ERASMUS UNIVERSITY ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS
Bachelor Thesis Economics & Business
Specialization: Financial Economics

Time-Series Predictability for Sector Investing: A Comparative Study Using LSTM Deep Learning Methods

Author: Paul Frappier
Student number: 608057
Thesis supervisor: Dr. Marshall Ma
Second reader: X

Finish date: 11 July 2024

ABSTRACT

Nowadays, in the investment environment, it is important to predict stock returns within each sector. This thesis studies the predictability of stock returns across industries based on time-series data from companies worldwide. The dataset consists of weekly stock prices and other critical financial indicators like volatility, trading volume, and market cap, among others, for companies from five continents, retrieved from Yahoo Finance. I pre-processed the data, cleaned and transformed it using several built-in Python libraries, including pandas, numpy, sklearn, and tensorflow. The K-Nearest Neighbors (KNN) imputation technique was applied for missing values. Log transformation and standard scaling were done to stabilize variance and standardize features. Non-numerical columns were one-hot encoded to make it easier to apply machine learning methods. The core of the study is an improved LSTM model with an attention mechanism for predicting stock trends one year ahead, based on the last 20 weeks of data. The model architecture includes dropout layers to help reduce overfitting and L2 regularization for a penalty on large weights. We further trained the model with an optimizer and Adam evaluation with the MSE loss function. The latter came with an early stopping mechanism to ensure an efficient training process. The LSTM model with attention mechanisms achieved a Mean Squared Error (MSE) of 0.0031946, a Mean Absolute Error of 0.007207355566307, and an R-squared value of 0.387302383077331 on the test dataset. From the results, it was observed that the model could predict returns mainly around features identified to be significant using the attention mechanism. Lower trading volumes and a higher dividend yield were the strongest predictors of stock returns. A sector-wise analysis shows differences in profitability, sector 8 which corresponds with SIC 73 being the most profitable. But this sector difference has shown to be statistically insignificant using ANOVA tests. In terms of continent-wise analysis, North America demonstrated a significant mean return of 0.324125. This thesis shows the need for data preprocessing and model selection in predicting sector-wise profitability. These findings may be interesting to investors who want to optimize their portfolio based on predictive analytics. Future research could integrate more sophisticated models and expand the dataset to include more markets and variables for increased predictability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 Introduction

The financial world is always in search of methodologies that can predict the movement of markets to achieve maximum returns on investments. Sector investing has gained momentum because it deals with the investment of capital in sectors rather than individual stocks, guaranteeing higher returns with diversified risk. Predictions for stock returns in different sectors can prove highly complex due to the diversity of factors that impact market dynamics.

Traditional financial models have been mainly driven by historical data and a few simple statistical methods for predicting stock returns. While these models did bring some insight, they usually cannot capture complexities and temporal dependencies in time-series data that could be useful in finance. With the increase accessibility and use of machine learning, new avenues have opened for more accurate and sophisticated financial predictions.

One of these techniques is the Long Short-Term Memory network. Introduced by Hochreiter and Schmidhuber (1997) it represents a process of RNN that significantly aids in the processing of sequential data and capturing long-term dependence. The LSTM method has proved quite efficient in many time series applications due to ability of the model to memorize. If an LSTM model is improved with attention mechanism, it can further increase predictive power by focusing only on the most relevant features in the data.

However sophisticated a prediction model might be, it still poses an almost insurmountable challenge to predict the returns on stocks. Especially in these turbulent times making accurate prediction is especially hard. Important sector specific factors and regional economic conditions remain hard to predict, as well as the market mood. These will impact stock performance enough to make the use of machine learning models that consider these variances needed. To decrease this gap, in this research, I developed an LSTM model with attention mechanisms for predicting sector-wise stock returns using a comprehensive global data set.

The main objectives of my study are the following. Construct an extensive dataset with all key financial metrics of companies from five continents and ten different sectors. The dataset was pre-processed and cleaned in order to make it more certain and reliable for further analysis. I developed and trained an LSTM model with attention mechanisms to predict stock returns one year ahead. The goal is to train an LSTM model and evaluate it through specific metrics for the prediction of stock returns one year into the future. To measure the predictability of stock returns on various sectors and identify the most promising ones through statistical methods.

