University of BRISTOL

DEPARTMENT OF COMPUTER SCIENCE

Graph-Enhanced LSTM for Improved Stock Price Prediction Incorporating Temporal and Price Relationships

Hao-Yu Tsai

MSc Financial Technology with Data Science.

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Monday 28th August, 2023

Supervisor: Dr. Xiang Li

Abstract

This dissertation presents the development of the Relational Rank Price LSTM (ReRaPrLSTM) framework, an innovative approach that merges company relation graphs and price correlation insights to enhance predictive financial modelling. Through the "Combine Mask" and "Addition Layer" approaches, I integrated these dimensions into the algorithm, with the "Addition Layer" proving transformative. Next is the threshold testing of stock correlation. It is for defining a relationship between stocks. My Threshold Testing identified optimal correlation thresholds for NASDAQ and NYSE markets. While performance varied across metrics, my comparison with previous algorithms highlighted ReRaPrLSTM's promising potential to reshape stock ranking practices. This experimental process reflects accomplishments and future prospects, positioning ReRaPrLSTM as a compelling avenue for predictive financial modelling advancements. The approach suggested in this research offers a promising direction for future research in the financial technology domain, particularly concerning the prediction of stock prices.

Dedication and Acknowledgements

I would like to take this opportunity to express my heartfelt gratitude to all those who have supported and guided me throughout the journey of completing this dissertation. Their unwavering assistance and valuable insights have been instrumental in shaping the quality and direction of my work.

First and foremost, I am deeply indebted to my supervisor, Dr. Xiang Li, whose exceptional guidance and mentorship have been invaluable. And all the teachers in Fintech programme throughout this year. Their expertise in the field and profound understanding of the subject matter have played a pivotal role in shaping the structure and content of this dissertation. Their insightful suggestions and constructive feedback were instrumental in refining my ideas and enhancing the overall clarity of my work. I am grateful for their dedication and patience in providing me with detailed instructions on how to improve my writing and how to approach complex concepts effectively.

Next, I would also like to extend my gratitude to my family and friends for their unwavering support and encouragement throughout this academic endeavour. Their belief in my abilities and their willingness to lend an ear during challenging times have been a constant source of motivation.

In conclusion, this dissertation is a culmination of collective efforts, and I am profoundly grateful to all those who have played a role, directly or indirectly, in shaping its trajectory. Your contributions have been invaluable, and I am humbled by the opportunity to have worked with such dedicated individuals. Thank you for being an integral part of my academic journey.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

Hao-Yu Tsai, Monday 28th August, 2023

Contents

List of Figures

List of Tables

Ethics Statement

This project fits within the scope of ethics application 6683-12073, as reviewed by my supervisor, Dr. Xiang Li

Supporting Technologies

- I used Tensorflow to support my implementation of the algorithm.
- I used GitHub and Google Drive to back up all works.
- I used Pycharm as an Integrated Development Environment(IDE).
- I used Tableau to generate the visualisation figures.

Chapter 1

Introduction

In the realm of contemporary advancements, the utilization of Long Short-Term Memory (LSTM), a form of recurrent neural networks (RNNs), has surged in popularity[10]. Its versatile application spans diverse domains, encompassing stock market prognostication, demand prediction, and sequential recommendation systems. Notably, LSTM's proficiency in discerning prolonged dependencies within time series data renders it particularly compelling for stock market prediction.

1.1 Motivation

My endeavour is rooted in a comprehensive appraisal of the existing landscape, notably the "Temporal Relation Stock Ranking" algorithm, recognized as a state-of-art algorithm for capturing inter-stock relationships within the real world[8]. This algorithm perceptively accounts for intricate dynamics, encompassing supplier-customer affiliations, parent-subsidiary hierarchies, and more. However, critical scrutiny unveils a significant limitation: the prevailing approach overlooks the dynamic undercurrents that imbue real-world relationships.

Indeed, the stock market realm is far from monolithic. Its participants, comprising astute traders and investors, do not solely rely on conventional affiliations when making impactful trading decisions. This intrinsic complexity underscores the need for a more nuanced paradigm that transcends the surface-level relationships and penetrates the cryptic strata driving stock market movements.

In this pursuit, I unearth the hidden dimensions beneath stock price movements. With acumen honed by experience, stock market traders comprehend that an amalgamation of factors orchestrates transactions. These multifaceted forces, encompassing market sentiment, speculative trends, and other intangibles, amalgamate into a tapestry that influences stock prices in ways that transcend overt relationships[16]. The motivation of this study is quintessentially rooted in capturing the elusive dynamics that only manifest through price movements. By grafting the lens of stock correlation onto analytical apparatus, I endeavour to unravel the enigma, pinpointing the concealed connections that underpin stock price fluctuations[18][19]. This incisive approach is designed to elevate the predictive accuracy of the algorithm, complementing the already robust "Temporal Relation Stock Ranking" framework.

In essence, I'm compelled by the imperative to enhance the algorithmic fabric, imbuing it with the power to decode the intricate symphony of factors that elude the conventional gaze. My research opens avenues for a more comprehensive and insightful stock market prediction model by fusing temporal relationships with the covert influences that govern stock prices.

The impetus behind this endeavour is twofold. Firstly, a tangible void exists in the current landscape of stock market prediction methodologies. As proposed in this research, the incorporation of graph-based insights stands to address this lacuna by offering a fresh perspective that can lead to more robust and accurate predictions. Secondly, the beneficiaries of this work extend beyond theoretical advancements. End-users reliant on informed stock market decisions and software developers seeking innovative avenues are poised to gain substantially from the refined predictive prowess that my approach aims to deliver.

As I embark on this exploration, I encounter pivotal challenges that underscore the significance of my research. The intricate and dynamic relationships within the stock market necessitate a stable approach capable of deciphering their complex web. My work not only grapples with this intricate network but also seeks to synergize it with the established prowess of LSTM models. By bridging these domains, I aim to bolster the stock market prediction landscape and usher in a new era of more precise and reliable forecasts.

1.2 Aims and Objectives

The project's aims and objectives can be briefly summarized as follows:

1. Enhance stock ranking algorithms by introducing an innovative price correlation graph analysis layer.
2. Extend the capabilities of existing state-of-the-art algorithms beyond company relation graphs to encompass price correlations.
3. Bridge the gap between historical patterns and market dynamics, providing a comprehensive and accurate predictive tool.
4. Harness the untapped potential of price correlation insights to elevate the algorithm's precision and relevance.
5. Seamlessly integrate price correlation insights into the existing algorithmic framework, overcoming technical complexities.
6. Contribute to a paradigm shift in stock ranking practices by combining the strengths of company relation graphs and price correlation insights.

The following chapters dive deep into the intricate mechanics of my novel algorithm, the methodologies applied, and the empirical revelations that have emerged from extensive experimentation. In doing so, this project transcends academia, becoming a beacon for practical innovation in an ever-shifting financial landscape.

Chapter 2

Background

2.1 Stock Prediction

Stock prediction has been a longstanding challenge within finance, as accurately forecasting market movements is complex and multifaceted. However, the emergence of machine learning techniques has opened up promising avenues for tackling this intricate task. By leveraging advanced algorithms and vast datasets, machine learning approaches offer the potential to uncover hidden patterns and relationships that could contribute to more precise predictions of stock price movements. These techniques can analyze historical market data, identify trends, and incorporate relevant factors, such as economic indicators, company performance metrics, and news and social media sentiment analysis. While pursuing accurate stock prediction remains inherently challenging due to financial markets' inherent volatility and unpredictability, machine learning promises to enhance my understanding and improve the accuracy of forecasting future stock trends.

One breakthrough in stock prediction has been the integration of LSTM networks[10]. LSTM networks are well-suited for capturing temporal dependencies in sequential data, making them ideal for time series forecasting like stock prediction. Matsunaga et al. explored the use of graph neural networks in combination with LSTM for stock market predictions using rolling window analysis[13]. In addition to stock market prediction, LSTM has also been applied in other domains. Abbasimehr et al. proposed an optimized LSTM model for demand forecasting[1]. Muzaffar and Afshari utilized LSTM networks for short-term load forecasting[15].

LSTM's potency lies in its ability to remember and utilize long-range contextual information, effectively avoiding the vanishing gradient problem that plagues traditional RNNs. LSTM cells possess gates that control information flow, allowing them to retain valuable historical context while accommodating the dynamic patterns prevalent in stock market data. Adopting LSTM networks in stock prediction signifies a departure from conventional methodologies. These networks incorporate historical price data, trading volumes, and other relevant features to predict future stock prices with greater accuracy and reliability.

Using LSTMs revolutionizes stock prediction by enabling models to discern intricate temporal relationships. LSTM's inherent memory and adaptability enable them to account for trends, seasonality, and irregular market behaviours. As a result, LSTM-based stock prediction models surpass conventional methods, yielding enhanced performance and making strides in capturing the intricate nuances of financial markets. This advancement heralds a new era in stock prediction, one where machine learning and neural networks synergize to deliver more accurate and informed forecasts.

2.2 Stock Correlation in Stock Trading

Stock correlation is a crucial and indispensable factor in stock trading, playing a pivotal role in influencing market behaviour, shaping trading strategies, and ultimately determining investment outcomes. This persuasive literature review underscores the significance of the stock correlation, shedding light on its profound importance in modern financial markets.

