Corporate Vitality Insight: Strategic Evaluation and Analysis Through NLP

by

Abrar Ahmed
21301309
Nusaiba Alam
21301274
Md. Tawhidur Rahman
21301308

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Abrar Ahmed

21301309

Nusaiba Alam

21301274

Md. Tawhidur Rahman

21301308

Approval

The thesis/project titled "Corporate Vitality Insight: Strategic Evaluation and Analysis Through NLP " submitted by

1. Abrar Ahmed (21301309)
2. Nusaiba Alam (21301274)
3. Md. Tawhidur Rahman (21301308)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on February 03, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque

Associate Professor
Department of Computer Science and Engineering
Brac University

Undergraduate Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Undergraduate Thesis Coordinator
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

The research paper introduces a methodology and analysis paradigm for evaluating and investigating a company's state through Machine Learning (ML) and Natural Language Processing (NLP). Going beyond traditional analytical limits, this technique provides a comprehensive assessment of a company's health. This approach enables corporations to make smarter choices, modify their strategies, and position themselves for long-term success. Additionally, the methodology assists Financial Advisors in offering insightful client counsel, eliminating the need for manual company research. In conclusion, this research paper uses Machine Learning (ML) and Natural Language Processing (NLP) to present company health assessments, fostering smarter decisions, adapting to evolving changes, and strategy optimization to sustain success in the corporate world as it navigates the complex terrain of understanding a company's health.

Keywords: Company Evaluation, Stock Market Analysis, Financial Insight, Natural Language Processing, Sentiment Analysis, Machine Learning

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Teacher Farig Yousuf Sadeque sir for his kind support and advice in our work. He helped us whenever we needed help.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

After Tax Roe: After Tax ROE indicates how effectively a company generates profits from shareholders' equity after accounting for taxes.

$$\text{After-Tax ROE} = \left(\frac{\text{Net Income After Taxes}}{\text{Shareholder's Equity}}\right) \times 100$$

Daily Range: The difference between the high and low prices for the day. Daily Range = High - Low

Daily Return : Calculated as the percentage change of the adjusted close price from one day to the next

$$DailyReturn = \left(\frac{\text{Adj Close}}{\text{Previous Day Adj Close}}\right) - 1$$

Debt-to-Equity Ratio: Debt-to-Equity Ratio compares a company's total liabilities to its shareholder equity, reflecting how much debt is used to finance its assets relative to equity.

$$\text{Debt-to-Equity Ratio} = \frac{\text{Total Debt}}{\text{Total Equity}}$$

Eval Accuracy: The metric calculates correctly identified values divided by the full dataset count.

Eval Loss: The error between predicted and actual values

Eval Precision: The proportion of true positive predictions out of all predicted positives.

Eval Recall: The proportion of true positive predictions out of all actual positives.

F1 Score: The harmonic mean of precision and recall by balancing false positives and false negatives.

Fine-tuning: When pre-trained models receive fine-tuning from reduced labeled datasets they are adjusted to perform specific task requirements.

Gross Margin: Gross Margin is the percentage of revenue that exceeds the cost of goods sold, indicating the profitability of the core activities of a company.

$$\text{Gross Margin} = \frac{\text{Net Sales} - \text{Cost of Goods Sold}}{\text{Net Sales}}$$

Median: A median represents the middle value from ordered data sets and results in the average value when the dataset has an even number of values.

Moving Average (MA30): The adjusted close prices averaged over a rolling 30-day period.

Net Profit Margin: Net Profit Margin is the percentage of net income generated from total revenue, showing how much profit a company makes for every dollar of revenue after all expenses.

$$\text{Net Profit Margin} = \frac{\text{Revenue} - \text{Cost}}{\text{Revenue}}$$

Pre-training: When models learn common patterns autonomously from large self-supervised data sets.