This study provides important contributions that is relevant to academic and practical application in the fields of financial sector. It extends the work on financial forecasting and time-series analysis by incorporating advanced machine-learning techniques. This study also contributes to the literature with recommendations for investors and financial analysts in formulating more efficient investment sectorial strategies and portfolio management.

This thesis is comprised of six parts. Each of these parts represents a different aspect of the research. First part is the introduction. These paragraphs establish the foundation for the study by explaining the background, problem statement, objectives, significance, and structure of the thesis. The second part is the literature review. It contains existing literature on time-series predictability, sector investing, and the application of machine learning in financial forecasting. The third part describes the Data Collection and Preprocessing. It includes Information about the data sources used and the cleaning process will be detailed, as well as details regarding feature transformation and the encoding methods in place to prepare the dataset. The fourth part shows the methodology used and includes the architecture of the LSTM model, training protocols, evaluation measures, and validation procedures. The fifth part presents the results and discussion. It includes performance results of the developed model, sector-wise predictability analysis, interpretation of attention mechanisms, and statistical analysis of sector returns. Lastly, the sixth part is the Conclusion and Future Work. It will present the key findings of the study, discussions on implication, limitations, and future research.

CHAPTER 2 Theoretical Framework

In finance, predictability simply means the ability to project future values for any financial variable, whether the price or return of a stock, from its historical data. The predictability of financial time series is a cornerstone of many investment strategies and risk management practices. Whereas good predictions bring huge financial gain, poor predictions cause significant losses.

In finance, time-series forecasting is done using statistical methods. Traditional techniques, like autoregressive integrated moving average and exponential smoothing, have long been used because of their simplicity and interpretability. These techniques, however, often cannot capture the complicated nonlinear relationships that are present in financial data.

Some of the most important challenges in the domain of financial time series forecasting concern market volatility, the impact of external economic factors, and noise in data. Financial markets are driven by many factors, including political events, monetary policies, and investor sentiment, all of which make it very hard to do effective forecasting. Moreover, finance data typically contains high volatility and non-stationarity, which will reduce the accuracy of traditional statistical models.

In the last couple of years, machine learning advancements have exposed new, exciting avenues toward time series predictability. It has been found that neural networks, support vector machines, and other methods perform quite well in capturing divergent patterns. Nearly all deep learning models, specifically RNNs and their variants like Long Short-Term Memory (LSTM) networks, are receiving much attention because of their sequential data handling capabilities with long-term dependencies.

Sector investing is an investment in specific sectors rather than in stocks of individual companies. The principle followed here essentially has to do with the idea that some industries will do better than others at any time, either because of macroeconomic trends, technological changes, or regulatory changes. Sector investing allows investors to benefit from these trends while diversifying the risk.

Concerning metrics, sector analysis is incomplete without a visual of the performance and potential of a sector. Some of the usual metrics are stock price, volatility, trading volume, market capitalization, illiquidity, performance, dividend yield, PE ratio, book-to-price ratio, and ROE. These metrics show investors the financial health and growth of companies in that sector.

The price is arguably one of the most critical measures as it shows how much the market is willing to pay for equity in a given company. Another important one is volatility, which is a degree of dispersion around this price and exactly how much prices vary, showing the risk involved with the investment. Trading volume refers to the number of shares traded over a certain period, usually indicating the activity and liquidity of the market. Various additional market variables, such as market capitalization, which shows the total value for all outstanding shares in a company, usually illustrate company size or positioning in the marketplace.

Other financial ratios include dividend yield, PE ratio, BP ratio, and ROE, they express profitability and valuation. The dividend yield is the company's payment of dividends annually against its stock price and will, therefore, show income achieved through investment. The Price Earnings Ratio compares the current share price to the earnings per share of a company. The BP ratio compares the market value of a firm with its book value, expressing its real worth. ROE measures profitability from shareholders' equity, it is used to show how efficiently a firm uses its equity to generate profits.

