

https://dspace.org/

Using graph neural networks to forecast the top 40 index in finance

Moepi, Johannah Reabetswe

2021
https://hdl.handle.net/10539/33158

Downloaded from a DSpace repository. https://dspace.org/

Using Graph Neural Networks to Forecast the Top 40 index in Finance.

Johannah Reabetswe Moepi
student number: 828301

Supervisor:
Dr Yudhvir Seetharam



A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

29 September 2021

Declaration

I, Johannah Reabetswe Moepi, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.



Johannah Reabetswe Moepi
student number: 828301
29 September 2021

Abstract

Context: After various researches and gauging the amount of work put in forecasting techniques, not much work is done in graph classification task where not only the stock market is predicted but individual stock prices and market index movements can be predicted together. In this thesis, the work will be based on capturing the effects of the South African Volatility Index (SAVI) and FTSE/JSE Top 40 index on the stock market volatility.

Aims:Is to develop a graph neural network model that forecasts the Top 40 index, to use a different technique and to evaluate what happens when the two forecasts are put together.

Method: The data is simulated in two steps. For SAVI, there is data period of 2007 February 3 to 2012 July 28. The relation data consisting of FTSE/JSE Top 40 constituents is from 2019 December 1 to 2020 March 31. The data was normalized through window sizing and LSTM model was implemented for stock market prediction. T-distributed stochastic neighbor embedding (t-SNE) algorithm is used for nodes representation. For prediction of SAVI, the volatility data is being stationarized first then LSTM model is applied to forecast it.

Results: The data set was divided into training and testing data. Using LSTM method and historical volatility models , the stock market was predicted with a dramatic fall when there was high volatility. A t-SNE map visualized the node representations and then classified the companies into their specific branches . The effect of using different relation data was evaluated and stock market was predicted to see if the data has any impact on it. SAVI data was stationarized and forecasted by LSTM technique.

Conclusion: Always normalize and balance the data to avoid biased incomparable results. The implemented LSTM model predicted the stock price and volatility.

For implementation of a hierarchical attention network, node representations of the relation data were required and implemented. With different relations in the data, the effect of using multiple relations in the stock market prediction was studied and poor prediction of the stock market was obtained. As such a hierarchical attention network was not able to be taken any further.

Acknowledgements

I would like to acknowledge my primary supervior Dr Yudhvir Seetharam for providing relevant guidance to the construction of my research proposal. For the invaluable advices and promptly answering the questions I had.

To the course lecture Dr Helen Robertson, thank you for the advices, patience and guidelines that led to a brilliant master piece of my research proposal.

To the platform that funds my studies: DST-CSIR National e-Science Postgraduate Teaching and Training Platform. Thank you for giving me a chance of fulfilling my studies. With your financial help I am able to study Master of Science in e-Science.

To my parents Johannes Morake Moepi and Gaamele Francinah Moepi, thank you for providing me with support, endless advices and motivations that kept me going. To my siblings: Thabe Phele, Tabea Phele, Rachel Moepi, Tshepiso Moepi, Hellen Moepi-Guebama-Affaga, Lentsho Moepi, Mmatsheko Moepi, Nnaniki Moepi, Lesego Moepi, Andries Moepi, Dineo Moepi and my nieces and nephews: Tshireletso Moepi, Mphasi Moepi, Dimpho Moepi and Reitumetse Moepi thank you for pushing me and occasionally telling me that i can do it. To Thabang Solomon Dolo, thank you for the motivation, for pushing and encouraging me. To my son Kgolofelo Mohau Letlotlo Moepi, thank you for giving me a reason to get up and complete my Masters.

Contents

Declaration i

Abstract ii

Acknowledgements iv

List of Figures viii

List of Tables x

1 Introduction 1
- 1.1 Problem Statement (or Research Hypothesis) . . . 1
- 1.2 Background . . . 1
- 1.3 Research Aims and Objectives . . . 3
  - 1.3.1 Research Aims and Objectives . . . 3
    - Objectives . . . 4
- 1.4 Limitations . . . 5
- 1.5 Contributions of the study . . . 5
- 1.6 Assumptions and Definitions . . . 5
- 1.7 Overview . . . 6

2 Literature Review 7
- 2.1 Theoretical Overview . . . 7
  - 2.1.1 The JSE Top 40 Index . . . 7
  - 2.1.2 The South African Volatility Index (SAVI) . . . 7
  - 2.1.3 Prediction Of Trends Using Traditional Methods And Artificial Intelligence . . . 8
  - 2.1.4 Node and Graph Classifications . . . 10
  - 2.1.5 How the key literature informed the model choice . . . 10

3 Research Methodology **11**
3.1 Research Design . . . 11
3.2 Data . . . 11
3.3 Stationarizing the data . . . 12
3.4 Applying Graph Neural Network . . . 12
3.4.1 Modules . . . 12
Feature Extraction Module . . . 12
Relational Modelling Module . . . 13
Task-Specific Module . . . 13
3.4.2 Hierarchical Attention Network . . . 13
3.5 Applying time series models . . . 14
3.5.1 One-step ahead prediction via averaging . . . 14
Standard averaging . . . 14
Exponential Moving Average . . . 14
3.5.2 Long Short Term Memory (LSTM) . . . 15
3.6 Analysis . . . 16
3.6.1 Evaluation Metrics . . . 16

4 Results and Discussion **17**
4.1 Stationarizing the data . . . 17
4.1.1 SAVI . . . 17
4.1.2 FTSE/JSE Top 40 . . . 19
4.2 Forecasting stock prices . . . 21
4.2.1 Individual Stock Prediction Task using LSTM . . . 21
4.2.2 Node Representation . . . 21
4.3 Effect of using relation data . . . 22
4.4 Forecasting SAVI . . . 24
4.4.1 Standard Averaging Prediction . . . 24
4.4.2 Exponential Averaging Prediction . . . 24
4.4.3 LSTM . . . 25
4.5 Summary . . . 27
4.5.1 Stock Prices forecasting . . . 27
Stationarizing the data . . . 27
Individual Stock price prediction . . . 29

Node representation 29
Effects of using relation data 30
4.5.2 Volatility forecasting 30
Stationarizing the data 30
Historical volatility models 31
4.5.3 Comparing the two forecasts together 31

5 Conclusions and Future Work **33**
5.1 Future Work 34

Bibliography **36**

List of Figures

1.1 Credit: Kim et al. (2019) 4

2.1 JSE Top 40 compared with SAVI. Credit: The image is generated from instrument comparison in inetbfa financial database. 8

3.1 Credit: Kim et al. (2019) 14
3.2 An LSTM model visualized by Nguyen, Tran, and Nguyen (2019). . . 15

4.1 Visualizing the SAVI data for any anomalies or irregular patterns. . . 17
4.2 Decomposing the data into observed, trend, seasonal and residual plots. 18
4.3 Checking the stationary test of the SAVI data 18
4.4 Visualizing the FTSE/JSE Top 40 data for any anomalies or irregular patterns. 19
4.5 Checking the stationary test of the FTSE/JSE Top 40 data. 20
4.6 Detrending approach to stationarize the FTSE/JSE Top 40 data. . . . 20
4.7 Differencing approach to stationarize the FTSE/JSE Top 40 data. . . . 20
4.8 Detrending and differencing approaches in one plot in order to stationarize the FTSE/JSE Top 40 data. 21
4.9 Training and validation loss versus epochs. 22
4.10 Individual Stock Prediction Task using LSTM on FTSE/JSE Top 40 data 23
4.11 A t-SNE map of the FTSE/ JSE Top 40 companies on a specific day represented on a 2 dimensional space. 24
4.12 A t-SNE map of the FTSE/ JSE Top 40 companies from different branches that each company fall under. The representation of the map on a specific day is shown on a 2 dimensional space. 25
4.13 Standard averaging method applied on a specific day using Top 40 companies close data. 26

4.14 Exponential Moving Average method applied on a specific day using Top 40 companies close data. 26
4.15 Standard Averaging prediction of volatility. 27
4.16 Exponential Averaging prediction of volatility. 27
4.17 Average loss and mse per epoch. 28
4.18 Training and validation loss versus epochs. 29
4.19 Exponential Averaging prediction of volatility. 30
4.20 Forecasting of FTSE/ JSE Top 40 index using the date ranges that produced Figure 4.19 32

List of Tables

3.1 Statistical significance values expected when data is stationarized. . . 12

4.1 Statistical significance values obtained from testing the stationary of the SAVI data . . . 19
4.2 Statistical significance values obtained from the approaches taken to stationarize the FTSE/JSE Top 40 data . . . 21
4.3 Branches assigned to the companies in the JSE Top 40 . . . 23
4.4 Historical volatility models of the indices along with their mean squared error. . . . 24
4.5 Historical volatility models of the volatility along with their mean squared error. . . . 31

Chapter 1

Introduction

1.1 Problem Statement (or Research Hypothesis)

Volatility forecasting is an important input in portfolio management, market regulations, option pricing, investments, risk management, security valuations and monetary policy making. For investment purposes, it is a crucial tool as it can forecast future returns of the stock market. Given that investors hold long positions tend to credit themselves when achieving positive returns, this can sometimes cause overconfidence in the stock market (Gervais and Odean (2001), Odean (1999), Soll and Klayman (2004)). Correspondingly, a "fear" can grip the market when volatility rises. It is important to therefore be able to forecast both indices, and to then compare the accuracy against traditional metrics as well as each other. In other words, can a stock index and the SAVI be forecasted, along with being inversely proportional to each other?

1.2 Background

While much research has been conducted in forecasting returns in the stock market, there still remains a question of the predictive accuracy of these models. Given that Graph Neural Networks (explained later) are relatively new, this method aims to enhance prediction by using graphical methods (and deep learning) to identify any patterns or trends in the data. This is particularly appealing to financial markets, as many traders use charting techniques (also known as technical analysis) to predict directional movements in stock prices.

Given a set of returns, volatility is the rate at which the price of a security fluctuates, whether it is increasing or decreasing. A low volatility of a security can therefore be described as occurring when prices are stable. It's converse of high volatility implies that highs and lows occur quite often, making prediction that much more difficult. If one is able to predict prices (returns) with greater accuracy, one can manage the associated volatility of returns, thereby resulting in stable returns over a time period.

Volatility is also sometimes used as a proxy for the emotional state of investors. Starting off in the US, a volatility index was created for the SP 500 to describe the average level of market volatility. If volatility increased, one would observe a decrease in the SP500 stock index level. This would signal to traders and investors that there is an increase in uncertainty, causing some market participants to withdraw from the market to safer assets (such as cash or gold). As a result, these volatility indices are informally used to gauge "fear" and "greed" in the stock market. The South African Volatility Index (SAVI) was introduced in 2007 and is used to estimate market risk in the country. It is also seen as a fear and market sentiment gauging model for investors that produces a volatility index given the mental "malfunctioning" of the investors (who can sometimes be overconfident). SAVI values are utilized by investors and research analysts to measure the level of fear and stress before committing themselves to any investment decision.

If one can forecast both the stock prices (or a stock index) and market volatility, it would be interesting to compare these forecasts to each other, based on their historical negative relationship. Often, forecasts of either index are based on historical values of those indices (the forecasted volatility index is compared against it's historic values). By introducing a further comparison of the volatility index forecast against the stock index forecast, one can also see if the forecasting methods produce results that conform to empirical observation (the negative correlation between the indices). There are different volatility forecasting methods, which range from simple to complex. Implied volatility models exist, such as an exponential weighted moving average and standard moving average. These are the simplest and perhaps the most effective from a trader's perspective. From time series econometrics, one can use auto regressive and heteroskedastic models, which include the

Auto Regressive Integrated Moving Average (ARIMA) model and the Generalized Auto Regressive Conditional Heteroskedasticity (GARCH) family of models. With technology advancing and more methods explored, Artificial Neural Networks are being studied as well. Each method has its advantages and disadvantages with each forecasting accuracy tested using the stock indices (Ladokhin, 2009).