Relative Strength Index (RSI30): A momentum oscillator that tracks the rate and change of price fluctuations over a 30-day period.

$$RSI = 100 - \left(\frac{100}{1 + \left(\frac{\text{Avg Gain}}{\text{Avg Loss}} \right)} \right)$$

ROA: ROA measures a company's efficiency in generating profit from its total assets, showing how well a company uses its assets to produce earnings.

$$\text{ROA:} = \frac{\text{Total Income}}{\text{Total Assets}}$$

Rolling Volatility (VT30): The standard deviation of daily returns over a 30-day rolling period.

Ticker Symbol: A stock exchange identifies publicly traded companies through unique symbolic codes known as ticker symbols.

Chapter 1

Introduction

1.1 Introduction

Studying corporate health has been of significant concern in managing the business and companies for a long time. In the past, this valuation largely relied on financial measures such as the overall sales revenue, the profit made, or even fluctuations in share prices. These measures provide information relating to the state of a company at a given time and most of the time do not give a full picture of organizational wellbeing. Specifically, at the start of the 20th century, financial analysis was limited to balance sheet and income statement analysis. This technique focuses on the analysis of a firm's balance sheet, its managers, and its competitive strengths. However, simple financial ratios are not sufficient to understand the health of a company in today's fast-paced and highly computerized world. In the late 20th century the emergence of computers and advancements in information technology revolutionized financial analysis. With the adoption of technical analysis, analysts were able to forecast future stock prices based on historical market data. However, they were only confined to financial models and were not very efficient in qualitative data analysis, like public opinion and social news influence.

There are so many books and articles published on this topic. This work extends several prior studies that have explored employing NLP and ML for sentiment analysis in news articles, stock market analysis, and finance data and text analysis. This paper is devoted to the enhancement of company vitality analysis with the help of technical and fundamental analysis complemented by public opinions as the market behavior depends on them to a great extent [7]. Similarly, we have also seen that the financial state has a relationship with the news sentiment which shows that qualitative data plays a significant role in the financial analysis [8]. The usefulness of sentiment analysis for financial state analysis was established by a predictive model that used real-time news analytics and public opinions to show an important connection between news sentiment, public sentiment, and the financial state of a company [5]. Subsequent investigation into transformer-based NLP models, such as BERT, indicated the ability to determine the financial state of a company through sentiment analysis of news and public opinions, providing a more nuanced view of market dynamics [18]. Applications of neural networks, particularly the backpropagation algorithm, have shown how secondary technical and primary fundamental analysis can be incorporated to provide comprehensive financial assessments for an-

alyzing financial performance of a company [3].

In this case, we will develop a model that provides different units of a company and an evaluation of their state and performance, as well as an overall score that incorporates financial ratios, sentiment from company reviews, stock market analysis, and news analysis.

1.2 Motivation

Understanding a company's performance, reputation, and market dynamics comprehensively is necessary before making an investment in it. Traditionally, investment decisions were made using metrics such as financial data, stock performance, public reviews, and company news. Even though these elements are significant, they are often taken under consideration separately, which gives investors an incomplete idea. To the best of our knowledge, there is no widely recognized model that integrates these diverse data sources into a single framework to holistically evaluate a company's overall health. This apparent gap motivates us to explore and develop such a model. We intend to create a scoring system that incorporates financial indicators, public sentiment, and stock market data using Machine Learning (ML) and Natural Language Processing (NLP) approaches. This research will be an invaluable guide for investors, allowing them to make well-informed and strategic decisions by identifying the most promising companies or industries.

1.3 Structure of Report

The report starts with Chapter **1**, introducing the study, its motivation, and the structure of the report, directs the research statement and objectives. Chapter **2** provides a literature review, covering the survey methodology and related works on stock markets, financial metrics and text analysis. Chapter **3** includes a detailed work strategy to deliver the research objectives through scheduled activities systematically. The background section in Chapter **4** delivers fundamental understanding about news content. The research examines news and company review analysis in addition to stock analysis and finance elements along with their relevant models in the study. The analysis of company reviews and news appears in Chapter **5** together with financial data and stock information. The study methodology section appears in Chapter **6** where the authors describe their data processing stages. The work deals with dataset preprocessing and subsequent model development before shifting to score calculation processes and finishing up with the conceptual framework architecture. The analysis section of Chapter **7** assesses how well the models perform in various sectors. The report ends in Chapter **8** through a summary of the observed outcomes, addressing limitations, and future work.