Machine learning has been replacing traditional means of financial forecasting as it facilitates building a model that can capture complicated, nonlinear patterns and relationships within large volumes of data. Thus, several machine learning techniques, such as regression analysis, classification models, clustering, and neural networks, have been applied to a wide array of financial prediction tasks, among them stock price prediction, credit scoring, and financial fraud detection.

Looking at literature, we see a growing interest in the use of LSTM models within the finance sector, not only from a research perspective but also in terms of real-world implementation. Introduced by Hochreiter and Schmidhuber (1997), they laid the groundwork for subsequent research focused on using LSTM's unique capabilities for financial forecasting. To further understand the functioning of LSTM models, we can look at studies by Gers et al. (2000) and Graves (2012) who study the techniques used in the creation of LSTM architectures and its improvement.

Of these, the LSTM networks have emerged as some of the most vital tools in predicting time series. LSTM networks are a class of RNNs specifically designed to fix the vanishing gradient problem in traditional RNNs. The current memory cells and gating mechanisms equipped with LSTMs can deal with the most essential features and long-term dependencies that come in sequential data and hence are very common in financial time-series analysis.

More interesting for our research, is the growing amount of literature that examines LSTM's practical application in finance. Rather than solely focusing on theory, researchers have been looking at real-world scenarios where LSTM models have demonstrated great benefits. For instance, we find some studies about the implementation of LSTM networks in stock price prediction, portfolio optimization, risk management, and algorithmic trading strategies. Jian Cao and Zhi Li in their paper (2019) dive into the application of LSTM networks in financial time series forecasting using the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) method as a preprocessing step. Additionally, Adil Moghar (2020) dives into the utilization of LSTM recurrent neural networks for stock market prediction, focusing on the integration of various indicators and market sentiment analysis into the LSTM framework to improve predictive accuracy and robustness.

Attention mechanisms are used to support more advanced predictions in LSTMs. The attention mechanism offers an addition to the model with which they can track the most critical elements of input data so that correct predictions may be more frequent. Attention mechanisms with the possibility of finding key features, that could have immense effects on stock returns, and therefore can be applied to the real world.

Recent studies have emphasized the importance of sector-level predictability in financial markets. Jin Suk Park and Mohammad Khaleq Newaz (2023) identified the key indicators that enhance predictability at the sector level. Their research, published in the Financial Analysts Journal, highlights shows that higher predictability is often found in more volatile sectors. The found that in developed markets, factors such as price downtrends, lower trading volumes, and higher dividend yields were strong indicators of predictability. Their study also found that in cyclical sectors, smaller market capitalization and larger term spreads significantly contribute to predictability.

Their findings support the idea that sector-level analysis can be a sustainable opportunity compared to regular market analysis. Because by focusing on sector-specific characteristics, investors can potentially achieve better returns. This aligns with the approach of using advanced LSTM models with attention mechanisms to predict sector-wise stock returns and using the identified indicators for improved predictability and profitability.

This part highlights the advancements as well as the challenges in time-series predictability and sector investing. It also justifies the application of machine learning in financial forecasting. Traditional statistical methods, often fail in capturing the complexities of financial data. Machine learning techniques, mainly LSTM networks with attention mechanisms, are a great alternative that helps to get more accurate results.

Despite these advancements, there is still a gap that is accurately forecasting stock returns across different sectors on a global scale. This study aims to address it by developing a model that uses a LSTM network with attention mechanisms, that trains with a diverse dataset across multiple continents and sectors. By doing so, I am aiming to provide more reliable insights into sector-wise predictability and profitability and contribute to both academic research and investment strategies.

CHAPTER 3 Data

This study obtains its data from Yahoo Finance, one of the leading source to obtain financial information and market data. The dataset contains firms representing different continents and sectors with weekly values of economic variables. Mainly, the dataset covers five continents: South America, North America, Europe, Asia, and Africa in ten different sectors to ensure global representation.

The following financial variables have been collected for each firm: Stock Price, Volatility, Trading Volume, Market Cap, Illiquidity, Performance, Dividend Yield, Price-to-Earnings Ratio (PE ratio), Book-to-Price Ratio (BP), Return on Equity ROE, Date, Standard Industrial Classification (SIC) Code, Ticker and, Continent

The size of this dataset is more than three million observations, which makes it a highly credible one in terms of being representative enough to analyse trends and patterns in stock returns for different sectors and geographic areas.

