

LB/TH/41/2025
TH5999

SHARE PRICE ACTION ANALYSIS USING NATURAL LANGUAGE PROCESSING

D M G C M Nalinga

219373D

Master of Science in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

April 2025

SHARE PRICE ACTION ANALYSIS USING NATURAL LANGUAGE PROCESSING

D M G C M Nalinga

219373D

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

April 2025

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

8th July 2025

Signature:

Date:

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Prof. G I U S Perera

Name of Supervisor:

Signature of the Supervisor:

Date:

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to all those who provided support to make my research on “SHARE PRICE ACTION ANALYSIS USING NATURAL LANGUAGE PROCESSING” successful.

First, I would like to express my gratitude to my project supervisor Prof. Indika Perera, Senior Lecturer, Department of Computer Science and Engineering. I am highly indebted to him for his guidance and for providing necessary information regarding the project and for his support in completing the project successfully.

I am sincerely thankful to Dr. Chathuranga Hettiarachchi, Senior Lecturer, Department of Computer Science and Engineering for the support given throughout the project period. Further, I would like to extend my gratitude to Dr. Shehan Perera for participating in meetings and providing me with very useful guidance in the beginning to make my research successful.

Finally, I wish to thank the academic and non-academic staff of the Department of Computer Science and Engineering and colleagues for the support and encouragement given.

ABSTRACT

Stock price prediction has been a widely researched topic, primarily through technical and fundamental analysis. While technical analysis relies on historical stock data and mathematical indicators, its effectiveness diminishes in illiquid stock markets such as the Colombo Stock Exchange (CSE) due to low trading volumes and irregular price movements. Fundamental analysis, on the other hand, focuses on intrinsic company value but does not fully capture short-term market reactions to external events.

This research explores an alternative approach by applying Natural Language Processing (NLP) techniques to conduct an event study analysis. The study examines how news articles influence stock price movements in the CSE by transforming textual data into numerical representations using Large Language Model (LLM)-based embeddings. The extracted feature vectors are then analysed using machine learning algorithms to identify correlations between news representation and stock price fluctuations.

By leveraging NLP-based vectorization and predictive modelling, this research provides new insights into price action analysis in illiquid stock markets, where traditional prediction methods often fail. The findings contribute to the field of financial analytics by demonstrating the feasibility of using textual data to enhance stock price forecasting in under-researched market conditions.

Keywords: Stock Market Prediction, Natural Language Processing, Event Study Analysis, Colombo Stock Exchange, Machine Learning

TABLE OF CONTENT

| | |
|--|-----|
| DECLARATION | i |
| ACKNOWLEDGEMENTS | ii |
| ABSTRACT | iii |
| LIST OF FIGURES | vii |
| LIST OF TABLES | ix |
| 1 INTRODUCTION | 1 |
| 1.1 Background and Motivation | 1 |
| 1.2 Problem Statement | 1 |
| 1.3 Research Challenges | 2 |
| 1.4 Research Objectives..... | 2 |
| 1.5 Significance of the Study | 3 |
| 1.6 Structure of the Thesis | 3 |
| 2 LITERATURE REVIEW | 5 |
| 2.1 Introduction..... | 5 |
| 2.2 Stock Price Prediction Methods (General Background)..... | 5 |
| 2.3 Event Study Analysis..... | 10 |
| 2.4 Stock Price Analysis using Natural Language Processing (NLP) | 12 |
| 2.5 Text Vectorization and Natural Language Representation..... | 17 |
| 2.5.1 Bag of Words (BoW)..... | 17 |
| 2.5.2 Term Frequency-Inverse Document Frequency (TF-IDF) | 17 |
| 2.5.3 Doc2Vec | 18 |
| 2.5.4 Global Vectors for Word Representation (GloVe) | 18 |
| 2.5.5 Large Language Models in Vectorization..... | 19 |
| 2.6 Machine Learning Models for Stock Price Action Analysis | 21 |
| 2.7 Graph Neural Networks for Stock Price Analysis Problem..... | 23 |
| 2.8 Research Gap and Contribution..... | 25 |
| 3 RESEARCH METHODOLOGY | 28 |
| 3.1 Introduction to Methodology | 28 |
| 3.2 Data Collection | 28 |
| 3.3 Data Preprocessing..... | 29 |
| 3.3.1 Preprocessing of News Articles | 29 |