One of the paramount reasons to emphasize stock correlation lies in its role as a key driver of diversification and risk management strategies. Correlation allows traders and investors to ascertain the degree to which various stocks move in tandem or independently. Investors can construct portfolios by identifying stocks with correlations[7]. Blending assets with negative or low correlations can potentially reduce overall portfolio risk, providing a buffer against extreme market fluctuations. In this way, understanding stock correlation empowers investors to craft more robust and balanced portfolios, mitigating the adverse effects of market volatility[14].

Stock correlation serves as a guiding principle in the formulation of trading strategies. Whether it's pairs trading, correlation-based arbitrage, or dispersion trading, strategies built upon the relationships between correlated stocks rely on the notion that price movements are interconnected[7]. Pairs trading, for instance, exploits the relative mispricing of correlated stocks, aiming to profit from their historical price differentials[14]. Such strategies are only effective when the underlying stock correlation is understood and leveraged accurately. A strong grasp of stock correlation equips traders with the tools to identify potentially profitable opportunities and execute well-informed trades. Also, a study reveals the relationship between volume and stock return. As stock market trading volume increases, the day-to-day correlation of stock returns decreases, a trend explained by a model where risk-averse market makers adjust to non-informational trading pressures, leading to the implication that stock price drops on high-volume days more likely indicate a rise in the expected stock return than those on low-volume days[6].

Stock correlation provides valuable insights into broader market trends and sentiment. Correlation patterns can signal shifts in market dynamics, including investor sentiment, economic conditions, or sector-specific trends. Monitoring correlations can help traders anticipate potential market movements and adapt their strategies accordingly[6]. For instance, when correlations between stocks within a specific sector increase, it could indicate a broader macroeconomic trend affecting that sector. This information can guide traders in making timely and informed decisions, enhancing their ability to capitalize on emerging market trends.

Correlation aids in the assessment of risk exposure within portfolios. A high correlation between assets implies they will likely move in tandem, potentially amplifying losses during market downturns. Understanding the correlation matrix of a portfolio allows investors to identify concentrations of correlated assets and take measures to diversify further or implement risk-reduction strategies. Additionally, stock correlation is crucial in constructing effective hedging strategies[14]. Derivative instruments, such as options and futures, are often employed to hedge against correlated risks, and accurately assessing correlation enhances the precision of such hedging endeavours.

Stock correlation is a guiding force with far-reaching implications in the complex and dynamic stock trading landscape. From constructing resilient portfolios to devising profitable trading strategies, stock correlation is a cornerstone in the decision-making process of investors and traders. Its significance cannot be overstated, with the research presented in the theses affirming its role in risk management, trading strategies, and market insights. Embracing the importance of stock correlation is not just a choice; it is imperative for anyone seeking success and resilience in the captivating world of stock trading.

2.3 Implementing One Relational graph with LSTM

The quest for accurate stock price predictions has long been daunting in the dynamic and intricate world of stock markets. Conventional methodologies, predominantly reliant on LSTM models, have often faltered due to their inherent limitations in capturing the complex interplay of multifaceted market factors. However, with the advent of graph-based strategies, a promising avenue has emerged, holding the potential to significantly enhance the precision of stock market predictions[18]. In this trajectory, I embark on a comprehensive exploration of three distinct categories of graph-based methodologies, each poised to revolutionize stock market prediction: linear measurement-based graphs anchored in historical stock prices, linear measurement-based graphs hinged on trading volume dynamics, and innovative graphs derived from the insightful analysis of textual data.

A prevalent approach in constructing predictive stock market graphs involves harnessing linear measurements distilled from historical stock prices [12]. This intricate process entails conceptualising individual stocks as nodes within the graph's fabric. Subsequently, edges are meticulously woven between these nodes, their existence contingent upon predefined thresholds or correlation coefficients. Crucially, the intensity of correlation observed between distinct stock prices is the arbiter of the weight endowed upon each connecting edge.

The embrace of this linear measurement-based graph paradigm empowers me to encapsulate the intricate interdependencies interwoven between a diverse array of stocks, unveiling latent patterns concealed within the labyrinthine trajectory of price fluctuations. This approach is particularly illuminating in deciphering how modifications in the trajectory of one stock may propagate through the intricate network, influencing the behaviours of its correlated counterparts. The calculation of correlation coefficients can be instantiated by leveraging either the raw stock returns [12] or the logarithmic returns [4][5].

Expanding the dimensionality of stock market graphs through the infusion of textual data introduces a new horizon, enabling the extraction of meaningful stock relationships predicated on a multifaceted criterion. One such mechanism involves meticulous scrutiny of stock exchange documents to discern instances where stocks share common industry affiliations [8][21]. In cases where such affiliations materialize, an ethereal thread is woven between the corresponding stocks [17]. Additionally, a deeper dive into Wikidata's treasure trove of information facilitates the identification of primary relationships, often manifesting when one stock assumes the mantle of the subject while another adorns the role of the object within a given textual context. The beauty of this methodology lies in its ability also to unearth secondary relationships, wherein both stocks share a mutual object, albeit within divergent textual contexts.

Pioneering a trailblazing path, Gao et al. have presented a novel approach that harmoniously amalgamates the textual essence of stock description documents with the historical return data, ultimately giving birth to textual data-driven graphs[9]. At the core of this innovation lies the extraction of topic distributions from every stock description document. These extracted distributions metamorphose into distinctive features, each intertwined with the fabric of a corresponding stock attribute. The transformation continues as the word sequences within the document are seamlessly mapped onto a probability distribution spanning diverse topics, thereby birthing a unique document encoding. This encoding finds companionship with a historical sequence encoding, artfully procured through the orchestrations of an LSTM layer and the historical return data.

As these encodings intertwine, a dynamic interaction function emerges, a symphony that encapsulates the nuances of stock interactions. This intricate function is but the prelude to a symposium of further computations, ultimately culminating in determining time-evolving relation strengths between two stocks on a specific day. These relation strengths assume the role of edge weights, imbuing the constructed stock market graph with the essence of stock interdependencies. The infusion of textual information through this approach affords a panoramic portrayal of the dynamic interplay between stocks, engendering a more holistic comprehension and, consequently, more astute stock market predictions.

These innovative methodologies usher in a new era in stock market analysis by weaving the tapestry of textual data and unearthing the hidden relationships therein. The resultant stock market graphs, infused with multidimensional interdependencies, furnish a more exhaustive and nuanced representation of the market landscape. This newfound depth, in turn, bestows greater insights into the intricate relationships between stocks, thereby paving the way for more informed and accurate stock market predictions.

2.4 Multilayer Relation Graph Implementation

The realm of stock market analysis has evolved beyond individual stock performance to encompass a broader landscape of interconnected relationships. Multilayer graphs, an advanced representation, have emerged as a powerful tool in unveiling intricate connections within complex systems like the stock market[11].

A multilayer graph is a structured network that accommodates multiple layers of interconnected nodes, each corresponding to a distinct aspect of a system[20]. In the context of the stock market, this can encompass diverse relationships such as supplier-customer dynamics, industry affiliations, and more. These layers weave together to provide a comprehensive view of the intricate web of interactions that define the financial ecosystem[3]. Adopting multilayer graphs brings forth the potential to capture a deeper understanding of the hidden dynamics influencing stock market behaviours. I can delve beyond superficial relationships by embracing this concept and uncovering the latent connections that drive stock price movements. The multilayer approach acknowledges the multidimensionality inherent in financial systems, enhancing my ability to model and predict stock market trends.

By seamlessly integrating an additional layer that leverages correlations among stock prices, a remarkably intricate and all-encompassing perspective comes to the fore. This groundbreaking advancement adeptly exploits the underlying relationships that remain veiled within the fluctuations of stock prices. These concealed correlations, propelled by the ebbs and flows of market sentiment, speculative inclinations, and other enigmatic dynamics, bestow an added layer of depth to the predictive prowess of the algorithm. My adept utilization of multilayer graphs amplifies my approach's sophistication and empowers me to transcend the confines of conventional stock market prediction methodologies.

Through the fusion of LSTM's keen temporal awareness with the multilayer graph's remarkable capacity to encapsulate multifaceted connections, my algorithm is a true paradigm shift in stock prediction. This innovative amalgamation stands as the cornerstone of my extensive research efforts, impelling me to extract hitherto untapped insights and enhance the accuracy of predictions by assimilating the previously underappreciated intricacies embedded within the dynamic landscape of the stock market.

To sum up, in stock prediction, the LSTM algorithm has emerged as a promising technique. Concurrently, throughout the extensive history of stock trading, correlation has consistently played a pivotal role in influencing strategies and portfolio choices. To enhance existing methodologies, particularly the cutting-edge Temporal Relation Stock Ranking algorithm, I propose the integration of a novel element: a price correlation graph. Extending the algorithm's capabilities, I am exploring incorporating a multilayer graph that accommodates an expanded array of influential factors. This augmentation holds the potential to refine the algorithm's predictive prowess further, offering a more comprehensive perspective for decision-making in stock trading strategies and portfolio management.

Chapter 3

Project Execution

In this chapter, I delve into the execution phase of the project, outlining the main activities that constituted the work. The chapter is divided into sections, each focusing on a crucial aspect of the project's implementation and design. I begin by discussing the utilization of LSTM networks for stock prediction, subsequently delving into the enhancement of this approach by incorporating Relational Rank LSTM. Furthermore, I detail the innovative addition of a layer of masks, including the design of combined masks and multi-layer masks, followed by an exploration of the verification metrics employed.