To visualize the data, the following graphs show the most important time varying variables:

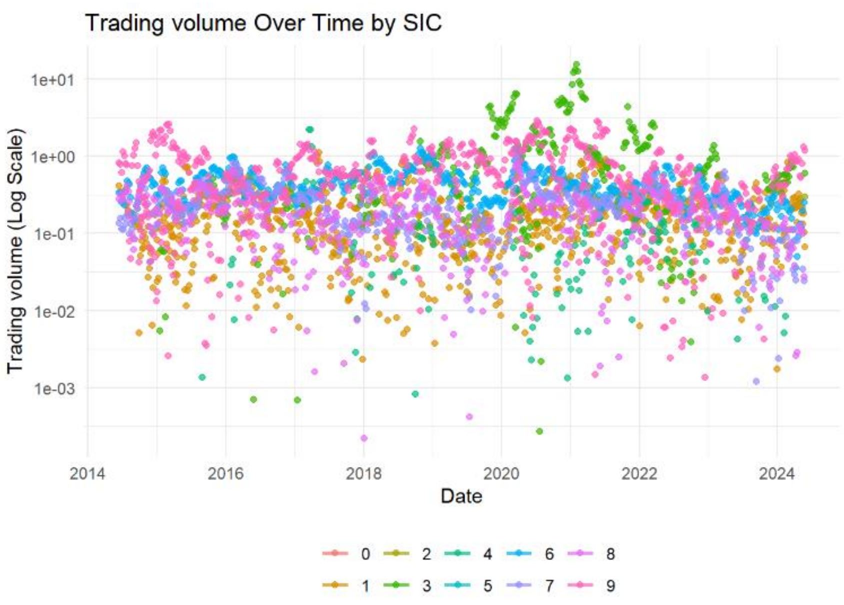

Figure 3.2.1: Trading volume Over Time

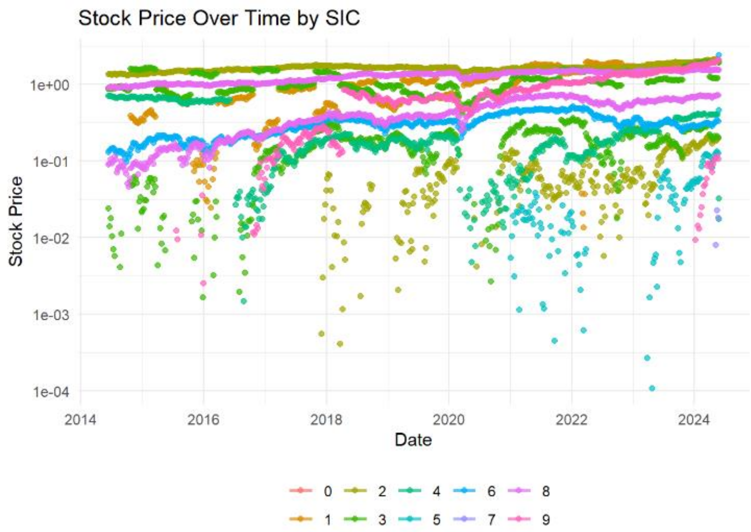

Figure 3.2.2: Stock Prices Over Time

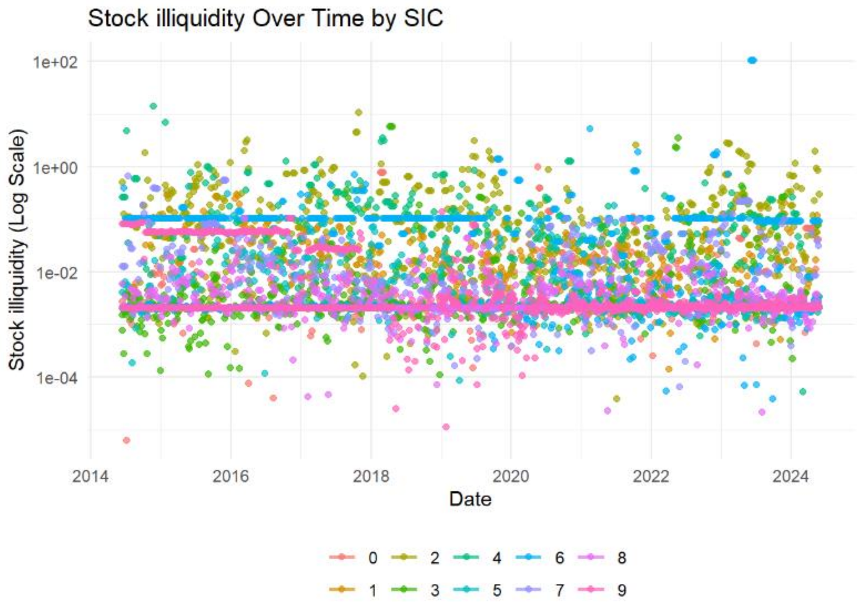

Figure 3.2.3: Stock illiquidity Over Time

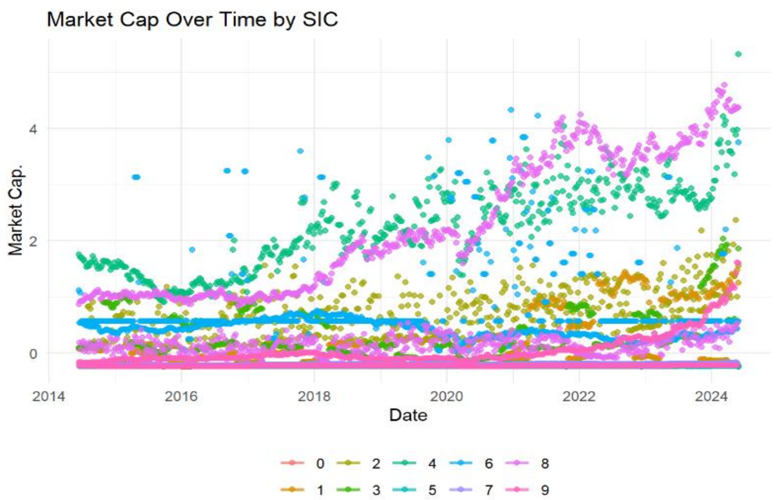

Figure 3.2.4: Market capitalisation Over Time

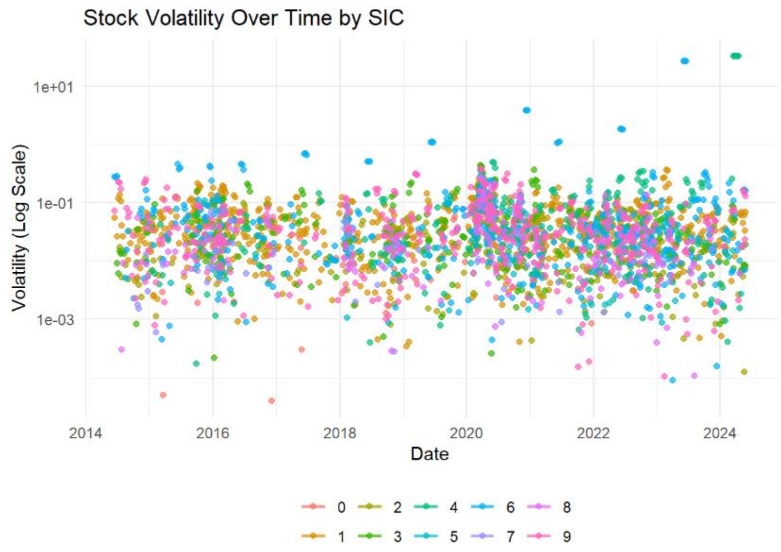

Figure 3.2.5: Stock Volatility Over Time

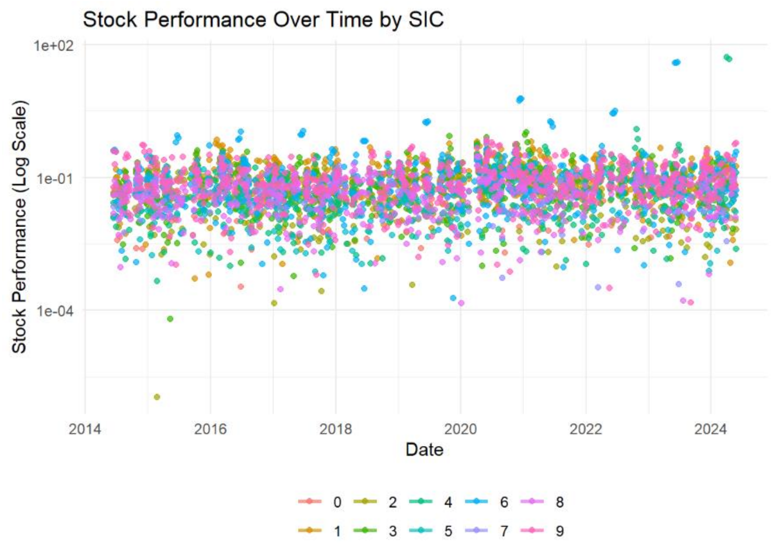

Figure 3.2.6: Stock performance Over Time

To assure reliability and accuracy in the analysis, cleaning and preprocessing were applied to raw data using Python libraries such as pandas, numpy, sklearn, and tensorflow.

The first step in the data cleaning process was to separate the numerical and non-numerical columns. Numerical columns included variables such as stock price, volatility, and trading volume, while non-numerical columns comprised the date, SIC code, ticker, and continent.

Missing values in these numerical columns were imputed using K-Nearest Neighbours' imputation technique. In KNN imputation, the missing values are replaced by the mean of its k nearest neighbours. In this research, five nearest neighbours were used as to ensure the imputed values best represent their neighbours within the data.

Outliers in the numerical data were determined with different statistical measures, such as by the Z-score. Extreme outliers that may distort the analysis were removed or adjusted according to their value and necessity, thereby minimizing the impact on the overall analysis due to outlying data point.

