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

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background and Motivation

Stock price prediction has long been a crucial area of research in financial markets, with investors relying on technical and fundamental analysis to make informed decisions. Technical analysis utilizes historical stock prices and mathematical indicators to predict future movements, while fundamental analysis assesses a company's intrinsic value based on financial statements, industry trends, and macroeconomic conditions. However, these methods have limitations, particularly in illiquid markets like the Colombo Stock Exchange (CSE), where low trading volumes and irregular price movements reduce the reliability of traditional approaches.

In such markets, event study analysis which examines how specific external events (e.g., news announcements, policy changes, political involvements) influence stock prices can provide deeper insights into market behaviour. The increasing availability of textual data from financial news presents an opportunity to enhance stock price analysis by leveraging Natural Language Processing (NLP) and Machine Learning (ML) techniques.

1.2 Problem Statement

Traditional stock price prediction models often fail to capture the immediate market response to news, especially in illiquid or emerging markets like the Colombo Stock Exchange (CSE). Approaches such as technical analysis, which rely on historical price patterns and technical indicators, lack sensitivity to real-time information flows or sentiment changes. Similarly, fundamental analysis focuses on a company's intrinsic value over the long term and is not well-suited for detecting short-term fluctuations driven by external events, rumours, or news sentiment.

This gap is particularly critical in the CSE, where market reactions to news can be delayed, sentiment-driven, and sparse due to limited liquidity and fewer institutional actors. Despite the growing relevance of financial news and unstructured data in influencing price movements, there has been limited research on integrating textual analysis with short-term stock prediction models within this context.

Moreover, traditional textual representation techniques such as TF-IDF and Doc2Vec fall short in capturing contextual and domain-specific semantics in financial language. This limits their effectiveness in modelling the nuanced relationship between news and stock prices. With recent advances in financial-domain language models (e.g., FinBERT, FinGPT) and semantic encoders like SBERT, there is an opportunity to represent financial news in a way that better reflects its latent meaning and predictive power.

Despite this, most existing models treat text and price data independently or use linear associations, ignoring the complex temporal and relational structure between news, sentiment, and price. There remains a lack of research that applies graph-based

neural architectures to this problem—particularly those that integrate semantic similarity, sentiment polarity, and temporal dependencies into a unified predictive model.

1.3 Research Challenges

Stock price action analysis in the Colombo Stock Exchange (CSE) presents several unique challenges, particularly due to its illiquid nature and limited digital resources. One primary issue is the scarcity of online financial news sources, making it difficult to gather sufficient textual data for analysis. Even when news articles are available, preprocessing becomes complex due to the need of filtering out irrelevant content, such as HTML tags and CSS content, before extracting meaningful financial insights.

Another significant challenge is data incompleteness, as both stock price data and daily news articles often contain missing values in irregular patterns, requiring robust handling methods. Applying unsupervised learning techniques for data filtration is essential to separate relevant financial information from noise, yet this remains a difficult task due to the unstructured nature of textual data. Vectorization of financial news also poses a challenge, as most general-purpose NLP models create broad representations rather than capturing domain-specific financial contexts, which can impact model accuracy.

Furthermore, price action stays unchanged over a considerable time making it complicated for model building. Finally, unpredictable market behaviours where stock price fluctuations do not always follow discernible patterns, or identifiable causes make it challenging to develop reliable predictive models. Addressing these issues requires a specialized approach combining NLP, machine learning, and robust data preprocessing techniques to improve the accuracy and effectiveness of stock price action analysis in the CSE.

1.4 Research Objectives

This research aims to achieve the following objectives,

- Develop an NLP-based Analytical Framework: Establish a robust Natural Language Processing (NLP)-based framework to systematically analyze and quantify the impact of daily news articles on stock price movements within the Colombo Stock Exchange (CSE), an illiquid market.

- Evaluate Text Vectorization Techniques: Conduct an extensive comparative analysis of text vectorization methods, including TF-IDF, Doc2Vec, and advanced Large Language Model (LLM)-based embeddings, specifically SBERT, FinGPT, and FinBERT, to effectively extract meaningful, context-rich features from financial news data.

- Integrate News Features with Stock Price Movements Using Deep Learning: Apply sophisticated deep learning methodologies, particularly Graph Neural Networks (GNN), to elucidate and model complex relationships between news-derived sentiment, semantic context, and stock price dynamics.

- Assess Model Efficacy in Illiquid Market Contexts: Evaluate the robustness, accuracy, and generalizability of the proposed predictive framework within the unique constraints and challenges posed by illiquid stock markets.

- Generalize Framework for Broader Applicability: Create a generalized analytical framework capable of being adapted to other stock indices, especially those in markets with limited liquidity and sparse prior research.

1.5 Significance of the Study

This research introduces a novel methodology for stock price analysis in illiquid markets, where conventional approaches often fail to provide accurate predictions. By integrating NLP-driven event study analysis, this study enhances stock market forecasting through non-company specific news analysis, offering a new perspective on market reactions to public events in illiquid stock exchanges. Additionally, it will be the foundation for further research in NLP-based financial analytics, contributing to the development of more effective tools for understanding stock price movements in markets with limited liquidity.

1.6 Structure of the Thesis

This thesis is organized into Seven main chapters. Following the comprehensive Introduction that outlines the background, significance, and objectives of the research, Chapter two presents the Literature Review, which begins by introducing core concepts in stock price prediction and event study analysis before examining more advanced topics such as natural language processing (NLP) for financial text and various machine learning techniques. It concludes by identifying the key research gaps and contributions of this study.

Next, Chapter three, titled Research Methodology, explains the plan of action, including the methods of data collection, the preprocessing of both news articles and stock trade data, and the feature engineering steps required to transform raw information into meaningful inputs. A significant portion of this chapter focuses on the Graph Neural Network (GNN) architecture, describing in detail how it is constructed, the hyperparameters it uses, how it is trained, and how it will be evaluated. Baseline models are also introduced here for comparison, and the tools, libraries, assumptions, and limitations relevant to the methodology are discussed.

Building on this, Chapter four, Implementation, demonstrates how the methodology was put into practice. It begins by outlining the training and evaluation

setup including data splitting and scaling routines and the practical process of model training before moving on to the calculation of performance metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). The results of each model are then presented by an analysis and discussion that interprets the findings in the context of the research goals and highlights any notable trends or anomalies in Chapter five, Research Evaluation.

Finally, the thesis concludes with Chapter six the Conclusion, and Chapter seven the References which are all the sources referred throughout this study.

2 LITERATURE REVIEW

2.1 Introduction

The field of stock price prediction has traditionally relied on technical and fundamental analysis, with varying degrees of success across different market conditions. However, in illiquid markets such as the Colombo Stock Exchange (CSE), where low trading volumes and irregular price movements limit the effectiveness of conventional forecasting methods, alternative approaches are necessary. This literature review explores the integration of Natural Language Processing (NLP) techniques with machine learning (ML) algorithms to analyse the impact of financial news on stock price movements, particularly through event study analysis.

A thorough review of existing literature is essential to understand the strengths and limitations of previous approaches to stock price prediction. By examining the prior research, this study identifies gaps in conventional stock prediction methods and explores how NLP and ML can enhance price action analysis for an illiquid stock market through event study analysis. Literature review will be discussed through the topics as mentioned below,

- Stock Price Prediction Methods (General Background); Analysing traditional approaches, including technical and fundamental analysis, and their limitations in illiquid markets.
- Event Study Analysis; Exploring how external events such as news announcements and policy changes influence stock prices.
- NLP in Stock Price Analysis Research problem; Reviewing different text vectorization techniques (e.g., TF-IDF, Doc2Vec, LLM embeddings) and their role in relevancy, semantic, and sentiment analysis and stock forecasting.
- Text Vectorization and Natural Language Representation
- Machine Learning in Stock Market Prediction; Investigating ML models used in financial forecasting, including regression models, neural networks, and combined learning methods.
- Research Gap and Contribution

2.2 Stock Price Prediction Methods (General Background)

Stock market price predictions often rely on technical indicators, particularly in highly liquid markets, due to their frequent trading activity providing ample data for predictive analysis. A substantial number of studies have investigated the efficacy of technical indicators combined with machine learning and deep learning techniques. Technical analysis generally focuses on historical price data, including opening prices, closing prices, highs, lows, and trading volume, to detect market trends, momentum, and possible price reversals. Common technical indicators fall into distinct categories and those will be categorized as below,

- Trend Indicators:
  - Moving Averages (MA)
  - Moving Average Convergence Divergence (MACD)
  - Average Directional Index (ADX)
- Momentum Indicators:
  - Relative Strength Index (RSI)
  - Stochastic Oscillator
- Volatility Indicators:
  - Bollinger Bands
  - Average True Range (ATR)
- Other Indicators:
  - Trading Volume
  - Fibonacci Retracement

Using these indicators with considered price index movements will be modelled using machine learning approaches to forecast the future price movements. [1] In this paper, 10 regressors and over 100 technical indicators have been examined on data of the last 13 years of Apple Company and the results have been investigated by error-based evaluation criteria. They have concluded that choosing the right indicator or indicators would have a particular impact on the accuracy of the model. The study has used the AAPL company's shares and was collected from Yahoo Finance. For the feature selection study has used Sequential Forwards Selection (SFS) and Sequential Backwards Selection (SBS). For model building study has experimented with Linear regression, Ridge regression, Lasso regression, Decision Tree regression, K-Nearest Neighbour regression, Multilayer Perceptron regression, Support Vector Regression, AdaBoost Regression, Gradient Boosting Regression and Random Forest Regression and found out the MLP Regression using the Sequential Forwards Selection and the gave the best performance.

In addition to the technical indicators and stock price features, other studies [2] add Google Trends data as a feature to make stock predictions. Using the Google service they have obtained a rough estimate of how many people are talking about a topic at any given moment. They have assumed before buying and selling shares, investors will look for more information about the stock market as decision support, affecting the increase in search volume on Google. The study analysed the predictability of stock prices using historical data from the Indonesian Stock Exchange. Six stocks namely, BBCA (Bank Central Asia), HMSP (PT Hanjaya Mandala Sampoerna), TLKM (PT Telkom Indonesia Tbk), BBRI (PT Bank Rakyat Indonesia Tbk), ASII (Astra International Tbk), and UNVR (Unilever Indonesia Tbk and were randomly selected from the top ten performing stocks as of February 2022. Historical price data for these stocks was sourced from Yahoo Finance, while search trend data was collected from Google Trends. To predict future stock prices, the researchers applied several machine learning techniques, including Support Vector Regression (SVR), Multilayer Perceptron (MLP), and Multiple Linear Regression. The performance of these models

was evaluated using the Mean Absolute Percentage Error (MAPE) metric. Among the tested methods, the SVR model demonstrated the strongest predictive performance, yielding predictions closely aligned with actual market prices during the testing phase. This study highlights the effectiveness of combining historical stock data and internet search trends as predictive indicators in stock market analysis.

[3] Teixeira and Barbosa (2025) conducted a comprehensive analysis to investigate the efficacy of various deep learning and machine learning techniques in predicting stock prices. They extensively evaluated models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), and the gradient boosting framework XGBoost. Their research utilized a diverse dataset comprising multiple global market indices alongside key technical indicators, primarily focusing on moving averages. The central objective of the study was to identify complex, nonlinear patterns and temporal dependencies within stock market data that traditional forecasting models often fail to capture effectively. The findings indicated that deep learning methods, notably LSTM and GRU, demonstrated superior predictive accuracy due to their capability to capture sequential and temporal relationships within the financial time series data. Furthermore, models like CNN and XGBoost provided robust results by effectively extracting features and handling data dimensionality. The researchers concluded that integrating deep learning models significantly enhances forecasting precision over conventional statistical approaches, providing investors and market analysts with more reliable predictive tools. This study contributes substantially to the ongoing discourse on employing advanced neural network architectures to enhance predictive accuracy in stock market forecasting.

A complementary study [4] published in Expert Systems with Applications (2025) explored the efficacy of various advanced machine learning models including XGBoost, Random Forest, Support Vector Regression (SVR), and LSTM in determining the most influential technical indicators for stock market forecasting. The study aimed to systematically evaluate the impact of integrating traditional technical indicators with sophisticated predictive algorithms. Researchers identified that by strategically combining indicators such as moving averages, RSI, MACD, and trading volume data, predictive accuracy markedly improved across all tested algorithms. Notably, XGBoost and Random Forest algorithms were particularly effective due to their powerful ensemble methodologies and robust feature-selection capabilities. The LSTM model also provided promising results by capturing temporal dependencies inherent in market data. The key conclusion drawn from this study was that leveraging a well-integrated set of technical indicators alongside powerful machine learning methods significantly enhances prediction performance, highlighting the practical benefit of hybrid analytical frameworks for informed investment decisions.

[5] This study conducted targeted research on predicting stock price trends within the context of an emerging market the Vietnamese stock market using Long Short-Term Memory (LSTM) algorithms. The authors specifically chose technical indicators including the simple moving average (SMA), moving average convergence divergence (MACD), and relative strength index (RSI) as input features. Their model

aimed to exploit LSTM's strength in modelling sequential financial data by effectively capturing historical trends and short-term fluctuations. The study achieved a notable predictive accuracy rate of approximately 93%, underscoring the suitability of LSTM networks for analysing sequential dependencies prevalent in emerging stock markets. This remarkable accuracy highlights the potential applicability of deep learning methodologies in emerging economies, providing strong evidence of LSTM's robustness in handling volatile market conditions. Tran et al. concluded that LSTM models, complemented by carefully chosen technical indicators, are highly effective tools for predicting market trends and assisting investors in decision-making processes in less stable and developing financial markets.

Another recent advancement [6] in stock market prediction methodologies is introduced in a study published by EPJ Data Science (2024). This research proposed an innovative hybrid predictive framework that incorporates both periodic (seasonal or cyclical data) and non-periodic features (such as news-driven market influences and unpredictable economic events). The authors employed robust algorithms including Random Forest, Support Vector Machines (SVM), and Gradient Boosting Decision Trees (GBDT), assessing their capacity to exploit diverse feature sets to enhance prediction performance. Their hybrid methodology showed significant improvement over single-source predictive models, demonstrating how combining various feature types strengthens predictive capabilities, particularly when addressing extensive and volatile datasets. By integrating periodic features, which capture cyclical market behaviours, with non-periodic elements reflecting instantaneous market changes, the proposed model achieved greater accuracy and reliability. The researchers concluded that a multi-faceted feature engineering approach substantially improves stock prediction outcomes, particularly in challenging market environments characterized by frequent volatility. This study contributes notably by advocating the integration of diverse data sources and hybrid methodologies for more comprehensive financial forecasting.

Fundamental analysis involves examining a company's financial statements, such as the income statement, balance sheet, and cash flow statement, to assess its financial health and performance. The most used financial indicators can be listed as below,

- Price-to-Earnings (P/E) Ratio: Widely cited in research as a primary indicator of valuation, often linked to future stock price movements.
- Price-to-Book (P/B) Ratio: Frequently used as an indicator of company valuation and stability, especially in value investing research.
- Return on Equity (ROE): Commonly analysed in academic work because it effectively measures profitability and management efficiency.
- Earnings per Share (EPS): Regularly used in studies as it directly reflects profitability trends, which are critical for prediction tasks.
- Debt-to-Equity (D/E) Ratio: Often included in research to assess financial stability, leverage risk, and vulnerability to economic fluctuations.

These are used to evaluate a company's liquidity, leverage, profitability, and valuation. The outcome of using these assessments is to determine the intrinsic value of a stock. This value indicates what a particular stock is "truly" worth, and they believe that the market will eventually realize this intrinsic value and adjust the stock price accordingly. This method, however, involves processing huge amounts of data and is subject to individual interpretation. Thus, even if the data suggests a favourable movement in the stock, it may take a long time for the rest of the market to realize or catch up. This method, thus, is suitable for long-term predictions and growth.

[7] Noel (2023) examined the role of fundamental analysis in predicting long-term stock market behaviour, highlighting its distinct advantages despite some inherent challenges. The research emphasized that fundamental analysis, which involves a meticulous evaluation of company-specific financial information such as earnings, valuation metrics, and debt ratios, can effectively identify intrinsic stock values. Unlike technical analysis, which primarily focuses on short-term price trends and patterns, fundamental analysis offers deeper insights into the financial health and sustainable performance of companies over extended periods. Although this method necessitates substantial historical financial data and is subject to interpretation and subjective judgment by analysts, Noel argues its utility lies in the identification of stocks whose market prices diverge significantly from their intrinsic values. The core contribution of this research is the recognition that the market eventually corrects itself by aligning stock prices with their intrinsic valuations, thus providing an opportunity for informed, strategic investments. In conclusion, Noel suggests that, despite its complexity and subjective elements, fundamental analysis remains invaluable for long-term investment strategies and for predicting gradual adjustments toward the intrinsic value of stocks.

Phan and Chang [8] investigated the effectiveness of integrating traditional fundamental analysis with advanced machine learning methods to improve the accuracy of stock trend predictions. They employed comprehensive financial indicators such as financial ratios (Price-to-Earnings, Debt-to-Equity, and Earnings per Share) along with the Discounted Cash Flow (DCF) model to assess a company's potential profitability and financial strength. By combining these traditionally robust financial metrics with predictive machine learning algorithms including Logistic Regression, Long Short-Term Memory (LSTM), and one-dimensional Convolutional Neural Networks (1D CNN) the researchers evaluated which methodology offered superior predictive capabilities. The study found that Logistic Regression, a comparatively simpler and computationally efficient model, outperformed both CNN and LSTM models, achieving average accuracies of 74.66% and 72.85% across the prediction tasks. This finding is particularly significant because it demonstrates that simpler, transparent models can effectively leverage fundamental financial information to produce robust predictions. Phan and Chang’s study notably contributes by validating the practicality and effectiveness of merging fundamental analysis with accessible machine learning techniques, thus providing valuable insights to both academia and investment professionals who prioritize clear interpretability and high predictive accuracy in their models.

The study [9] conducted extensive research exploring the integration of fundamental financial analysis with advanced machine learning algorithms to predict stock market performance. Their comprehensive study utilized financial data spanning 22 years, incorporating essential financial ratios and indicators as predictive inputs. Specifically, they applied three distinct machine learning approaches: Random Forest (RF), Multilayer Perceptron (MLP), and Adaptive Neural Fuzzy Inference Systems (ANFIS). Recognizing the complexity and high dimensionality of financial data, the researchers emphasized feature selection to enhance model performance, demonstrating that eliminating irrelevant or redundant financial indicators significantly improved predictive accuracy. Their findings revealed that among the individual algorithms tested, the Random Forest model consistently outperformed both the MLP and ANFIS models due to its ability to handle nonlinearities, interactions, and feature importance ranking effectively. Furthermore, Huang et al. introduced an innovative aggregated model, which combined the predictive strengths of these algorithms. This ensemble approach substantially outperformed all individual baseline models and traditional market indices, notably surpassing the predictive accuracy of the Dow Jones Industrial Average (DJIA) during the evaluation period. Their study contributes profoundly to stock prediction literature by highlighting the advantages of feature selection techniques combined with machine learning models. It underscores the practical value of machine learning-driven fundamental analysis in enhancing investment decisions and portfolio performance in realistic financial market settings.

2.3 Event Study Analysis

Event studies are statistical analyses used to assess the impact of specific events on the value of a firm. By examining stock price movements around the time of an event, researchers can infer the event's economic significance. This methodology is widely applied to evaluate the effects of corporate announcements, regulatory changes, and macroeconomic policies on stock performance.

The Event Study Analysis in other words how market events impact stock prices, is useful where an unmatured stock exchange like CSE having a smaller number of transactions per given time and almost unfeasible to predict stock price actions with historical data and mathematical methodologies. Such markets are volatile to external factors such as stability of the country, inflation, foreign currency rates/investments, local currency rate, natural disasters, political involvements/movements, governing rules and policies, etc.

[10] The study shows that understanding these factors is essential for investors, financial analysts, and policymakers who aim to navigate the complexities of the market and make informed decisions. It has explained the external impacts by Economic Indicators such as a growing Gross Domestic Product (GDP) typically which signals a healthy economy, lower interest rates which reduce borrowing costs for companies, boosting investment and stock prices, high employment levels which indicate economic strength and consumer spending power. Moreover, it explains the understanding of market sentiment like positive news about a company or the economy can boost

investor confidence and drive stock prices up, while negative news can have the opposite effect. Also, geopolitical events such as elections and changes in government policies can affect market confidence and investor sentiment. Tariffs, trade agreements, and international trade disputes can influence corporate profits and stock prices. Finally, External Shocks such as earthquakes, hurricanes, and other natural disasters can disrupt economic activity and affect stock markets health crises, such as the COVID-19 pandemic, can cause widespread economic disruption and significant stock market fluctuations [11].

For instance, [12] a study on the Chilean stock market, characterized by lower liquidity, utilized an event study analysis to investigate the serial correlations of illiquid stocks' price changes. The research highlighted how event studies could effectively capture the unique behaviours of stock prices in illiquid markets, providing valuable information on how specific events influence stock performance under such conditions. Conclusively, event study analysis serves as a valuable method for examining stock price actions in illiquid markets, enabling researchers and investors to understand the effects of particular events on stock performance when traditional analysis methods may be less effective due to limited trading activity.