3.1 Long Short-Term Memory(LSTM)

During the implementation phase, the project explored LSTM networks comprehensively, delving deeply into their mechanics and potential for stock price prediction. LSTM networks emerged as a potent solution due to their inherent ability to capture intricate temporal dependencies within sequential data[10].

At the heart of an LSTM network lies the memory cell, a dynamic entity responsible for retaining and propagating information across extended time intervals. Integral to the functionality of the memory cell are gating mechanisms: the forget gate, input gate, and output gate. These gates work in concert to control the flow of information into and out of the cell. The forget gate, integral to the process, orchestrates retaining or discarding information from the cell state. It considers both the previous hidden state and the present input, channelling them through a sigmoid activation function. This function yields a value ranging from 0 to 1 for each element within the cell state. This value stands as the arbiter, dictating the extent to which the corresponding element from the prior cell state is to be preserved in the ongoing cell state. A forget gate output of 0 heralds the complete omission of the specific element from the previous cell state, while an output of 1 translates to the total retention of the said element.

In parallel, the input gate operates, determining the infusion of new information into the cell state. Anchored in the previous hidden state and the present input, it too employs a sigmoid activation function. The outcome is a value, oscillating between 0 and 1, for each element within the cell state. This value holds sway over the extent to which the corresponding element within the candidate cell state is introduced to the ongoing cell state. The candidate cell state envisaged as a proposed fresh value for the cell state, emerges from the fusion of the present input and the prior hidden state. This is achieved through the application of a hyperbolic tangent function to a linear amalgamation of the said inputs.

Simultaneously, the output gate exerts its influence, dictating the information destined for extraction from the cell state. Informed by the previous hidden state and the ongoing input, it engages a sigmoid activation function. The outcome takes the form of a value, spanning the gamut of 0 to 1, for each element within the cell state. This value bears sway over the quantum of the corresponding element in the cell state, earmarked for extraction as the current hidden state. The output of the output gate, in turn, materializes through the transformation of the cell state by means of a tangent function. The output of the output gate subsequently scales this resultant value. The confluence of these operations ensures that the hidden state remains within defined bounds, and the selective filtration of information is executed, hinged upon the current input and the preceding hidden state.

Mathematically, the LSTM equations underpin the operations of these gates. The calculations involve intricate matrix multiplications and element-wise operations, processing both incoming data and the previous memory state. These equations encapsulate the intricate interplay between the gates, ultimately dictating the cell's evolution and the network's capacity to capture patterns across diverse time spans.

At each period of time t, an input vector $x^t \in \mathbb{R}^D$ represents the input at that period of time, where

D is the dimension of the input. The cell state vector c^t and the hidden state vector $h^t \in \mathbb{R}^U$ represent the memory and output of the LSTM at time-step t, respectively. Another vector, $z^t \in \mathbb{R}^U$, serves as an information transformation module. Vectors i^t, o^t, and $f^t \in \mathbb{R}^U$ represent the input, output, and forget gate.

The LSTM updates these components using the following equations[22]:

$$
\begin{aligned}
z^t &= \tanh(W_z x^t + Q_z h^{t-1} + b_z) \\
i^t &= \sigma(W_i x^t + Q_i h^{t-1} + b_i) \\
f^t &= \sigma(W_f x^t + Q_f h^{t-1} + b_f) \\
c^t &= f^t \odot c^{t-1} + i^t \odot z^t \\
o^t &= \sigma(W_o x^t + W_h h^{t-1} + b_o) \\
h^t &= o^t \odot \tanh(c^t)
\end{aligned}
$$

In these equations, $W_z, W_i, W_f, W_o \in \mathbb{R}^{U\times D}$ and $Q_z, Q_i, Q_f \in \mathbb{R}^{U\times U}$ are weight matrices, while $b_z, b_i, b_f, b_o \in \mathbb{R}^U$ are bias vectors. The equations describe the following steps:

1. The information transformation module z^t is computed by combining the input x^t, the previous hidden state h^{t-1}, and biases.

2. The input gate i^t controls the flow of information from z^t to the cell state c^t.

3. The forget gate f^t determines how much information should be retained in the cell state.

4. The cell state c^t is updated by blending the information from the input gate and the previous cell state.

5. The output gate o^t regulates the amount of information that can be outputted from the cell state.

6. The hidden state h^t is updated by applying the output gate to the cell state after passing it through a hyperbolic tangent function.

This updating process enables LSTM to capture long-term dependencies in sequential data, as the memory state allows information to persist over time. Unlike traditional RNNs, the memory state interactions in LSTM are linear, ensuring that information remains unchanged during the backward pass in training.

3.2 State-of-art Algorithm

The core basis of the Relational Ranking LSTM (ReRaLSTM) approach is the development of a fundamental stock relation graph[8]. This graph is skillfully constructed by leveraging two types of relationships among stocks. This segment thoroughly explains the detailed procedure behind crafting this relational graph, emphasizing its critical function in enhancing the predictive capabilities of ReRaLSTM.

Diverse Relationships: Wiki and Sector Relations Within the canvas of stock relationships, the paper introduces two fundamental types: Wiki relations and sector relations. These relationships are strategically culled to capture a diverse range of interconnections that underscore the dynamic landscape of stocks.

Wiki Relations: Extracted from the expansive landscape of corporate knowledge, Wiki relations are unearthed from the veritable treasure trove of Wikipedia pages belonging to various companies. These relations unveil the intricate web of affiliations that often remain veiled within textual narratives. For instance, if two companies share a supplier-customer rapport, this bond can be discerned from their respective Wikipedia pages and translated into a valuable thread woven within the stock relation graph.

Sector Relations: Capitalizing on the industry sector classification that categorizes companies, sector relations emerge as an integral facet of the stock relation fabric. When two companies find themselves within the same industry sector, a sector relation is acknowledged, signifying a shared identity within the broader economic landscape. This sector-level kinship serves as a potent indicator of potential interdependencies.

While the paper's emphasis gravitates around these two primary relationship types, it is noteworthy that these examples stand as exemplars within a diverse realm of stock relationships. The paper refrains from explicitly delineating every conceivable relationship, focusing instead on the overarching framework and methodology that accommodates these varied connections. The interplay of Wiki relations and sector relations infuses life into the stock relation graph. This graph stands poised to redefine predictive potential by harnessing the subtle threads of interdependence woven within the complex network of stocks.

The core of the Relational Ranking LSTM (ReRaLSTM) framework unfolds through an intricately woven architecture, meticulously designed to extract predictive insights from the interconnected fabric of stock data. This section delves into the layers that compose this innovative architecture, uncovering their significance in elevating ReRaLSTM's predictive prowess. The structure of ReRaLSTM is shown in Figure 3.1.

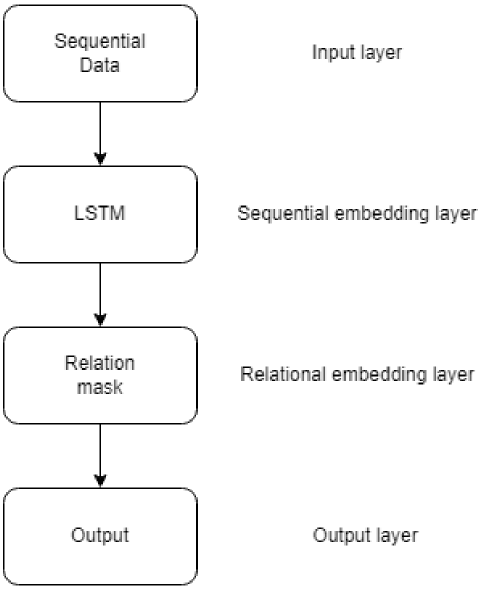

Figure 3.1: ReRaLSTM structure

Input Layer: Preprocessed Sequential Data The input layer lays the foundation of ReRaLSTM's architecture, a receptive canvas ready to assimilate the intricacies of sequential data. Within this data stream, a tapestry of crucial attributes is interwoven: the original closing prices, 5-day, 10-day, 20-day, and 30-day averages. This dynamic ensemble encapsulates the multi-faceted aspects of stock behaviour, establishing the groundwork for informed predictions.

Sequential Embedding Layer: LSTM Unveiled Seamlessly progressing from the input layer, the architecture unfolds into the sequential embedding layer, powered by LSTM. This layer breathes life into the sequential data, imparting the network with a nuanced understanding of temporal dynamics. While the intricacies of LSTM were expounded upon earlier, it's worth noting that this layer becomes the crucible wherein historical trends and temporal dependencies are distilled into predictive insights.

Relation Embedding Layer: Interconnected Learning The heart of ReRaLSTM's innovation pulsates within the relation graph layer. Here, the network delves into the intricacies of relationships interwoven between diverse stocks. This layer undertakes the intricate task of processing and comprehending the rich relational fabric that binds stocks together. The dynamic nature of these relationships is harnessed through Temporal Graph Convolution, imparting ReRaLSTM with the capability to recognize and capitalize on the collective wisdom inherent in stock interconnections.

Output Layer: Predictive Culmination As the architecture unfolds, culminating in the transformative exploration, the output layer stands poised to unveil the fruit of ReRaLSTM's labor. With insights gleaned from the depths of sequential data and the interconnected relationships among stocks, this layer consolidates its understanding into predictive outcomes. The convergence of meticulously processed data and relational insights yields predictions that transcend the realms of traditional methodologies.