1.3 Research Aims and Objectives

1.3.1 Research Aims and Objectives

The aims of the research is:

- To develop a graph neural network model that forecasts the Top 40 index.
- To use a different technique and forecast the SAVI.
- To evaluate what happens when the two forecasts are put together.

Given that there is a growing interest when it comes to utilizing graph structured data, Kim et al. (2019) developed a Hierarchical Attention network for stock price prediction while utilizing relational data for predicting the stock, see Figure 1.1. The framework in Figure 1.1 is based on various stock market prediction methodologies that utilizes the corporate relational data. It is divided into three modules; the feature extraction, relational modelling and the prediction layer. The feature extraction module represents the current state of individual stocks based on historical movement patterns. The relational modelling module is a function for node updating where neighboring nodes exchange information. Using the same technique, nodes will be created from a feature selection module and fed into a task-specific layer.

Second, as previous studies have not been able to forecast volatility, a different model will be applied which is a time series model. Last, after obtaining results of the two forecasts, an analysis of how each behaves given the other will be observed.

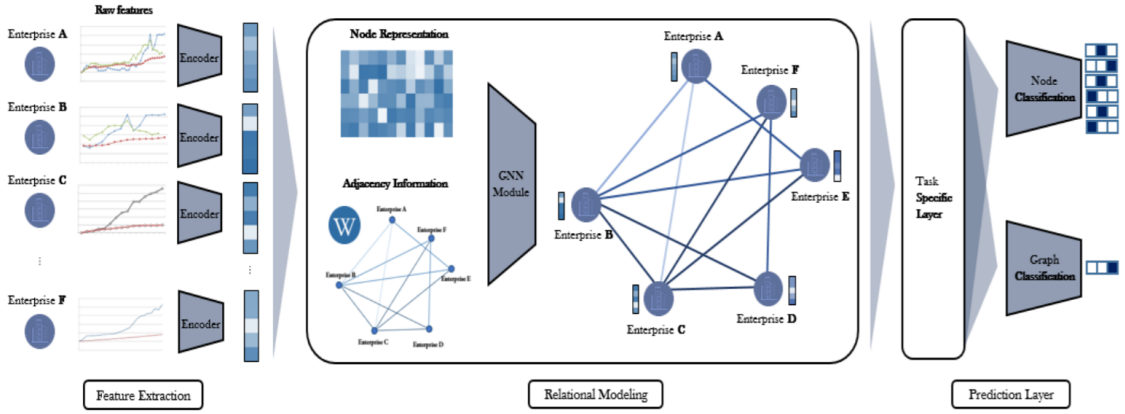


Figure 1: General framework of stock prediction using relational data.

FIGURE 1.1: Credit: Kim et al. (2019)

Most forecasting techniques that use neural networks focused on node classification tasks as well as individual stock price prediction. In terms of a graph classification task, little to no work has been done from a stock price prediction perspective, causing a gap in the literature. As such the research aims to explore graph neural networks as they are the information exchange chain between neighbouring nodes. By utilizing these networks, the study will visually classify and forecast stock index movements and volatility relating to equity market risk in South Africa. Harrilall and Seetharam (2016) stated that the SAVI is extremely difficult to forecast and it should be seen as an investment opportunity since it is a fear gauging tool of investor sentiment. With a different forecasting tool, we will explore the difficulty of forecasting the SAVI.

Objectives

The study will use a graph neural network to forecasting prices of the FTSE/JSE Top 40 Index. The model requires inputs of the FTSE/JSE Top 40 index and each top 40 company. The advantage of using a graph neural network framework is that it requires graph-structured data. A Hierarchical Attention Network will assist in this regard. The study also proposes time series models in order to predict volatility. In terms of time series models, the LSTM, standard averaging and exponential moving average approaches will be used to compare forecasting ability.

1.4 Limitations

From the literature review, there are not many studies done using graph classifications in finance. While this study will therefore add to the literature, it is also problematic as there cannot be a comparison of results to other studies (apart from comparing the results to other classification methods).

When implementing the graph neural network, the injective function are required.They map a multi-set of embeddings into new embeddings. A mean function can be applied however this provides limitations as it can give off the same embedding and not the average leading to no injective function.

1.5 Contributions of the study

Graph neural network is relatively new in enhancing prediction through graphical methods that identify any patterns/ trends in the data and if one can forecast both the stock prices (or a stock index) and market volatility, it would be interesting to compare these forecasts to each other, based on their historical negative relationship. In financial markets, many traders use technical analysis to predict stock price movements, and knowing the effects of SAVI and FTSE/ JSE Top 40 index on the stock market volatility, it will be a new different approach in the field. Previously SAVI was not able to be forecasted and being a fear gauging tool, it will contribute much in the field if it can be forecasted.

1.6 Assumptions and Definitions

SAVI cannot be forecasted as it is just an investor sentiment gauging index. With its inclusion in the study, we will factor in the impact of expected future volatilities in the South African market on the FTSE/JSE Top 40.

1.7 Overview

The thesis is sectioned into 5 chapters:

- Chapter 1: Its introduction section with background, research aims and objectives of the study.
- Chapter 2: Literature review chapter reflecting on the previous studies conducted on the same research topic.
- Chapter 3: This section explains all the methods to be used in the study.
- Chapter 4: The results section gives a visual output of what the intended study was on and then discusses the achieved results .
- Chapter 5: This section gives a brief summary of what the results mean, their link to the literature review and future work.

Chapter 2

Literature Review

2.1 Theoretical Overview

2.1.1 The JSE Top 40 Index

Liu et al. (2020a) identified stock indexes as social and economic indicators that are important and are able to broadly reflect the stock market's overall trend and performance. In this study, the focus is on the performance of the South African stock market, which is proxied by the FTSE/JSE Top 40 Index. This index is regarded as the primary index on the JSE. From all the shares that are listed in the JSE, above 80% of the total market capitalisation is captured by the FTSE/JSE Top 40 (Liu et al., 2020a).

2.1.2 The South African Volatility Index (SAVI)

The SAVI is an index that was launched in 2007 with the aim of measuring the market's expectation of the three month period of implied market volatility. In 2009, the JSE updated the SAVI in terms of measuring expected volatility. The SAVI is based on the FTSE/JSE Top 40 index and is determined using at-the-money volatilities and the volatility skew. As such the SAVI is a "fear" gauge index given that it is determined using the volatility skew which is expected to incorporate the markets expectations of a crash (De Kock et al., 2015).

Figure 2.1 below shows both the JSE Top 40 index along side the SAVI. It is typical for there to be a negative correlation between the two indices. As fear increases

in the market, one would expect the Top 40 to have a lower price as investors would exit their positions in favour of more safety, usually in the form of cash.

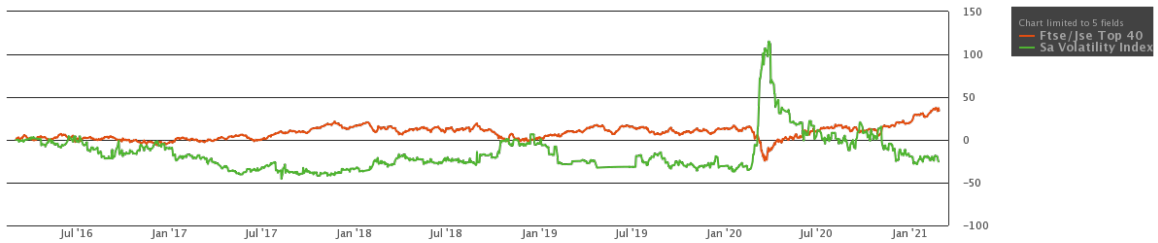


FIGURE 2.1: JSE Top 40 compared with SAVI. Credit: The image is generated from instrument comparison in inetbfa financial database.

2.1.3 Prediction Of Trends Using Traditional Methods And Artificial Intelligence

Harrilall and Seetharam (2016) investigated the ability of different forecasting techniques in forecasting the SAVI. A Time-Delay Neural Network's (TDNN) ability to forecast the South African Volatility Index was compared to other traditional models. The daily price levels of the SAVI were used as the dataset dating from 2007 February to 2017 December and a comparison of the residual errors of the tools used in the study suggested that the historical average model and the time-delay neural network have a forecasting ability that is higher than traditional forecasting models. However, the study suggested that the SAVI cannot be predicted, it is merely a gauging tool of investors sentiments.

Chaudhuri and Ghosh (2016) constructed Artificial Neural Network (ANN) models which are based on various back propagation algorithms in order to predict volatility in the Indian stock market. The study used the volatility of NIFTY returns and volatility of gold returns with several predictor variables like the Indian VIX, CBOE VIX, volatility returns of crude oil, the Dow Jones Industrial Average (DJIA), DAX, Hang Seng and Nikkei for the 2013 to 2014 period. The efficacy of the ANN was examined in terms of volatility prediction and provide satisfactory results for 2015. However, the forecasting of market volatility dating back to 2008

was less effective.

Ibrahim and Ramu (2016) examined the accuracy of neural networks in volatility forecasting. Using Apple stock returns, simpler models produced poor results whilst more complex models produced better results. While accuracy is important, the need to ensure simplicity with neural networks is equally important to ensure that the results can be practically actioned by the investor. With ANNs being easy to implement and efficient, it can provide a good initial benchmark of volatility forecasting, but the estimation should not solely rely on ANNs as opposed to other models as well.

Min (2020) uses a Long-Short Term Memory (LSTM) network to forecast trends in the market proxied by the Nasdaq 100 index. Comparing against an ANN, Support Vector Regression (SVR) and a Recurrent Neural Network (RNN), the author found that the LSTM performs better with weekly and not daily data.

With any stock price movement, research has incorporated non-fundamental factors, relating to sentiment to improve on the accuracy of the prediction. Data sourced from the news and social media were utilized in this regard, sometimes leading to improved accuracy (Liu et al., 2020b). Some of these existing methods stem from the artificial intelligence and computer science literature. For example, the Dirichlet Mixture Model derives sentiment from news that is then used to construct a sentiment time series and regressed with a stock index in order to predict the market (Si et al., 2013). The Topic Sentiment Latent Dirichlet Allocation (TSLDA) model obtains features which simultaneously captures topics and their sentiments in order to have a model that predicts movements of stock prices using social media (Nguyen and Shirai, 2015). The Hybrid Attention Network (HAN) was designed to predict trends in stock prices based on the sequence observed in related, recent news (Hu et al., 2018). The HAN however has been found to not account for the different influences of events. Hence Liu et al. (2020b) derived a Multi-Element Hierarchical Attention (MEHA) capsule network consisting of two components: generically named a former component and latter component. The former component uses the weights assignment process to quantify the importance of any valuable information in social media and multiple news channels. The latter

component utilises the vector representation in the hidden layer to learn more information from the events. To maintain the complementarity between social media and news, a combined data set was constructed and their model showed improvement in prediction accuracy through quantifying of different influences of events.

2.1.4 Node and Graph Classifications

Kim et al. (2019) employed a Hierarchical Graph Attention Network (HGAN) for stock price prediction (an index in their study). The method aggregates information based on different relational types (which can loosely be thought of as exogenous variables) and then combines them with company-specific information. The feature extraction module is utilized for node representations which is then fed into a task-specific layer. The study showed that the performance varies depending on the relational data used. Li et al. (2020) employed a hierarchical graph attention network with semi-supervised node classification in order to learn node features. This was found to be better than the Kim et al. (2019) model as it achieved superior results according to goodness of fit criteria.