In 2019 [13] investigated how news events, such as company or macroeconomic announcements, contribute to the pre- and post-event jump dynamics of stock prices. They introduced a non-parametric framework to statistically examine these effects, considering intraday seasonality of news and jumps. Utilizing data from the S&P 500 index ETF (SPY) and Nasdaq Nordic Large-Cap stocks, the study provided strong evidence that non-scheduled company announcements and certain macroeconomic announcements lead to significant jumps following the releases. Additionally, some pre-jumps preceding scheduled information releases suggested potential non-gradual information leakage. The study concluded that unexpected information releases have a pronounced impact on stock price volatility, emphasizing the need for investors to account for such events in their trading strategies.

Budennyy et al. [14] examined the influence of clinical trial announcements on the stock prices of pharmaceutical companies. They developed a predictive pipeline incorporating a Bidirectional Encoder Representations from Transformers (BERT) - based model for sentiment analysis of announcements, a Temporal Fusion Transformer for forecasting expected returns, a graph convolution network for capturing event relationships, and gradient boosting for predicting price changes. Analysing a dataset of 5,436 clinical trial announcements from 681 companies over five years, the study found that negative announcements had a more substantial impact on stock prices than positive ones. The research highlighted the importance of drug portfolio size and network effects of related events in determining stock price reactions. The study concluded that integrating advanced machine learning models can effectively predict stock market reactions to clinical trial outcomes, aiding investors in making informed decisions.

[15] investigated the occurrence of extreme events (EEs) in stock prices caused by fundamental parameters, technical setups, and external factors. They employed the Hilbert-Huang transformation (HHT) to identify EEs based on high instantaneous

energy concentrations. Additionally, support vector regression was used to predict stock prices during EEs, with close price data providing the most accurate inputs. The study achieved prediction accuracies of up to 95.98% for one-step forecasts and 94.09% for two-step forecasts. The research concluded that monitoring factors leading to EEs is crucial for developing effective entry or exit strategies, as these events can result in significant capital gains or losses for investors.

Moreover in [16] authors analysed the interaction between stock prices of major companies in the USA and Germany during crises, using Granger Causality and recurrence analysis. They proposed that increased pair-wise Granger causality interactions during crises result from simultaneous market responses to external stimuli, rather than actual causal relationships between stock prices. The study modelled stock price patterns by incorporating an exogenous term representing external factors into the geometric Brownian motion model. The research concluded that significant crises, such as the 2007/2008 mortgage crisis and the COVID-19 outbreak, induce collective market behaviours driven by external events, highlighting the importance of considering external stimuli in financial modelling during crises.

These studies provide valuable insights into how various external events, including news announcements, political developments, and crises, influence stock prices. Integrating advanced analytical methods and machine learning techniques enhances the understanding and prediction of stock market reactions to such events, offering practical implications for investors and policymakers.

2.4 Stock Price Analysis using Natural Language Processing (NLP)

Recent stock price prediction using the NLP research area varies considerably. Differences exist in the computational models employed, the stock market data analysed, and the text data sources used, which can range from financial news online to social media discussions. Consequently, the best predictive model for a given study is highly dependent on the specific goals and resources available, allowing for continuous improvement of current models. Furthermore, datasets vary based on the time period and geographic location they cover. These factors significantly impact the research context and initial parameters, creating opportunities to explore diverse datasets.

[17] This study has created the predictive model by integrating a sentiment analysis module on Twitter data to correlate the public sentiment of stock prices with the market sentiment. Authors conducted predictive analysis on stock market data, specifically focusing on India's NIFTY 50 index. The researchers obtained daily stock price data spanning from January 2, 2015, to June 28, 2019, sourced from Yahoo Finance. Using this historical price data, they generated six derived variables to enhance predictive modelling. Among these was “close_norm”, a numeric variable representing the standardized value of the percentage change in the closing prices between consecutive trading days. Furthermore, the researchers expanded their predictive framework by incorporating Twitter-based sentiment analysis, leveraging social media sentiment data associated with specific stocks during the studied period. This hybrid

approach, combining traditional financial indicators with sentiment analytics, aimed to improve the accuracy and reliability of stock price prediction models.

That was done using previous week closing values to predict stock price movement for the next week. This work has been carried out as an extra work to the related methodology and mainly has focused on several approaches to stock price and movement prediction on a weekly forecast using eight regression and eight classification methods using historical price data.

Back in 2016 Wang and Wang [18] explored the use of NLP techniques to enhance the prediction of stock market movements by examining relationships between textual information in news headlines and stock prices. The authors developed and compared several NLP-based models, including convolutional neural networks (CNN), recurrent neural networks (RNN), and transformer-based approaches. They integrated sentiment scores extracted from these models with historical stock prices to train predictive algorithms. The findings revealed a substantial correlation between news sentiment and stock price variations, with the most effective predictions achieved by the Gated Recurrent Unit (GRU) model combined with historical stock prices and sentiment indicators. The research significantly contributed to stock forecasting by validating the efficacy of incorporating news sentiment analysis into predictive models, ultimately concluding that NLP-derived sentiment combined with historical data greatly enhances stock price prediction accuracy.

[19] the same authors used data from “Sina Weibo”, China’s largest and most widely used social media site and the Support Vector Machine (SVM) algorithm for stock price prediction and proved that stock price action has a correlation with the social sentiment. In their 2016 research, they explored how social media mining technologies can enhance stock price predictions by leveraging market sentiment data. They utilized textual data from social media platforms primarily microblogs and online forums as their data source. They implemented sentiment analysis techniques to interpret investor attitudes and opinions expressed online. Machine learning methods, including Support Vector Machines (SVM) and neural networks, were employed to integrate this sentiment information with traditional financial indicators. Their study concluded that models combining sentiment from social media with conventional market data substantially outperformed predictions made by relying solely on traditional financial data, thus validating the effectiveness of integrating natural language processing with financial modelling.

[20] The stock market datasets are collected from reddit news headlines (Top 100 online news of each day) and the Dow Jones Industrial Average (DJIA). Extracted news sentiments (positive, negative and neutral) were trained with DJIA values on multiple models including Random Forest, AdaBoost, Gradient Boosting, XGBoost and A Multilayer Perceptron (MLP) neural network. Proposed work shows that MLP Classifier has provided the most accurate trend prediction suggesting that MLP is the optimal algorithm for highly volatile and non-static data. Conclusively authors found that,

- Neural Network (MLP) (Accuracy: 81.216%)
  The Multilayer Perceptron (MLP) is a deep learning approach that utilizes multiple layers of interconnected neurons to learn complex patterns within data. Due to its ability to model non-linear relationships effectively, the MLP achieved the highest accuracy among the tested algorithms. This makes it especially suitable for predicting intricate behaviours like stock market movements.

- XGBoost (Accuracy: 70.899%)
  XGBoost (Extreme Gradient Boosting) is an advanced tree-based ensemble method that incrementally builds strong predictive models by optimizing weak decision trees. Its robust regularization capabilities and efficient handling of missing values and outliers contribute to reliable predictions, placing it second-best in performance.

- Random Forest (Accuracy: 65.608%)
  Random Forest is a widely used ensemble learning algorithm combining multiple decision trees trained on random subsets of data. It reduces overfitting through averaging results from individual trees, offering solid accuracy and stability.

In the 2021 [21] study, investigated stock price prediction for "Meinian Health" a Chinese company. Authors mention that they selected this company because it ranks second among all listed medical companies in total profit and its investors' comments and official news are very active and at the same time, "Meinian Health" went public in 2005, so they could collect all documents after 2010. They proposed a new stock price prediction model (Doc-W-LSTM) based on deep learning technology, which integrates Doc2Vec, Stacked Auto Encoder (SAE) to reduce the dimension of text vectors to avoid a serious imbalance between text features, wavelet transform to generate denoised stock price timeseries data and Long Short-Term Memory (LSTM) model. They develop this approach by incorporating textual data from investor comments and company news sources alongside traditional stock financial indices. Finally, authors have proven, extracted sentiment data along with traditional analysis has outperformed the baseline models and given MAE = 0.019, RMSE = 0.110, R2 = 0.957.

In this study [22], stock price predictions were made using multiple modelling approaches, including time series techniques (ARIMA and Facebook Prophet), neural networks (RNNs), and a hybrid approach combining neural networks with financial news sentiment analysis. The authors observed a significant correlation between financial news articles and stock price movements, with Recurrent Neural Networks (RNNs) yielding the best predictive outcomes. Specifically, they employed an RNN-LSTM model trained on historical stock prices along with sentiment polarity scores extracted from news articles related to S&P500 (large-cap companies trading on the American stock exchanges) companies. Textual polarity was obtained using the Natural Language Toolkit (NLTK), considering only positive and negative sentiments. However, model performance notably declined during periods characterized by low

stock prices or high volatility. They tried the approach with Apple, Facebook, American Airlines, Amazon and Microsoft stocks data in their related methodology.

In 2021, [23] research proposed a cutting-edge method for predicting stock prices. This approach combined the power of Natural Language Processing (NLP) with advanced machine learning techniques. They utilized a specialized version of Google's Bidirectional Encoder Representations from Transformers (BERT), a machine learning algorithm developed by Google for the NLP model, called FinBERT, which was specifically trained on financial news articles. FinBERT was used to analyse news sentiment of the news and the headlines for the company Apple Inc, listed on the National Association of Securities Dealers Automated Quotations (NASDAQ) and extract the overall sentiment (positive, negative, or neutral) expressed within them. Then they incorporated traditional technical indicators like moving averages, Bollinger Bands, and Relative Strength Index (RSI) which are commonly used by traders to analyse price trends with extracted sentiment. Finally, they integrated these sentiment values and technical indicators into a sophisticated machine learning model called a Generative Adversarial Network (GAN). GANs are known for their ability to generate realistic data, and in this case, they were used to predict future stock prices. Here they pointed out GAN has outperformed traditional methods like LSTM, GRU, and ARIMA, which are commonly used for time series forecasting.

In their research, Cheng and Chen (2021) [24] introduced an advanced approach to sentiment analysis specifically designed for financial texts. Their method combined FinBERT the version of the well-known BERT model specially fine-tuned with financial language to generate meaningful contextual embeddings. These embeddings were subsequently fed into a Bidirectional Long Short-Term Memory (BiLSTM) network, allowing the model to capture the contextual dependencies and relationships between words from both forward and backward directions in the text. Additionally, they incorporated an attention mechanism, which improved the model's ability to selectively emphasize crucial sentiment-bearing words and phrases. By blending these technologies, the authors enhanced the accuracy and reliability of sentiment predictions in financial contexts, outperforming conventional sentiment-analysis methods. In the context of Cheng and Chen's (2021) sentiment analysis approach,

- Bidirectional (BiLSTM) refers to a type of recurrent neural network (RNN) that processes text sequences in two directions forward (from start to end) and backward (from end to start). By considering both past and future contexts simultaneously, a BiLSTM can better capture complex patterns and nuances in financial texts, resulting in a richer and more accurate understanding of the overall sentiment.

- The Attention mechanism is a computational technique that allows the model to selectively focus on specific words or phrases within the text, assigning higher importance to sentiment-critical words. Rather than treating all words equally, attention dynamically weighs each word based on its relevance to predicting sentiment. This makes the model more effective at identifying and emphasizing meaningful information from financial documents.

Mane and Kandasamy [25] provided a comprehensive survey on recent advancements in NLP and machine learning techniques specifically aimed at predicting stock market movements. Their study systematically reviewed and categorized contemporary research articles, highlighting emerging trends, methodologies, and prevalent practices within the domain. By examining various NLP techniques, including sentiment analysis and semantic analysis, along with their integration into machine learning models, the authors illuminated key approaches that researchers increasingly adopt for stock prediction. The survey notably underscored significant progress in using NLP for market forecasting, concluding that NLP combined with machine learning offers substantial potential for improving stock market predictions. This work serves as an extensive resource, guiding future researchers towards promising directions in NLP-driven financial analytics.

The study [26] investigates how news articles about innovations affects the returns of illiquid stocks. The researchers utilized a dataset comprising information on 2,000 United States based companies over six years, analysing approximately 1.4 million news articles from the investor platform "Benzinga." They employed machine learning techniques, including the BERT and Top2Vec (an algorithm for topic modelling and semantic search) models, to assess news sentiment and identify key topics discussed by investors. Their findings show that information on product innovations has a significant effect on the returns of illiquid stocks, while other types of innovation-related news do not exhibit the same impact. Additionally, the study suggests that under conditions of uncertainty, innovation-related publications do not affect the returns of illiquid stocks. The analysis also reveals that narratives related to important corporate announcements positively influence the returns of illiquid stocks.

In their study [27], authors explored the effectiveness of advanced NLP and machine learning models—including FinBERT, GPT-4, and Logistic Regression—for sentiment analysis and stock market prediction. Utilizing textual data from financial news and market index data, specifically the Nigerian Stock Exchange All-Share Index, the authors compared the accuracy and reliability of these models through various statistical metrics like accuracy, precision, recall, F1-score, and the area under the ROC curve. Their experimental findings revealed that, while traditional logistic regression provided a solid baseline, advanced NLP models like FinBERT and GPT-4 demonstrated superior performance in accurately capturing market sentiment and predicting stock index movements. The study contributed valuable insights into the comparative advantages of sophisticated NLP techniques, ultimately suggesting that the integration of advanced AI-driven NLP methodologies substantially enhances predictive performance in financial contexts.

In 2024 a study [28] proposed an innovative multimodal approach to predict stock price trends by integrating specialized NLP models, particularly BERT-based methods, along with multiple data modalities such as publicly available news articles. Unlike conventional methods focusing primarily on absolute stock prices, their approach utilized stock percentage changes to better capture meaningful market movements. The experimental results demonstrated that even smaller NLP models could effectively predict general stock market trends. Additionally, their findings

emphasized the importance of leveraging targeted data features and sector-specific information to significantly boost prediction accuracy. This research advanced the literature by proving that combining NLP techniques with tailored data significantly improved stock price forecasting, suggesting that multimodal NLP frameworks hold great promise for financial analytics.

2.5 Text Vectorization and Natural Language Representation

Text vectorization is the process of converting unstructured text data into numerical representations that can be used by machine learning algorithms. Since computers cannot process raw text directly, vectorization transforms words, sentences, or entire documents into numerical formats while preserving their meaning and relationships. Different vectorization techniques vary in how they capture word importance, semantic similarity, and contextual relationships in each text corpus. There are several text vectorization methods the study should be considered and those will be illustrated in below,

2.5.1 Bag of Words (BoW)

Bag of Words (BoW) [29] introduced to represent text data numerically by converting documents into fixed-length vectors based on word occurrences. This creates a vocabulary from a corpus and represents each document as a vector indicating the presence or absence (or frequency) of words from this vocabulary, disregarding grammar and word order. This is also a straightforward and computationally efficient, making it suitable for tasks like document classification and spam detection. While effective for certain tasks, BoW's simplicity limits its ability to capture semantic relationships and context within text.

2.5.2 Term Frequency-Inverse Document Frequency (TF-IDF)

[30] Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure that evaluates how important a word is to a document relative to a collection (corpus) of documents. It assigns weights to words based on their frequency in a document and their rarity across the entire corpus.

- Term Frequency (TF): Counts how often a word appears in a document.
- Inverse Document Frequency (IDF): Gives less importance to words that appear frequently in many documents (e.g., "the", "is") and more importance to rare words.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \qquad (1)$$

Figure 2-1 TF-IDF Equation

As a key finding, TF-IDF improves upon BoW by reducing the influence of common words, thereby highlighting terms that are more indicative of document content. This is a popular method for information retrieval and text mining, though it still lacks the ability to capture semantic meaning and context.

2.5.3 Doc2Vec

Another way of vectorization is Doc2Vec [31] which extends Word2Vec to represent entire documents (instead of individual words) as continuous vectors in a high-dimensional space. Doc2Vec extends the idea behind Word2Vec by shifting the focus from generating embeddings for individual words to producing vector representations for entire documents or paragraphs. These document embeddings are dense numerical vectors that efficiently capture the semantic meaning of texts, making them useful for various natural language processing tasks. Doc2Vec relies on a neural network architecture trained on large text corpus, where the model learns to predict target words using their surrounding context and a unique document identifier. It operates through two primary models: Distributed Memory (DM), which uses both context words and a document ID to predict a specific word, and Distributed Bag of Words (DBOW), which aims to predict the document ID from words within the document. These approaches enable Doc2Vec to effectively capture the contextual and semantic characteristics of longer text segments.

Authors have proven that doc2vec performs well in representing semantic meaning and semantic similarity between given documents or paragraphs.

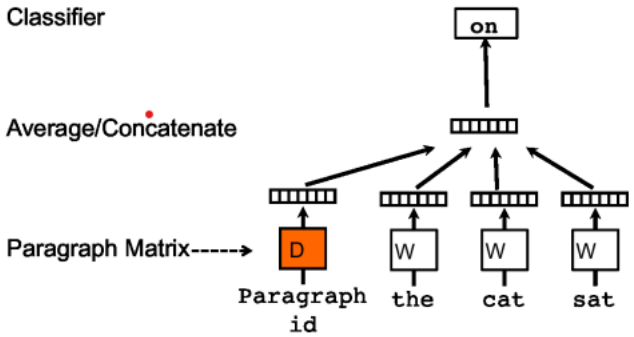


Figure 2-2 Predicting the words based on their surrounding context within a document, along with a document ID

Despite its advancements, Word2Vec treats each word as a single vector, failing to account for polysemy (multiple meanings) and lacking context sensitivity.

2.5.4 Global Vectors for Word Representation (GloVe)

The objective of GloVe [32] is to create word embeddings by capturing global statistical information from a corpus. It constructs a word co-occurrence matrix and factorizes it

to produce word vectors, effectively capturing both local context and global corpus statistics. This combined the strengths of matrix factorization and context window methods, providing a robust approach to generating word embeddings. As a finding GloVe embeddings demonstrate improved performance in tasks requiring semantic understanding and outperform other models in analogy tasks. While effective, GloVe shares the same limitations with Word2Vec regarding polysemy and context sensitivity.

2.5.5 Large Language Models in Vectorization

Traditional text vectorization methods, including Doc2Vec, TF-IDF, BoW, and GloVe, have notable limitations when utilized in financial news analysis for stock price prediction. Methods such as BoW and TF-IDF, despite their computational efficiency, largely neglect contextual semantics by treating words independently, thus failing to capture the nuanced meanings and sentiment within financial texts. Although Doc2Vec improves upon these methods by creating document-level embeddings, it still averages individual word vectors, potentially diluting essential sentiment cues, context-dependent financial jargon, and critical market-impacting expressions. Similarly, GloVe captures global semantic relations through word co-occurrence patterns but remains limited by static, context-insensitive embeddings, thus struggling with polysemy and dynamic sentiment shifts commonly present in financial news.

Conversely, advanced large language models such as FinGPT, FinBERT, and SBERT overcome these limitations by effectively capturing semantic complexity, sentiment nuances, and financial context relevance through deep contextual embeddings. FinBERT, specifically tailored for financial texts, leverages a transformer-based architecture to identify subtle sentiment differences relevant to stock market behaviour, improving sentiment polarity recognition even in complex, context-dependent scenarios. SBERT contributes further by providing accurate sentence-level embeddings optimized for semantic similarity tasks, making it highly effective at differentiating between closely related financial contexts. Moreover, FinGPT, as an open-source financial language model, extensively trained on financial corpora, excels at understanding intricate financial terminology, market-specific phrases, and the evolving sentiment inherent in economic discourse.

[33] Araci (2019) addressed the challenge of applying general-purpose language models to the financial domain, where the specialized vocabulary and nuanced context of financial texts often limit the effectiveness of standard models like BERT. To bridge this gap, the author introduced FinBERT, a pretrained language model specifically fine-tuned for financial communications, aiming to enhance performance in tasks such as sentiment analysis, classification, and entity recognition within financial texts.

The approach involved taking the original BERT (Bidirectional Encoder Representations from Transformers) architecture and fine-tuning it on a large corpus of Corporate Report, Earnings Call Transcripts, Analyst Reports totalling around 4.9 billion tokens. This adaptation allowed FinBERT to better grasp domain-specific language and context, which general models often fail to interpret accurately. The underlying technology behind FinBERT is based on transformer-based self-attention

mechanisms, which enable the model to understand relationships between words in a sentence regardless of their distance, providing deep contextual embeddings.

A key innovation in FinBERT is its domain adaptation strategy unlike typical models trained on open-domain data (e.g., Wikipedia, BooksCorpus), FinBERT was specifically trained on financial texts, making it more adept at interpreting sentiment polarity, financial jargon, and subtle cues in corporate communications. This domain-specific pretraining resulted in superior performance on financial sentiment classification tasks compared to baseline models like standard BERT, LSTM, and traditional machine learning classifiers.