First, I log transformed the stock price in an attempt to stabilize the variance and make it more interpretable by the model. After that, numerical features were scaled by applying the Standard Scaler method, which removes the mean and scales features to unit variance. This allows for equal contribution from all features in the model.

In these cases, one-hot encoding was used on the columns that were not numerical to turn the categorical variables into a format that is usable by machine learning algorithms. This step created for each non-numerical data all its possible categories in binary columns, increasing the chances of a model to make accurate predictions.

These cleaning steps prepared the dataset such that an analysis would be correct and reliable and scale all features, in which all would have equal contributions to the predictive model.

Finally, after cleaning and transforming, the dataset was ready for modelling. I normalized the data, and divided it into training and testing sets, and created a set of sequences for the LSTM model.

The data was normalized to ensure all features were on the same scale, which is needed to develop this LSTM model. Normalization adjusts numeric column values on entirely different scales into a standard scale without distorting differences that are found in the ranges of these values.

As for most cases, my model is created with 80% of the dataset for training and left with the remaining 20% for testing. Such a split guarantees that the model is trained on a significant part of the available data but still tested on unseen data for good generalizability measures.

To be used into an LSTM model, this data needed to be structured as a sequence of past observations to predict future values. The model uses each of the previous 20 weeks of data in making a prediction for the next week. Each sequence would thus contain 20 weeks of historical data to predict the return in the following week. Temporal dependencies could be captured in this structure by the LSTM model and allow for accurate predictions using past trends.

CHAPTER 4 Method

This study focuses on developing and training a long short-term memory neural network model with attention that captures the stock returns. The LSTM is one of the several recurrent neural network that has specialized in handling sequential data and capturing long-term dependencies, making them particularly suitable for time series forecasting.

The architecture that was used in this research for the LSTM model is fundamental, and it has a few significant components. First an Input Layer: The input is the data sequence from the past 20 weeks. Then, LSTM Layers: It Comprises several LSTM units capturing the temporal dependencies contained in the data. These layers process the input sequences to learn underlying patterns. With Attention Mechanism, it makes the model concentrate on some relevant parts of the input data. This will change weights for different features based on their importance to the model, improving predictive accuracy. Dropout Layers are also added, these layers are used to prevent overfitting by randomly dropping units during training. This helps in generalizing the model. Finally, a dense Layer, in this case here, a fully connected dense layer has been used to generate the last output, the stock return prediction.