2.1.5 How the key literature informed the model choice

Given that there is a growing interest when it comes to utilizing graph structured data, Kim et al. (2019) developed a Hierarchical Attention network for stock price prediction while utilizing relational data for predicting the stock. Most forecasting techniques that use neural networks focused on node classification tasks as well as individual stock price prediction. In terms of a graph classification task, little to no work has been done from a stock price prediction perspective, causing a gap in the literature. As such the chosen model is to explore a different technique: graph neural networks as they are the information exchange chain between neighbouring nodes. By utilizing these networks, the study will visually classify and forecast stock index movements and volatility relating to equity market risk in South Africa because previously studies have not been able to forecast volatility, a different model will be applied.

Chapter 3

Research Methodology

3.1 Research Design

The broad research design is shown in Figure 3.1 which is a hierarchically attention mechanism used for stock prediction. The mechanism is designed to select meaningful information at each level. It consists of the state attention layer that considers the neighboring nodes and then select important meaningful information from the same types of relations. After gathering the information, it is fed into the relation attention layer consisting of the relation information vector and the relation types. In order to apply the hierarchical attention network for stock prediction (HATS) method, a few modules must be applied first. These are feature selection module and task specific layer.

3.2 Data

Data will be sourced from IRESS BFA, which is accessed by WITS students. The IRESS Business Financial Analysis (BFA) is a technology company which provides continents including Africa with the software to the financial services industry. Three datasets will be used. The first is the JSE Top 40 index. This contains the top 40 JSE listed firms by market capitalization. The data consists of High, Low, Open and Closing values, along with volume. The second is the SAVI, while the third are the JSE Top 40 constituents. The JSE Top 40 Index prices and the SAVI levels are sourced daily over the period 01 December 2019 to 31 March 2020 as this is the period when COVID19 started spreading globally, leading to the first lockdown

of South Africa. This is also an interesting time period to examine the accuracy of the Graph Neural Network as it is a period of increased uncertainty and volatility.

3.3 Stationarizing the data

In order to forecast, it is required that the data should be stationary. Stationary data have statistical properties (mean and variance) that do not change much over time. Given that the data is not stationary, three approaches will be taken to stationarize the data which is detrending, differencing and the addition of the two. On the detrending approach, the method removes the entire underlying trend. On the differencing approach, the method removes underlying seasonal or cyclical patterns. Lastly, both the approaches are observed on the same plot.

From the critical values and test statistic values obtained per approach, if the test statistic values are greater than the critical value 1% obtained as this shows 99% confidence level then the data will not be successfully stationarized and vice versa. See Table 3.1.

Statistical significance	Test Stationary	Detrend	Difference	Detrend + Difference
Test statistic	<1%	<1%	<1%	<1%
Critical Value: 1%	<1%	<1%	<1%	<1%
Critical Value:5%	<1%	<1%	<1%	<1%
Critical Value:100%	<1%	<1%	<1%	<1%

TABLE 3.1: Statistical significance values expected when data is stationarized.

3.4 Applying Graph Neural Network

3.4.1 Modules

Feature Extraction Module

Having individual stocks as node classification tasks means that each node feature can be used to understand each company's current state given the movement of

their prices. Fischer and Krauss (2018) explain how node features are extracted from raw price data and they utilized the LSTM network along with the Gated Recurrent Units (GRU) network. From their experience of implementation difficulty, the LSTM was used for individual stock price prediction and GRU for index movement prediction.

Relational Modelling Module

In relational modelling, collections of inter-related tables are used to represent data and relationships. In this study, relational modelling is used as a node updating function that updates the representations of the node. It collects information obtained from relation types and different nodes and combines them to represent the relationships.

Task-Specific Module

The module is task-specific as different tasks can be assigned node representations. Two graph-based learning tasks are performed - the first is individual stock price prediction which is similar to a node classification task and the second is market index prediction, which predicts the index movement given the current state of a company.

3.4.2 Hierarchical Attention Network

This network is designed to express the importance of different types of relations and neighbouring nodes in the data. There is an attention mechanism that assigns different weights to the selected information which is based on the neighbouring node's current state. The HATS model is key to performance improvements as it only selects meaningful information at each level of model construction.

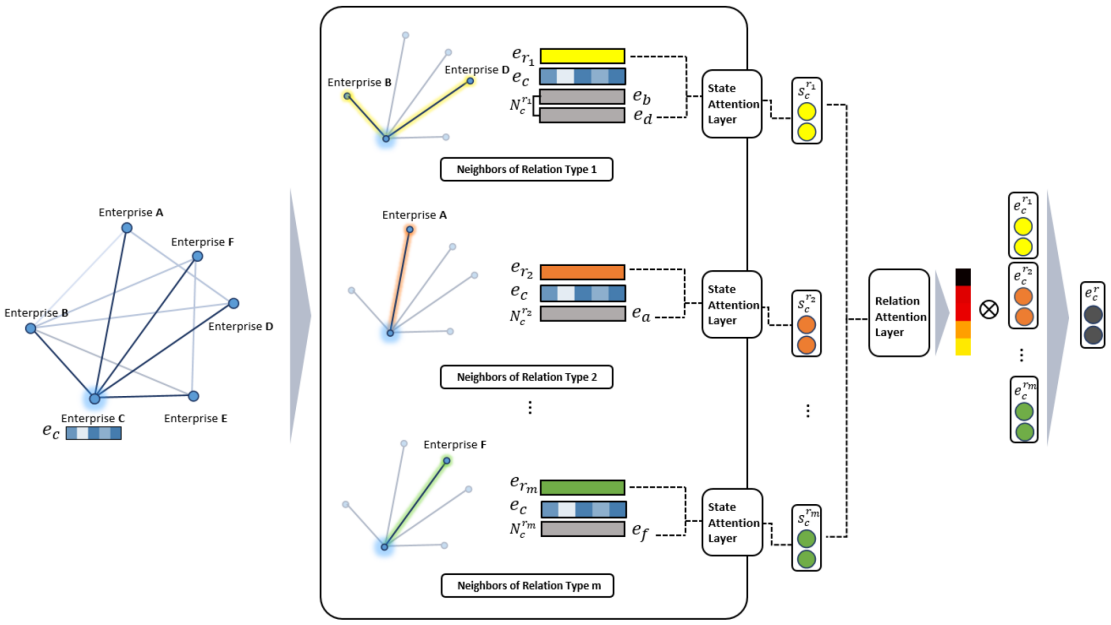


FIGURE 3.1: Credit: Kim et al. (2019)

3.5 Applying time series models

3.5.1 One-step ahead prediction via averaging

This is an averaging mechanism that represents the future stock prices as a mean of past observed stock prices in order to predict one step ahead.

Standard averaging

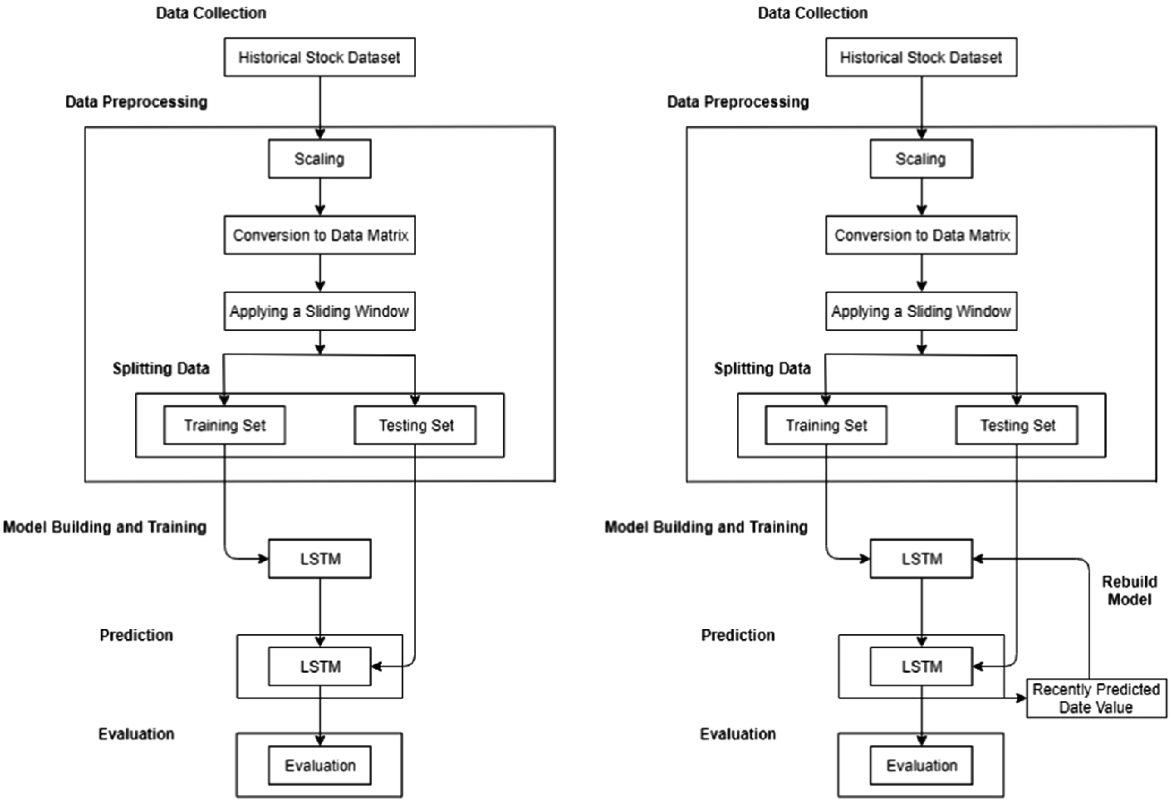
This is a historical volatility model that predicts future stock market prices as a mean of the previous stock market prices that are observed within a window size that is fixed.

Exponential Moving Average

The exponential averaging prediction method uses time steps as one step ahead prediction.

3.5.2 Long Short Term Memory (LSTM)

LSTM is a recurrent neural network that learns long term dependencies in the data. It is a model that can predict a number of steps in the future. Nguyen, Tran, and Nguyen (2019) visualized two types of LSTM models (Figure 3.2), the static and dynamic models where the dynamic model is continuously retrained using new augmented data produced after prediction. The LSTM models shows that they take in historical stock data and applies 5 essential components that allows long and short term modelling of the data. Under data pre-processing, hyperparameters and window sizes are defined and the data is then divided into training and testing data sets. Then the model will be run to predict the stock prices for various epochs.



(a) Static LSTM method (b) Dynamic LSTM method

FIGURE 3.2: An LSTM model visualized by Nguyen, Tran, and Nguyen (2019).

3.6 Analysis

In the analysis of the results, 4 components will be evaluated: data understanding, data preparation, modelling and evaluation.

Visualisations of the raw data will be produced in order to have a feel of how the data is structured, flows and if there are any irregular patterns. This will help in understanding the data. Data is then prepared by normalizing it using a window. This assists in removing outliers. Modelling is done through the Hierarchical Attention Network model, which consists of the individual stock prediction layer and the graph pooling layer. For the individual stock prediction task, the LSTM will be implemented and for the graph pooling layer, the GRU model will be implemented.

Finally, Python will be used with training, evaluation and testing of the data, which will be divided into three classes of upward, downward and neutral movements. The dataset will be divided into smaller subsets that goes through the different phases above. The data will be optimized, hyperparameters will be tuned to a certain range and the performance of the models will be measured based on each period's evaluation set.

3.6.1 Evaluation Metrics

Mean squared errors will be evaluated per epoch using equation below and the training and validation loss plots will be produced.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i^{'})^2 \tag{3.1}$$

MSE = Mean Squared Error
n = number of data points
Yi = observed values
Y'i = predicted values

Chapter 4

Results and Discussion

4.1 Stationarizing the data

In order to forecast, the data should be tested for stationary and stationarized first.