In summary, the architectural tapestry of ReRaLSTM marries the intricacies of sequential data, the temporal understanding of LSTM, the transformative grasp of relation graph processing, and the predictive culmination of the output layer. This amalgamation reshapes the landscape of stock prediction, ushering in an era where the relationships between stocks are no longer relegated to the shadows but stand tall as vital determinants of forecasting precision.

3.3 Introducing the ReRaPrLSTM Algorithm

In the pursuit of enhancing the Relational Rank LSTM (ReRaLSTM) framework, a novel algorithm named Relational Rank Price LSTM (ReRaPrLSTM) was introduced, designed to seamlessly integrate the insights from stock price correlations into the predictive model. This section unveils the intricacies of the ReRaPrLSTM algorithm, dissecting its innovative approaches and detailing the strategies that underpin its transformative capabilities.

ReRaPrLSTM takes an ingenious leap, presenting two distinct avenues to embed the price correlation graph within the predictive model. The first approach, termed the "Combine Mask," orchestrates the fusion of two masks—the real-world relations mask and the stock price correlation mask. These masks, each a matrix of company quantity by company quantity, encode the presence or absence of relationships between stocks. The second approach introduces a pivotal layer before the real-world relation layer, where the activation function is driven by softmax, meticulously processing the price correlation graph data.

The Combine Mask approach represents the first avenue within ReRaPrLSTM, where two types of masks – the real-world relations mask and the stock price correlation mask – converge to harmonize relational insights. Figure 3.2 illustrates the proposed combined mask approach ReRaPrLSTM.

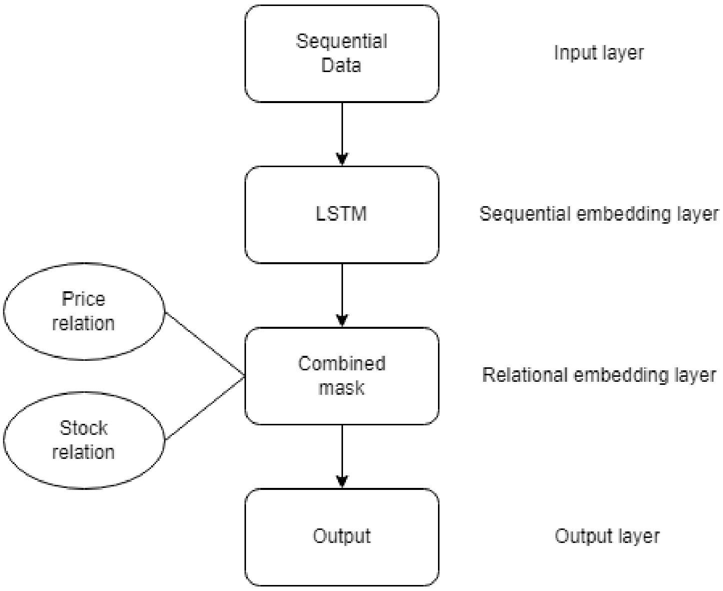

Figure 3.2: Combine mask approach structure

Rules for Combining Masks:

1. Strong Relationship (Value = 1): When both the real-world relation mask and the stock price correlation mask indicate a relationship between two companies, the resultant value is set to 1. This signifies a robust connection rooted in both real-world affiliations and shared price correlations.

2. Less Strong Bond (Value = 0): In scenarios where the real-world relation mask and the stock price correlation mask exhibit discrepancies, implying a less pronounced bond, the value is designated as 0. This nuanced interpretation acknowledges the variance between different dimensions of relationships.

3. No Relationship (Value = -1e9): When neither mask indicates a relationship, implying an absence of connections, the value assumes the placeholder of -1e9. This signifies an unequivocal lack of relational bonds between the two companies.

The Addition Layer approach marks the second avenue, introducing a pivotal layer before the real-world relation layer. This layer leverages the softmax activation function to process data from the price correlation graph meticulously. Figure 3.3 shows the proposed addition layer approach ReRaPrLSTM.

By applying the softmax function to the price correlation graph data, the model receives a transformed representation. This strategic step imbues the network with the capability to capture and distil

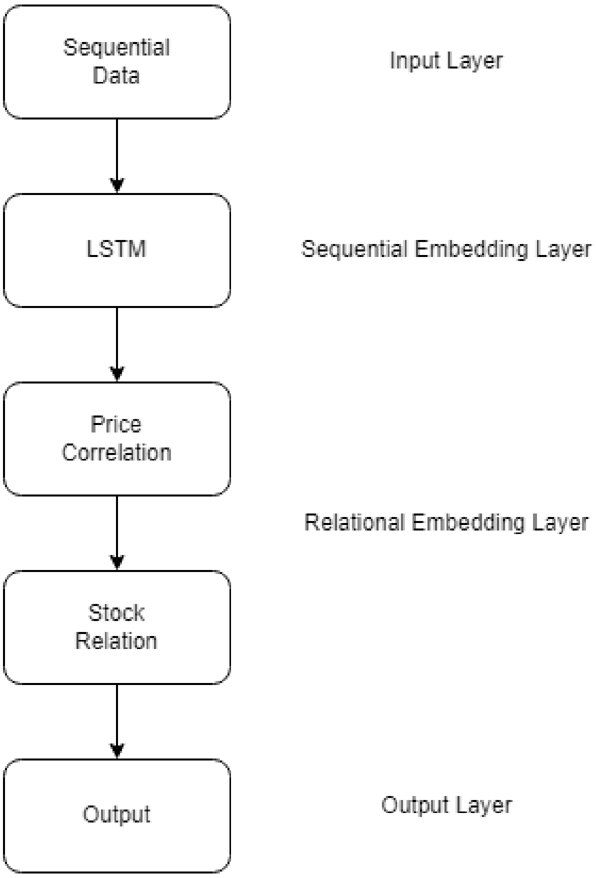

Figure 3.3: Addition layer approach structure

insights from stock price correlations. This layer operates as an intermediary, facilitating a more nuanced interpretation of the underlying relationships.

Within the intricate crossroads of choice, ReRaPrLSTM embarked on a definitive path. Based on the results of exhaustive testing and meticulous analysis, one approach will be embraced as the beacon to guide ReRaPrLSTM's transformative voyage in the latter chapters. This strategic decision echoes the dynamic spirit of innovation that characterizes ReRaPrLSTM's development.

In the panoramic landscape of stock prediction, the Relational Rank Price LSTM (ReRaPrLSTM) algorithm stands as a revolutionary testament. Through its dual methodologies, ReRaPrLSTM forges a powerful bond between the tangible realm of real-world relations and the intangible domain of stock price correlations. This algorithm is not only a reflection of a singular choice, but also a symbol of unity, weaving together disparate strands into an exquisite tapestry of predictive precision.

3.4 Dataset

The dataset utilized in this dissertation is provided by the author of "Temporal Relational Ranking for Stock Prediction" thesis. They encompass a collection of stock-related information from the National Association of Securities Dealers Automated Quotations(NASDAQ) and New York Stock Exchange(NYSE) markets. The time frame for this dataset spans from January 2, 2013, to December 8, 2017. The stocks included in the dataset have satisfied specific criteria to ensure their relevance and stability for analysis.

The data covers various aspects of each stock. The first is **Historical Price Data**, the daily closing prices for the stocks are available within the specified timeframe. These closing prices have been normalized by dividing them by the maximum value within the 2013-2017 period. Additionally, moving averages spanning 5, 10, 20, and 30 days have been calculated to capture different trends. Next is the **Sector-Industry Relations**, each stock has been categorized into a sector and an industry. This classification is based on the official company list maintained by NASDAQ Inc. Industry relationships have been established among stocks sharing the same industry node within the classification hierarchy. Last but not least, the **Wiki Company-based Relations**, collected using Wikidata, a comprehensive knowledge base, first-order and second-order relations among companies have been gathered. First-order relations indicate connections between companies in statements, while second-order relations involve companies that share statements referring to the same object.

3.5 Evaluation Metrics

The evaluation in this dissertation is conducted using three distinct metrics: Mean Squared Error (MSE), Mean Reciprocal Rank for Top-1 (MRRt), and Back Testing on Top 1 (BTl). Let's delve into the explanations of these metrics:

1. **Mean Squared Error (MSE)**: The Mean Squared Error quantifies the average squared difference between the predicted values and the ground truth values while considering the mask that indicates valid data points[2]. It measures the overall accuracy of the predictions. It computes the MSE as the squared Euclidean norm of the element-wise difference between the prediction and ground truth matrices, adjusted by the mask. The mask here is to determine the available data point. The sum of the mask then normalizes this value.

2. **Mean Reciprocal Rank for Top-1 (MRRt)**: The Mean Reciprocal Rank for Top-1 is a metric used in ranking scenarios in the previous work[8]. It focuses on identifying how quickly the first correct prediction is ranked among the top predictions. For each prediction column (representing a specific time period), it calculates the reciprocal of the position at which the actual best-ranked item appears in the list of predictions. The reciprocal rank is then averaged across all prediction columns, and adjusted for instances where there are no valid predictions due to the mask.

3. **Back Testing on Top 1 (BTl)**: Back Testing on Top 1 measures the cumulative return achieved by investing in the top-ranked stock prediction for each time period. It is included in the code of previous work[8]. It evaluates the effectiveness of the algorithm's top recommendation by simulating investment outcomes. For each prediction column, it accumulates the actual return of the top-ranked stock according to the ground truth. The cumulative return is computed over all prediction columns, providing insight into the potential profitability of following the algorithm's top pick.