[34] Yang, Liu, and Wang (2023) introduced FinGPT, an open-source large language model specifically designed for financial text analysis. Recognizing the growing demand for specialized NLP solutions in finance, the authors developed a domain-adapted Large Language Model trained specifically on financial texts, capable of effectively capturing financial terminologies and context. The model underwent fine-tuning on extensive financial datasets, including market news, analyst reports, and economic commentary. Experimental results demonstrated FinGPT's superior performance in financial-specific natural language tasks, such as sentiment analysis, text classification, and forecasting market movements, compared to general-purpose language models. The authors concluded that FinGPT significantly enhances financial text modelling, providing an accessible foundation for future research and practical financial analytics applications. Unlike the other models, FinGPT takes a data-centric approach, providing researchers and practitioners with accessible and transparent resources to develop their FinLLMs. In other words, FinGPT represents an innovative open-source framework designed specifically for applying LLMs within the financial domain.

[35] Reimers and Gurevych introduced Sentence-BERT (SBERT), a modification of the widely used BERT model designed specifically to generate meaningful sentence embeddings efficiently. Unlike the traditional BERT approach, which is computationally intensive for tasks involving sentence-pair comparisons, SBERT utilizes a Siamese network architecture to produce semantically rich embeddings directly. Their experiments demonstrated that SBERT achieved state-of-the-art performance with significantly improved computational efficiency on a variety of NLP tasks, including semantic textual similarity, clustering, and classification. Ultimately, the authors concluded that SBERT provided high-quality, task-agnostic sentence representations suitable for diverse NLP applications, outperforming many existing embedding methods both in terms of accuracy and computational speed.

LLMs demonstrated higher accuracy and better generalization in financial NLP applications, proving especially effective in tasks like predicting market sentiment and analysing investor communications. Its advantages include improved understanding of financial terminology, greater prediction accuracy, and better handling of complex sentence structures commonly found in financial documents. Overall, these offer a robust and efficient solution for researchers and practitioners seeking to apply natural language processing in finance, setting a new benchmark for domain-specific language models.

2.6 Machine Learning Models for Stock Price Action Analysis

The integration of machine learning (ML) techniques into stock market prediction has revolutionized financial forecasting by enhancing the analysis of complex market dynamics. Traditional statistical models often struggled with the non-linear and volatile nature of financial markets, prompting the adoption of more sophisticated ML approaches. These include regression models, neural networks, and ensemble learning methods, each offering unique advantages in processing and interpreting financial data. Machine learning (ML) has transformed stock price analysis by leveraging vast amounts of structured and unstructured data to identify patterns and make predictions. Traditional models, such as fundamental and technical analysis, rely on historical financial data, but ML enables dynamic learning from past trends, financial news, and market sentiment.

There are supervised, unsupervised, deep learning and reinforcement learning models when it comes to models that are frequently used for stock price action analysis tasks.

In general, under the supervised learning, Regression Analysis serves as a foundational tool in financial forecasting, aiming to identify relationships between dependent and independent variables to predict stock prices. Linear regression models, for instance, have been employed to forecast stock prices based on historical data. However, their linear assumptions often limit their effectiveness in capturing the complex, non-linear relationships inherent in financial markets. To address these limitations, advanced regression techniques such as Support Vector Regression (SVR) have been utilized. SVR can model non-linear relationships by mapping input features into high-dimensional spaces, thereby enhancing predictive performance. Studies have demonstrated that SVR outperforms traditional linear models in stock price prediction tasks by effectively capturing complex patterns in financial data.

Under unsupervised learning models which is useful when we don't have labelled outputs as in classification problems. It can be used to investigate unknown class labels and most importantly data filtration tasks. K-Means, DBSCAN, and Hierarchical Clustering are popular clustering models which are related to stock price action analysis problems. Importantly, to reduce higher dimensional data we can use dimensionality reduction models such as Principal Component Analysis (PCA) models is helpful because when it comes to event study we have to deal with large features as input data.

Then there are deep learning models, especially recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) Networks, Convolutional Neural Networks (CNNs), and Transformer-Based Models (BERT, FinBERT, GPT-3 for Financial Text Analysis) which have been highly effective in stock price action analysis problems. Neural networks, particularly deep learning architectures, have gained prominence in financial forecasting due to their ability to model intricate patterns and dependencies within data. Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, are adept at processing sequential data, making them suitable for time-series forecasting in stock markets. LSTMs can capture long-term

dependencies and temporal patterns, which are crucial for predicting stock price movements. Research has shown that LSTM models achieve superior predictive accuracy compared to traditional models by effectively learning from historical stock price data.

Ensemble learning combines multiple base models to improve predictive performance and robustness. Techniques such as bagging, boosting, and stacking have been applied in stock market prediction to mitigate the weaknesses of individual models. For example, an ensemble approach integrating LSTM and Autoregressive Integrated Moving Average (ARIMA) models has been proposed to leverage both linear and non-linear patterns in financial data. This hybrid model demonstrated enhanced forecasting accuracy by capturing a broader spectrum of market dynamics.

Natural Language Processing (NLP) has enabled the extraction of valuable insights from unstructured textual data, such as financial news articles, which significantly influence stock market behaviour. By analysing the sentiment and content of news articles, NLP techniques can assess market sentiment and predict stock price movements. For instance, transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) have been employed to analyse news headlines and extract sentiment scores, which are then used alongside historical stock prices to predict future stock movements. Studies have demonstrated that incorporating news sentiment analysis enhances the accuracy of stock price predictions, highlighting the importance of textual data in financial forecasting.

In addition to that there are reinforcement learning models which are used to optimize stock price action analysis problems by trial-and-error approach. It is useful since it can adaptively make decisions based on present market conditions.

Using each or multiple models, recent studies have been able to capture sequential dependencies in stock price movements, identify patterns in candlestick charts and technical indicators, process textual financial data as news, social media, and earnings reports and enhance price action prediction by integrating sentiment analysis.

Conclusively, the application of machine learning models in stock market prediction has significantly advanced financial forecasting capabilities. Regression models, neural networks, and ensemble learning methods each contribute uniquely to understanding and predicting market trends. The integration of NLP techniques further enriches these models by incorporating insights from financial news, thereby providing a more comprehensive analysis of factors influencing stock prices. As financial markets continue to evolve, the synergy between advanced ML models and NLP is expected to play an increasingly pivotal role in developing robust and accurate stock market prediction systems.

Since the focus is to build a robust model that can be used to analyse stock price action for an illiquid stock exchange as CSE using daily business-related news articles, which must be addressed separately because of the lack of effectiveness in traditional approaches. It is required to dive in to look at the challenges doing event study for the illiquid stock market price analysis. In this study it is used daily news articles preprocessed, filtered, vectorized in a way to capture each news article’s, contextual semantic and sentiment.

2.7 Graph Neural Networks for Stock Price Analysis Problem

Graph Neural Networks (GNNs) have emerged as a powerful tool for modelling non-Euclidean data structures, such as the interconnected and dynamic nature of financial markets. Unlike traditional time series models or deep learning architectures like LSTM and CNN, GNNs allow for the integration of both temporal and relational information, which is particularly useful for understanding inter-stock dependencies and external influences like news or macroeconomic signals.

Yao et al. [36] introduced a novel application of Graph Convolutional Networks (GCNs) to the task of text classification, demonstrating that graphs can represent textual data more effectively than traditional sequential or bag-of-words models. The central objective of the study was to model both word-word co-occurrence and document-word relationships by constructing a heterogeneous graph that connects documents and words within a corpus. In this graph, each document and word were treated as a node, and edges were formed based on term frequency-inverse document frequency (TF-IDF) and co-occurrence statistics in a fixed-size sliding window.

The authors applied a two-layer GCN over this graph to propagate label information and learn global document representations. The graph structure allowed GCNs to leverage long-distance semantic dependencies across documents and words, which are often missed in sequential models like RNNs or CNNs. Experimental results across multiple standard benchmarks (including 20-Newsgroups and R8) showed that this graph-based approach significantly outperformed conventional methods, including fastText and TextCNN.

The paper's primary contribution is the TextGCN framework, which reframes the text classification problem as node classification in a document-word graph. This paradigm enables better incorporation of global corpus-level statistics while maintaining the ability to model word semantics. The authors concluded that GCNs are particularly well-suited for tasks requiring global context aggregation in textual data and laid the foundation for later work involving graph-based models in NLP.

Huang et al. [37] proposed the Text-Level Graph Neural Network (Text-Level GNN) to enhance text classification by dynamically constructing task-specific graphs for each input text, as opposed to using a global graph across the entire corpus. The authors argued that previous methods like TextGCN, while effective, used static, corpus-level graphs which lacked adaptability and limited scalability for online or dynamic tasks. Their model addressed this limitation by constructing individual graphs per text instance, capturing fine-grained word dependencies within each document.

In their approach, words in a sentence were treated as nodes, and edges were established based on syntactic dependencies and proximity-based co-occurrence. The graph structure was then fed into a Gated Graph Neural Network (GGNN) that aggregated node features across multiple propagation layers. Additionally, the authors applied an attention mechanism over the graph to emphasize more informative nodes for the classification task.

Empirical results on several benchmark datasets (including MR, SST-2, and TREC) demonstrated that Text-Level GNN consistently outperformed both global-

graph methods (like TextGCN) and sequential models (like BiLSTM and CNN), particularly in shorter texts. The key contribution of the study is the introduction of a localized and flexible graph construction strategy that tailors the graph to each specific input, enhancing the model's ability to capture nuanced intra-document relationships. The authors concluded that instance-level GNNs hold promise for improving both accuracy and interpretability in text classification tasks

Wu et al. [38] presented a comprehensive survey on the application of Graph Neural Networks (GNNs) in various Natural Language Processing (NLP) tasks. The objective of this survey was to systematically review existing GNN methodologies within NLP and categorize them effectively. The authors proposed a structured taxonomy covering three principal aspects: graph construction techniques, methods for graph representation learning, and models employing graph-based encoder-decoder frameworks. Their systematic analysis identified the versatility of GNNs in handling complex linguistic structures and relationships inherent within textual data. The survey contributed by offering clear categorizations, summaries of benchmark datasets, evaluation standards, and open-source implementations. It concluded by addressing current challenges and future research avenues, particularly highlighting the extensive potential of GNNs in advancing NLP tasks.

In their systematic review, Li et al. [39] aimed to consolidate existing research on the application of Graph Neural Networks in stock market forecasting. By thoroughly analysing various methodologies, datasets, graph modelling techniques, and evaluation criteria, the authors identified critical patterns and insights across different prediction tasks. The review's primary finding emphasized that GNN-based methods significantly enhanced predictive accuracy due to their unique ability to capture complex market interdependencies effectively. The review's main contribution was the development of a generalizable framework guiding the use of graph-based methods in financial prediction, alongside providing clear directions for future research. Ultimately, the study concluded that GNN approaches offer considerable promise for addressing the intricate, interconnected nature of stock markets.

Xiang et al. [40] introduced a Temporal and Heterogeneous Graph Neural Network (THGNN) specifically tailored for predicting financial time series. The objective of their research was to model dynamic relationships within stock price movements effectively. The methodology involved creating dynamic graphs reflecting daily relationships among stocks and utilizing transformer encoders to embed temporal patterns of stock movements. A heterogeneous graph attention mechanism was integrated to enhance the predictive accuracy further. The experimental evaluations, conducted on datasets from U.S. and Chinese stock markets, demonstrated the THGNN model's superior performance over conventional baselines. The significant contribution of this work lies in its innovative combination of temporal and heterogeneous graph structures, which effectively captured evolving and diverse stock relationships. The study concluded that the THGNN approach substantially improves financial forecasting accuracy by explicitly modelling these temporal and relational dynamics.

Qian et al. [41] developed the Multi-relational Dynamic Graph Neural Network (MDGNN) framework to predict stock investments more effectively. Their research

aimed to overcome limitations associated with traditional predictive methods by capturing multifaceted and dynamic relationships between stocks over time. The researchers constructed discrete dynamic graphs reflecting multiple types of stock relationships, leveraging a transformer-based encoder to capture temporal variations. Results indicated that MDGNN consistently outperformed established methods across benchmark financial datasets. A key contribution of this study is its ability to model complex, evolving interdependencies between stocks explicitly. The authors concluded that the MDGNN methodology provides an advanced approach to stock prediction by effectively incorporating dynamic, multi-relational structures, ultimately enhancing the reliability and accuracy of investment forecasting.

Zhang et al. [42] addressed the limitations of existing graph-based text classification methods by proposing an inductive learning framework known as TextING. Their research aimed to capture the contextual relationships between words within individual documents effectively. Unlike traditional methods relying on global graphs, TextING constructs unique graph representations for each document, applying Graph Neural Networks to generate detailed and contextually rich word embeddings. Extensive testing on multiple benchmark text classification datasets demonstrated that TextING consistently outperformed other state-of-the-art classification models. The study's key contribution was developing a methodology that supports inductive learning allowing efficient generalization to new, unseen documents and words. The authors concluded that their approach significantly enhances text classification by effectively leveraging the local contextual structures inherent within each document.

Together, these studies showcase the growing effectiveness of GNNs in financial prediction tasks. By representing stocks and related information sources as nodes in a graph, and learning from their interactions, GNNs can capture both structural and temporal market dependencies. This literature provides a strong foundation for integrating textual and quantitative data through graph-based models, particularly in under-researched, illiquid markets where traditional methods may underperform.

2.8 Research Gap and Contribution

While considerable research efforts have focused on stock price prediction and market trend analysis, most of such studies target highly liquid markets characterized by substantial trading volumes, frequent transactions, and consistent price variability. These well-established markets benefit from ample historical trading data, facilitating the development of predictive models that offer high levels of accuracy, reliability, and robustness. Conversely, illiquid markets, such as the Colombo Stock Exchange (CSE), remain underrepresented in financial research literature. These markets exhibit unique characteristics, including infrequent transactions, limited trading volumes, sparse and inconsistent data availability, and irregular price movements. Such conditions diminish the effectiveness of conventional prediction methods that rely predominantly on rich, regular datasets and stable market dynamics, thus creating a clear research gap demanding innovative analytical strategies.

Moreover, the existing body of research in this field primarily emphasizes company-specific or stock-specific data sources such as corporate financial disclosures, earnings statements, analyst reports, investor sentiment through social media platforms, and direct feedback mechanisms. Although these resources can be valuable in assessing individual stock performance, they often fail to adequately capture broader macroeconomic events or general daily news trends that significantly influence market sentiment and drive cross-sector stock price actions, particularly in smaller, less active markets. The oversight of daily news-driven market sentiment creates limitations, especially in capturing subtle shifts in investor perception and market dynamics that could impact stock indices broadly, beyond specific stocks or sectors.

Addressing these research limitations, this study proposes a hybrid analytical framework designed explicitly for illiquid market environments. It employs advanced large language models (LLMs - namely SBERT (Sentence-BERT), FinBERT, and FinGPT to develop enriched, context-aware vector representations from daily financial news articles. SBERT generates high-quality sentence-level embeddings that preserve semantic meaning, enabling precise evaluations of semantic similarities across diverse financial news events, which are crucial for interpreting subtle variations in market context. In contrast, FinBERT, a transformer-based language model specifically fine-tuned on financial text corpora, excels at nuanced sentiment analysis, accurately distinguishing between positive, neutral, or negative financial sentiments embedded in news narratives. Its strength lies in effectively discerning how subtle sentiment signals within financial news directly correlate with subsequent market behaviour, a capability essential for navigating sentiment-driven market dynamics. FinGPT complements these capabilities as an advanced financial domain-specific language model extensively trained on diverse and comprehensive financial textual data. It robustly interprets complex financial jargon, economic indicators, industry-specific terminology, and evolving market narratives, thus significantly enhancing contextual accuracy and reliability of news representation compared to traditional vectorization methods.

Collectively, these large language models overcome the shortcomings of traditional methods by providing deeper contextual insights, advanced sentiment recognition, and enhanced semantic interpretation capabilities. Their integration into this hybrid model ensures a thorough understanding of the multifaceted influences of daily news on stock price movements, particularly in markets suffering from low liquidity and sparse transactional data.

Furthermore, the academic investigation into illiquid market behaviours remains scarce, and the availability of generalized prediction frameworks that can be broadly applied across different stock indices is notably limited. For instance, existing literature, such as the study by explained in [26], predominantly focuses on specific case scenarios such as the influence of product innovation-related news sentiment on illiquid stock returns within U.S.-based companies. However, this study relied exclusively on sentiment polarity derived from approximately 1.4 million news articles sourced solely from a specific investor platform, "Benzinga," highlighting limitations in generalizability and contextual comprehensiveness.

In contrast to prior studies that narrowly concentrate on single industries or selected companies, this research explicitly aims to develop a versatile and scalable analytical framework applicable across multiple stock indices and various market conditions, irrespective of market liquidity levels. By systematically correlating daily financial news embeddings derived from advanced LLMs with observed stock price actions, the proposed model provides a universally adaptable forecasting tool that can accommodate distinct market environments. Such adaptability is particularly beneficial for investors, financial analysts, and stakeholders operating within under-explored markets like the CSE, where traditional predictive methodologies grounded in technical and fundamental analysis frequently prove inadequate due to inherent data constraints. Thus, the proposed research not only addresses critical gaps identified in current academic literature but also offers tangible analytical tools enhancing practical decision-making capabilities within challenging and under-researched financial market contexts.

3 RESEARCH METHODOLOGY

3.1 Introduction to Methodology

This research aims to analyse stock market price action using Natural Language Processing (NLP) and Graph Neural Networks (GNNs). Unlike liquid markets, where price movements are often driven by well-established patterns in technical and fundamental analysis, illiquid markets present unique challenges. Low trading volume, irregular price movements, and limited available data make traditional methods unreliable for predicting price action. This research explores an alternative approach by leveraging news-based insights to determine whether daily financial news has a measurable correlation with stock price fluctuations in illiquid markets.

To achieve this, financial news articles are scraped from online sources, while historical stock prices are collected from official stock market platforms. The raw text data undergoes preprocessing and vectorization using advanced language models such as FinGPT, FinBERT and SBERT, which convert unstructured text into meaningful numerical representations. These vectorized news embeddings, along with historical stock price changes, are then used to construct a graph-based representation of market dynamics. A GNN is trained on this structured data to capture complex relationships between news meaning and sentiment with stock price actions.

The key motivation behind this approach is the inapplicability of traditional stock prediction models in illiquid markets. Fundamental analysis relies on company financials, which may not accurately reflect short-term price movements, while technical analysis depends on historical patterns that are often unreliable due to low trading frequency and high volatility. By integrating news-based features into a graph learning framework, this research seeks to uncover hidden patterns that influence stock price behaviour in illiquid markets, offering a data-driven approach to price prediction where traditional methods fail.

The following sections detail the data collection, preprocessing, feature extraction, graph construction, model training, and evaluation methods used in this study.

3.2 Data Collection

Data collection is carried out by creating the dataset of news articles with its corresponding stock market trading data. Key points to highlight in collection news data as follows,

- Daily news articles were collected from the archives of the Daily Mirror (https://www.dailymirror.lk/) from January 1, 2010, to November 1, 2024.
- Due to the absence of a dedicated API, data extraction was performed using web scraping techniques.
- A total of over 250,000 news articles were collected.

- The dataset was split into two sets as training dataset used for model development and test set used for model evaluation.
  The collection of stock trading data was done as follows,
- Daily stock price movements for leading and highly active stocks were obtained from the official Colombo Stock Exchange website.
- To address missing data points, stock price data from Investing.com - Stock Market Quotes & Financial News (https://www.investing.com/) was also utilized.

While pre analysing the datasets there was trading data present for weekends, despite the CSE operating only on business days. Moreover, missing data entries were there, with no identifiable pattern (e.g., missing data on business days, missing data on multiple consecutive days). After merging the trade data obtained from both resources, to mitigate the impact of missing data, a combination of forward and backward filling techniques was employed to fill the missing values within the period.

Then to integrate news and stock trading data, an inner join was performed on the date column. This resulted in a merged dataset and was supposed to have observations for business days. To account for potential weekend news influence on Monday's stock prices, an attempt was made to accumulate news articles from Saturdays and Sundays with Monday's news articles. However, due to significant missing data on Thursday and Friday stock prices, this approach has become infeasible and formed incorrect data accumulation patterns, so avoided. Consequently, the merged dataset has the days covering the whole week (not only business days), its news articles and stock trading details for a particular stock. There was replication of the same dataset aligning with different active stocks in CSE.

3.3 Data Preprocessing

3.3.1 Preprocessing of News Articles

The preprocessing of financial news articles is a critical step in ensuring that the textual data used for stock price analysis is clean, structured, and free from noise. The scraped news articles often contain Hypertext Markup Language (HTML) tags, metadata, special characters, and redundant formatting elements, which can introduce inconsistencies in the data representation. To address these challenges, a multi-step text cleaning pipeline was implemented.