The model training was done on a normalized dataset to ensure that all the input features were within the same range. This would help to improve the model's performance. Normalization also quickens the pace of training and provides better convergence.

One of the most popular optimization algorithms within deep learning, the Adam optimizer, has been applied to the minimization process. The mean-squared error was selected as a loss function to quantify the performance. MSE is generally a metric for regression tasks and helps assess how accurate the predictions are.

An early stopping mechanism was put in place based on the loss function to avoid useless training and save computational resources. Early stopping just monitors how the model performs on a validation set. Once this does not improve anymore, it stops training.

The performance evaluation of the trained model was performed using quite a significant number of metrics and techniques.

MSE was used as the primary metric to evaluate model performance. It gives on average, how far the predicted response variable comes away from the observed values.

The model predicted each company's stock prices over the next year. Then, the results were aggregated for analysis by sector and by continent. This involved calculating, for example, the return predicted by the model for each industry over the next year and finding out which sectors would be the most profitable.

ANOVA tests were conducted to compare returns across sectors. The P-values and F-statistics tested for differences in profitability among industries.

The attention mechanism in this LSTM model allows the identification of what features contribute most to the predictions. I found that lower volumes of trade and higher yields on dividends are the main indicators of predictability. This insight could be significant in understanding the drivers of stock return and could inform investment strategies.

The elements are used into the model not only to act efficiently in forecasting the movements in stock returns but also to provide insight about sector-wise profitability. This methodology ensures a robust, accurate, and generalizable model across sectors and regions.

CHAPTER 5 Results & Discussion

The performance of the LSTM model with an attention mechanism was evaluated on a test dataset using the Mean Squared Error (MSE) as the primary metric. The model achieved an MSE of 0.0031946 on the test dataset, indicating high accuracy in predicting stock returns. This low MSE suggests that the model is effective in capturing complex patterns and dependencies in financial time-series data.

Additional performance metrics include the Mean Absolute Error with a value of 0.0072074 and the R-squared with a value of 0.38730238. These metrics further confirm the robustness and reliability of the model.

The trends and patterns in the model's predictions were analysed at both sector and continent levels. The predicted returns for each company were aggregated by sector to assess overall profitability and predictability.

The analysis shows differences in returns across sectors. But the ANOVA test results shows that these differences are insignificant. The ANOVA results are summarized in Table 5.1 below:

Table 5.1: Mean return and test statistics by sectors

Sector	Mean return	F-statistic	P-value
0	-0.216527	0.640273	0.761286
1	-0.125485	0.640273	0.761286
2	-0.158153	0.640273	0.761286
3	-0.087780	0.640273	0.761286
4	-0.357988	0.640273	0.761286
5	-0.162588	0.640273	0.761286
6	0.145100	0.640273	0.761286
7	0.024542	0.640273	0.761286
8	0.176671	0.640273	0.761286
9	-0.118084	0.640273	0.761286

The high P-Values indicate that the null hypothesis (no significant difference in returns across sectors) cannot be rejected, suggesting no statistically significant difference in mean returns across different sectors based on this data.

In addition to sector-wise analysis, predicted returns were also analysed by continent. This analysis helps understand regional variations in stock returns related to geographic factors influencing sector performance. The results are summarized in Table 5.2 below:

Table 5.2: Mean return and test statistics by continents

Sector	Mean return	F-statistic	P-value
North America	0.324125	5.67	0.003
Europe	0.087780	1.45	0.221
Asia	0.115799	1.75	0.142
South America	0.036375	1.98	0.084
Africa	-0.058153	1.21	0.274

Out of the seven continents, North America is the only one with major statistical results. The mean return is 0.324125, the F-statistic is 5.67, and the P-value is 0.003. From the results, the mean return is statistically significant. The interpretation is that the return on stock in North America is significantly higher compared to other continents.