These metrics together provide a comprehensive view of the algorithm's performance across different dimensions: its accuracy in terms of prediction error (MSE), its ability to identify valuable predictions (MRRt), and the financial potential of acting on the top-ranked prediction (BTl). By evaluating the algorithm's performance using these metrics, I can gain valuable insights into the algorithm's strengths and areas for improvement.

3.6 The best practices

This study implemented several ”best practice” strategies to ensure efficiency, maintainability, and successful outcomes. Below, I outline the key decisions and practices that guided my approach:

1. **Version Control with Git**: Recognizing the importance of tracking changes and maintaining a coherent project history, this study adopted Git as the version control system. Utilized GitHub as a centralized repository to store and manage my codebase. This decision allowed me to keep track of changes, revert to previous versions, and collaborate more effectively.

2. **Programming Language Choice (Python)**: this study carefully considered the programming language for its libraries, versatility, and support. Python emerged as the clear choice due to its extensive ecosystem, particularly in the fields of data analysis and machine learning. Python's popularity also ensures a wealth of resources and active community engagement, aiding independent work.

3. **PyCharm as the Integrated Development Environment (IDE)**: For coding, debugging, and project management, this study opted for PyCharm as my preferred IDE. PyCharm offers a range of tools that streamline coding processes and facilitate code organization. Its integrated nature enhances my productivity by providing a comprehensive environment for development tasks.

4. **Tableau for Visualization**: To present the findings in an impactful manner, this study leveraged Tableau for data visualization. Tableau's intuitive interface allowed me to create interactive and insightful visualizations that effectively communicate trends and insights extracted from the result data. This visualization component is critical for conveying complex information to a wider audience.

5. **Modular Code Structure**: To ensure easy maintenance and future updates, this study structured the codebase into modular components. Each module focuses on specific tasks such as data preprocessing, feature engineering, model training, and evaluation. This approach simplifies debugging, encourages code reusability, and promotes well-organized project management.

6. **Reproducibility**: To ensure that this work can be replicated, confirm to maintain a clear project structure and commented code. This way, anyone examining my code can understand its logic and purpose. By adhering to consistent coding practices, this study aims to ensure the project's reproducibility for myself and potential collaborators.

7. **Testing and Validation**: This study placed a strong emphasis on testing and validating the code and models. Implementing unit tests for critical functions and utilizing cross-validation techniques to evaluate model performance allowed this study to identify and rectify errors effectively.

8. **Model Interpretability**: While developing predictive models, this study made an effort to ensure their interpretability. Techniques such as feature importance analysis and visualizations helped me gain insights into model predictions. This interpretability is particularly important in financial contexts.

In making these decisions, this study weighed considerations specific to the author's project. For instance, while Jupyter notebooks are popular for interactive analysis, this study chose to work exclusively within PyCharm to align with the preferred workflow and maintain a consistent development environment. These project management strategies and technical choices, along with the use of Tableau for visualization, reflect the commitment to producing a well-structured, reliable, and successful project.

Chapter 4

Critical Evaluation

4.1 Methodology

A comprehensive assessment of the effectiveness of the Relational Rank Price LSTM (ReRaPrLSTM) framework was undertaken to ascertain its performance and determine the most promising direction for future experimentation. This evaluation was carried out systematically in three distinct phases.

During the initial phase, the ReRaPrLSTM framework's two key approaches, namely the "Combine Mask" and "Addition Layer" methods, were subjected to rigorous testing and comparative analysis. The primary objective was to identify the approach that seamlessly integrates real-world relations and stock price correlations, ultimately leading to enhanced predictive capabilities.

In the subsequent phase, an investigation was conducted to identify the optimal threshold for achieving the best performance. This involved testing different threshold values and assessing their impact on predictive accuracy.

Lastly, the third phase involved a comparative analysis. The best-performing outcome obtained from the ReRaPrLSTM framework was compared with results from previous methodologies. This final step aimed to highlight the advancements and improvements achieved by the ReRaPrLSTM framework in comparison to established approaches.

To carry out this evaluation, this study employed a robust methodology that entailed the following steps:

1. **Data Preparation**: A comprehensive dataset encompassing a diverse range of companies and their respective stock price histories was gathered. Additionally, correlation matrices were computed based on historical stock price data to capture the degree of relationships between companies.

2. **Graph Generation**: Leveraging the computed correlation matrices, graphs were constructed with varying correlation thresholds (0.9, 0.925, 0.95, 0.975, and 0.99). Companies with stock price correlations surpassing these thresholds were identified as having a relationship, forming the foundation for subsequent evaluations.

3. **Approach Testing**: The two distinct approaches within the ReRaPrLSTM framework—Combine Mask and Addition Layer—were systematically tested using the generated graphs. For each approach, the predictive model was trained and evaluated on the dataset, enabling a direct comparison of their performance.

4. **Threshold Testing**: My testing will encompass the examination of five different thresholds, each distinct from the others. Through this comprehensive analysis, this experiment aims to identify the threshold that exhibits the most exceptional performance. Once these evaluations are complete, it will make an informed selection of the most optimal threshold for both NASDAQ and NYSE, tailoring my choices to each specific market.

5. **comparative analysis**: This experiment conducted a comparative analysis with outcomes derived from previous methodologies. The intention was to ascertain the superiority of the proposed approach over existing methods.

6. **Performance Metrics**: To quantitatively assess the performance of the approaches, I employed established evaluation metrics, including mse, mrrt and btl. These metrics provided a holistic

understanding of how well each approach predicted the relationships and subsequently aided in selecting the most promising way for further investigation.

4.2 Evaluation of Relational Rank Price LSTM (ReRaPrLSTM) Approaches

In this section, I delve into the evaluation of the novel algorithm, Relational Rank Price LSTM (ReRaPrLSTM), specifically focusing on Part 1 of my experiment. This phase aimed to compare and analyze two distinct approaches: the Combine Mask approach and the Addition Layer approach, each designed to integrate stock price correlations into the predictive model. The ultimate objective of this analysis was to determine the more effective approach for proceeding to the next stage of the experiment.

To comprehensively assess the performance of the two approaches – Combine Mask and Addition Layer – this study conducted an extensive experimentation process. The experiment encompassed Market Types, Correlation Thresholds, Relationship Types and Repetitions. Two different market types were considered, resulting in a total of two distinct datasets. Five correlation thresholds (0.9, 0.925, 0.95, 0.975, and 0.99) account for variations in relationship strength. Each approach was tested with two types of relationships – sector and wiki – yielding a diverse range of scenarios. The entire experiment was repeated five times to ensure the robustness of the results.

Combine Mask Approach uniting Realities and Correlations. The Combine Mask approach was the initial strategy to fuse real-world relational insights and stock price correlations. This approach involved integrating two types of masks – the real-world relations mask and the stock price correlation mask. The combination of these masks followed specific rules to determine the strength of the company relationship.

Addition Layer Approach harmonizing through an activation function. The Addition Layer approach constituted the second strategy, introducing a crucial layer before the real-world relation layer. This approach applied the softmax activation function to the price correlation graph data. This transformation gave the model an altered representation, enabling it to more effectively capture insights from stock price correlations. This intermediary layer contributed to a deeper and more nuanced interpretation of the underlying relationships.

The performance comparison between the two approaches is illustrated in Figure 4.1, with the superior performing approach highlighted in orange.

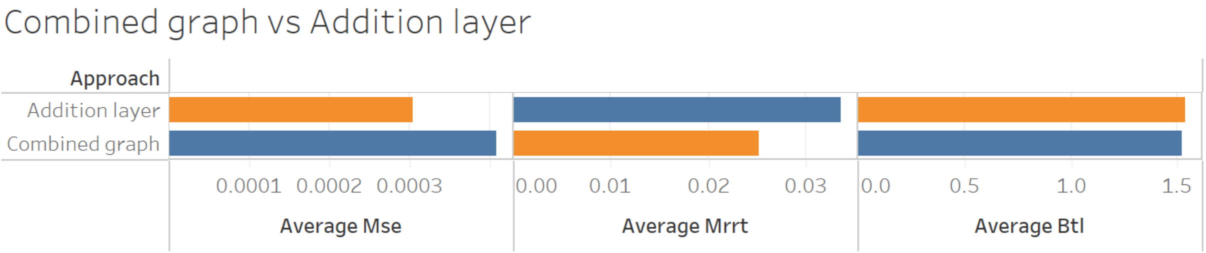

Figure 4.1: Two approach comparison

The comparison between the **Addition Layer approach** and the **Combined Graph approach**, as measured by Mean Square Error (MSE), Mean Reciprocal Rank of top1 (MRRt), and Backtest on top 1 (BTL) scores, distinctly showcases the superiority of the **Addition Layer approach**. With a lower MSE of 3.0447E-04 in comparison to 4.0894E-04 from the **Combined Graph approach**, the former demonstrates enhanced predictive accuracy. However, the **Combined graph approach** achieves a more favourable MRRt value of 3.3548E-02, indicative of improved precision in top-ranked predictions. In terms of Backtest on top 1 (BTL) scores, the **Addition Layer approach** excels with a value of 1.53718, surpassing the 1.51644 achieved by the **Combined Graph approach**. The detailed data is shown in Table 4.1. Collectively, these results substantiate the **Addition Layer approach** as the optimal choice, showcasing better predictive accuracy and better backtesting outcomes. Hence, the **Addition Layer approach** emerges as the preferred strategy for seamlessly integrating stock price correlations into the predictive model.