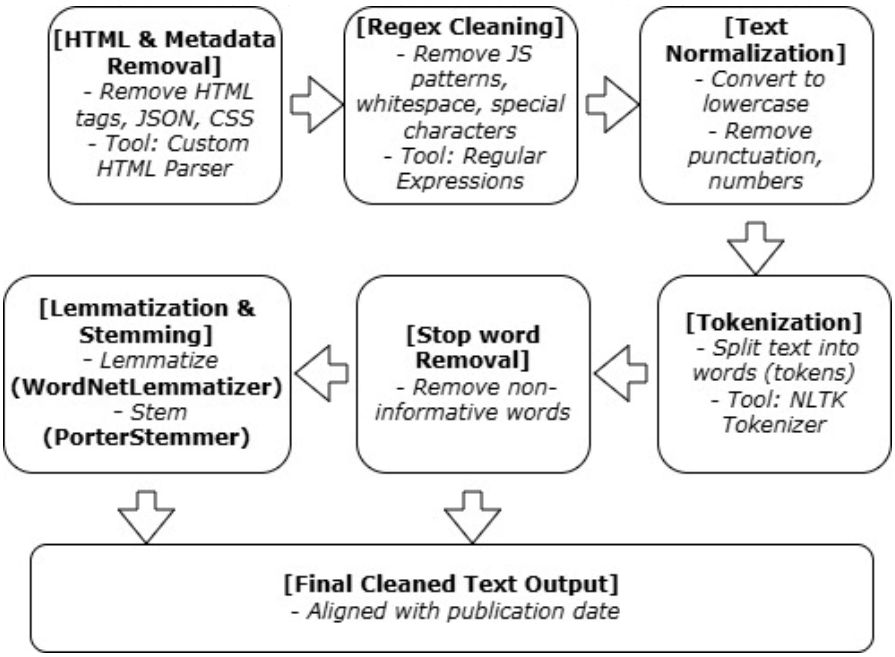
Initially, HTML tags and unwanted metadata (such as JSON objects, CSS styles, and special character encodings) were stripped from the raw text using a custom HTML parser. Additionally, regular expressions were employed to remove JavaScript-style patterns, redundant whitespace, and unnecessary newline characters, ensuring a well-formatted textual representation. After this, the text was converted to lowercase to maintain uniformity and prevent case-sensitive inconsistencies.

Further text normalization steps were applied to enhance the quality and consistency of the data. This included removing punctuation, numerical values, and special characters while preserving only alphanumeric words. The processed text was

then tokenized using the Natural Language Toolkit (NLTK) to break it into meaningful units. To reduce dimensionality and redundancy, stopwords (common non-informative words like "the," "is," and "in") were removed, ensuring only contextually significant words were retained.

Following stopword removal, lemmatization and stemming techniques were applied. Lemmatization by using WordNetLemmatizer ensures that words are converted into their root form while preserving their original meaning (e.g., "running" → "run"). Similarly, stemming by using PorterStemmer reduces words to their base morphology to unify variations of the same word (e.g., "flies" → "fli"). The final cleaned and processed text was stored in a structured format, aligning each article with its respective publication date (added_date) for further analysis.

This comprehensive text preprocessing pipeline ensures that the extracted financial news articles are clean, semantically meaningful, and structured, making them suitable for vectorization using FinGPT and SBERT. The refined textual representations serve as input features for the subsequent Graph Neural Network (GNN) training, enabling the analysis of correlations between news content and stock price action.



3.3.2 Preprocessing of Stock Trade Data

The Colombo Stock Exchange (CSE) typically operates only on business days, yet the initial dataset contained irregularities such as missing values on certain Thursdays, Fridays, and Mondays, as well as erroneous weekend entries. Although the plan was to accumulate weekend news and treat it as a partial impact of Monday's trading data, discrepancies in the dataset prevented a reliable implementation of this approach. Instead, a general preprocessing workflow was followed to handle missing values and outliers, ensuring consistency in the stock price data.

First, all null entries in the dataframe were set to zero (*fillna(0)*) and then replaced with *np.nan* to facilitate forward/backward fill operations. Using *fillna(method='ffill').fillna(method='bfill')* propagates the last known non-null price to subsequent missing days and, if necessary, the next known price backward. This prevents abrupt gaps in the time series while retaining valid market trends. Next, the daily price return was calculated using the first difference (*.diff()*), and any remaining NaN rows introduced by this operation were dropped.

To mitigate the impact of extreme outliers, the Interquartile Range (IQR) method was employed. Quartiles (Q1 and Q3) were computed for the 'price' column, and an IQR-based cutoff was used to define lower and upper bounds. Rows with prices falling outside 1.5 × IQR were considered outliers and filtered out. This IQR filtering ensures a more robust distribution of stock prices, making the dataset suitable for subsequent analysis and modeling. By applying these cleaning and filtering steps, the final stock price data reflects reliable time-series patterns without introducing artificial trends due to missing or erroneous entries.

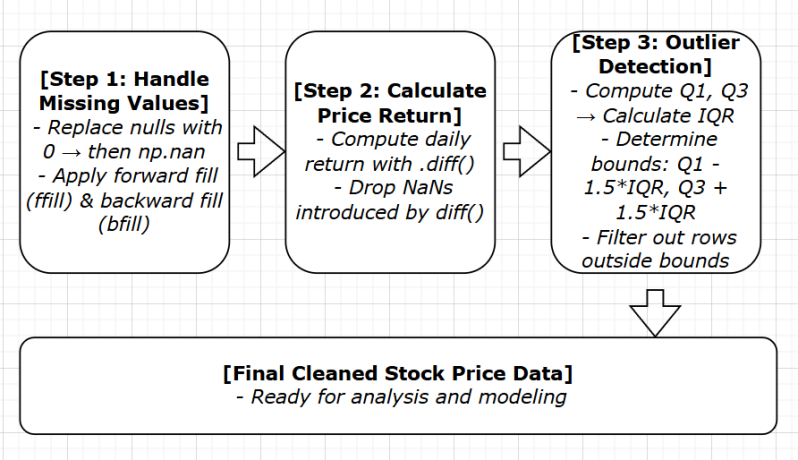


Figure 3-2 Flowchart Representing Stock Price Data Preprocessing Workflow

3.4 Feature Engineering

3.4.1 Feature Engineering and Vectorization of News Articles

To justify the use of FinGPT, FinBERT, and SBERT over traditional vectorization methods like TF-IDF and Doc2Vec, the proposed study to make a sufficient argument based on both qualitative reasoning (theoretical justification) and quantitative evidence (empirical performance).

3.4.1.1 Theoretical Justification

FinBERT and FinGPT are pretrained on financial texts, giving them a deeper understanding of financial terminology, sentiment, and nuanced language in news articles while SBERT offers sentence-level embeddings using semantic similarity, which TF-IDF and Doc2Vec cannot capture effectively.

TF-IDF is a bag-of-words model but consider no context, no word order, no semantic similarity. On the other hand, Doc2Vec captures some semantic meaning but performs poorly on short texts and lacks fine-grained contextual understanding. In contrast, FinGPT/SBERT/FinBERT generate embeddings based on full sentence meaning, not just word frequency.

3.4.1.2 Empirical Comparison

To visually assess the semantic quality of various text embedding models, the study employed Uniform Manifold Approximation and Projection (UMAP) to reduce high-dimensional news vectors to two dimensions. This allowed to observe how news articles with similar financial impact (measured via stock price movement) cluster in semantic space. In this work UMAP applied to four distinct embedding types as below,

- TF-IDF
- Doc2Vec
- SBERT
- FinGPT

Each point in the resulting plots represents a news article or a daily aggregated news embedding, color-coded by market return direction (Up or Down).

To establish a strong baseline for textual feature extraction, it is implemented using TF-IDF. It is a well-known, interpretable method that assigns higher weight to terms that are particularly relevant within a given document, making it suitable for capturing the most distinguishing words in financial news. To identify the optimal parameter configuration, a Grid Search was conducted over various hyperparameters. The final TF-IDF model included English stopword removal, a maximum document frequency of 0.9, a minimum document frequency of 5, sublinear term frequency, and L2 normalization, with 1-to-2-gram analysis. To manage computational complexity, the maximum feature dimension was set to 200.

Subsequently, the study continued with a Doc2Vec approach to capture richer semantic meaning from the text. Again, Grid Search guided the selection of critical hyperparameters, representing in a vector size of 200, a minimum word count of 2, 40 training epochs, and four worker threads. Both configurations balanced depth with computational efficiency, ensuring that each method produced meaningful document embeddings.

The UMAP projection of TF-IDF vectors resulted in a diffuse and loosely scattered plot, where "Up" and "Down" return labels were extensively intermixed. This outcome is expected, as TF-IDF merely encodes word frequency and lacks the ability to understand contextual or semantic nuances.

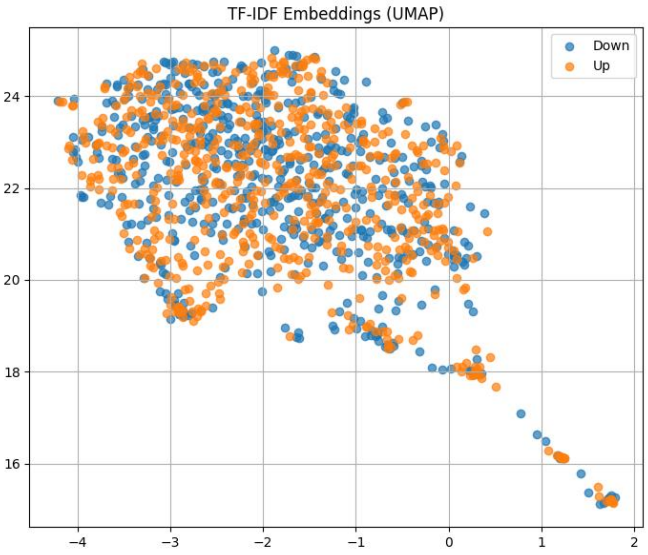


Figure 3-3 TF-IDF Embeddings (UMAP)

Doc2Vec showed slightly improved compactness but still failed to produce meaningful clusters aligned with return labels. The representation remained semantically shallow, as Doc2Vec struggles with short-text variability and domain-specific terminology.

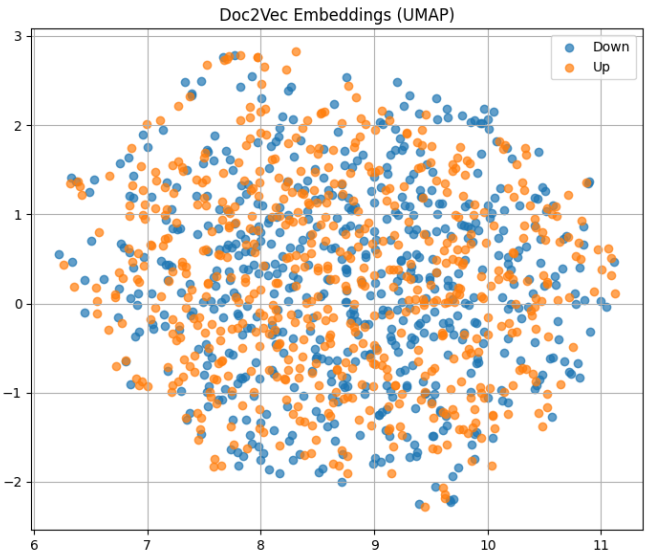


Figure 3-4 Doc2Vec Embeddings (UMAP)

SBERT embeddings began to show denser clustering in UMAP space. Although Up/Down separation was not perfect, the structure was visibly more coherent. This

reflects SBERT's strength in capturing general sentence-level semantics, suggesting it recognizes similarities in financial narratives, albeit in a domain-agnostic manner.

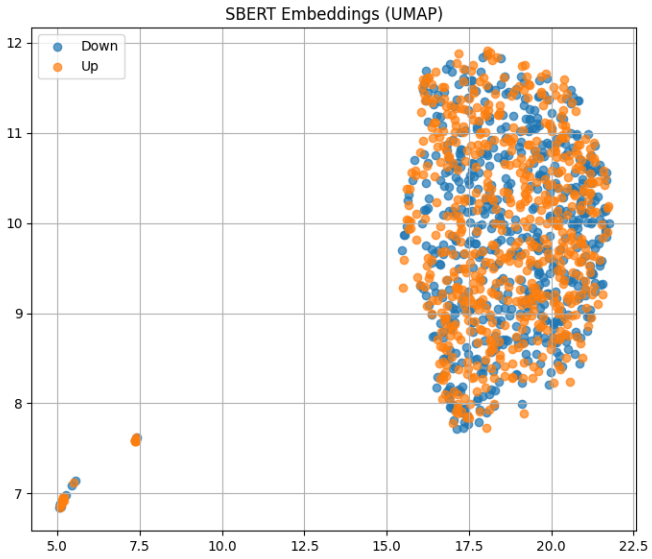


Figure 3-5 SBERT Embeddings (UMAP)

The UMAP projection of FinGPT embeddings revealed the most structured and spatially distinct distribution. Certain regions of the embedding space showed emerging patterns of return label grouping, suggesting that FinGPT effectively captures domain-specific semantic features relevant to stock movement. The shape of the clusters also suggested underlying thematic or sentiment structures within the financial news data.

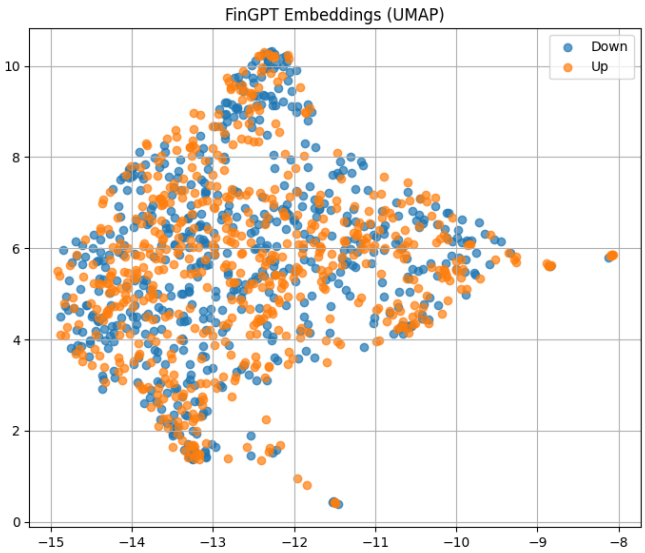


Figure 3-6 FinGPT Embeddings (UMAP)

This visual analysis supports the hypothesis that transformer-based embeddings, particularly those trained or fine-tuned on financial data (like FinGPT),

produce richer, more meaningful representations of news text. These embeddings exhibit better alignment with stock price movement, as evidenced by more structured clustering and subtle separation of return directions in UMAP space.

Thus, FinGPT and SBERT outperform traditional vectorization techniques (TF-IDF, Doc2Vec) in their ability to encode the kind of semantic and domain-specific signals that are essential for predictive modelling in financial markets.

3.4.1.3 Feature Engineering with FinGPT, FinBERT and SBERT

FinGPT is an open-source financial Large Language Model specifically designed for the finance domain. It's trained on vast amounts of financial news, reports, and market data, enabling it to understand industry-specific context. By converting entire news articles into numerical representations (embeddings), FinGPT captures subtle relationships, sentiment, and domain-specific nuances, which are often missed by generic language models. For this work, FinGPT - "*cm309/distilroberta-base-finetuned-Financial-News-Superior"* an architecture designed to represent news articles in 768-dimensional embedding was used.

FinBERT is a BERT-based model fine-tuned on financial text, with a particular emphasis on sentiment analysis in finance. It works at detecting positive, negative, or neutral sentiment within text, which is important for evaluating how the market might react to specific news events. In practice, FinBERT transforms each news article into an embedding vector that encodes sentiment polarity providing a way to measure the market's emotional response to daily events. FinBERT - "*yiyanghkust/finbert-tone"* is the architecture used to extract the financial sentiment classification since it is best suited for extracting sentiment-oriented embeddings or direct sentiment labels (positive, negative, neutral).

SBERT (Sentence-BERT) is a general-purpose model that generates high-quality sentence embeddings via a Siamese network structure. While it is not trained exclusively in the finance domain, SBERT is known for producing semantically rich representations that outperformed in similarity related tasks. It is especially useful when determining how closely related different news contents might be, providing another dimension of textual insight for price action analysis. In this research work it is used Sentence-BERT (SBERT) - "*all-MiniLM-L6-v2"* an architecture, designed to output 384-dimensional embedding for each news article.

- FinGPT - Contextual & Generative Understanding

FinGPT will be used to extract the context-aware representations of the news which are potential for feature extraction, trend summarization, financial jargon understanding. It will also be helpful to identify hidden patterns, and financial insights via embeddings. It bridges gaps between semantic meaning and financial context, potentially providing richer features.

Limitations:

Heavier computational cost and may need fine-tuning related to the context that will be applying for best performance.

- FinBERT - Financial Sentiment Extraction:

It will be used to extract sentiment features: positive, negative, neutral scores and used for identifying market sentiment, investor tone, or emotional weight of the news.
FinBERT - Vectorization:
Converts raw text into sentiment vectors, providing a directional signal (e.g., "market may react positively").
Limitations:
Focused only on sentiment, not on the semantic meaning or topic relevance.
May miss contextual nuances outside sentiment (e.g., cause-effect relationships).

- SBERT (Sentence-BERT) - Semantic Representation

SBERT will be used to extract the semantic meaning of entire sentences/news articles. It enables similarity comparisons (e.g., which news articles are semantically close) while capturing topic relevance and contextual depth.
SBERT - Vectorization
Outputs dense vector embeddings that represent context, relevance, and relationships.
Limitations:
It does not analyse sentiment. It understands "what is said" but not "how it's said emotionally".

Table 3-1 LLMs used for Feature Extraction Explanation

Model	Feature Extracted	Use Case	Output Dimension
FinGPT	Financial Contextual Embedding	Forecasting, Insight Generation	768
FinBERT	Sentiment (Pos/Neg/Neu)	Sentiment Analysis	3
SBERT	Semantic Context	Similarity, Relevance	384

By combining Language models explained above the methodology covers both general-purpose semantic embeddings (SBERT) and financially specialized representations (FinGPT, FinBERT) as this research aimed to capture both domain-specific details and semantic insights that simpler methods like TF-IDF or Doc2Vec may overlook.

3.4.2 Feature Engineering of Trade Data and Selection Rationale

While many stock prediction studies rely on price returns (percentage change) instead of raw prices, this research adopts closing stock prices as the primary target variable.

Using price returns has certain advantages, as they inherently normalize stock movements, making it easier for a model to learn patterns across stocks with different absolute price levels. The price return is calculated as follows:

$$\frac{p_t - p_{t-1}}{p_{t-1}} \tag{2}$$

By using returns, the model avoids bias toward stocks with higher absolute prices, which might otherwise dominate the learning process due to scale differences. However, a key limitation of stock returns is that they are small and tightly clustered around zero, resulting in a low-variance target variable. This can lead to models (especially linear models or simpler architectures) predicting values close to the mean return for all samples, producing near-constant outputs with little predictive value. This limitation was observed during the research concluding that choosing stock prices instead of returns as the main predictive target will be helpful to capture variance in target values.

Instead of predicting returns, this study directly models stock prices, allowing the model to learn actual market trends rather than relying on subtle percentage changes. However, returns are still incorporated in the model structure by using them as edge weights in the proposed Graph Neural Network (GNN), which will be further explained in the Graph Construction section. This hybrid approach enables the model to capture both absolute price trends and relative price movements, improving its ability to generalize across different stocks.

The stocks selected for this research Hatton National Bank (HNB), John Keells Holdings (JKH), and Browns Investments Limited (BIL)—were carefully chosen based on their prominence and representation in the Colombo Stock Exchange (CSE). These equities were selected from the official CSE website due to their high trading volumes, market capitalization, and significant influence within different sectors, including banking, diversified holdings, and investment sectors. This diverse sector representation helps ensure the generalizability of the proposed model across varying market segments and conditions. Additionally, these companies have substantial visibility in local financial news coverage, providing ample textual data necessary for analysing the relationship between news sentiment and stock price movements. By selecting stocks that are well-established, actively traded, and frequently discussed in financial media, the study ensures sufficient liquidity, data availability, and relevance, thereby enhancing the robustness and applicability of the research findings.

3.4.3 Dimensionality Reduction

To manage the high dimensionality resulting from combining FinGPT, SBERT, and FinBERT embeddings, this study implemented a Stacked Autoencoder (SAE) for

dimensionality reduction. The merged feature vector, though rich in semantic, contextual, and sentiment information, resulted in a high number of features, which could negatively impact computational efficiency and lead to overfitting during model training. The SAE, a multi-layer neural network trained in an unsupervised manner, was employed to compress these high-dimensional vectors into a 256 - dimensional latent space. This was done using two iterations, one is to reduce SBERT embedding and the other one is to reduce the FinGPT embedding. This bottleneck layer was selected after iterative testing to balance information retention and model performance it preserved meaningful patterns while ensuring efficient processing in downstream models.