4.1.1 SAVI

From the plot in Figure 4.1, trends and anomalies can be examined to determine if the data are stationary. After visualizing the data, a more detailed view of the trend was obtained in Figure 4.2.

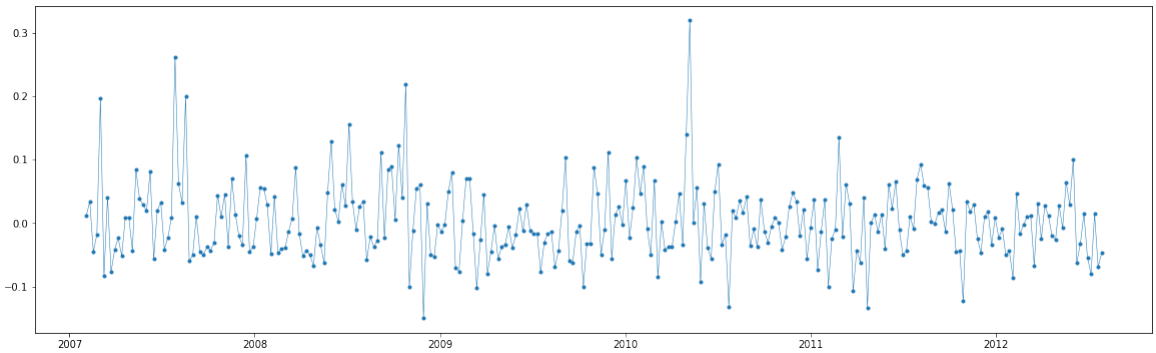


FIGURE 4.1: Visualizing the SAVI data for any anomalies or irregular patterns.

The observed plot shows the actual values of the volatility, any upwards and downwards patterns, seasonality, variation per year and the residuals. Figure 4.3 checks the stationary state of the data. From the figure, it can be seen that the data is stationary with the rolling mean and standard deviations not much changing over

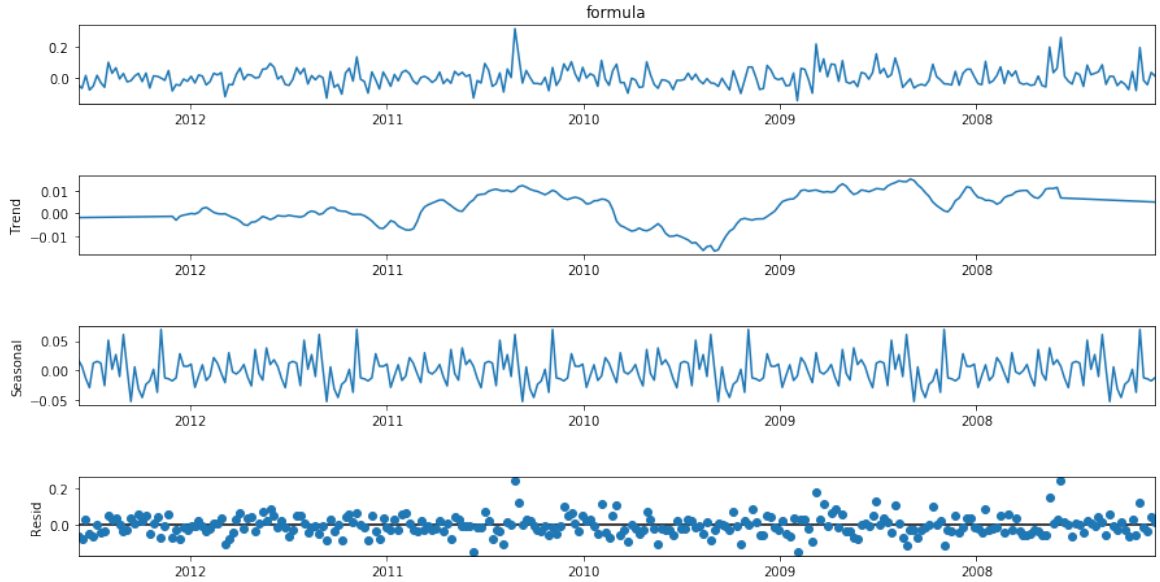


FIGURE 4.2: Decomposing the data into observed, trend, seasonal and residual plots.

time.

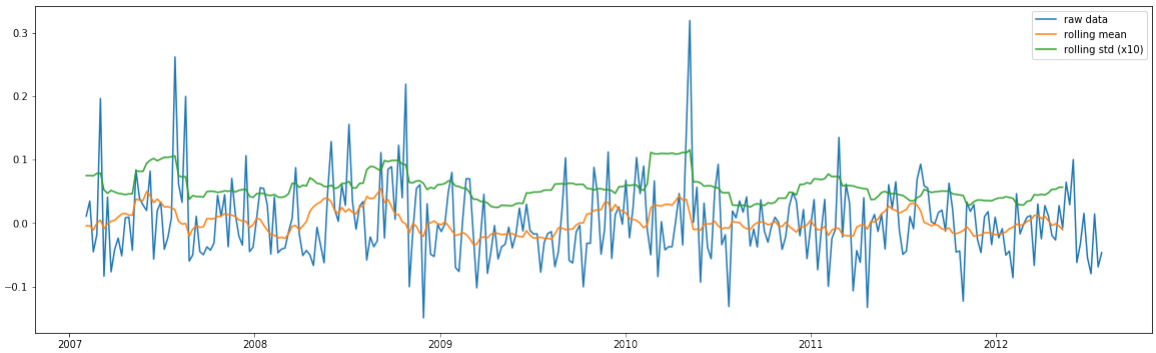


FIGURE 4.3: Checking the stationary test of the SAVI data

Given that the data is stationary, there is no need for further transformations. From the critical values and test statistic values obtained when testing for stationarity, it can be seen that the data is stationary as the test statistic values are lower than the critical values. The statistical values obtained can be found in Table 4.1.

Statistical significance	Test Stationary
Test statistic	-15,839
P-value	0,0
Critical Value: 1%	-3,453
Critical Value:5%	-2,872
Critical Value:100%	-2,572

TABLE 4.1: Statistical significance values obtained from testing the stationary of the SAVI data

4.1.2 FTSE/JSE Top 40

From Figure 4.4 and Figure 4.5, it can be seen that the data starts off stationary and then becomes non-stationary as the rolling mean and standard deviations starts off with not much change over time and then increase.

Given that the data is not stationary, three approaches were taken to stationarize the data which is detrending, differencing and the addition of the two. On the detrending approach, the method removes the entire underlying trend as seen in Figure 4.6. On the differencing approach, the method removes underlying seasonal or cyclical patterns as seen in Figure 4.7. In Figure 4.8, both the approaches are observed on the same plot. From the critical values and test statistic values obtained per approach, it can be seen that the data was successfully stationarized as the test statistic values are lower than the critical value. The statistical values obtained per approach can be found in Table 4.2.

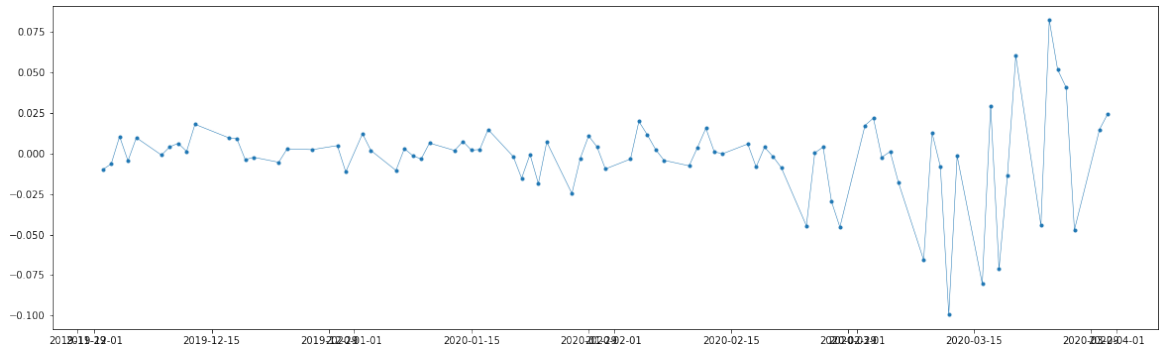


FIGURE 4.4: Visualizing the FTSE/JSE Top 40 data for any anomalies or irregular patterns.

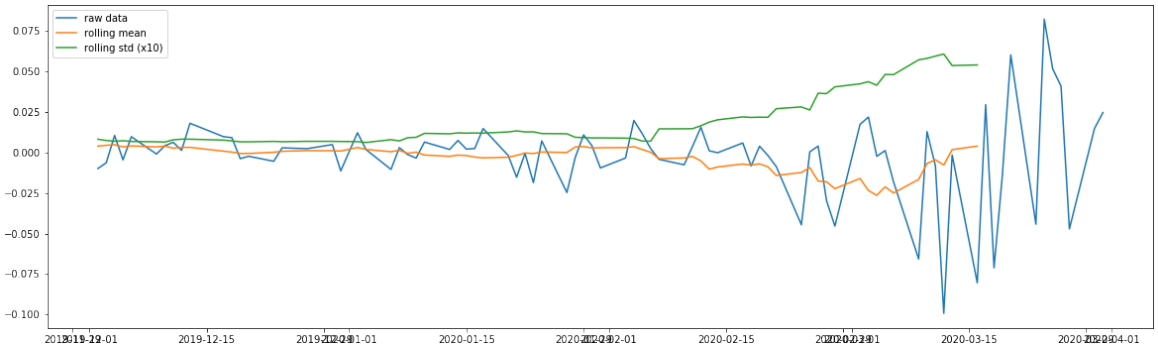


FIGURE 4.5: Checking the stationary test of the FTSE/JSE Top 40 data.

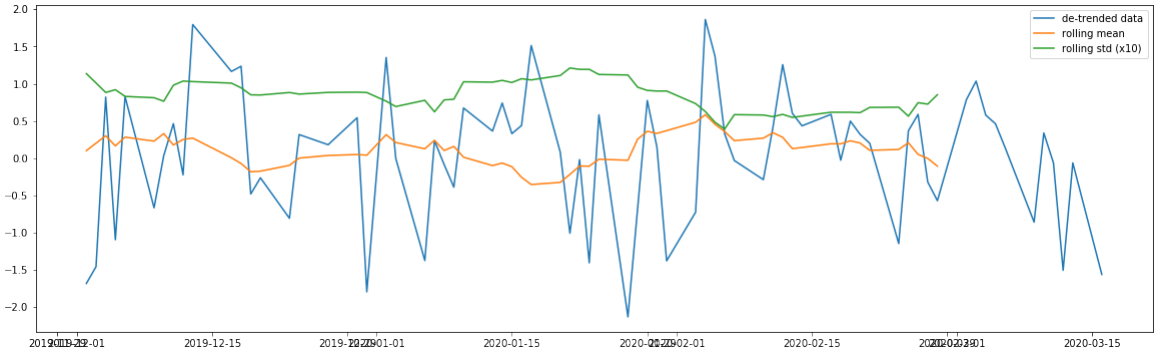


FIGURE 4.6: Detrending approach to stationarize the FTSE/JSE Top 40 data.