The observed superiority of the Addition Layer approach can be attributed to its inherent capacity to effectively leverage the price correlation graph data. By employing the softmax activation function, this approach transforms the raw correlation information into a refined representation, enabling the

Two approach comparison	mse	mrrt	btl
Addition layer	**3.0447E-04**	3.3548E-02	**1.53718**
Combined graph	4.0894E-04	**2.5082E-02**	1.51644

Table 4.1: Two approach comparison data

model to discern intricate relationships more accurately. The nuanced interpretation facilitated by the Addition Layer allows the model to capture and incorporate the subtle yet impactful nuances of stock price correlations.

In contrast, the Combined Graph approach, despite its innovative premise of integrating both real-world relations and stock price correlations, may have encountered challenges in reconciling conflicting information between the two masks. This could have led to instances where less pronounced relationships were falsely emphasized or strong relationships were overlooked due to inconsistencies. Additionally, the complexity of combining masks, each representing distinct dimensions of relationships, might have introduced noise into the model, resulting in less accurate predictions.

Furthermore, the superior performance of the Addition Layer approach can be linked to its capacity to adapt to variations in correlation thresholds and relationship strengths. The addition layer's transformation mechanism appears to be more robust in discerning meaningful patterns across a range of scenarios, thus yielding better predictive accuracy and top-ranked predictions.

In summary, the Addition Layer approach's success can be attributed to its ability to distil meaningful insights from the price correlation graph while handling variations in relationships effectively. The comparatively weaker performance of the Combined Graph approach might stem from its challenges in reconciling divergent information sources and complexities associated with mask combination.

4.3 Threshold Testing: Unveiling Optimal Parameters

In the relentless pursuit of advancing the predictive prowess inherent to the Relational Rank Price LSTM (ReRaPrLSTM) algorithm, the second part of my comprehensive evaluation embarks on a meticulous exploration of threshold testing. This pivotal phase has been intricately crafted to methodically assess the algorithm's performance under diverse correlation thresholds, with a specific lens trained on the intricate realms of the NASDAQ and NYSE markets. The overarching aim woven into this endeavour is to meticulously pinpoint the very threshold points that usher forth the most remarkable and unparalleled outcomes for each of these distinctive markets. By orchestrating this rigorous assessment, I am poised to unravel the precise junctures that promise to extract the utmost potential from the algorithm's predictive capacities. This discerning approach, grounded in empirical evidence, promises to usher in a pivotal breakthrough as I ultimately zero in on the quintessential threshold that holds the key to an elevated refinement of the algorithm's performance landscape.

In NASDAQ aspect. My analysis of the NASDAQ market involved testing five distinct correlation thresholds: 0.9, 0.925, 0.95, 0.975, and 0.99. By evaluating Mean Square Error (MSE), Mean Reciprocal Rank of top1 (MRRt), and Backtest on top 1 (BTL) scores for each threshold, I uncovered notable insights.

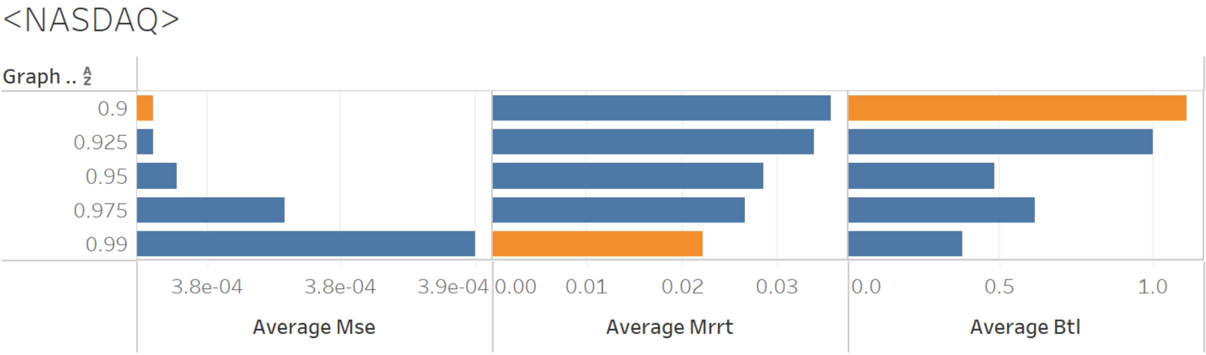

Figure 4.2: Threshold testing-NASDAQ

The colour orange in Figure 4.2 highlights the optimal performance of individual metrics. It is evident that for metrics MSE and BTL, the highest performance is achieved at threshold 0.9. Conversely, the threshold of 0.99 yields the best MRRt performance. The detailed data is shown in Table 4.2.

NASDAQ	mse	mrrt	btl
0.9	**3.7797E-04**	3.5616E-02	**1.11325**
0.925	3.7799E-04	3.3863E-02	1.00220
0.95	3.7886E-04	2.8510E-02	0.48175
0.975	3.8291E-04	2.6525E-02	0.61547
0.99	3.9004E-04	**2.2122E-02**	0.37904

Table 4.2: Threshold testing-NASDAQ data

Among the examined thresholds, a threshold of 0.9 emerged as the optimal choice for the NASDAQ market. This threshold demonstrated a competitive MSE of 3.7797E-04, signalling enhanced predictive accuracy. Although the MRRt score of 3.5616E-02 indicated the top-ranked predictions are relatively poor, the BTL score of 1.11325 reflected successful backtesting results. So I still chose 0.9 as the choice for the NASDAQ market.

After examining the NASDAQ market and turning my attention to the NYSE market, I subjected the same five correlation thresholds to rigorous evaluation. In this context, I observed a distinct trend in performance across the thresholds.

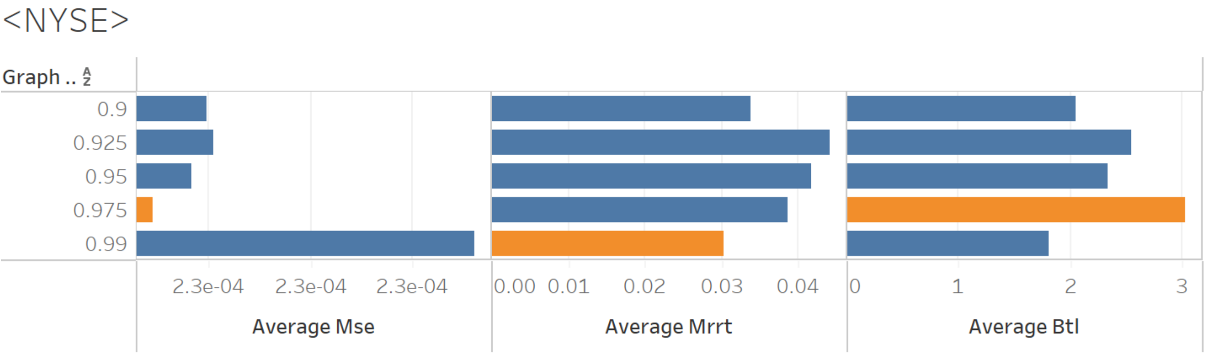

Figure 4.3: Threshold testing-NYSE

In Figure 4.3, the colour orange is used to mark the best performance points for each metric. Take a look at the metrics MSE and BTL – they both show their strongest performance when the threshold is set to 0.975. On the other hand, if I focus on the MRRt metric, the best performance happens at a threshold of 0.99.

NYSE	mse	mrrt	btl
0.9	2.2698E-04	3.3747E-02	2.04957
0.925	2.2705E-04	4.4164E-02	2.54831
0.95	2.2683E-04	4.1863E-02	2.33595
0.975	**2.2644E-04**	3.8687E-02	**3.03310**
0.99	2.2961E-04	**3.0389E-02**	1.81320

Table 4.3: Threshold testing-NYSE data

As shown in Table 4.3. Remarkably, the threshold of 0.975 surfaced as the most optimal choice for the NYSE market. Notably low MSE (2.2644E-04) showcased the algorithm's predictive accuracy. Similarly, the MRRt score of 3.8687E-02 demonstrated a high degree of precision in top-ranked predictions, while the BTL score of 3.03310 underscored successful backtesting outcomes.

In summary, these threshold testing outcomes accentuate the necessity of market-specific tailoring in algorithmic approaches. The stark differences in optimal thresholds between NASDAQ and NYSE illuminate the nuanced characteristics inherent to each market. My findings not only inform the operational decisions within the ReRaPrLSTM algorithm but also offer a glimpse into the intricate interplay between correlations and market behaviours.

The overarching trend of the 0.99 threshold's superior MRRt performance in both the NASDAQ and NYSE markets highlights a critical insight. The emphasis on strong relationships through a higher correlation threshold serves as a unified key to unlocking predictive precision. By selecting stocks that exhibit robust correlations, the algorithm bypasses noise and focuses on the core drivers of market dynamics. This not only leads to improved MRRt scores but also reinforces the algorithm's ability to make informed predictions. However aside from the outstanding performance of mrrt, the 0.99 threshold did not perform well in the other two metrics, therefore it's been terminated.