Given that FinBERT outputs sentiment polarity scores (positive, negative, neutral), this study employed a strategic approach to maintain the interpretability and impact of sentiment features. Specifically, the high-dimensional SBERT and FinGPT embeddings were first passed through the SAE for compression and then concatenated, while FinBERT's sentiment scores were appended to the reduced concatenated vector after dimensionality reduction process. Only positive and negative polarities were used because neutral sentiment typically reflects no significant market-moving information, so its impact is minimal or hard to quantify. This method allowed the semantic and contextual embeddings to be compressed effectively while preserving the explicit sentiment signals from FinBERT. The final feature vector, composed of two 128 reduced dimensions from SAE and 2 sentiment scores, provided a 258 - dimensional input for machine learning models, ensuring a comprehensive yet efficient representation of financial news articles.

Compared to traditional dimension reduction methods like Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE), the SAE offered several advantages. Unlike PCA, which is limited to linear transformations, the SAE captured non-linear relationships within the financial news data, enabling better preservation of complex semantic and sentiment features. Similarly, while t-SNE and UMAP are effective for data visualization, they are less suitable for feature reduction in predictive modelling tasks, as they do not consistently preserve global structures or scale well to large datasets. The SAE's flexibility in architecture and its ability to learn task-specific feature representations made it an ideal choice for this research.

The selection of 256 as the reduced feature size was guided by experimental evaluation smaller dimensions (e.g., 64, 128) led to information loss and decreased prediction accuracy, while higher dimensions (e.g., 512, 1024) offered minimal performance improvement at the cost of increased computation. Thus, 256 dimensions were found to offer an optimal trade-off between accuracy and efficiency, ensuring that the model retained sufficient information for reliable stock price action analysis while remaining computationally tractable.

Table 3-2 Feature Selection and Reduction by LLM

Model	Reduced/Selected Features
FinGPT	128 – Reduced with SAE
FinBERT	2 – Selected
SBERT	128 – Reduced with SAE

3.5 Machine Learning Model - Graph Neural Network

The approach of using Graph Neural Networks (GNNs) for stock price forecasting is a strong choice, especially when there are large, high-dimensional data and interdependent features. GNN are better at capturing complex relationships in high-dimensional data and generalizing well. As discussed in Literature Review Chapter, several studies show that how GNN can outperform traditional neural networks like LSTM, Convolution Neural Network (CNN) for text classification problem with the characteristics such as better representation of non-sequential relationships, aggregation of contextual information and specially reduced overfitting.

Conclusively there are studies reveal that the results of graph-based models are better than traditional models like CNN, LSTM, and Bi-LSTM for classification problems. They have shown the reason behind the better results should be the characteristics of the graph structure. Graph structure allows a different number of neighbour nodes to exist, which enables word nodes to learn more accurate representations through different collocations.

The Graph Neural Network (GNN) approach was selected in this study as the machine learning model because it provides an effective way to model relationships between news articles and capture their impact on stock prices. By leveraging semantic and sentiment similarities between news articles, GNNs can build a structured graph representation that connects related pieces of information. Additionally, the data used in this research is high-dimensional and considerably harder to understand, with nonlinear patterns that traditional models struggle to observe. GNNs are well-suited for handling such higher-dimensional data and nonlinear relationships, making them an optimal choice for this task. This approach allows for a more comprehensive analysis of how daily news content influences stock price movements.

3.5.1 Graph Neural Network (GNN) Architecture

The core model used in this research is a Graph Neural Network (GNN), chosen for its ability to capture complex relationships between daily news articles and stock prices. The graph was constructed using SBERT and FinGPT embeddings to represent news articles as nodes and daily stock prices as additional nodes. The edges were defined

based on cosine similarity between news vectors and scaled stock price returns as edge weights.

The GNN architecture was built using GraphSAGE, a variant of GNN that effectively aggregates neighborhood information using sampling. The model consisted of three layers, with hidden dimensions of 128, 64, and 32, each followed by ReLU activation and dropout (30%) to reduce overfitting. Batch normalization was applied to the output before passing it through a linear layer to predict the final stock price. The model was compiled with the mean squared error (MSE) loss function and optimized using the Adam optimizer with an exponentially decaying learning rate.

The graph's node features included concatenated and reduced FinBERT, SBERT and FinGPT vectors, resulting in 258 - dimensional feature vectors for each news node. Stock nodes included scaled stock price values and scaled returns as additional node attributes.

3.5.2 Graph Construction

The graph was constructed with a focus on capturing both semantic similarity and sentiment polarity from financial news articles, integrated with stock price dynamics. The construction process involved the following steps,

- News Nodes: Each daily news article was represented as a node. The features of the news nodes consisted of reduced-dimensional embeddings obtained from SBERT and FinGPT models, capturing semantic meaning and contextual relevance. Additionally, each news node retained its FinBERT positive and negative sentiment scores for edge weighting purposes.

- Stock Nodes: Each trading day was represented by a stock node, with features including the normalized closing price and scaled price return.

- News-News Edges: Cosine similarity was computed between reduced SBERT and FinGPT embeddings, and edges were created between semantically similar news articles. A tuned threshold of 0.8 was used to ensure that only highly similar articles were connected.

- News-Stock Edges: Each news node was connected to the corresponding stock node for the same trading date. The edge weight was computed using a hybrid formula that integrated both market movement and sentiment polarity.

Edge Weight = tanh(scaled_return) x FinBERT (Positive score – negative score) (3)

This formulation ensured that both market response (price_return) and news sentiment intensity contributed to the influence weight of the article on stock price action. It also helped the model learn how sentimentally strong news impacts stock movements differently based on market context.

- Stock-Stock Edges: Temporal continuity was modelled by connecting stock nodes chronologically. The edge weight between consecutive stock nodes was set to the tanh of the scaled return on the earlier day, capturing market momentum and continuity.

Using tanh is often a convenient way to rescale continuous values into the [-1, 1] range, which can be particularly helpful for edge weights in a graph. It ensures that both positive and negative returns are captured while limiting extreme values that might otherwise dominate the training process. FinBERT outputs probability scores for positive and negative sentiment, both in the range [0, 1] and subtracting negative from positive creates a single sentiment polarity score ranging from -1 (strongly negative) to +1 (strongly positive). This enables the edge to reflect whether the news is expected to have a positive or negative influence on the stock, based on its content.

Combining return and sentiment scores allows the edge weight to represent both the market's actual response and the news article's predicted impact. This makes the GNN aware of sentiment-driven price movements and allows it to learn the strength of news influence based on how the market reacted historically.

The resulting graph was converted to a StellarGraph object, and a train-test split was applied, selecting 20% of the stock nodes for testing.

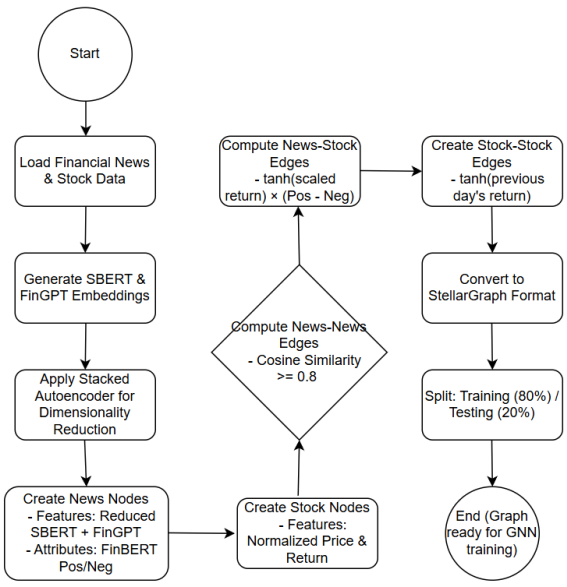


Figure 3-7 Graph Construction Visualization

3.5.3 Hyperparameter Tuning

Hyperparameter tuning was performed using a combination of Grid Search and manual experimentation. The key hyperparameters explored included,

- Cosine similarity thresholds for news-news edges: values between 0.2 and 0.9 were tested.

- GraphSAGE hidden dimensions: [64, 128, 256], [32, 64, 128], [128,256,512], [16,32,64,128]
- Learning rate: [0.01, 0.003, 0.005, 0.001, 0.0005, 0.0001]
- Dropout rate: [0.2, 0.3, 0.5]

The best-performing configuration used a cosine similarity threshold of 0.8, hidden dimensions of [32,64,128], a learning rate of 0.001, and a dropout rate of 0.3.

3.5.4 Training Procedure

The model was developed using 80% of the stock nodes for training and 20% for testing. An additional 20% of the training nodes were held out as a validation set during training. 18% preprocessed merged data kept unseen for the final evaluation purposes. The *Adam* optimizer with an exponentially decaying learning rate was used, along with early stopping to prevent overfitting. The training loop ran for a maximum of 100 epochs, with a patience of 10 epochs for early stopping.

Normalization was applied to the target values (stock prices) using a *StandardScaler*, and the predictions were later inverse transformed to recover the original price scale. The training process was run on a GPU to speed up computation.

3.5.5 Evaluation Metrics

The performance of the models was evaluated using the following metrics:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual stock prices.
- Mean Absolute Error (MAE): Calculates the average absolute difference between predictions and actual values, making it easier to interpret in the context of stock prices.
- R^2 Score: Indicates how well the model explains the variance in the target variable, with values closer to 1 representing better performance.

3.5.6 Baseline Models and Architecture

3.5.6.1 Deep Residual Multilayer Perceptron (MLP) - Sequential Model

The Deep Residual Multilayer Perceptron (MLP) model leverages residual connections, a state-of-the-art deep learning technique popularized by the ResNet architecture, which effectively addresses training issues such as vanishing gradients. The model processes high-dimensional embeddings derived from NLP methods specifically, semantic embeddings from SBERT and financial-contextual embeddings from FinGPT into an expressive representation capable of capturing complex relationships between news and stock prices. Its design includes multiple dense layers enhanced with batch normalization and dropout regularization, facilitating stable training and improved

generalization. Residual connections in the network allow gradients to bypass layers, enabling the efficient training of deeper and more sophisticated architectures without loss of performance. Recent literature validates the superiority of residual-connected MLPs in NLP-driven financial forecasting, underscoring their ability to accurately model intricate, nonlinear relationships inherent in market data. Thus, employing this Deep Residual MLP as a benchmark is methodologically sound, providing a competent baseline to objectively evaluate and highlight the effectiveness of this proposed Graph Neural Network model for stock price prediction using NLP inputs. The best trial informed the final MLP design, which was then retrained on reduced, and scaled input features and scaled target prices. Following hyperparameter tuning, a specific MLP architecture was chosen for the final training and evaluation. The data was split into 80% training and 20% testing sets, and both X (features) and y (target) were scaled using *StandardScaler* to stabilize training. The network began with an input layer (258 neurons), followed by multiple hidden layers (e.g., 192, 64, 32 units) each using *ReLU* activation, *LeakyReLU* operations, *Batch Normalization*, and *L2 Regularization*. *Dropout* at 20% was applied to reduce overfitting, and a linear output layer produced the final price prediction. The model was compiled using the *Adam* optimizer at a *Learning Rate* of 0.000149 with mean squared error (MSE) as the loss function. Trained for up to 400 epochs with a batch size of 32, this tuned MLP served as the core baseline for comparing stock price prediction performance alongside other approaches.

3.5.6.2 MLP-Attention-BiLSTM Model

To capture potential time-based dynamics in stock price predictions, the Multilayer Perceptron (MLP) – Attention – Bidirectional Long Short-Term Memory (BiLSTM) model was designed and then optimized through a grid search. This model was selected due to its proven effectiveness in recent literature for time-series forecasting tasks involving textual news data. It is particularly competent as a baseline because it incorporates a hybrid neural design that combines multilayer perceptron (MLP) layers, bidirectional Long Short-Term Memory (BiLSTM) networks, and an attention mechanism, which collectively provide robust representation, sequential modelling, and dynamic feature weighting capabilities. Specifically, the baseline integrates pre-trained semantic embeddings (SBERT), financial-domain embeddings (FinGPT), and sentiment indicators from FinBERT, forming a comprehensive and informative input representation of news data. This model is constructed beginning with an MLP-based dimensionality reduction (512 units, ReLU activation, L2 regularization), followed by stacked BiLSTM layers (256 and 128 units respectively, bidirectional, with *LeakyReLU*, dropout rate of 30%, and batch normalization). A self-attention mechanism, combined via residual connections, is incorporated to emphasize the most critical temporal information dynamically. The final layer is a smaller BiLSTM (64 units, bidirectional), culminating in a linear dense layer that outputs the predicted stock price. The chosen parameters including embedding dimensions (258), number of hidden units, activation functions (LeakyReLU), dropout rate (0.3), batch normalization, and regularization (L2 = 0.001)—reflect best practices identified

through empirical research, thereby making this model a competent and reliable baseline against which more advanced architectures, such as the proposed *GraphSAGE* model, can be benchmarked. These baseline results will be compared with the GNN model using MSE, MAE, and R^2 metrics to assess the relative performance.

3.5.7 Implementation Tools and Libraries

This research leverages a variety of Python-based libraries and frameworks to implement data processing, graph construction, and machine learning pipelines.

Pandas and *NumPy* are used for data manipulation and numerical computations, while *NetworkX* facilitates the creation and management of graph-based data structures. *StellarGraph* powers the Graph Neural Network workflows, providing specialized generators and layers such as *GraphSAGE* and *MeanAggregator*. For text processing, libraries like *NLTK* (including stopwords, *WordNetLemmatizer*, and *PorterStemmer*) and re (regular expressions) help clean and tokenize news articles. *Transformers* from Hugging Face enable the use of pretrained models (e.g., AutoTokenizer, AutoModel, SBERT, FinGPT, FinBERT) for advanced text embeddings.

The *torch* package (*PyTorch*) and *torch.nn* modules allow for custom deep learning architectures, whereas *tensorflow.keras* is used for building and training neural networks (e.g., *Dense*, *BatchNormalization*, callbacks, optimizers, etc.). Additional libraries such as *scikit-learn* provide *MinMaxScaler*, *QuantileTransformer*, and various metrics (MSE, MAE, R^2) and model selection tools (*train_test_split*, *Grid Search*). *Gensim* (including Doc2Vec) offers another perspective for document embeddings, and *TfidfVectorizer* supports classic bag-of-words text vectorization. Lastly, *matplotlib* and *pyplot* aid in visualizing data and model performance, and *joblib* is employed to serialize models and scalers for later use.

3.5.8 Limitations and Assumptions

This study acknowledges several limitations. First, the data availability was a significant challenge. While the initial dataset of 250000 became approximately 48,000 daily business news articles after cleaning and filtering, and aggregation by date reduced this to 2,993 data points. Although the intended timeframe was 2010-2024, data inconsistencies, including erroneous and missing values made it even less than 10 years of a dataset.

Particularly in historical stock trade data included with trade data for weekends which is erroneous and misleading. The inflationary environment in Sri Lanka, causing a significant gap between older and recent stock prices, has to be concerned with limiting the data to post-2010. Furthermore, the daily aggregation of stock data, while the finest granularity available from the CSE official website, prevents capturing potential intraday volatility. Ideally, hourly stock data would have allowed for a more precise mapping of news timestamps to market activity. In this study it assumes an immediate market reaction to news, potentially overlooking lagged or shifted impacts.

The shifted or lagged impact is yet a larger research area that could continue with a pattern recognition approach. Regarding the model, the Graph Neural Network (GNN) assumes specific graph structures (e.g., daily news and stock nodes), which may not fully represent real-world market complexities because this impact from news articles are mainly influenced on the general or occasional traders, not much the professionals.

Finally, while this approach is hypothesized to generalize to other stocks or markets, its efficacy may be strongest within specialized financial domains or for companies analysing the impact of news directly related to their own stock trading data, such as investor commentary, social media reactions or company-specific news. These limitations highlight areas for future research and refinement of the methodology.

4 IMPLEMENTATION

4.1 Training and Evaluation Setup

4.1.1 Data Splitting and Scaling

First, the feature matrix ***X*** and target variable ***y*** are converted to NumPy arrays with a *float32* data type for compatibility with neural networks. The data is then split into training and testing sets using *train_test_split* with an 80/20 split and a random state of 42 for reproducibility. Following the split, both the features (***X***) and the target variable (***y***) are standardized using *StandardScaler*. This involves fitting the scaler on the training data only to learn the mean and standard deviation and then transforming both the training and testing sets using the same fitted scaler. This ensures that the test data is scaled in the same way as the training data, preventing data leakage. The fitted scalers for both ***X*** and ***y*** are then saved using *joblib* for later use during model evaluation and prediction on new data.

4.1.2 Model Training

The final versions of each model; GNN, MLP, and MLP-A-BiLSTM were trained using the hyperparameters determined in the Methodology. All training was conducted on a NVIDIA-RTX4060 GPU with Python 3.8 and TensorFlow 2.9 (for GNN/MLP/LSTM). Each baseline model ran for up to 400 epochs and GNN model up to 100 epochs, with early stopping triggered if validation loss failed to improve for 10 consecutive epochs. Throughout training, batch size was set to 32, and a learning rate scheduling strategy (e.g.,*ReduceLROnPlateau*, *ExponentialDecay*) adjusted the learning rate according to validation performance. Logs detailing epoch durations, validation loss, and training time were recorded for each run. These practical steps ensured that the final model versions closely aligned with the methodology's design while accommodating computational resources.

4.2 Metric Calculation

To evaluate and compare model performance, three primary metrics were used: MSE (Mean Squared Error), MAE (Mean Absolute Error), and R^2 Score. MSE, defined as the average of the squared differences between predicted and actual values, emphasizes large errors by squaring residuals; MAE captures the average absolute deviation of predictions, providing an easy to interpret measure in the same units as the target variable. The R^2 Score, or coefficient of determination, indicates how much of the variance in the target is explained by the model, with values closer to 1 indicating higher explanatory power. These metrics together provide a robust overview of each model's accuracy, sensitivity to outliers, and overall explanatory capability. As an additional

check, RMSE could be considered for interpretability in the original target scale, though MSE, MAE, and R^2 are the principal metrics reported in this work.

4.3 Results Presentation

The outcomes of the GNN, MLP, and MLP-A-BiLSTM models are summarized in this section, which reports each model's final MSE, MAE, and R^2 on the test set. In addition, Figures will illustrate a sample of actual vs. predicted stock prices for selected test periods, offering a visual gauge of each model's accuracy. The GNN consistently exhibited lower MSE and higher R^2 scores compared to MLP and BiLSTM, suggesting that capturing relationships between news nodes through a graph structure provided a notable advantage.

Table 4-1 Model Performance Comparison Using Evaluation Metrics - HNB

Model	Stock	MSE	MAE	R^2
Residual MLP Model	HNB	701.5548	18.9324	0.5322
BiLSTM Model		769.5390	18.7846	0.4869
Proposed Model		210.0240	11.0811	0.8600

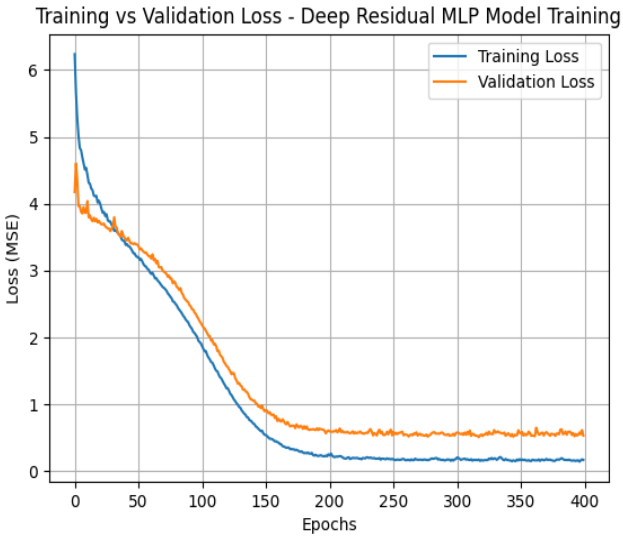


Figure 4-1 Training vs Validation Loss during Residual MLP Model Training - HNB

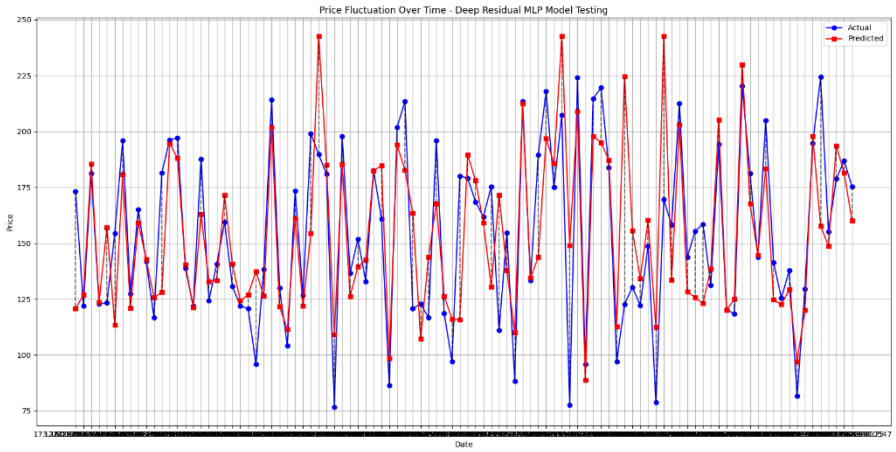


Figure 4-2 Actual vs Predicted over Time with Residual MLP Model Testing - HNB

In the *figure 4-2*, the predicted prices (red) show significant deviation from the actual prices (blue) across several points on the timeline. While the model captures the general direction of price movements in some areas, it fails to consistently align with the volatility and abrupt price shifts of the actual stock. The inconsistency in tracking sharp peaks and troughs suggests that the Deep Residual MLP model (Multilayer Perceptron) lacks the temporal memory or context-awareness necessary for modelling time-series financial data. This behaviour is also supported by the relatively high MSE (701.55) and lower R^2 (0.5322), reflecting moderate accuracy with limited generalization in a volatile, illiquid market setting.