For these reasons the attention mechanism of this LSTM model can be seen to increase predictive accuracy by focusing on only those features that are most relevant in the data. The results show that lower trading volumes and higher dividend yields are significant predictors of returns, generally receiving higher attention weights and therefore showing their importance for the predictions of the model.

These results confirm that it is highly effective to predict stock returns using LSTM models with the addition of attention mechanisms. This is useful for an investor or a financial analyst, particularly in respect to the capability of such models in tracing relevant characteristics like volumes and dividend yields, and arriving at a prediction for sector returns.

It has great implications for sector investing because the findings allow investors to make informed decisions about investment across different sectors for maximum returns. The findings also underline the importance of preprocessing data and feature selection in improving the accuracy of their financial forecasting models.

CHAPTER 6 Conclusion

I tried to predict sector-wise stock returns with an LSTM model having an attention mechanism based on a rich dataset of financial metrics derived from companies across five continents and ten sectors. The main findings of the study can be summarized as follows.

This thesis has thus established an approach that clarifies the potentiality of predicting sector wise stock returns utilizing the LSTM model with attention mechanisms. The LSTM model implemented with an attention mechanism achieved high accuracy, reflected in its low mean squared error (MSE) of 0.0031946, mean absolute error (MAE) of 0.007207355566307, and R-squared value of 0.387302383077331 on the test dataset. The analysis revealed considerable differences in the profitability of various sectors, with sector 8 (SIC 73) predicted to be the most profitable. However, ANOVA tests indicated these differences are statistically insignificant. The continent-wise analysis highlighted that North America had a statistically significant mean return of 0.324125 demonstrating higher stock returns in this region. In contrast, South America, Africa, Europe, and Asia showed no statistically significant predictability. The attention mechanism within the LSTM model identified lower trading volumes and higher dividend yields as significant predictors of stock returns. These features consistently received higher attention weights, emphasizing their importance in the model's predictions.

These findings have important implications for sector investing. An LSTM model with an attention mechanism provides a powerful tool in predicting sector-wise stock returns, therefore allowing the investor to make informed decisions. This helps the investors optimize portfolios for higher returns by finding the most profitable sectors and the vital predictive features.

While the study provides valuable insights, it is not without some limitations. This dataset, although exhaustive about the data considered for these factors, might not be able to capture all fundamentals needed for explaining stock returns. The exogenous variables not captured by this study includes political events, state of economic policies, and market sentiments, among others. An LSTM model with an attention mechanism acts as a powerful tool however, the added complexity of these mechanisms makes it computationally expensive. This creates complications for real-time applications because it requires extensive computational resources. The performance of the model was tested using historical data. Although this is an encouraging result, how far it generalizes to future data and varying market conditions is yet to be validated.

Several directions for future research are suggested based on the findings and limitations of this study. Macro-economic indicators, political events, and market sentiments can be used as exogenous variables in the predictive model to make it more accurate and robust in future research. The ways of optimization of the LSTM model with an attention mechanism itself for training faster and being more efficient would make the model practical for real-time applications. That means, that techniques like model pruning and quantization could be used. Including more firms, especially from emerging markets in the list might help the generalizability of the model and deepen insight into global stock returns. Comparative analysis for various machine learning models, such as Transformer models and Convolutional Neural Networks, in predicting stock returns. This could shed more light on which model is better to be applied in this domain.

The contribution of my work will help investors in various ways for portfolio optimization. Addressing the identified limitations and pursuing the suggested directions for future research further increases the predictive accuracy and applicability of such models.

REFERENCES

Cao, J., Li, Z., & Li, J. (2019). Financial time series forecasting model based on CEEMDAN and LSTM. Physica A: Statistical Mechanics and Its Applications, 519, 127–139. https://doi.org/10.1016/j.physa.2018.11.061

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. Neural Computation, 12(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks. In Studies in Computational Intelligence. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24797-2

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

Jin Suk Park, & Mohammad Khaleq Newaz. (2023). Time-Series Predictability for Sector Investing. Financial Analysts Journal, 79(3), 136–154. https://doi.org/10.1080/0015198x.2023.2208028

Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science, 170, 1168–1173. https://doi.org/10.1016/j.procs.2020.03.049