| | | |
|-------|--|----|
| 3.3.2 | Preprocessing of Stock Trade Data..... | 30 |
| 3.4 | Feature Engineering..... | 31 |
| 3.4.1 | Feature Engineering and Vectorization of News Articles..... | 31 |
| 3.4.2 | Feature Engineering of Trade Data and Selection Rationale..... | 37 |
| 3.4.3 | Dimensionality Reduction..... | 37 |
| 3.5 | Machine Learning Model - Graph Neural Network..... | 39 |
| 3.5.1 | Graph Neural Network (GNN) Architecture..... | 39 |
| 3.5.2 | Graph Construction..... | 40 |
| 3.5.3 | Hyperparameter Tuning..... | 41 |
| 3.5.4 | Training Procedure..... | 42 |
| 3.5.5 | Evaluation Metrics..... | 42 |
| 3.5.6 | Baseline Models and Architecture..... | 42 |
| 3.5.7 | Implementation Tools and Libraries..... | 44 |
| 3.5.8 | Limitations and Assumptions..... | 44 |
| 4 | IMPLEMENTATION..... | 46 |
| 4.1 | Training and Evaluation Setup..... | 46 |
| 4.1.1 | Data Splitting and Scaling..... | 46 |
| 4.1.2 | Model Training..... | 46 |
| 4.2 | Metric Calculation..... | 46 |
| 4.3 | Results Presentation..... | 47 |
| 4.3.1 | Result Presentation for Unseen Data..... | 57 |
| 5 | RESEARCH EVALUATION..... | 60 |
| 5.1 | Analysis and Discussion..... | 60 |
| 5.1.1 | Performance Evaluation of the Proposed Model..... | 60 |
| 5.1.2 | Robustness on Unseen Data..... | 61 |
| 5.1.3 | Technical Strength and Innovation..... | 61 |
| 5.1.4 | Comparative Limitations of Traditional Models..... | 62 |
| 5.1.5 | Contributions and Practical Implications..... | 62 |
| 5.1.6 | Limitations and Future Directions..... | 63 |
| 6 | RESEARCH CONCLUSION..... | 65 |
| 6.1 | Research Summary..... | 65 |
| 6.2 | Methodological Contribution..... | 65 |
| 6.3 | Achievement of Research Objectives..... | 66 |

| | | |
|-----|---|----|
| 6.4 | Practical Value and the Limitations | 67 |
| 6.5 | Future Work | 68 |
| 7 | REFERENCES | 69 |

LIST OF FIGURES

| | |
|---|----|
| Figure 2-1 TF-IDF Equation..... | 17 |
| Figure 2-2 Predicting the words based on their surrounding context within a document, along with a document ID..... | 18 |
| Figure 3-1 Flowchart Representing Financial News Preprocessing Workflow..... | 30 |
| Figure 3-2 Flowchart Representing Stock Price Data Preprocessing Workflow..... | 31 |
| Figure 3-3 TF-IDF Embeddings (UMAP)..... | 33 |
| Figure 3-4 Doc2Vec Embeddings (UMAP)..... | 33 |
| Figure 3-5 SBERT Embeddings (UMAP)..... | 34 |
| Figure 3-6 FinGPT Embeddings (UMAP)..... | 34 |
| Figure 3-7 Graph Construction Visualization..... | 41 |
| Figure 4-1 Training vs Validation Loss during Residual MLP Model Training - HNB | 47 |
| Figure 4-2 Actual vs Predicted over Time with Residual MLP Model Testing - HNB | 48 |
| Figure 4-3 Training vs Validation Loss during BiLSTM Model Training - HNB | 49 |
| Figure 4-4 Actual vs Predicted over Time with BiLSTM Model Testing - HNB | 49 |
| Figure 4-5 Training vs Validation Loss during Proposed Model Training - HNB..... | 50 |
| Figure 4-6 Training vs Validation Loss during Proposed GNN Model Training - HNB | 50 |
| Figure 4-7 Training vs Validation Loss during Residual MLP Model Training - JKH | 51 |
| Figure 4-8 Actual vs Predicted over Time with Residual MLP Model Testing - JKH | 51 |
| Figure 4-9 Training vs Validation Loss during BiLSTM Model Training - JKH | 52 |
| Figure 4-10 Actual vs Predicted over Time with BiLSTM Model Testing - JKH | 52 |
| Figure 4-11 Training vs Validation Loss during Proposed Model Training - JKH..... | 53 |
| Figure 4-12 Actual vs Predicted over Time with Proposed Model Testing - JKH..... | 53 |
| Figure 4-13 Actual vs Predicted over Time with Residual MLP Model Testing - BIL | 54 |
| Figure 4-14 Actual vs Predicted over Time with Residual MLP Model Testing - BIL | 55 |

| | |
|--|----|
| Figure 4-15 Actual vs Predicted over Time with BiLSTM Model Testing - BIL | 55 |
| Figure 4-16 Actual vs Predicted over Time with BiLSTM Model Testing - BIL | 56 |
| Figure 4-17 Actual vs Predicted over Time with Proposed Model Testing - BIL..... | 56 |
| Figure 4-18 Actual vs Predicted over Time with Proposed Model Testing - BIL..... | 57 |
| Figure 4-19 Actual vs Predicted over Time with Residual MLP Model for Unseen Data - HNB | 57 |
| Figure 4-20 Actual vs Predicted over Time with BiLSTM Model for Unseen Data - HNB... | 58 |
| Figure 4-21 Actual vs Predicted over Time with Proposed Model for Unseen Data - HNB .. | 58 |

LIST OF TABLES

| | |
|---|----|
| Table 3-1 LLMs used for Feature Extraction Explanation | 36 |
| Table 3-2 Feature Selection and Reduction by LLM..... | 39 |
| Table 4-1 Model Performance Comparison Using Evaluation Metrics - HNB..... | 47 |
| Table 4-2 Model Performance Comparison Using Evaluation Metrics – JKH | 50 |
| Table 4-3 Model Performance Comparison Using Evaluation Metrics - BIL..... | 54 |
| Table 4-4 Proposed Model vs Baseline Models Performance Comparison Using Evaluation Metrics for Unseen Data - HNB | 58 |