1.4 Research Statement

In this research, we will analyze the health of corporations and provide a detailed assessment of their situation using natural language processing (NLP) and machine

learning (ML). Through our computational model, we aim to make it easier for corporations and financial advisors to manage the challenges of the corporate state and to make the right decisions without the need for manual analysis.

1.5 Research Objectives

- Design and implement Machine Learning and Natural Language Processing techniques to inspect corporate health.
- Develop a clear framework that assesses various units within the company and provides specific information about the condition and performance of each unit.
- Design a model to evaluate the health of a company considering factors from various aspects of corporate health to get an overall assessment.
- Define the most significant financial metrics that are crucial for a company's performance and apply machine learning and natural language processing to analyze and interpret the company's health.
- Gather and analyze company information from newspapers, and company reviews, to understand the public's sentiment and how it affects the view of the current status of the company.
- Publish the findings and methods of the conducted research in prominent academic journals or conferences concerning natural language processing, machine learning, or business analytics to contribute to the advancement of knowledge in these fields.

Chapter 2

Literature Review

2.1 Survey Methodology

To prepare for our research on corporate vitality insight, we have performed a literature review by obtaining relevant papers from well-known databases such as IEEE Xplore, Google Scholar, Semantic Scholar, and Science.org. Such sources offered a plethora of material related to our goals in the form of stock analysis, financial performance evaluation, and company state analysis based on the news and reviews on social media. Thus, restricting the analysis to papers published between 1995 and 2022 allows us to ground the research in both historical context and up-to-date methodologies while ensuring reliability and relevance. Indeed, the chosen literature matches the objective of examining companies from multiple perspectives, including the use of sophisticated methodologies such as machine learning, data mining, and sentiment analysis. For example, IEEE Xplore Google Scholar sources provided information concerning current trends in stock market prediction and the models of financial assessment; semantic scholars science. gov offered valuable insights into how news and social media affect corporate reputation. This way, our further analysis will be exhaustive, relevant, and sensitive to the changes in modern academic and practical approaches to corporate analysis. Through combining these diverse sources and methodologies, our study is expected to lay solid groundwork for assessing the state of companies correctly and to give insights into their healthfulness and performance from multiple perspectives.

2.2 Related Works

2.2.1 Stock Market

The stock market can be described as an active market composed of shares in public limited companies within the reach of the people to invest in by buying shares. Hence, for a long time, it has been used to determine the nature of the performance of a company because stock prices and trading volumes are always real-time information on the market condition of a firm. By analyzing and observing trends in the stock market, one can determine the profitability of a company or organization in addition to the stability and prospects of the business. Therefore, including computational models like Artificial Intelligence in analyzing the stock market is a promising field in the present world.

Based on the work of Golan and Ziarko (1995) [1], who apply rough set theory (RST) to stock market analysis, their research aims to provide a methodical analysis and identification of patterns in financial data. Their work stems from the suggested framework for building a knowledge discovery method that aims to achieve high forecast accuracy of stock markets while improving the intelligibility of the results. The first step in the application of this methodology therefore entails the gathering of data, and arguably, this is one of the most critical stages of data collection. Discretization of discrete data: This method converts the continuous data to symbolic data to pinpoint patterns in the data that may be of importance. There are several key methods of discretization outlined in the paper to create the best rules for making predictions. The paper of authors Golan and Ziarko is the most representative of employing mathematical knowledge such as the rough set with variable precision models (VPRS) to produce both deterministic and probabilistic rules. Their methodology further accounts for shocks and fluctuations in the stock market due to the stochastic nature of the data as an extension of this integration. The checking of rules using expert analysis also ensures that the generated patterns are correct and accurate in their practical application to making investment decisions. According to Golan and Ziarko, this methodology offers the best approach to the analysis of the stock market. They also claim that in their methodology, less monitoring is required with the rough-set products. In conclusion, the paper of Golan and Ziarko gives a clear and rather useful approach to the methods of analyzing the stock market with the help of mathematical analysis and describes how the rough set theory works, as well as how the stock market can be enhanced further.