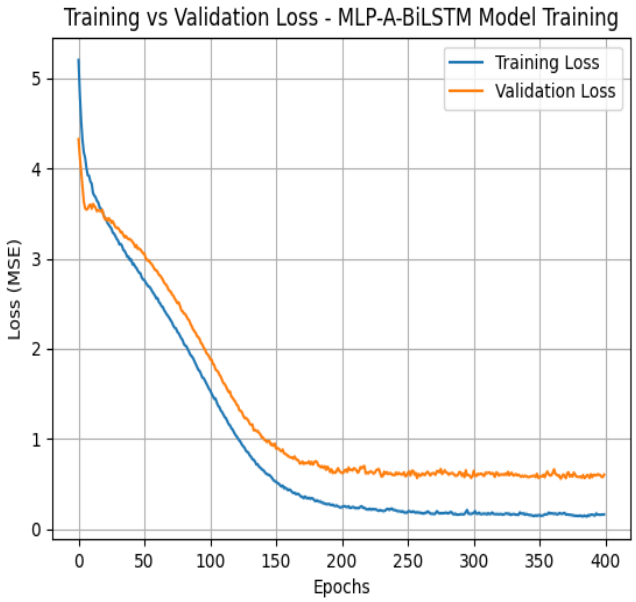


Figure 4-3 Training vs Validation Loss during BiLSTM Model Training - HNB

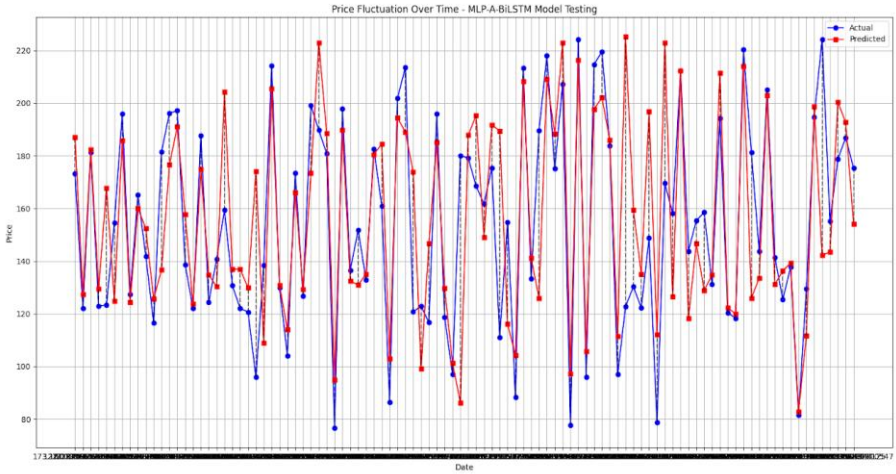


Figure 4-4 Actual vs Predicted over Time with BiLSTM Model Testing - HNB

The *figure 4-4* demonstrates improved tracking of price trends compared to the Residual MLP model. The BiLSTM model (Bidirectional Long Short-Term Memory) is better equipped to understand temporal sequences and shows smoother alignment with actual price movements. However, the predictions still exhibit noticeable divergence during sharp price changes and high-volatility periods. While it captures short-term dependencies well, it occasionally misaligns with steep rises and falls in the price curve. This is reflected in slightly better but still limited performance metrics (MSE: 769.53, R^2: 0.4869), suggesting that although BiLSTM accounts for some historical patterns, it is still insufficient for complex interdependencies found in illiquid markets.

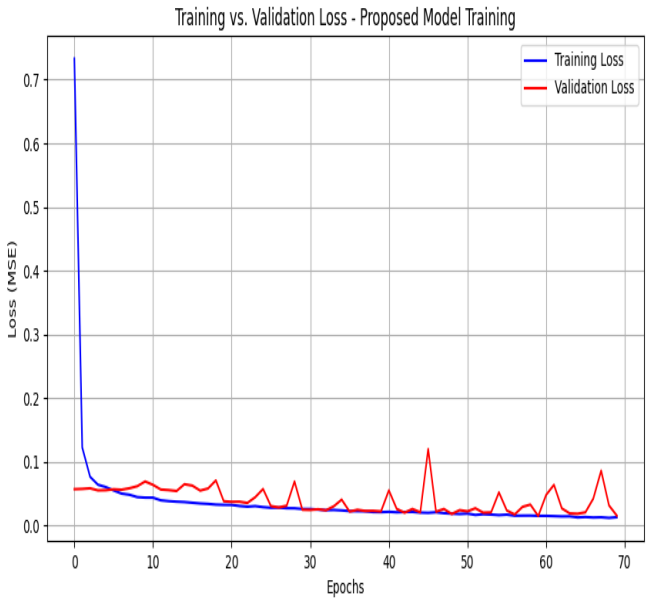


Figure 4-5 Training vs Validation Loss during Proposed Model Training - HNB

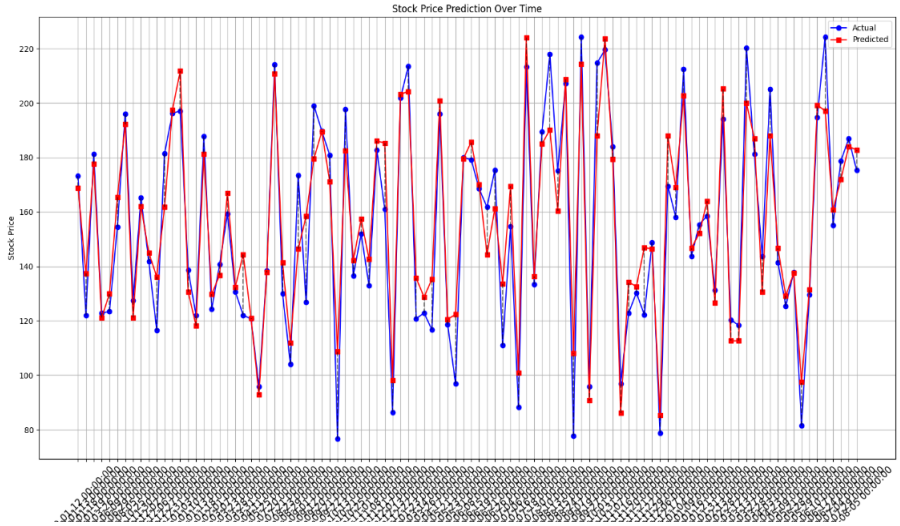


Figure 4-6 Training vs Validation Loss during Proposed GNN Model Training - HNB

The *figure 4-6* represents the output of the Proposed Graph Neural Network (GNN) model. This graph shows a much closer alignment between the predicted (red) and actual (blue) prices throughout the entire timeline. The GNN model effectively mirrors the overall trend and volatility of the stock, including rapid fluctuations, peaks, and dips. There are fewer and smaller deviations, indicating the model's capacity to capture both temporal dependencies and contextual influences such as sentiment and semantic features from financial news. This is strongly supported by the significantly improved metrics (MSE: 210.02, R^2: 0.8600), validating the model's suitability for handling illiquid market behaviour and complex, event-driven stock price movements.

Table 4-2 Model Performance Comparison Using Evaluation Metrics – JKH

Model	Stock	MSE	MAE	R^2
Residual MLP Model	JKH	488.9472	15.6901	0.3832
BiLSTM Model		470.2347	15.3693	0.4068
Proposed Model		223.2572	11.9391	0.7884

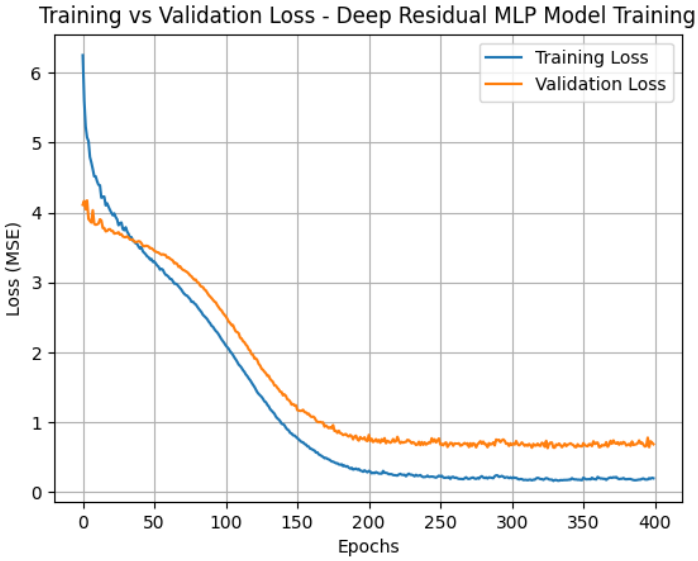


Figure 4-7 Training vs Validation Loss during Residual MLP Model Training - JKH

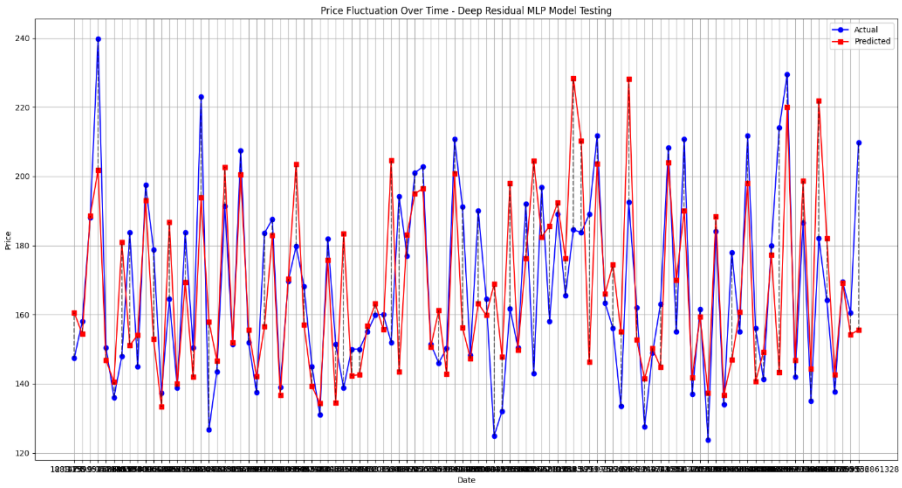


Figure 4-8 Actual vs Predicted over Time with Residual MLP Model Testing - JKH

The *figure 4-8* visualizes the performance of the Deep Residual MLP model on JKH stock. Here, the actual prices (blue) and predicted prices (red) exhibit noticeable discrepancies throughout the timeline. While the model captures the general trend in some stable regions, it frequently misses sharp upward or downward movements. Particularly during volatile phases, the predictions tend to lag or flatten, indicating the

model's struggle with dynamic price behaviour. The fluctuations in price are not closely mirrored, reflecting the model's limited capacity to handle temporal and contextual dependencies. This is consistent with its quantitative performance (MSE: 488.94, R^2: 0.3832), revealing that although the model provides a baseline prediction, its precision and trend sensitivity are lacking for more volatile and irregular datasets like JKH.

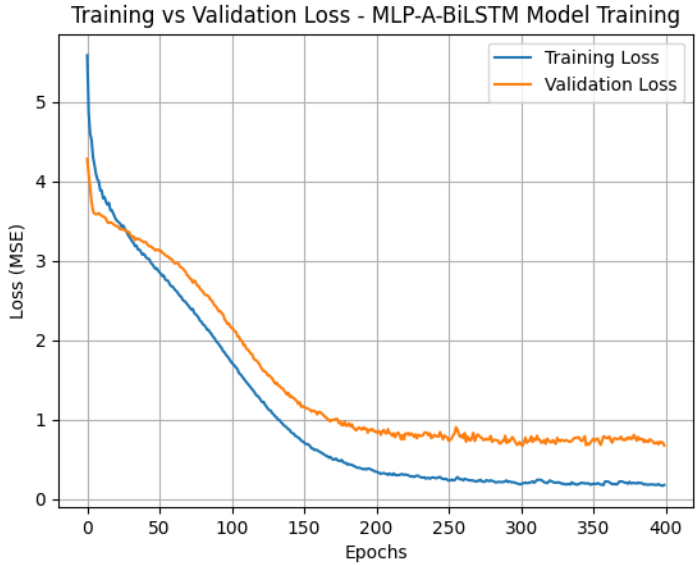


Figure 4-9 Training vs Validation Loss during BiLSTM Model Training - JKH

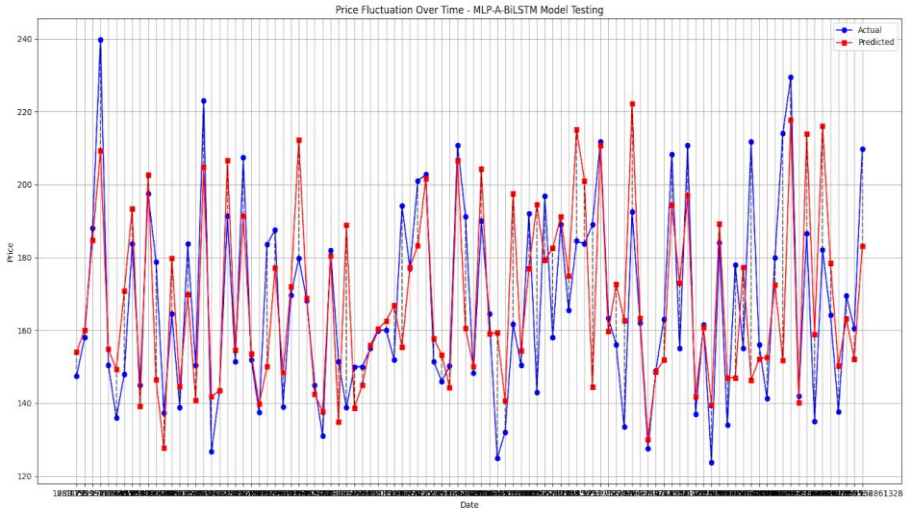


Figure 4-10 Actual vs Predicted over Time with BiLSTM Model Testing - JKH

In the *Figure 4-10*, graph displays the BiLSTM model's prediction for JKH. Compared to the Residual MLP model, the BiLSTM exhibits a slightly improved ability to follow the trajectory of actual prices, especially in moderately fluctuating regions. However, its prediction line still diverges significantly during sudden spikes or drops, and some

predictions are overly smoothed, missing the extremes. This reveals the model's limited ability to generalize temporal dependencies during high-volatility periods or abrupt news-driven events. The results are reflected in its evaluation metrics (MSE: 470.23, R^2: 0.4068), which indicate that the BiLSTM, while better equipped for time series, still underperforms due to a lack of contextual awareness and sensitivity to sudden market changes.

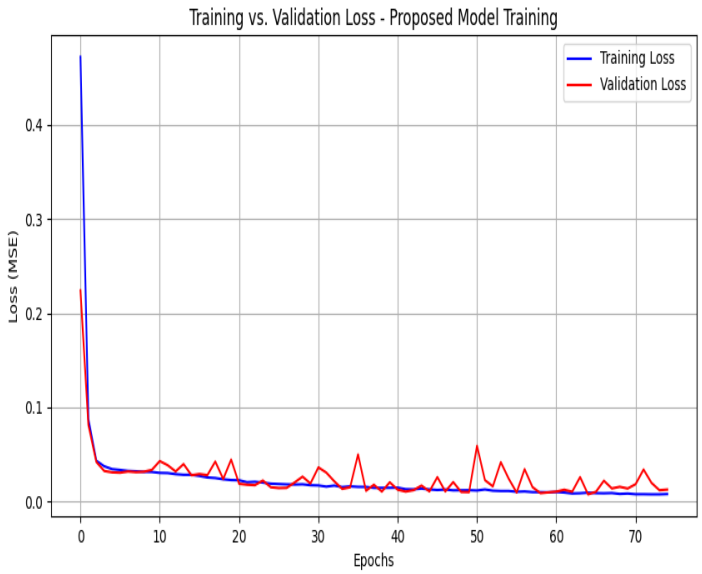


Figure 4-11 Training vs Validation Loss during Proposed Model Training - JKH

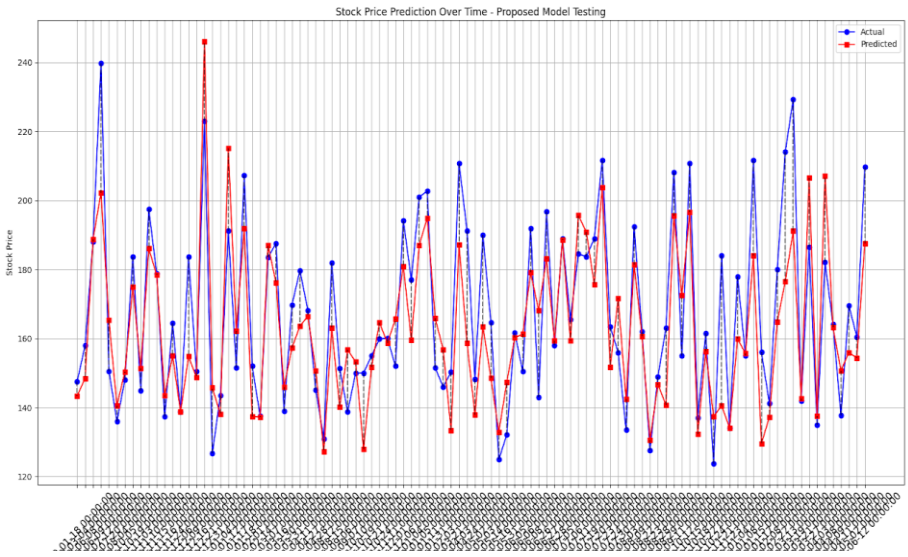


Figure 4-12 Actual vs Predicted over Time with Proposed Model Testing - JKH

Figure 4-12 represents the performance of the Proposed GNN model. Unlike the baseline models, the predicted price curve aligns far more closely with the actual prices. The GNN not only captures the general directional trend but also responds effectively

to short-term price fluctuations. It tracks local peaks and troughs with greater fidelity and exhibits fewer deviations even during erratic movements, which are common in illiquid market stocks like JKH. The model's ability to integrate semantic, sentiment, and temporal dependencies from financial news contributes to this superior performance. This is supported by a significantly improved MSE (223.26) and a strong R^2 value of 0.7884, reflecting both accuracy and reliability in prediction. The smoother but responsive pattern in the graph demonstrates the model's strength in adapting to noisy, sparse market data and showing dependencies that affect stock price movement.

Table 4-3 Model Performance Comparison Using Evaluation Metrics - BIL

Model	Stock	MSE	MAE	R^2
Residual MLP Model	BIL	1.6809	0.8495	0.5733
BiLSTM Model		1.7527	0.8190	0.5550
Proposed Model		0.6914	0.6436	0.8245

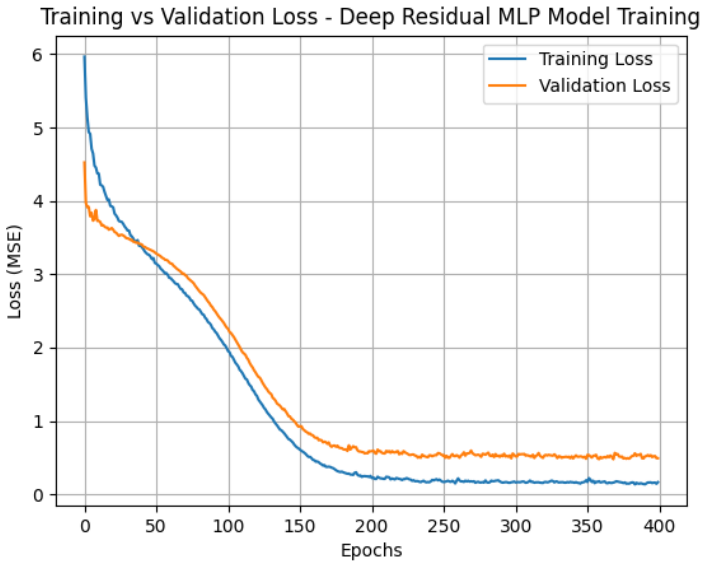


Figure 4-13 Actual vs Predicted over Time with Residual MLP Model Testing - BIL

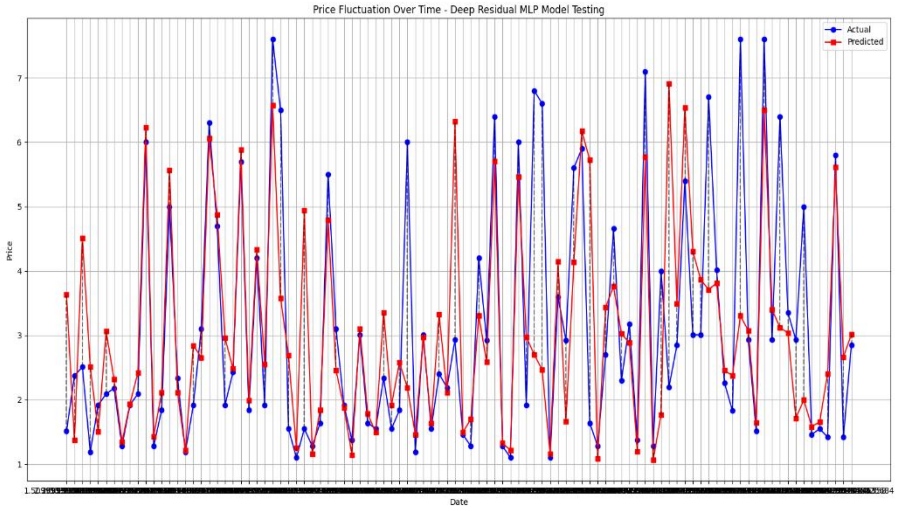


Figure 4-14 Actual vs Predicted over Time with Residual MLP Model Testing - BIL

The *figure 4-14* represents the Residual MLP Model prediction for BIL. The predicted prices (red) show moderate alignment with the actual prices (blue) in flat or less volatile periods but deviate substantially during price spikes. Given BIL's lower price range, even small deviations result in relatively high errors. The model appears to struggle with sudden fluctuations and cannot capture sharp peaks and troughs, leading to underfitting in volatile intervals. The performance metrics reflect this, with an MSE of 1.6809 and R^2 of 0.5733, indicating limited effectiveness for precision modelling of low-priced, high-volatility data.