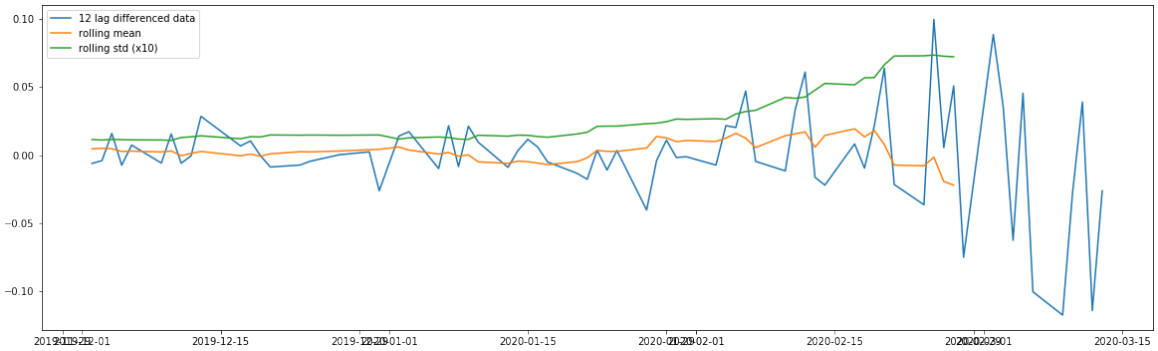


FIGURE 4.7: Differencing approach to stationarize the FTSE/JSE Top 40 data.

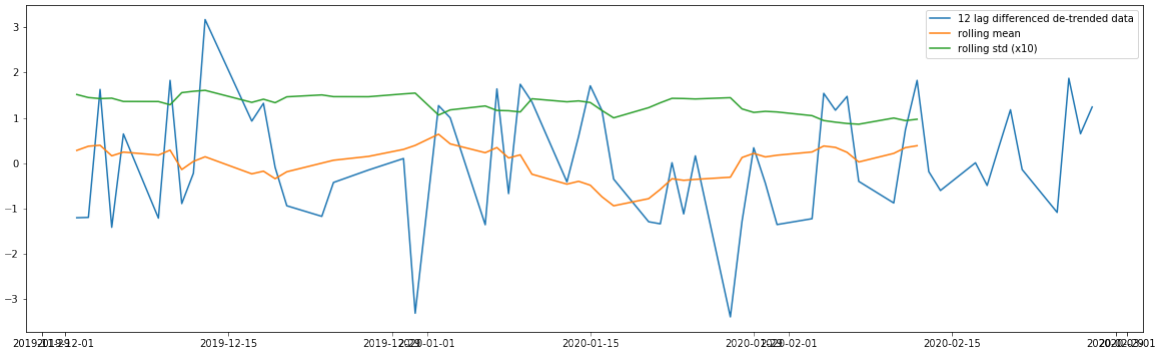


FIGURE 4.8: Detrending and differencing approaches in one plot in order to stationarize the FTSE/JSE Top 40 data.

Statistical significance	Test Stationary	Detrend	Difference	Detrend + Difference
Test statistic	-1,655	-7,826	-2,985	-3,739
P-value	0,455	0,0	0,036	0,004
Critical Value: 1%	-3,525	-3,526	-3,546	-3,575
Critical Value:5%	-2,903	-2,903	-2,912	-2,924
Critical Value:100%	-2,589	-2,589	-2,594	-2,600

TABLE 4.2: Statistical significance values obtained from the approaches taken to stationarize the FTSE/JSE Top 40 data

4.2 Forecasting stock prices

4.2.1 Individual Stock Prediction Task using LSTM

Using a LSTM network, the data set consisting of 83 data points should be scaled first for better performance. The data was scaled between 0 and 1 and time steps of 2 were created. The training data set consists of 62 data points and the testing set consists of 21 data points. An LSTM layer was added as well as a dropout layer which prevents overfitting. In terms of optimizing the data, the Adam optimizer was used. With 100 epochs (see Figure 4.9), the model was compiled. Figure 4.10 was produced after compiling the LSTM.

4.2.2 Node Representation

For node representations, the FTSE/JSE Top 40 companies closing data was used. The period of the data is from 2019-December-01 to 2020-March-31. The selected

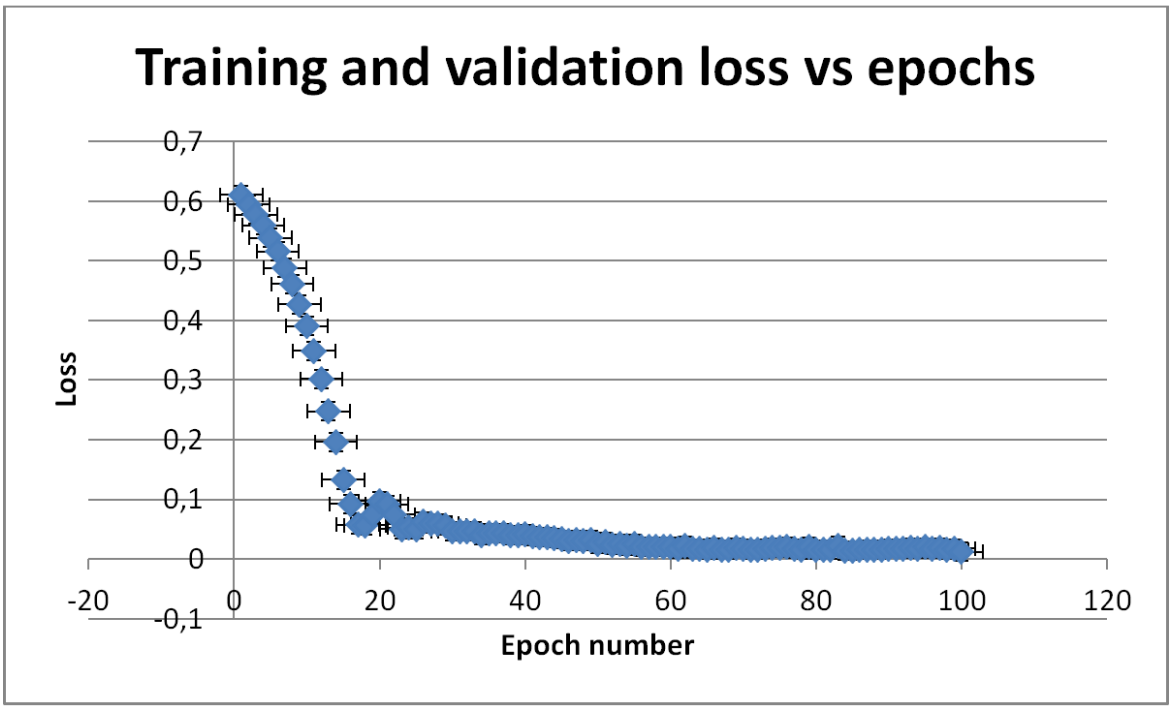


FIGURE 4.9: Training and validation loss versus epochs.

data time frame is due to the start of COVID-19 leading to the first lockdown in South Africa. The data was first normalized then a t-SNE algorithm was used to map each of the FTSE/ JSE Top 40 companies on a specific day to a 2 dimensional space as seen in Figure 4.11.

As each company has a specific branch that it falls under, each branch was assigned a value from 1 to the length of the companies, see Table 4.3. Node representation of the branches of the companies was produced as seen in Figure 4.12.

4.3 Effect of using relation data

In order to investigate the impact of different relationships applied on the data for stock price predictions, Figure 4.13 and Figure 4.14 were produced with the standard averaging method and exponential moving average method respectively. Different MSE values were obtained after plotting the prediction of the Top 40 company returns. As seen in Table 4.4, the exponential moving average had a lower

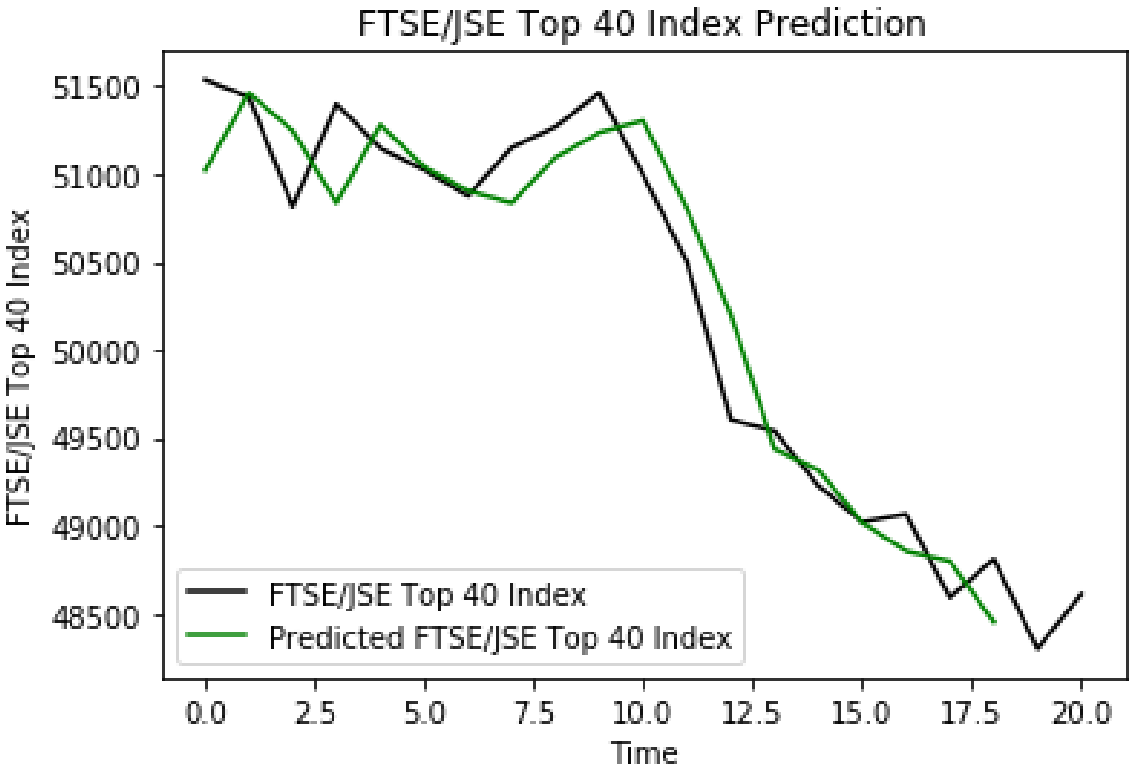


FIGURE 4.10: Individual Stock Prediction Task using LSTM on FTSE/JSE Top 40 data

Company	Branch
Anglo American	Mining
FirstRand Limited	Bank
Discovery Ltd	Finance
Exxaro Resources	Chemistry
Growth point Prop Ltd	Property
Impala Platinum Hlds	Holdings
Mr Price Group	fashion retail chain
MTN Group	Telecommunication
Naspers	Media
Shoprite	Retail Trade

TABLE 4.3: Branches assigned to the companies in the JSE Top 40

MSE of 0,00081. On Figure 4.14, the prediction is not as badly predicted as the one on Figure 4.13 but it does not show better prediction of the stock market.

