.
* **Real-Time Decision Making**: The ability to process and analyze data in real-time enables traders and analysts to make faster and more informed decisions. This can be crucial in high-frequency trading where milliseconds can make a significant difference.
* **Risk Management**: Big data analytics helps in identifying and mitigating risks by analyzing patterns and predicting potential market disruptions. This proactive approach to risk management can protect investments and enhance returns.
* **Algorithmic Trading**: Big data enables the development of sophisticated trading algorithms that can process vast amounts of data and execute trades based on predefined criteria. These algorithms can adapt to market changes and optimize trading strategies.

## Real-Time Data Processing

# Chapter 3: Objectives

## Specific Goals of the Project

## Expected Contributions and Outcomes

# Chapter 4: Requirements Analysis

## Data Requirements

## Hardware and Software Requirements

## User Requirements

A diagram of a data warehouse

Description automatically generated

Figure 1 ELT Model

# Chapter 5: Methodology

## Data Collection Process (Real-Time Data Flow)

## LSTM Model Development

## Data Visualization Techniques

# Chapter 6: System Design & Implementation

## Architecture of the Big Data System

## Integration of LSTM with Real-Time Data

## Implementation of Data Visualization

# Chapter 7: Results and Discussions

## Evaluation of the LSTM Model

## Analysis of Real-Time Data Handling

## Discussion of Findings

# Chapter 8: Future Work

## Potential Enhancements

## Future Research Directions

# REFERENCES

[1] Chand, Eswar. “Amazon Musical Instruments Reviews.” *Kaggle*, https://www.kaggle.com/datasets/eswarchandt/amazon-music-reviews. Accessed 10 June 2022.

[2] “What is ELT? Understanding the difference between ELT and ETL.” *Stitch Data*, https://www.stitchdata.com/resources/what-is-elt. Accessed 10 July 2022.

[3] *Introduction - react*, https://react.i18next.com/. Accessed 11 June 2022.

[4] “Axios tutorial - GET/POST requests in JavaScript with Axios.” *ZetCode*, 19 June 2022, https://zetcode.com/javascript/axios/. Accessed 11 June 2022.

[5] “How to Use Chart.js?” *W3Schools*, https://www.w3schools.com/ai/ai\_chartjs.asp. Accessed 10 June 2022.

[6] Bracey, Kezz. “What is Figma?” *Web Design*, 26 November 2018, https://webdesign.tutsplus.com/articles/what-is-figma--cms-32272. Accessed 18 June 2022.

# APPENDIX