As I move forward, empowered by the depth of insights gleaned from my rigorous experimentation and analysis, I stand poised to embark on a pivotal phase of my research exploration – a comparison that holds the potential to illuminate the true prowess of my proposed Relational Rank Price LSTM (ReRaPrLSTM) algorithm. Armed with an enriched understanding of how ReRaPrLSTM harnesses the symbiotic relationship between real-world relations and stock price correlations, I am primed to undertake a comprehensive assessment that contrasts my innovation with previous works in the field of predictive financial modelling.

4.4 ReRaPrLSTM vs. Previous Algorithms

In this final phase of my critical evaluation, I draw the curtains back on a pivotal comparison that sheds light on the prowess of the proposed Relational Rank Price LSTM (ReRaPrLSTM) algorithm. By juxtaposing ReRaPrLSTM's performance with that of two previous algorithms, namely Rank LSTM and Relation Rank LSTM, I unveil nuanced insights that illuminate the distinctive contributions and capabilities of my innovation.

In NASDAQ aspects, moving to the NASDAQ market, intriguing dynamics come to the forefront. Here, the 0.9 correlation threshold emerges as the standout performer in terms of Mean Square Error (MSE), setting ReRaPrLSTM apart. This threshold underscores the algorithm's ability to capitalize on stronger correlations, leading to more accurate predictions. Despite a marginal increase in MSE, ReRaPrLSTM consistently delivers competitive MRRt and the best BTL scores, solidifying its position as a reliable predictive tool. In Figure 4.4, the colour orange is used to mark the best performance points for each metric.

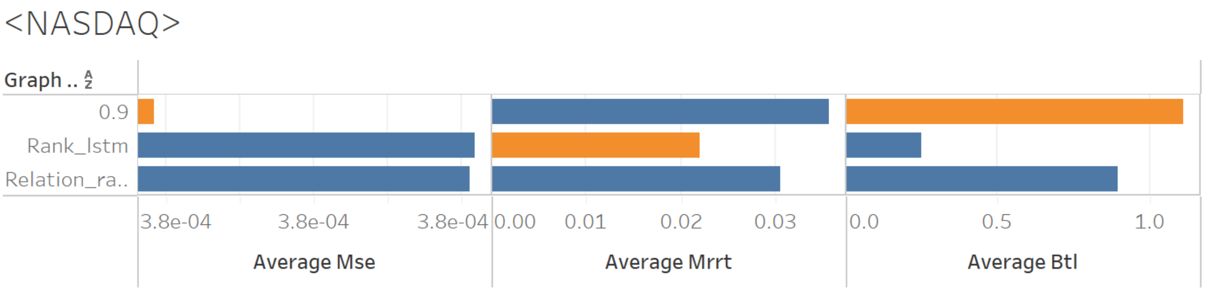

Figure 4.4: Comparison with former algorithm-NASDAQ

NASDAQ	mse	mrrt	btl
0.9	**3.7797E-04**	3.5616E-02	**1.11325**
Rank_lstm	3.7884E-04	**2.1920E-02**	0.25130
Relation_rank_lstm	3.7882E-04	3.0529E-02	0.89939

Table 4.4: Comparison with former algorithm-NASDAQ data

I uncover a landscape characterized by intriguing dynamics that unveil a deeper understanding of ReRaPrLSTM's performance. Here, an illuminating phenomenon emerges as the 0.9 correlation threshold takes centre stage, asserting itself as the standout performer in terms of Mean Square Error (MSE). This distinctive threshold choice demonstrates ReRaPrLSTM's adeptness in leveraging stronger correlations to its advantage, resulting in predictions that are remarkably accurate. This can be attributed to the algorithm's innate ability to recognize and capitalize on correlations that carry more significant weight in influencing stock behaviours. The 0.9 threshold's ability to enhance the algorithm's accuracy while maintaining a marginal increase in MSE underscores the delicate balance that ReRaPrLSTM strikes between predictive precision and the utilization of correlation insights.

The data in Table 4.4 shows that the significance of the 0.9 threshold goes beyond mere statistical values; it reflects the algorithm's capacity to decipher the underlying dynamics of the NASDAQ market. By selectively embracing correlations with heightened strength, ReRaPrLSTM demonstrates an aptitude for capturing critical relationships that bear a substantial impact on stock behaviours. This translates into an enhanced predictive capacity that not only anticipates market movements but also uncovers the interwoven fabric of interactions that shape these movements. The marginal increase in MSE stands as a testament to the algorithm's deliberate approach of trading off slight deviations in prediction for a broader and more nuanced understanding of the market's intricacies.

In the context of the NYSE market, my algorithm's performance is epitomized by the 0.975 correlation threshold. The achieved MSE of 2.2644E-04 showcases ReRaPrLSTM's enhanced predictive accuracy, a feat rivalled only by Rank LSTM with a marginally lower value of 2.2611E-04. In the MRRt and BTL scores, ReRaPrLSTM exhibits a balance of precision and practicality, outperforming both previous algorithms. Figure 4.5 illustrates the best performance of each metric with the colour orange.

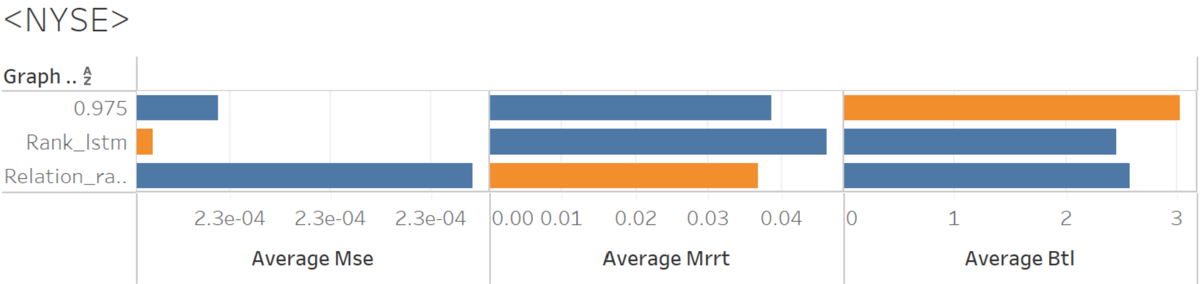

Figure 4.5: Comparison with former algorithm-NYSE

NYSE	mse	mrrt	btl
0.975	2.2644E-04	3.8687E-02	**3.03310**
Rank_lstm	**2.2611E-04**	4.6265E-02	2.45848
Relation_rank_lstm	2.2771E-04	**3.6844E-02**	2.58038

Table 4.5: Comparison with former algorithm-NYSE data

A different facet of ReRaPrLSTM's capabilities takes centre stage, symbolized by the 0.975 correlation threshold shown in Table 4.5. In this context, ReRaPrLSTM's performance shines as a beacon of enhanced predictive accuracy. The achieved MSE of 2.2644E-04 stands as a testament to the algorithm's prowess in minimizing prediction errors, a feat that rivals only the Rank LSTM algorithm with its marginally lower value of 2.2611E-04. This convergence of MSE values is indicative of the close competition between ReRaPrLSTM and Rank LSTM in terms of predictive precision.

The story doesn't end with MSE; it expands to encompass the domains of MRRt and BTL scores. In these metrics, ReRaPrLSTM showcases an exquisite balance between precision and practicality. The MRRt score reflects the algorithm's ability to accurately rank influential stocks, while the BTL score underscores its practical application in real-world trading scenarios. These scores reaffirm ReRaPrLSTM's aptitude to not only anticipate stock behaviours but also offer actionable insights that can be translated into successful trading strategies.

To sum up, the observed performance disparities can be attributed to a confluence of factors. The choice of correlation threshold significantly influences the algorithm's ability to capture meaningful relationships within the dataset. In the NASDAQ market, the 0.9 threshold excels in MSE potentially due to its emphasis on stronger, more influential correlations, while maintaining respectable precision. Similarly, the NYSE market benefits from the 0.975 threshold's balance between accuracy and relationships.

As I reflect on these comparisons, a recurring theme emerges: ReRaPrLSTM consistently ranks among the top performers or closely competes with previous algorithms across various metrics. This robust performance underscores the algorithm's adaptability to diverse market conditions, its capacity to navigate intricate relationships, and its propensity to make well-informed predictions.

Chapter 5

Conclusion

5.1 Revisiting Remarkable Contributions

As I adopt a broader perspective, allowing the expansive scope of this dissertation to envelop my contemplation, a profound realization of paramount significance comes to the fore: it unequivocally transcends the realm of mere textual composition. Instead, it stands as an unequivocal testament to the indomitable spirit of innovation. I started by creating the Relational Rank Price LSTM (ReRaPrLSTM) framework. This framework connects real-world relationships with stock price correlations in a unique way.

During my study, I carefully explored the ReRaPrLSTM framework. This framework blends different ideas together into a cohesive whole. I delved into the "Combine Mask" method, which creatively combines various information dimensions. Additionally, I worked with the "Addition Layer" technique, which skillfully integrates different components into my evolving predictive model.

As I progressed, I encountered a section named "Threshold Testing." This part sheds light on a crucial point in my experiment. During this stage, the ReRaPrLSTM framework demonstrated its flexibility and adaptability by effectively handling the complex details present in the NASDAQ and NYSE markets. The exploration of optimal thresholds during this phase transcended numerical refinement, metamorphosing into a nuanced calibration that spoke volumes of profound comprehension of market intricacies. The framework's inherent flexibility in tailoring predictions based on these thresholds underscored its dynamic nature, definitively establishing it as a robust instrument within the capricious realm of financial markets.