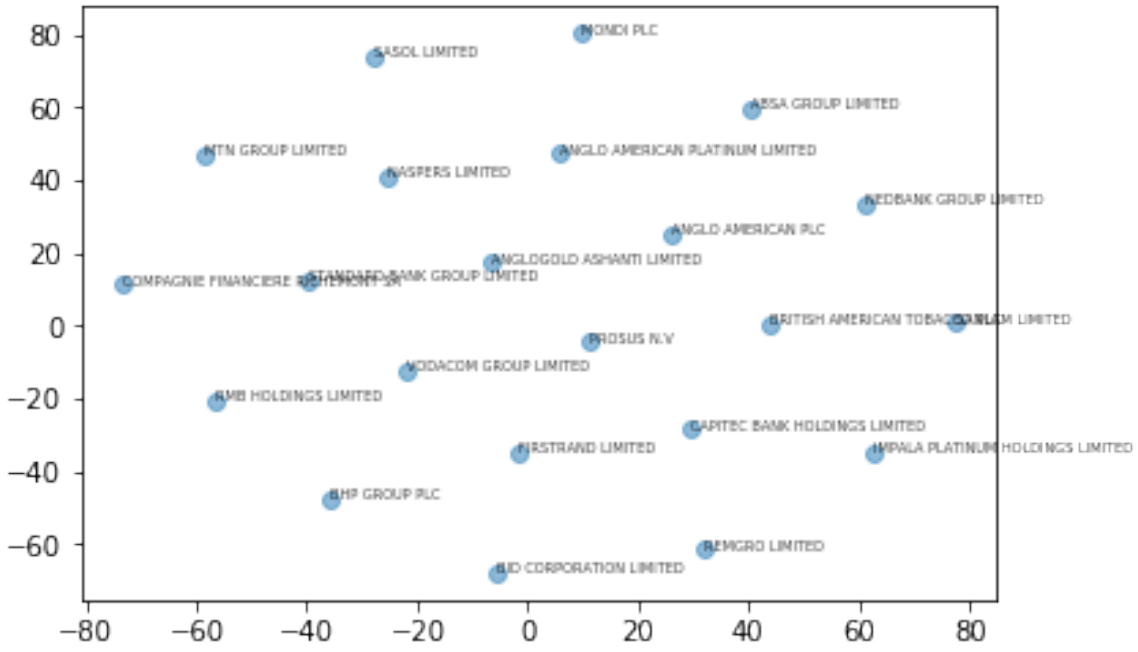


FIGURE 4.11: A t-SNE map of the FTSE/ JSE Top 40 companies on a specific day represented on a 2 dimensional space.

Historical Volatility Models	Index	MSE
Standard Averaging	Top40 Companies	0,00118
Exponential Moving Average	Top40 Companies	0,00081

TABLE 4.4: Historical volatility models of the indices along with their mean squared error.

4.4 Forecasting SAVI

4.4.1 Standard Averaging Prediction

The standard averaging prediction method predicts future values as a mean of the previous values that are observed within a fixed window size. Setting window sizes of 10, Figure 4.15 was produced.

4.4.2 Exponential Averaging Prediction

The exponential averaging prediction method uses time steps as one step ahead prediction. Figure 4.16 was produced for the SAVI.

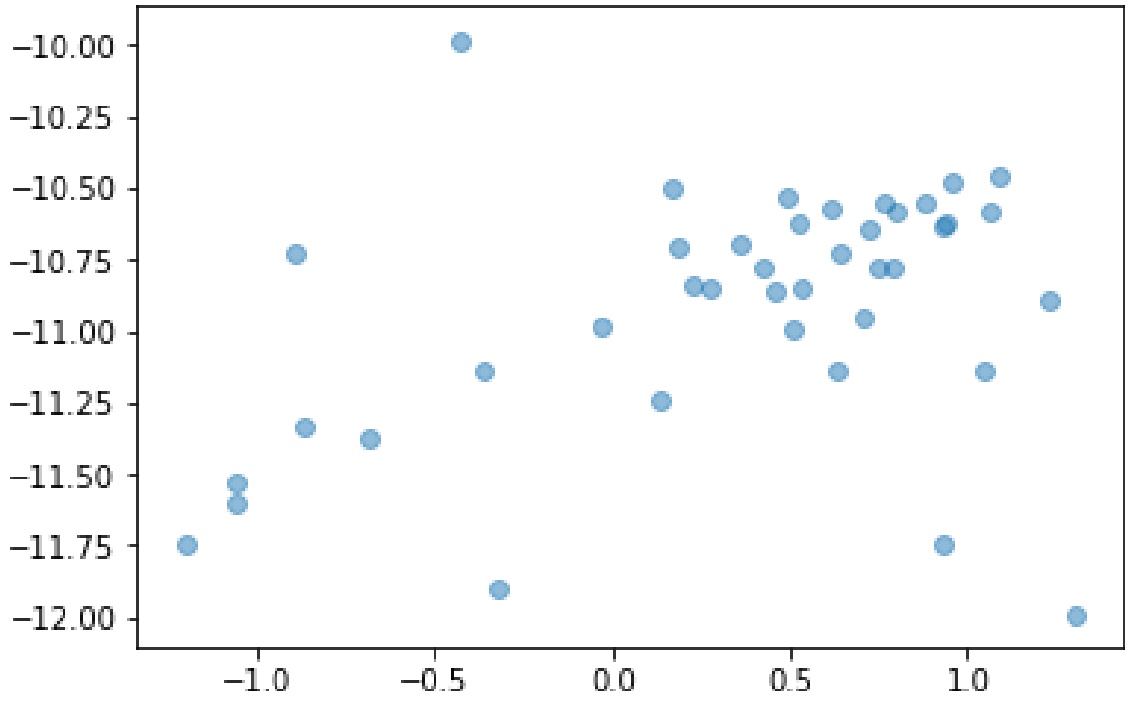


FIGURE 4.12: A t-SNE map of the FTSE/ JSE Top 40 companies from different branches that each company fall under. The representation of the map on a specific day is shown on a 2 dimensional space.

The progression of the loss and the mean squared error is shown in Figure 4.17 with error bars. As the average loss decreases, the mean squared errors also decrease.

4.4.3 LSTM

Using an LSTM network, the data should be scaled first for better performance and in the case of the SAVI, the data was scaled between 0 and 1 and time steps of 10 were created. An LSTM layer was added as well as dropout layer which prevents over fitting. In terms of optimizing the data, the Adam optimizer was used. With 100 epochs (see Figure 4.18), the model was compiled. Figure 4.19 was produced after compiling the LSTM.

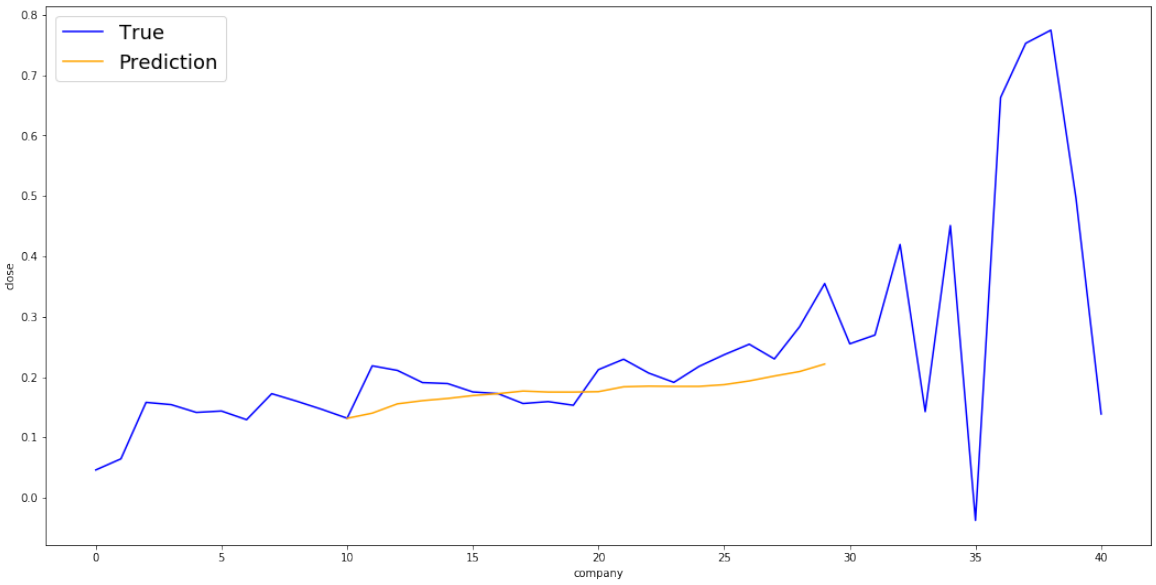


FIGURE 4.13: Standard averaging method applied on a specific day using Top 40 companies close data.

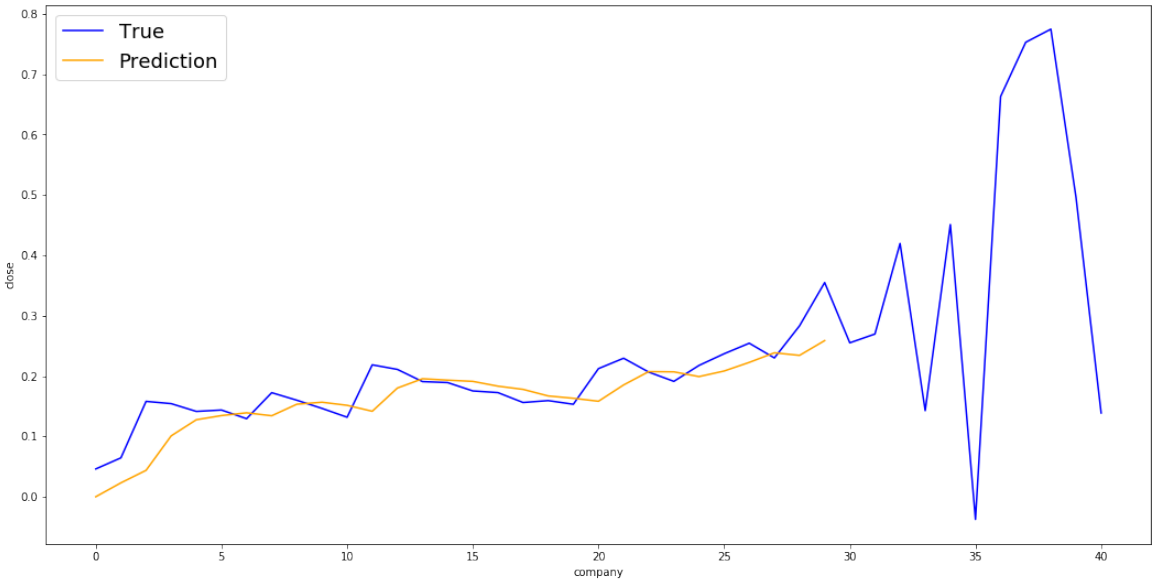


FIGURE 4.14: Exponential Moving Average method applied on a specific day using Top 40 companies close data.

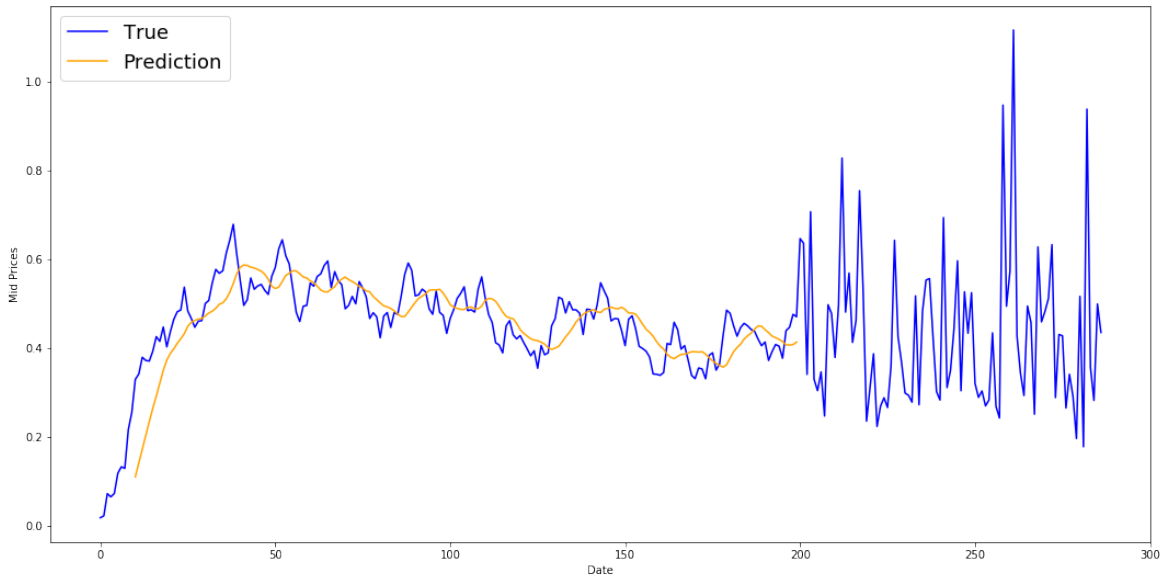


FIGURE 4.15: Standard Averaging prediction of volatility.