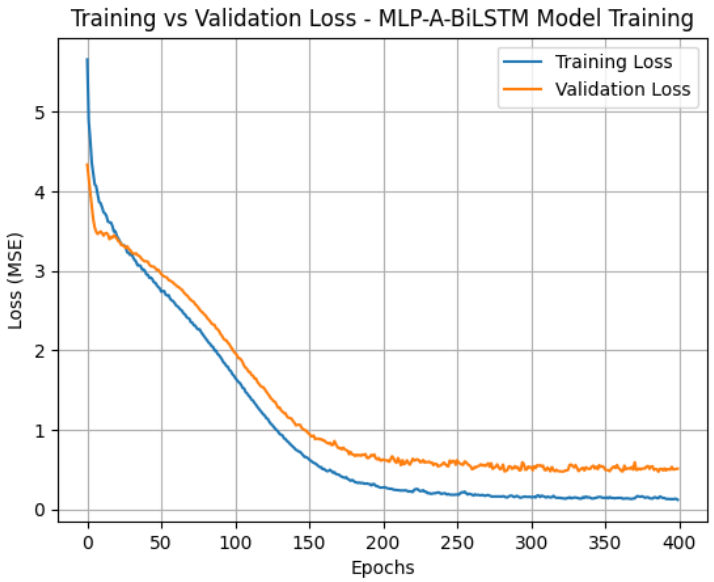


Figure 4-15 Actual vs Predicted over Time with BiLSTM Model Testing - BIL

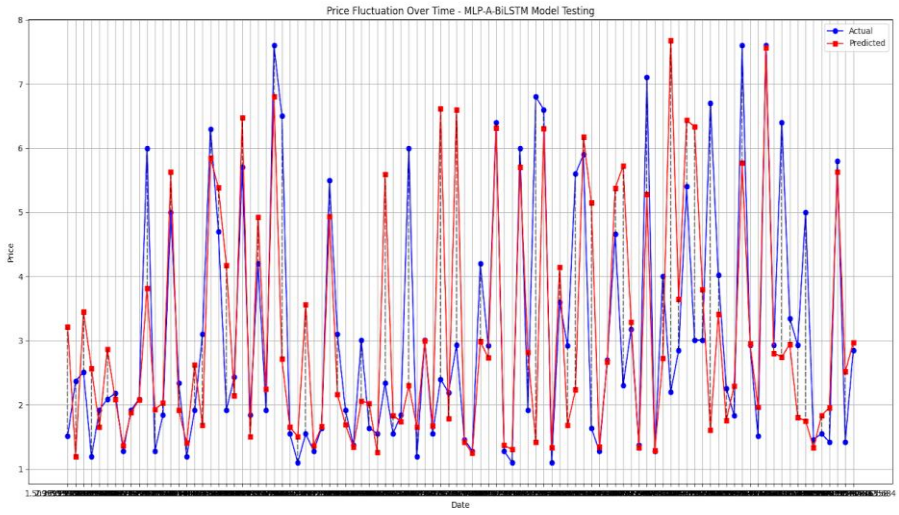


Figure 4-16 Actual vs Predicted over Time with BiLSTM Model Testing - BIL

Figure 4-16 shows the BiLSTM model's performance. While slightly better than the Residual MLP model in following short-term trends, the BiLSTM still fails to fully capture extreme price swings. Its predictions are often delayed or smoothed out, reducing its responsiveness to abrupt market-driven movements. Although it tracks gradual transitions better, it still lacks the sensitivity to handle sharp movements in BIL's price accurately. This aligns with the quantitative metrics (MSE: 1.7527, R^2: 0.5550), suggesting a marginal improvement but still not robust enough for reliable stock forecasting in such granular data conditions.

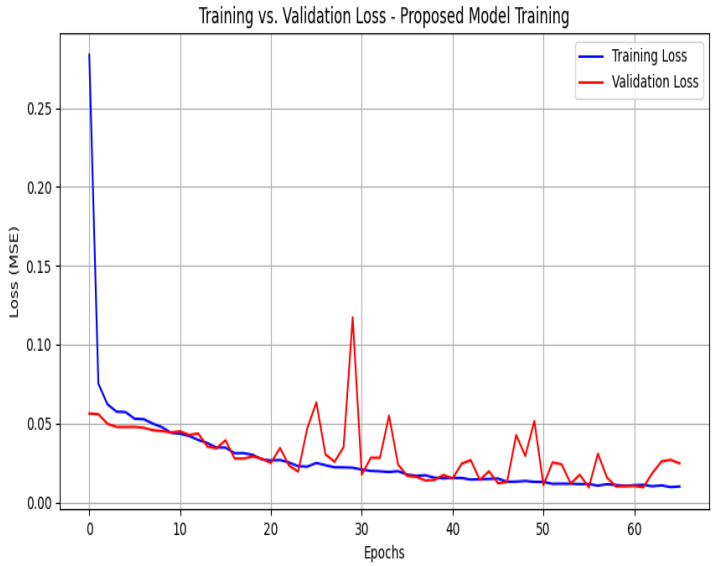


Figure 4-17 Actual vs Predicted over Time with Proposed Model Testing - BIL

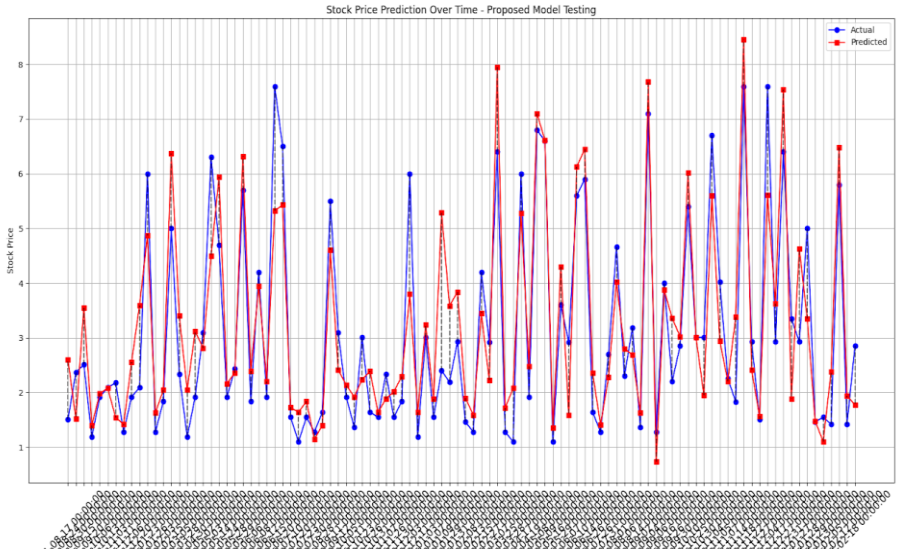


Figure 4-18 Actual vs Predicted over Time with Proposed Model Testing - BIL

In the *figure 4-18*, graph showcases the Proposed GNN model, which demonstrates a remarkable improvement in prediction accuracy. The GNN predictions (red) closely follow the actual stock price curve (blue), including quick peaks and drops. This model effectively adapts to the volatility and maintains precision despite the small numerical range of the stock price. Its ability to learn from interconnected features like semantic sentiment from financial news, temporal dependencies, and historical returns allows it to outperform traditional models. With an MSE of 0.6914 and a strong R^2 of 0.8245, the GNN exhibits high precision and low variance, making it highly suitable for forecasting low-value, high-sensitivity stocks like BIL.

4.3.1 Result Presentation for Unseen Data

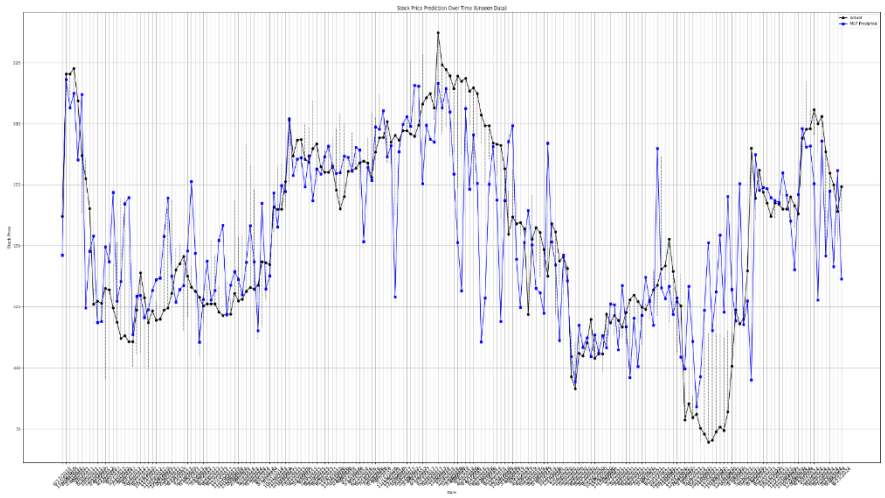


Figure 4-19 Actual vs Predicted over Time with Deep Residual MLP Model for Unseen Data - HNB

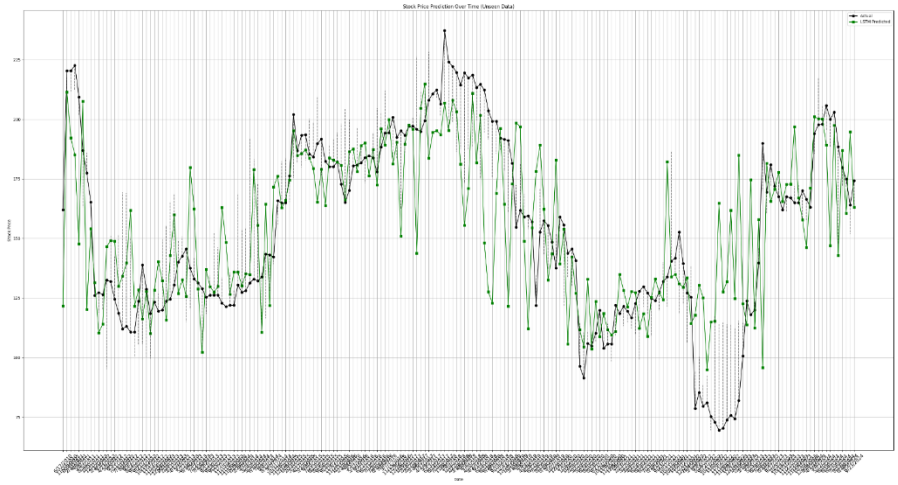


Figure 4-20 Actual vs Predicted over Time with BiLSTM Model for Unseen Data - HNB

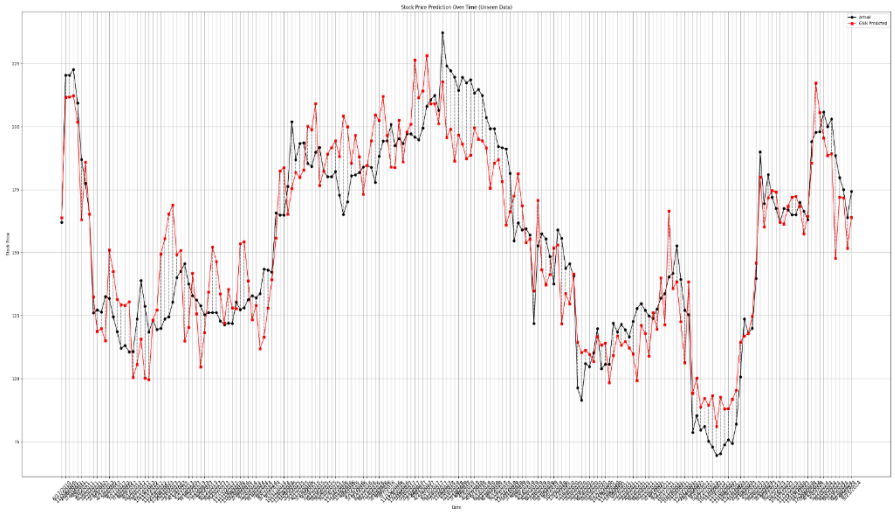


Figure 4-21 Actual vs Predicted over Time with Proposed Model for Unseen Data - HNB

Table 4-4 Proposed Model vs Baseline Models Performance Comparison Using Evaluation Metrics for Unseen Data - HNB

Model	MSE	MAE	R^2
Residual MLP Model	781.4211	19.0208	0.4811
BiLSTM Model	804.3276	20.6822	0.4530
Proposed GNN	434.9702	16.4557	0.7127

The *figure 4-19* depicts the performance of the Deep Residual Model (MLP) on previously unseen data. The actual stock prices (black) and the predicted values (blue) reveal considerable inconsistency throughout the prediction period. While the model occasionally tracks the general trend, it often fails to capture the magnitude of fluctuations, leading to significant gaps between the predicted and actual lines, especially in volatile sections. This erratic behaviour underscores the model's limited generalization ability in unfamiliar market scenarios, a common shortcoming in MLPs that lack temporal memory and context awareness. These visual observations align with the low R^2 score of 0.4811 and high MSE of 781.42, signalling poor predictive reliability under real-world conditions.

The *figure 4-20* illustrates the MLP-A-BiLSTM model's predictive performance on unseen data. The predicted line (green) follows the direction of the actual price movements (black) slightly better than the Deep Residual MLP Model, particularly in more stable regions. However, the model still struggles during high volatility and trend reversal zones. Sharp transitions in price are often smoothed out or lagged in the predictions, suggesting that although the BiLSTM captures some sequential dependencies, it fails to adapt to unexpected shifts or rare events not present in the training data. The performance metrics reflect this limitation, with an MSE of 804.32 and an R^2 of 0.4530, indicating marginal improvement but insufficient for robust forecasting in illiquid markets.

The *figure 4-21* displays the Proposed GNN model's performance on the same unseen dataset. Unlike the baseline models, the GNN predictions (red) demonstrate strong alignment with actual stock movements (black) across the timeline. The model successfully tracks both gradual trends and abrupt fluctuations, capturing the overall shape and direction of the price curve with minimal deviation. This visual performance is consistent with the significant improvements in evaluation metrics MSE of 434.97, MAE of 16.45, and a notably higher R^2 score of 0.7127. These results indicate that the GNN model generalizes well to unseen data, thanks to its ability to incorporate semantic and sentiment cues from financial news, along with temporal and structural dependencies learned from graph relationships.

5 RESEARCH EVALUATION

5.1 Analysis and Discussion

This research aimed to develop a robust, innovative approach for analysing and predicting stock price actions, specifically within the context of an illiquid market such as the Colombo Stock Exchange (CSE). The study compared the performance of a proposed Graph Neural Network (GNN) model against traditional predictive models, namely the Multilayer Perceptron (MLP) and Multilayer Perceptron – Attention - Bidirectional Long Short-Term Memory (MLP-A-BiLSTM) networks. The evaluation encompassed three representative stocks Hatton National Bank (HNB), John Keells Holdings (JKH), and Brown Investments PLC (BIL) to ensure generalizability and practical relevance of the results.

5.1.1 Performance Evaluation of the Proposed Model

The evaluation of the two baseline models Multilayer Perceptron (MLP), MLP-A-BiLSTM, and the proposed Graph Neural Network (GNN) revealed significant differences in their predictive performance on stock price data from an illiquid market. The results, summarized in Table *4-1*, *4-2* and *4-3* show that the proposed GNN model substantially outperformed the baseline models across all evaluation metrics.

The experimental evaluation of the three selected stocks HNB, JKH, and BIL clearly demonstrates the superiority of the proposed Graph Neural Network (GNN)-based model over the baseline models, used the Deep Residual MLP Model (Multilayer Perceptron) and the MLP-A-BiLSTM Model. Across all three stocks, the GNN achieved significantly lower Mean Squared Error (MSE) and Mean Absolute Error (MAE), while consistently obtaining higher R^2 scores, indicating a better fit to the true stock price data.

For HNB stock, the GNN model achieved an impressively low Mean Squared Error (MSE) of 210.0240 and Mean Absolute Error (MAE) of 11.0811, coupled with an R^2 score of 0.8600. In comparison, the Deep Residual MLP Model and BiLSTM demonstrated significantly lower predictive power, with MSE values of 701.5548 and 769.5390, and MAE scores of 18.9324 and 18.7846 respectively. These differences underscore the GNN's enhanced ability to capture and predict stock price movements, reflecting its effectiveness in modelling complex market dynamics.

The analysis further validated this trend with JKH stock data, where the GNN achieved an MSE of 223.2572, a notable improvement compared to the baseline models' MSE of 488.9472 (Deep Residual MLP) and 470.2347 (BiLSTM). The GNN's MAE was equally impressive at 11.9391, underscoring its precision in capturing average prediction deviations. Its high R^2 value of 0.7884 indicates superior explanatory power, capturing nearly 79% of the variability in the stock prices compared to approximately 36%-42% by baseline models.

For the BIL stock, characterized by narrower price fluctuations, the GNN maintained its predictive consistency and accuracy, evidenced by the lowest MSE (0.6914) and MAE (0.6436). The superior R^2 score of 0.8245 further confirmed its efficiency in handling even subtle market variations, a crucial feature for financial models operating in constrained and less volatile market scenarios.

5.1.2 Robustness on Unseen Data

Critical to assessing the practical utility of any financial prediction model is its ability to generalize effectively to unseen market data. The proposed GNN model excelled in this regard when evaluated on unseen data from HNB stock (which has elaborated in *figure 4-19*, *4-20*, *4-21* and *table 4-4*), maintaining robust performance metrics with an MSE of 434.9702 and MAE of 16.4557. Its R^2 score of 0.7127 substantially outperformed the baseline models, which recorded R^2 values of 0.4819 (Residual MLP) and 0.4662 (BiLSTM). And the prediction pattern of the proposed model is following the actual trend with least number of variations compared to Deep Residual MLP and BiLSTM. This substantial difference highlights the model's strength in dealing with unfamiliar, highly volatile market conditions typical of illiquid markets.

5.1.3 Technical Strength and Innovation

The profound performance of the GNN model can be attributed primarily to its innovative architecture and effective utilization of diverse data sources. Unlike traditional sequential and fully connected networks, the GNN employs a multi-dimensional approach integrating semantic meaning, sentiment, and temporal relationships. This analysis reveals that the baseline models struggled to generalize during periods of high market volatility or unusual trading patterns, common in illiquid markets. In contrast, the GNN demonstrated greater robustness and stability during these periods, likely due to its ability to incorporate richer contextual information through node features and edge relationships. The key strength of the GNN model lies in its ability to capture complex relationships between financial news and stock price movements, which are not easily modelled by traditional sequential models.

Specifically, the model captures semantic and sentiment similarities through cosine similarity edges linking news articles, enabling nuanced comprehension of market-driving information. The temporal dimension is encapsulated by creating sequential edges between consecutive trading days, allowing the model to track evolving market trends accurately. Moreover, the employment of scaled return values as edge weights, transformed through hyperbolic tangent functions, effectively moderates the impact of extreme values and ensures balanced learning from both positive and negative market movements. This unique combination enables the GNN to discern complex patterns, facilitating deeper insights into the intricate dynamics' characteristic of illiquid markets.