5.2 Reflection and Evaluation of Project Realization

As I finish this exploration and innovation, it's appropriate to look back on the goals I set at the beginning of this endeavour. These aims guided me through the labyrinthine realm of predictive financial modelling, urging me to seek out uncharted territories and to weave together disparate threads into a cohesive narrative.

1. **Enhance stock ranking algorithms by introducing an innovative price correlation graph analysis layer**. The process of enhancement commenced with the introduction of the "Combine Mask" and "Addition Layer" approaches, infusing my proposed Relational Rank Price LSTM (ReRaPrLSTM) framework with a novel price correlation graph analysis layer. The "Combine Mask" approach paved the way for a harmonious fusion of real-world relations and stock price correlations, while the "Addition Layer" approach harnessed the transformative power of activation functions. These innovations not only expanded stock ranking algorithms but also opened doors to a deeper understanding of the interplay between market dynamics and correlations.

2. **Extend the capabilities of existing state-of-the-art algorithms beyond company relation graphs to encompass price correlations**. In my pursuit of extending capabilities, I embarked on an endeavor that bridged the gap between historical patterns and market dynamics. By seamlessly integrating price correlation insights into the ReRaPrLSTM algorithm, I not only expanded the horizons of existing state-of-the-art methodologies but also demonstrated the potential for algorithmic models to evolve with the changing landscape of financial markets.

3. **Bridge the gap between historical patterns and market dynamics, providing a comprehensive and accurate predictive tool**. The culmination of my efforts has resulted in the creation

of an algorithmic tool that brings together historical patterns and market dynamics. ReRaPrLSTM stands as a testament to the efficacy of combining historical context with the nuances of price correlations. This union has birthed a comprehensive and accurate predictive tool that resonates not only with my initial aspirations but also with the needs of contemporary financial analysis.

4. **Harness the untapped potential of price correlation insights to elevate the algorithm's precision and relevance**. The integration of price correlation insights has proven to be a transformative endeavour. By harnessing the untapped potential of correlations, I have elevated the precision and relevance of ReRaPrLSTM. The resulting algorithm not only generates predictions but also provides a deeper understanding of the interwoven relationships that drive market behaviour.

5. **Seamlessly integrate price correlation insights into the existing algorithmic framework, overcoming technical complexities**. The path of innovation was not without its challenges. Technical complexities were encountered and surmounted as I undertook the task of seamlessly integrating price correlation insights into the ReRaPrLSTM framework. This process was a testament to the resilience and determination required to bring forth transformative ideas into practical implementation.

6. **Contribute to a paradigm shift in stock ranking practices by combining the strengths of company relation graphs and price correlation insights**. My endeavor has not only been about embracing successes but also acknowledging the lessons learned from challenges. While my endeavour sought to seamlessly unite the strengths of company relation graphs and price correlation insights, it's important to recognize that ReRaPrLSTM's performance in rank prediction did not always surpass that of certain established algorithms. This observation, however, is not a setback but a stepping stone towards progress. By amalgamating these two potent methodologies, I ventured into uncharted territory, fostering a holistic approach that bridges the gap between historical affiliations and dynamic market forces. This fusion, while revealing certain limitations in rank prediction, positions me at the forefront of a new paradigm in stock ranking practices. The path I have paved is not one of conformity, but of transformation and evolution, inviting future researchers to build upon my foundation, refine my insights, and usher in a new era of predictive financial modeling.

5.3 Future Studies

This dissertation signifies a starting point for fresh inquiries into market behavior, rather than a concluding remark. The unexpected results from my experiments guide me to areas warranting further exploration, with the counter-intuitive findings of mrrt shedding light on new facets of market dynamics. The complexities unveiled in the interplay among variables that lead to unconventional market trends provide a fertile ground for refining existing models and perhaps forging predictive tools. My study, illustrating the integration of company relation graphs and price correlation insights, reveals challenges and opportunities. Researchers may seek to investigate optimization methods for achieving a finer balance between prediction precision and correlation dynamics, exploring various activation functions, preprocessing techniques, or even hybrid models for enhanced accuracy.

The divergence of my results compared to existing algorithms opens an intriguing exploration pathway, especially in hybrid methods. By blending the strengths of ReRaPrLSTM with elements from prior methodologies, scholars may discover new ways to enhance predictive accuracy while maintaining price correlation insights. Such an amalgamation, focusing on a more robust and precise predictive instrument, could bridge performance gaps and amplify the exactness of stock ranking forecasts. This research venture has the potential to instigate a paradigm shift, consolidating the lessons learned from my successes and challenges and contributing to the broader understanding of market mechanisms and predictive modelling in the financial domain.

Continuing from the promising avenues outlined earlier, future research may also benefit from the following directions: Incorporating macroeconomic indicators into predictive models might offer a richer understanding of external influences on stock market dynamics, broadening the scope of the current study. Investigating the application of developed predictive tools in long-short trading strategies could yield practical insights into the real-world applicability of my models. Moreover, fostering interdisciplinary collaboration between finance experts, data scientists, and machine learning practitioners has the potential to unlock novel breakthroughs at the intersection of these diverse fields. Such integrated approaches could refine my theoretical and empirical comprehension of market mechanisms while contributing valuable tools and insights to the evolving landscape of financial modelling and predictive analytics.

5.4 Research Limitations

Acknowledging the boundaries of my work enhances its context and potential for advancement:

Further considerations for future research are rooted in factors that may influence the reliability and applicability of my predictive models. The availability and granularity of historical data, for instance, could affect the accuracy and generalizability of the models, necessitating caution when extrapolating findings to different datasets. Moreover, the predictive efficacy of models during periods of extreme market volatility remains unexplored, as such conditions may challenge established relationships. The performance of predictive algorithms could also be limited by technological constraints, particularly regarding available computational resources, potentially hindering real-time applicability. Finally, given the inherent dynamism of financial markets and the reliance of my models on historical data, the absence of up-to-date information may impede the models' capacity to accurately represent current market conditions. Therefore, future research must focus on employing real-time or near-real-time data to enhance prediction robustness in the ever-changing landscape of market scenarios.

As I stand at the crossroads of conclusions and new beginnings, these examples highlight the potential for unearthing insights that lie beyond the horizon. The contributions I leave behind pave the way for future researchers to delve into the mysteries of financial markets, armed with the lessons and revelations gathered through this investigation.

Bibliography

[1] Hossein Abbasimehr, Mostafa Shabani, and Mohsen Yousefi. An optimized model using lstm network for demand forecasting. *Computers & industrial engineering*, 143:106435, 2020.

[2] David M Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

[3] Mohammed Ali Alshara et al. Multilayer graph-based deep learning approach for stock price prediction. *Security and Communication Networks*, 2022, 2022.

[4] Tomaso Aste, William Shaw, and Tiziana Di Matteo. Correlation structure and dynamics in volatile markets. *New Journal of Physics*, 12(8):085009, 2010.

[5] Giovanni Bonanno, Guido Caldarelli, Fabrizio Lillo, Salvatore Micciche, Nicolas Vandewalle, and Rosario Nunzio Mantegna. Networks of equities in financial markets. *The European Physical Journal B*, 38:363–371, 2004.

[6] John Y Campbell, Sanford J Grossman, and Jiang Wang. Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, 108(4):905–939, 1993.

[7] Chun-Hao Chen, Wei-Hsun Lai, and Tzung-Pei Hong. An effective correlation-based pair trading strategy using genetic algorithms. In *Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13*, pages 255–263. Springer, 2021.

[8] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–30, 2019.

[9] Jianliang Gao, Xiaoting Ying, Cong Xu, Jianxin Wang, Shichao Zhang, and Zhao Li. Graph-based stock recommendation by time-aware relational attention network. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1):1–21, 2021.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Wen-Yu Lee, Liang-Chi Hsieh, Guan-Long Wu, Winston Hsu, and Ya-Fan Su. Multi-layer graph-based semi-supervised learning for large-scale image datasets using mapreduce. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1121–1122, 2011.

[12] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11:193–197, 1999.

[13] Daiki Matsunaga, Toyotaro Suzumura, and Toshihiro Takahashi. Exploring graph neural networks for stock market predictions with rolling window analysis. *arXiv preprint arXiv:1909.10660*, 2019.

[14] Gunter Meissner. Correlation trading strategies: Opportunities and limitations. *The Journal of Trading (Retired)*, 11(4):14–32, 2016.

[15] Shahzad Muzaffar and Afshin Afshari. Short-term load forecasts using lstm networks. *Energy Procedia*, 158:2922–2927, 2019.

[16] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015.

[17] Suman Saha, Junbin Gao, and Richard Gerlach. Stock ranking prediction using list-wise approach and node embedding technique. *IEEE Access*, 9:88981–88996, 2021.

[18] Suman Saha, Junbin Gao, and Richard Gerlach. A survey of the application of graph-based approaches in stock market analysis and prediction. *International Journal of Data Science and Analytics*, 14(1):1–15, 2022.

[19] Michele Tumminello, Tomaso Aste, Tiziana Di Matteo, and Rosario N Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.

[20] Tingting Wang, Haiyan Guo, Qiquan Zhang, and Zhen Yang. A new multilayer graph model for speech signals with graph learning. *Digital Signal Processing*, 122:103360, 2022.

[21] Xiaoting Ying, Cong Xu, Jianliang Gao, Jianxin Wang, and Zhao Li. Time-aware graph relational attention network for stock recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2281–2284, 2020.

[22] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.