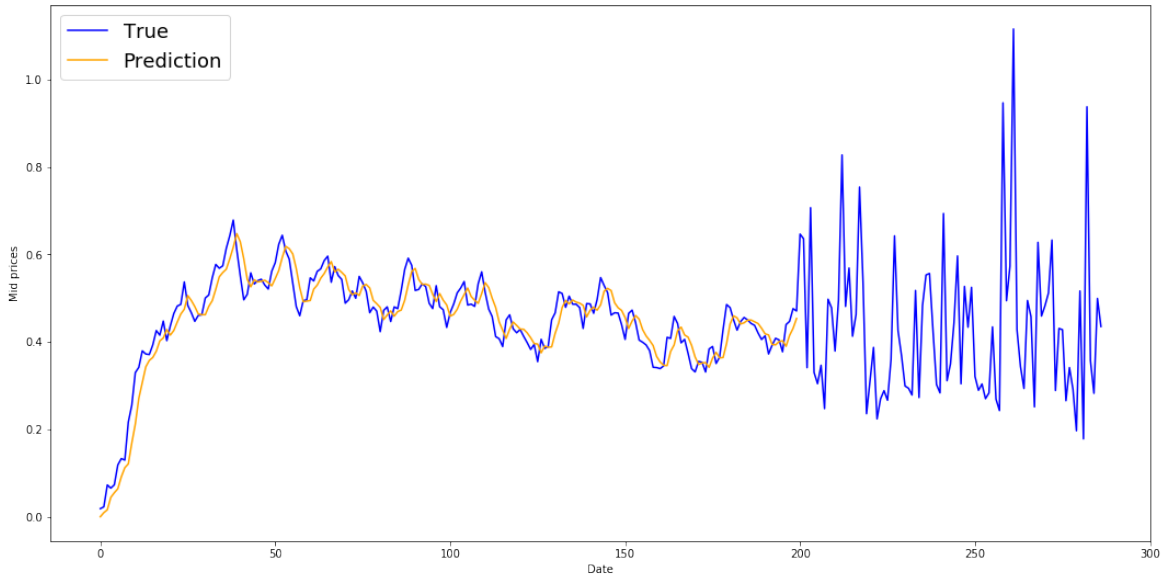


FIGURE 4.16: Exponential Averaging prediction of volatility.

4.5 Summary

4.5.1 Stock Prices forecasting

Stationarizing the data

In order to forecast FTSE/JSE Top 40, the data had to be processed first. The following equation was used to find the differences in FTSE/JSE Top 40:

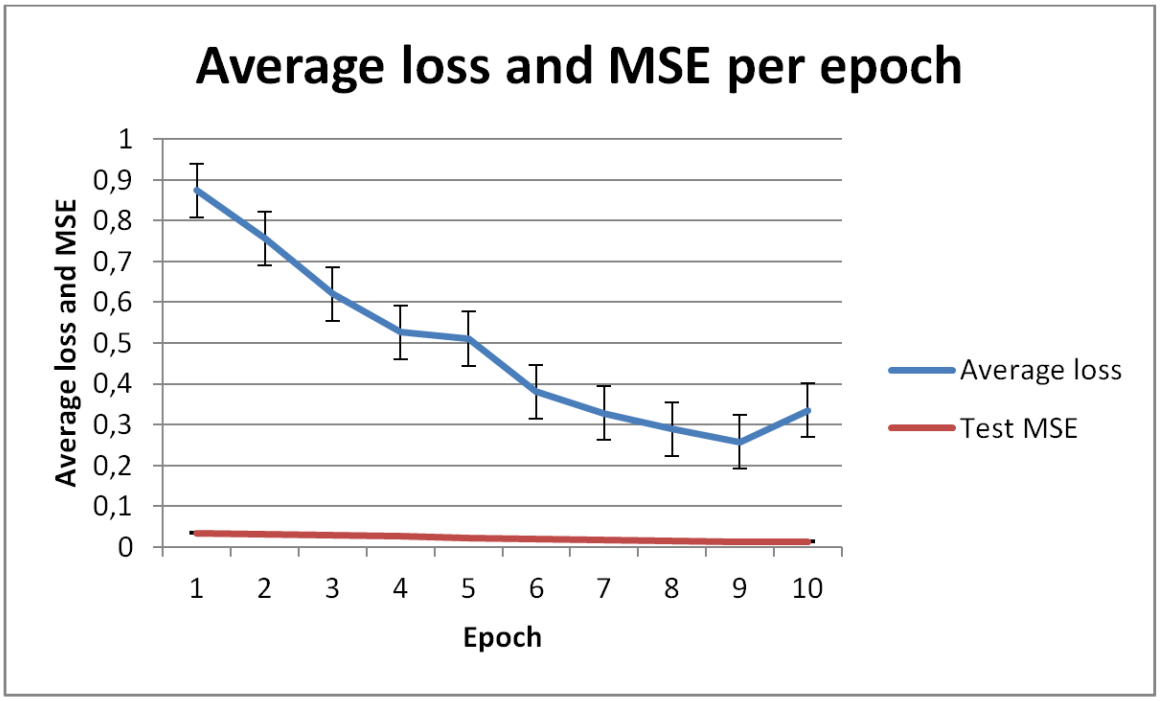


FIGURE 4.17: Average loss and mse per epoch.

$$\frac{newClosing - oldClosing}{oldClosing} \tag{4.1}$$

oldclosing = closing price of the recent period
newclosing = closing price of the day before the most recent period

After manipulating the data, the stationarity was tested on the FTSE/JSE Top 40 data and as seen in Figure 4.5, it was observed to not be stationary with rolling mean and standard deviation changing over time. To confirm its non-stationarity, the test statistic was found to be greater than the critical values that show the confidence levels.

Three approaches were then implemented on the data. After implementing the approaches, the FTSE/JSE Top 40 data was then stationarized. A detailed view of the values is in Table 4.2.

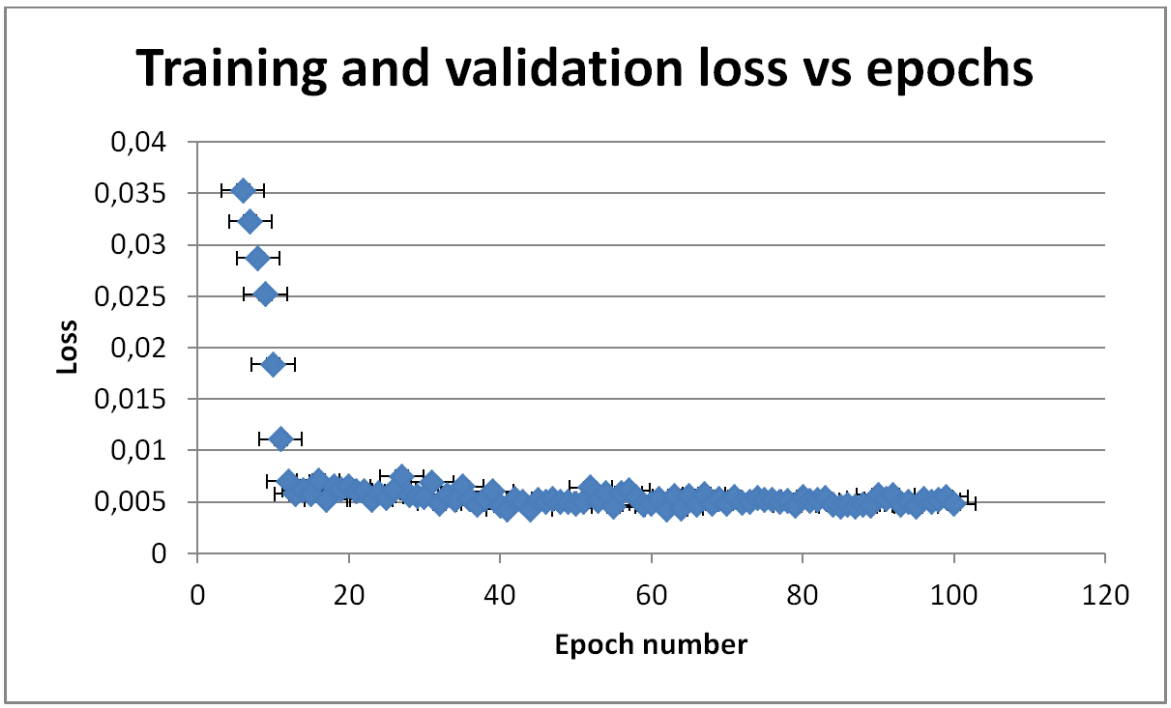


FIGURE 4.18: Training and validation loss versus epochs.

Individual Stock price prediction

Using daily data from 2019-12-01 to 2020-03-31, an LSTM model was compiled and visualized in Figure 4.10. From Figure 4.10, we can see different behaviours of stock prices over time, with a drastic fall during the first COVID19 case in South Africa till the first announcement of the lockdown. However, the model was predicted with a mean squared error of 0,0156.

Node representation

Node representation was obtained from using the t-SNE algorithm. The companies were first represented as nodes and further on they were classified according to their branches found in Table 4.3. As node classification tasks can be classified as individual stock prediction using company data and their closing values, Figure 4.11 shows poor prediction as the nodes are scattered.

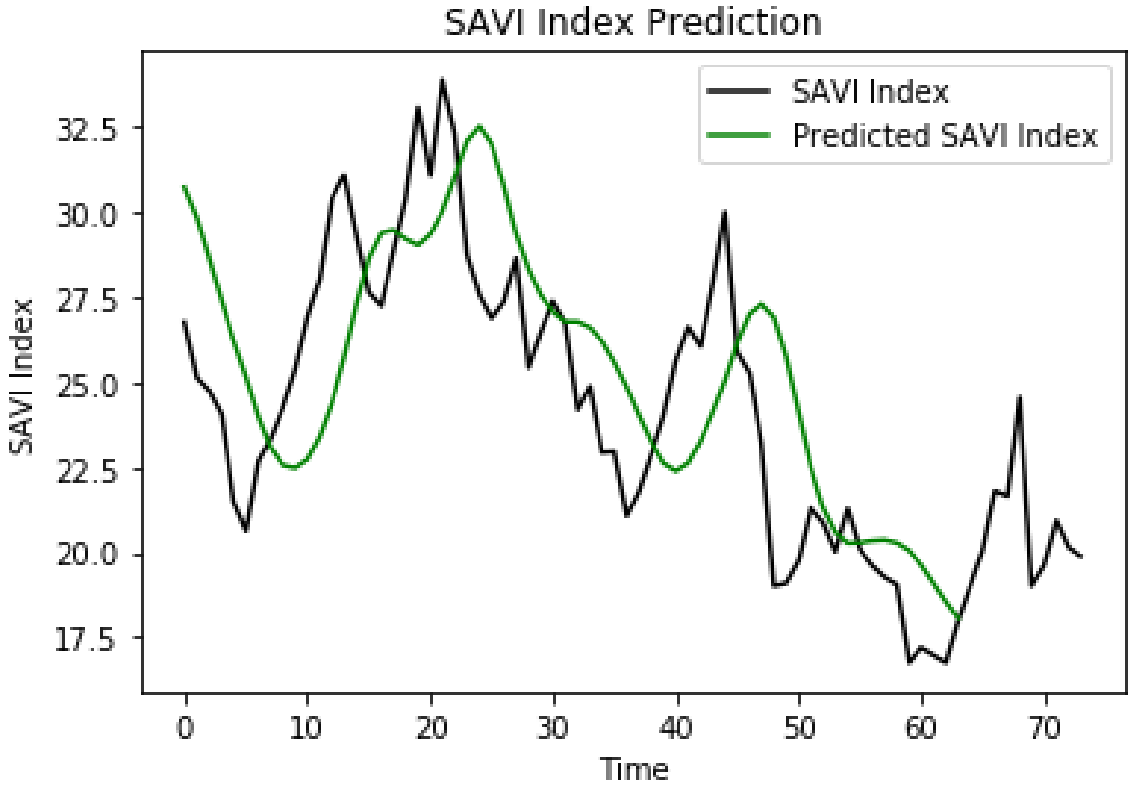


FIGURE 4.19: Exponential Averaging prediction of volatility.