Unlike Residual MLPs and BiLSTMs, which treat input features independently or sequentially, the GNN uses graph-based learning to model,

- Semantic and sentiment similarity between news articles through edge connections based on cosine similarity.
- Temporal relationships between stock price data, using edges between consecutive trading days.
- Weighted influences via graph edges, where the return values are encoded as edge weights, allowing the model to learn not just from isolated features, but also from interdependencies over time and across news content.

5.1.4 Comparative Limitations of Traditional Models

The evident underperformance of traditional models (Residual MLP and MLP-A-BiLSTM) highlights their intrinsic limitations in the context of illiquid stock market predictions. The Residual MLP's lack of temporal sensitivity and the BiLSTM's constraints in capturing complex relational data resulted in inferior performance, particularly during volatile market conditions. Both models demonstrated substantial limitations in accurately interpreting and integrating qualitative sentiment-driven data, a crucial factor in stock price movements, especially in less predictable markets.

5.1.5 Contributions and Practical Implications

This research significantly contributes to the field of financial predictive analytics by validating the effectiveness of advanced NLP integration and graph-based learning for stock price prediction. The ability of the GNN to leverage multiple data dimensions semantic, sentimental, and temporal presents a robust analytical framework that is highly relevant for investors, analysts, and policymakers navigating illiquid market conditions.

Practically, this model provides market participants with a powerful tool to better anticipate stock price fluctuations, offering actionable insights that can improve investment decisions and market strategies. Furthermore, the research underscores the necessity of sophisticated analytical techniques in understanding market behaviour beyond traditional methodologies.

This structure enables the GNN to better understand contextual and event-driven impacts on stock prices, especially critical in illiquid markets, where traditional patterns are unreliable.

Overall, the results confirm that the proposed GNN model is well-suited for predicting stock prices in illiquid markets and by incorporating semantic relevance, sentiment information, and temporal connections, the GNN offers a more comprehensive representation of the underlying data. The findings highlight the potential for graph-based learning to provide improved predictive accuracy and insights in financial applications, particularly for illiquid markets.

5.1.6 Limitations and Future Directions

Despite achieving significant advancements in predicting stock prices using the proposed Graph Neural Network (GNN), several limitations inherent to the current research framework warrant further attention. One primary limitation arises from the quality and reliability of news data. News articles utilized in this study included numerous inaccuracies, incomplete information, politically biased content, or articles shaped more by speculation and gossip than factual reporting. This type of unreliable or biased news can negatively impact model predictions, introducing noise and undermining the model's ability to accurately capture genuine market dynamics. Moreover, the dataset used in the study did not explicitly differentiate between authentic, impactful financial news and speculative or manipulative content, including politically motivated narratives, market manipulations, or insider trading scandals. Such limitations pose risks to predictive accuracy, as the model may inadvertently incorporate misleading signals from these flawed sources.

Similarly, the accuracy of price action data posed additional challenges. Despite the Colombo Stock Exchange (CSE) being operational only on business days, the dataset contained erroneous price actions for weekends, thus potentially distorting genuine market behaviour and impacting predictive performance. Additionally, despite initially processing around 250,000 data entries, the stringent filtering and preprocessing methods necessary to ensure data quality resulted in a significantly reduced dataset of approximately 48,000 data points. Such data density constraints potentially restrict the comprehensiveness and robustness of the analysis, particularly when dealing with complex models like GNNs, which require substantial and high-quality data for optimal performance.

An intriguing limitation emerged regarding the model's performance differences across stocks with varying price scales. Stocks with higher price values typically led to larger Mean Squared Error (MSE) values, whereas stocks with smaller price ranges displayed lower MSE values despite possibly being harder to accurately represent. This discrepancy highlights a potential challenge as models may favour price ranges, influencing the generalizability and adaptability of predictions across diverse market conditions. Moreover, the current model's design did not account for the lagged impact of external events on market prices. Real-world market responses to external stimuli, such as significant news announcements or policy changes, are not always immediate; instead, they might manifest after delays. Capturing such lagged impacts is beyond the scope of this study and requires separate, dedicated, and more sophisticated modelling approaches.

Recognizing these limitations, several avenues for future research have emerged. Future studies could focus on capturing and analysing the lagged effects of news and external events using advanced pattern recognition and time-series methodologies. Exploring delayed market responses would provide deeper insights into market dynamics and investor behaviour, especially in illiquid and less predictable markets. Additionally, enhancing the existing GNN architecture by incorporating more sophisticated graph structures with additional nodes and edges could further refine

predictive accuracy and explanatory power. An enriched graph might include more granular representations of market factors, broader data integration, or more sophisticated sentiment analysis methods.

Moreover, future research could explore ensemble modelling, combining GNNs with other predictive architectures like transformers or hybrid deep learning models, to leverage their complementary strengths. Such ensemble methods could yield even higher prediction accuracy and robustness across diverse market scenarios. Finally, investigating the influence of weekend news on subsequent Monday market movements could provide valuable insights into how non-trading day information might still influence trading behaviours and stock prices, refining model capabilities and enhancing predictive precision in real-world financial forecasting scenarios. As additional features there could be technical indicators such as Moving Averages (MA), Relative Strength Index (RSI), Bollinger Bands, On-Balance Volume (OBV) and etc (explained in section 2.2). Moreover, future work could be carried out finding the shifted or lagged patterns of news impact to the stock price action.

6 RESEARCH CONCLUSION

6.1 Research Summary

This research investigated the effectiveness of advanced Graph Neural Network (GNN) architectures combined with sophisticated Natural Language Processing (NLP) methodologies in forecasting stock price movements, particularly targeting illiquid markets exemplified by the Colombo Stock Exchange (CSE). The core objective was to enhance the predictive capability of traditional models by integrating comprehensive financial news data and capturing the inherent complex dynamics present in illiquid stock markets. To achieve this, the study utilized state-of-the-art NLP techniques, including SBERT and FinGPT, to extract semantically meaningful and contextually relevant embeddings from financial news articles. Additionally, the sentiment polarity derived from the FinBERT model provided nuanced insights into market sentiment dynamics.

6.2 Methodological Contribution

The core methodological contribution of the research was the innovative construction of a graph-based representation of market dynamics. This graph encompassed several critical elements: nodes representing individual news articles, enriched with semantic embeddings and sentiment polarity; nodes representing daily stock trading activities, equipped with normalized closing prices and scaled price returns; edges connecting news articles based on semantic similarity; and edges linking news articles to corresponding stock nodes, weighted by a novel formula combining sentiment polarity and price returns. Furthermore, temporal continuity was modelled by connecting sequential stock nodes with edges weighted according to the scaled returns from previous trading days.

By encapsulating semantic context, sentiment insights, and temporal dynamics within a unified GNN framework, the research demonstrated superior predictive performance over baseline models such as Deep Residual Multilayer Perceptron (MLP) and Multilayer Perceptron – Attention Bidirectional Long Short-Term Memory (BiLSTM) networks. Experimental evaluations across multiple stocks listed on the CSE confirmed the robustness and generalizability of the proposed GNN model, significantly outperforming traditional methods across multiple metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Moreover, the model's ability to maintain high predictive accuracy even when tested against unseen data highlights its robustness and practical applicability.

Overall, this research underscores the substantial potential of integrating NLP-derived financial news features with graph-based learning architectures, offering a more comprehensive understanding of market behaviours in illiquid financial contexts. This combined methodological approach not only enhances prediction accuracy but also

provides a nuanced perspective on the intricate interplay between external events, market sentiment, and stock price movements.

The foremost contribution of this research lies in the innovative integration of NLP-derived features within a graph-based predictive model specifically tailored for illiquid stock markets. By combining semantic embeddings from SBERT and FinGPT with sentiment polarity extracted via FinBERT, the proposed methodology significantly surpasses traditional Residual MLP and BiLSTM models in predictive accuracy, which is evidenced through consistently lower error metrics and enhanced correlation scores. Additionally, the application of advanced dimensionality reduction techniques, particularly the Stacked Autoencoder (SAE), significantly optimized the model's computational efficiency without compromising the richness and accuracy of predictive features.

6.3 Achievement of Research Objectives

- Development of NLP-based Analytical Framework

The research successfully developed an NLP-driven analytical framework, demonstrating effective extraction and utilization of contextual and sentiment-based features from daily financial news articles. The integration of sophisticated NLP models such as SBERT, FinGPT, and FinBERT enabled comprehensive analysis of semantic nuances and sentiment polarity in financial texts. This foundational framework provided the basis for correlating public news sentiment with tangible stock price movements, significantly enriching predictive analysis.

- Evaluation of Text Vectorization Techniques

A detailed comparative study of multiple text vectorization methods, including TF-IDF, Doc2Vec, and advanced LLM embeddings, revealed the superior performance and nuanced feature extraction capabilities of LLM-based techniques. The extensive experiments indicated that embeddings derived from SBERT and FinGPT provided the richest semantic representation, while FinBERT added critical sentiment insights, significantly enhancing model accuracy and prediction reliability compared to traditional vectorization approaches.

- Integration of News Features with Stock Price Movements Using Deep Learning

Through the adoption of Graph Neural Networks (GNN), the study effectively modelled the complex interdependencies between news-derived sentiment, semantic meaning, and stock price dynamics. The constructed graph captured semantic similarity between articles, sentiment polarity from financial texts, and temporal continuity between stock price nodes. This innovative integration allowed the model to achieve higher predictive accuracy and robustness, as evidenced by consistently superior evaluation metrics (MSE, MAE, R^2) when compared to traditional Residual MLP and BiLSTM models.

- Assessment of Model Efficacy in Illiquid Market Contexts

The research rigorously assessed the efficacy of the proposed GNN framework specifically tailored to the illiquid market conditions of the Colombo Stock Exchange. Experimental results across various stocks (HNB, JKH, and BIL) demonstrated that the GNN significantly outperformed baseline models, particularly during periods of irregular trading activity and high volatility, conditions characteristic of illiquid markets. This demonstrated the model's adaptability and effectiveness within challenging market scenarios.

- Generalization of Framework for Broader Applicability

The developed analytical framework was designed with generalizability in mind, allowing easy adaptation to other stock indices and financial markets exhibiting similar liquidity constraints. The methodological rigor and comprehensive feature extraction approach provide a scalable foundation for future applications, significantly contributing to research areas where data scarcity, limited liquidity, and sparse historical analyses present significant obstacles.

In conclusion, this study effectively addresses each stated objective, demonstrating the viability and superior predictive capabilities of integrating advanced NLP techniques with Graph Neural Networks. The outcomes highlight the framework's significant potential for broader applications, fostering improved analytical capabilities and informed decision-making in illiquid market environments.

6.4 Practical Value and the Limitations

The developed GNN-based model presents considerable practical value for various stakeholders within illiquid markets, including investors, analysts, and regulatory bodies. Its ability to reliably predict stock price movements provides a critical decision-support tool that enhances informed decision-making, risk management, and strategic investment planning. Such precise forecasting tools not only benefit individual investors but also potentially contribute to broader market stability and increased investor confidence, especially within volatile and less predictable financial environments.

Despite achieving significant advancements, the research was subject to several constraints that affected the comprehensiveness and precision of the findings. Key limitations included data inaccuracies, specifically related to unreliable or incomplete news articles, including politically biased information, misleading gossip, and market manipulation events. Additionally, stock price data sometimes inaccurately contained weekend trading information despite market closures. The research also faced challenges stemming from relatively low data density; although initially extensive, data processing significantly reduced available samples. Furthermore, capturing the delayed or lagged impacts of external events on market movements remained beyond the scope of this study, necessitating further targeted research. The model's predictive performance heavily relies on the quality and comprehensiveness of the news

embeddings (SBERT and FinGPT vectors); any biases or gaps in the original textual data may directly impact prediction accuracy. Then the model does not explicitly account for macroeconomic indicators or global financial events, which can significantly influence stock price movements beyond the captured sentiment in news articles. Also the current GNN architecture requires careful feature engineering and hyperparameter tuning, which could lead to complexity and reduced generalizability if not appropriately managed. Additionally, the model's evaluation was limited to three selected stocks (HNB, JKH, and BIL) within the Colombo Stock Exchange, potentially restricting the applicability of findings to other markets or less liquid securities. Finally, computational complexity and resource requirements associated with training GNNs on large-scale datasets might limit their practical deployment in resource-constrained environments. Future research should address these limitations by integrating additional data sources, testing across broader market conditions, and further optimizing the architecture for improved scalability and robustness.

6.5 Future Work

Future research could substantially benefit from exploring the lagged market impacts of external events through advanced time-series modelling and pattern recognition techniques. Additionally, improving the existing GNN architecture by introducing more extensive and detailed graph components could enhance model accuracy. Exploring ensemble modelling approaches that combine the strengths of GNNs with other state-of-the-art predictive algorithms could provide further advancements in forecasting performance. Furthermore, investigations into the specific impacts of weekend news on subsequent trading days could yield valuable insights, enriching the analytical capabilities of stock prediction models.

In conclusion, this study effectively demonstrates the considerable advantages of integrating NLP techniques into graph-based models to accurately predict stock price fluctuations within challenging and illiquid market environments. By offering a robust methodological framework, this research not only advances academic understanding but also presents significant practical tools for financial analytics. The comprehensive approach outlined here highlights extensive potential for future innovations and contributes substantially to enhanced financial forecasting and strategic decision-making capabilities.

7 REFERENCES

[1] Moodi, F., & Jahangard-Rafsanjani, A. (2023). Evaluation of feature selection performance for identification of best effective technical indicators on stock market price prediction. *arXiv preprint arXiv:2310.09903*.

[2] Hani'ah, M., Abdullah, M., Sabilla, W., Akbar, S., & Shafara, D. (2023). Google Trends and Technical Indicator based Machine Learning for Stock Market Prediction. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(2), 271-284. https://doi.org/https://doi.org/10.30812/matrik.v22i2.2287

[3] Teixeira, D. M., & Barbosa, R. S. (2025). Stock Price Prediction in the Financial Market Using Machine Learning Models. Computation, 13(1), 3. Available online: https://www.mdpi.com/2079-3197/13/1/3

[4] Key technical indicators for stock market prediction. (2025). Expert Systems with Applications. https://www.sciencedirect.com/science/article/pii/S2666827025000143

[5] Tran, P., Pham, T. K. A., Phan, H. T., & Nguyen, C. V. (2024). Applying Machine Learning Algorithms to Predict the Stock Price Trend in the Stock Market – The Case of Vietnam. Humanities and Social Sciences Communications, 11, 393. Available online: https://www.nature.com/articles/s41599-024-02807-x

[6] "A Hybrid Stock Prediction Method Based on Periodic/Non-Periodic Features." (2024). EPJ Data Science. Available online: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-024-00517-7

[7] Noel, D. (2023). Stock Price Prediction using Dynamic Neural Networks. *arXiv preprint arXiv:2306.12969*.

[8] Phan, J., & Chang, H.-F. (2024). Leveraging Fundamental Analysis for Stock Trend Prediction for Profit. arXiv preprint arXiv:2410.03913.

[9] Huang, S., Capretz, L. F., & Ho, D. (2022). Machine Learning for Stock Selection Based on Fundamental Analysis. Expert Systems with Applications, 202, 117206. https://doi.org/10.1016/j.eswa.2022.117395

[10] Thompson, O., (2024). Factors influencing stock market prices: A comprehensive analysis. Academy of Accounting and Financial Studies Journal, 28(4), 1-2.

[11] Sun, J., & Hong, Y. (2021). Analysis of Stock Pricing Factors. Open Access Library Journal, 8(11), 1-16

[12] Eisler, Z., & Kertész, J. (2020). Liquidity and correlation in stock price changes: An empirical study of the Chilean stock market. arXiv preprint. https://arxiv.org/abs/2008.06168

[13] Kanniainen, J., & Yue, Y. (2019). The arrival of news and return jumps in stock markets: A nonparametric approach. arXiv preprint arXiv:1901.02691. Available at: https://arxiv.org/abs/1901.02691

[14] Budennyy, S., Kazakov, A., Kovtun, E., & Zhukov, L. (2022). New drugs and stock market: How to predict pharma market reaction to clinical trial announcements. arXiv preprint arXiv:2208.07248. Available at: https://arxiv.org/abs/2208.07248

[15] Rai, A., Luwang, S. R., Nurujjaman, M., Hens, C., Kuila, P., & Debnath, K. (2022). Detection and forecasting of extreme event in stock price triggered by fundamental, technical, and external factors. arXiv preprint arXiv:2206.13860. Available at: https://arxiv.org/abs/2206.13860

[16] Zamani, M., Paekivi, S., Meyer, P., & Kantz, H. (2022). Collective behavior of stock prices in the time of crisis as a response to the external stimulus. arXiv preprint arXiv:2205.06677. Available at: https://arxiv.org/abs/2205.06677

[17] Sen, Jaydip & Mehtab, Sidra. (2021). A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing. 10.36227/techrxiv.15023361.v1.

[18] Wang, Y., & Wang, Y. (2016). Predicting stock market using natural language processing. Asian Journal of Business and Accounting, 9(2), 1-22. https://doi.org/10.1108/AJB-08-2022-0124

[19] Wang, Y. and Wang, Y. (2016), "Using social media mining technology to assist in price prediction of stock market", 2016 IEEE International Conference on Big Data Analysis (ICBDA), 2016, pp. 1-4, doi: 10.1109/ICBDA.2016.7509794

[20] Kameshwari, S., Kaniskaa, S., Kaushika, S. and Anuradha, R. (2021), "Stock trend prediction using news headlines", 2021 IEEE India Council International Subsections Conference (INDISCON), pp. 1-5, doi: 10.1109/INDISCON53343.2021.9582219.

[21] Ji, X., Wang, J. and Yan, Z. (2021), "A stock price prediction method based on deep learning technology", International Journal of Crowd Science, Vol. 5 No. 1, pp. 55-72

[22] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D.C. (2019), "Stock price prediction using news sentiment analysis", 2019 IEEE Fifth

International Conference on Big Data Computing Service and Applications (BigDataService), IEEE, pp. 205-208.

[23] Sonkiya, P., Bajpai, V. and Bansal, A. (2021), "Stock price prediction using BERT and GAN", ArXiv, abs/2107.09055

[24] Cheng, W. and Chen, S. (2021), "Sentiment analysis of financial texts based on attention mechanism of FinBERT and BiLSTM", 2021 International Conference on Computer Engineering and Application (ICCEA), pp. 73-78, doi: 10.1109/ICCEA53728.2021.00022.

[25] Mane, O., & Kandasamy, S. (2022). Stock market prediction using natural language processing: A survey. International Journal of Computer Applications, 174(9), 1-8. https://doi.org/10.5120/ijca2022922062

[26] Gursoy, G., & Cakici, N. (2022). The Impact of Innovation News Coverage on Illiquid Stocks: The Case of U.S. Market. *European Journal of Innovation Management*. https://doi.org/10.1108/ejim-07-2022-0387

[27] Olaniyan, O., Obembe, O., & Akanbi, O. (2023). Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4, and logistic regression: A data-driven approach. Journal of Financial Data Science, 5(4), 65-77. https://doi.org/10.3905/jfds.2023.1.054

[28] Taylor, K., & Ng, J. (2024). Natural language processing and multimodal stock price prediction. arXiv preprint arXiv:2401.01487.

[29] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

[30] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1), 11-21.

[31] Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning (ICML-14), 1188-1196.

[32] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[33] Araci, D. (2019). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:1908.10063.

[34] Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-Source Financial Large Language Models. arXiv preprint arXiv:2306.06031. Retrieved from https://arxiv.org/abs/2306.06031

[35] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Conference on Empirical Methods in Natural Language Processing.

[36] Yao, L., Mao, C., & Luo, Y. (2019). Graph Convolutional Networks for Text Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 7370-7377. https://doi.org/10.1609/aaai.v33i01.33017370

[37] Huang, L., Ma, D., Li, S., Zhang, X., & Wang, H. (2019). Text level graph neural network for text classification. arXiv preprint arXiv:1910.02356

[38] Wu, Y., Liu, Y., He, H., & Liu, S. (2020). Stock Movement Prediction with Graph Attention Networks. Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM), 1993–1996. https://doi.org/10.1145/3340531.3412146

[39] Li, Z., Yang, S., Chen, Y., & Liu, J. (2022). A Systematic Review on Graph Neural Network-based Methods for Stock Market Forecasting. Expert Systems with Applications, 206, Article 117915. https://doi.org/10.1016/j.eswa.2022.117915

[40] Xiang, S., Cheng, D., Shang, C., Zhang, Y., & Liang, Y. (2023). Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction. arXiv preprint arXiv:2305.08740. Retrieved from https://arxiv.org/abs/2305.08740

[41] Qian, H., Zhou, H., Zhao, Q., Chen, H., Yao, H., Wang, J., Liu, Z., Yu, F., Zhang, Z., & Zhou, J. (2024). MDGNN: Multi-Relational Dynamic Graph Neural Network for Comprehensive and Dynamic Stock Investment Prediction. arXiv preprint arXiv:2402.06633. Retrieved from https://arxiv.org/abs/2402.06633

[42] Zhang, Y., Yu, X., Cui, Z., Wu, S., Wen, Z., & Wang, L. (2020). Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 334–339. https://doi.org/10.18653/v1/2020.acl-main.31