Effects of using relation data

The effect of using relation data was explored on a specific window. The standard averaging approach and the exponential moving average were explored and the mean squared errors can be found in Table 4.4. However Figure 4.13 and Figure 4.14 showed poorly predicted prices.

4.5.2 Volatility forecasting

Stationarizing the data

In order to forecast SAVI, the data had to be processed first. The following equation was used to find the differences in SAVI:

$$\frac{newClosing - oldClosing}{oldClosing} \tag{4.2}$$

oldclosing = closing price of the recent period
newclosing = closing price of the day before the most recent period

After manipulating the data, it was tested for stationarity as seen in Figure 4.3. It was observed to be stationary. To conclude its stationarity, the test statistic was found to be lower than the critical values that show the confidence levels. A detailed view of the values is in Table 4.1.

Historical volatility models

After stationarizing the data, three historical volatility models were implemented on the SAVI data with two using one-step ahead prediction averaging. The data was normalized using window sizes of 10 to ensure as much information can be extracted from the data as possible. All the models successfully predicted volatility, however the exponential moving average approach performed better than standard averaging approach and the LSTM model with a mean squared error of 0,00072, see Table 4.5.

Historical Volatility Models	Index	MSE
Standard Averaging	SAVI	0,00188
Exponential Moving Average	SAVI	0,00072
LSTM	SAVI	0,01013

TABLE 4.5: Historical volatility models of the volatility along with their mean squared error.

4.5.3 Comparing the two forecasts together

In theory, the FTSE/JSE Top 40 and its volatility have a negative relationship. When the FTSE/JSE Top 40 increases, the SAVI decreases as it is a 'fear' gauging tool of investors. When fear is high, stock prices decrease as investors take decisions that will benefit them in the future.

Given that the two parts utilized data from different time frames, a comparison of the two forecasts could not be concluded. However another prediction of FTSE/JSE Top 40 index was done using SAVI dates (see Figure 4.20). From theory and from Figure 2.1, the two forecasts have an inverse correlation. In this case, the dates utilized are from the time there was global crisis, as such stock prices went down and since it was evident that the whole world is going through financial crisis, there was not much "fear" expressed by the investors since they could have easily taken quick decisions to save themselves. In that case, there is not much negative relationship shown between Figure 4.19 and Figure 4.20.

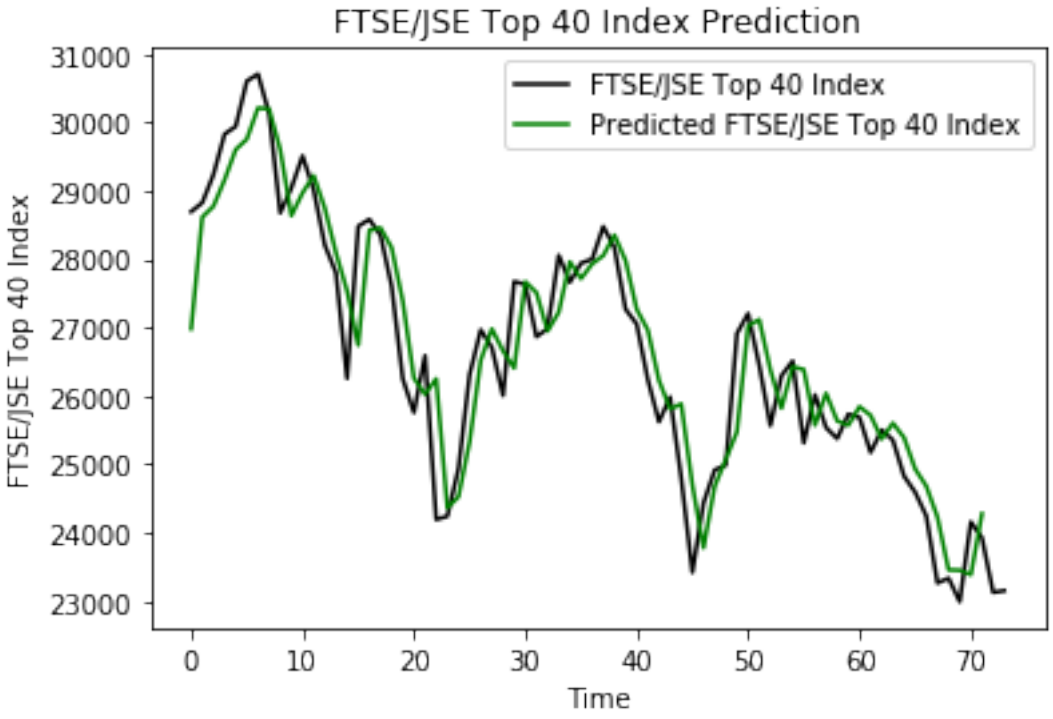


FIGURE 4.20: Forecasting of FTSE/ JSE Top 40 index using the date ranges that produced Figure 4.19

Chapter 5

Conclusions and Future Work

The main aim of the study is to develop a model that focuses on the South African Top 40 listed companies on the JSE in order to predict stock prices. As the prediction method requires relational data, the FTSE/JSE Top 40 data and Top 40 constituents will be the data sets in the research. In order to implement a graph classification algorithm, a hierarchical attention network was implemented. Further, an LSTM model was implemented to assist in this process. In order to implement a hierarchical attention network, node representations of the relation data were required. With different relationships in the data, the effect of using multiple relations in the stock price prediction was studied and from Figure 4.14, poor prediction of stock prices was obtained. As such a hierarchical attention network was not suitable for further analysis.

The objective of the study was to achieve a graph structured model that can forecast South African stock prices. There are quite a number of studies that only focused on node classifications (Chaudhuri and Ghosh (2016),Harrilall and Seetharam (2016),Ibrahim and Ramu (2016),Min (2020)) which are individual stock price predictions and few on graph classifications (Kim et al. (2019),Li et al. (2020)). Focusing on the Top 40 listed firms in the South African market, the results contribute to the literature in South Africa and abroad in terms of how the FTSE/JSE Top 40 index can be graph structured along with its relational data.

In the study, a graph structured model was not achieved, however nodes were represented using the t-SNE algorithm which visualized the FTSE/JSE Top 40 companies, as seen in Figure 4.11. The nodes are shown to be scattered but when the

companies are classified into specific branches shown in Table 4.3, a better clustering of node representations is obtained, as seen in Figure 4.12. As node representations is more like individual stock predictions, Figure 4.12 shows that there can be a group of clusters in the same branch that are found in any time phase. This shows that the prices of stocks from the same branch do not necessarily reflect same direction of the stock movement. Just like what Kim et al. (2019) obtained. For better results the data was divided into training data and testing data that are divided into three classes of upward, neutral and downward movement. For all the data to have meaningful value and contribution, the data sets were divided in window sizes. In terms of evaluation of the results, the data was normalized, optimized and hyper parameters were tuned to a certain range. The performance was measured based on each period's mean squared errors. (Kim et al., 2019). The data was divided into training and testing data. It was normalized and divided into window sizes. Using the LSTM model for individual stock prediction, Figure 4.10 shows a better prediction with MSE of 0,0156. From the plots we do see that with high volatility, there are dramatic falls in the stock market.

It has been seen from previous research studies that the SAVI is generally not predictable. With the significance of this index in the market space, there is both a practical need and economic need to forecast it. With the different methods used by Harrilall and Seetharam (2016), this particular study incorporated the SAVI in forecasting volatility. With the data being stationary, the forecasting of volatility was implemented through SA, EMA and LSTM approaches. From Table 4.5, EMA forecasted volatility better. From theory, like in Figure 2.1, the FTSE/JSE Top 40 index and its volatility have an inverse relationship towards each other. In the study, the last aim was fully concluded after FTSE/JSE Top 40 index was forecasted in the same time frame as SAVI and not much negative relationship was shown, however an inverse proportion relationship was expected.

5.1 Future Work

For future studies HATS will be implementable as it only gathers information from useful relations only. As the relational data needs to be re-evaluated where a feature selection technique will be applied to obtain attention scores the GNN was not able

to be taken further at this stage hence no injective functions were used. Graph classification is interesting to explore in finance and with fewer studies on it, there is scope. With difficulties experienced in forecasting the SAVI, more theoretical models are needed to solve this puzzle.

Bibliography

Chaudhuri, Tamal Datta and Indranil Ghosh (2016). "Forecasting volatility in Indian stock market using artificial neural network with multiple inputs and outputs". In: *arXiv preprint arXiv:1604.05008*.

De Kock, Andre et al. (2015). "Market timing on the JSE using the South African Volatility Index". PhD thesis. University of Pretoria.

Fischer, Thomas and Christopher Krauss (2018). "Deep learning with long short-term memory networks for financial market predictions". In: *European Journal of Operational Research* 270.2, pp. 654–669.

Gervais, Simon and Terrance Odean (2001). "Learning to be overconfident". In: *The Review of Financial Studies* 14.1, pp. 1–27.

Harrilall, Ushir and Yudhvir Seetharam (2016). "Forecasting changes in the South African volatility index. A comparison of methods". In: *EuroEconomica* 34.2.

Hu, Ziniu et al. (2018). "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction". In: *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 261–269.

Ibrahim, Hemavati and S Ramu (2016). "Analysis of Stock Market Volatility using Neural Network for Apple Stock Index". In: *IJSTE - International Journal of Science Technology Engineering,Volume 3,Issue 03*.

Kim, Raehyun et al. (2019). "Hats: A hierarchical graph attention network for stock movement prediction". In: *arXiv preprint arXiv:1908.07999*.

Ladokhin, Sergiy (2009). "Forecasting volatility in the stock market". In: *Unpublished Thesis, VU University Amsterdam, Faculty of Science*.

Li, Kangjie et al. (2020). "Hierarchical graph attention networks for semi-supervised node classification". In: *Applied Intelligence* 50.10, pp. 3441–3451.

Liu, Jianxu et al. (2020a). "Measurement of systemic risk in global financial markets and its application in forecasting trading decisions". In: *Sustainability* 12.10, p. 4000.

Liu, Jintao et al. (2020b). "Multi-element hierarchical attention capsule network for stock prediction". In: *IEEE Access* 8, pp. 143114–143123.

Min, Jonghyeon (2020). "Financial Market Trend Forecasting and Performance Analysis Using LSTM". In: *arXiv preprint arXiv:2004.01502*.

Nguyen, Duc Huu Dat, Loc Phuoc Tran, and Vu Nguyen (2019). "Predicting stock prices using dynamic lstm models". In: *International Conference on Applied Informatics*. Springer, pp. 199–212.

Nguyen, Thien Hai and Kiyoaki Shirai (2015). "Topic modeling based sentiment analysis on social media for stock market prediction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1354–1364.

Odean, Terrance (1999). "Do investors trade too much?" In: *American economic review* 89.5, pp. 1279–1298.

Si, Jianfeng et al. (2013). "Exploiting topic based twitter sentiment for stock prediction". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 24–29.

Soll, Jack B and Joshua Klayman (2004). "Overconfidence in interval estimates." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.2, p. 299.