Later on, Ramon Lawrence (1998) [2] explored this field by introducing neural networks to stock market analysis. The purpose of the study conducted by Ramon Lawrence was to improve stock market price prediction given that neural networks have been identified to be more helpful in the determination of patterns in a nonlinear system as compared to traditional methods. According to Lawrence, the neural networks are not much different from other techniques, for instance, technical, fundamental, and regression analysis. The data set used for the analysis contained several financial metrics, such as stock price, the volume of stock, and macroeconomic factors. To achieve the objective of the study, several neural network models were compared to their ability to predict. The suggested model by the neural network yielded an accuracy of 92%, while other conventional models such as Box-Jenkins got an accuracy of 60%, and multiple discriminant analysis (MDA) got an accuracy of 74%. The following methods have been applied in the research work: multilayer feed-forward networks, recurrent networks, and the integration of these with expert systems. The strategies that were experimented with were the recurrent neural network with windowing, which gave good return rates of about 50%, while the back-propagation and genetic algorithms gave returns of about 20%–30%. The worst results were obtained when constructing single-layer networks, which means that it is necessary to use even more advanced models of networks to describe the market. The findings of this study were more general and showed that neural networks were not only able to raise questions on the applicability of the Efficient Market Hypothesis (EMH) because they were proven to be predictive but also had better performance than all other conventional and regression statistical tools that

are used in the market.

Building upon the previous foundation, Vaisla Bhaat (2010) [4] added artificial neural networks (ANN) to the field of Stock Market analysis. Vaisla Bhatt [4] provided a worthy exploration of this topic. As we grasped the idea of using neural network-based learning procedures, one of the main advantages of this approach was found to be the ability to model relationships that are by their nature non-linear, which traditional statistical models could not manage. The authors' research analysis focused on the daily stock prices of companies that are listed on the NSE, RBI, and SEBI from April 2005 to March 2007. The authors used five hundred observations as the foundation for their samples in their analysis. They compared the results of the ANN with traditional statistical approaches to forecasting and used MAPE, MSE, and RMSE to compare the models. The accurately trained six-layer feed-forward neural network using the Predict software of NeuralWare could predict stock prices with high accuracy, according to Vaisla Bhatt's paper. This was due to the flexibility of the ANN to learn and adapt to the non-linearity of the input data set. On the other hand, although the statistical methods were more tolerable within a simpler context, they failed as the data got slightly more complicated. Vaisla and Bhatt's research pipeline ensured the stationarity of the acquired data using tools such as unit root tests, which are important to confirm an analytical model. According to their study, the ANN was more effective in managing the complex relationships between stock prices, using technical analysis as a strong rival to conventional prediction models. According to the authors, the strength of their model is that it works for a non-linear model, as stock accuracy is more accurate in a non-linear model than in a linear model. In conclusion, their study not only demonstrated the capability of the ANN to forecast financial data but also provided new directions to enhance predictive techniques, as well as integrating computational intelligence with quantitative analysis to develop new effective approaches for financial market understanding.

In 2015, Attigeri et al. [7] focused on the possibility of using machine learning and sentiment analysis for the stock market forecast and combining them with technical and fundamental analysis to improve the accuracy of their forecast. The study adopts historical data on stock prices and uses data from social media and news feeds to establish the mood. There are other technical analysis tools, that use certain aspects of machine learning methods to analyze the results; Linear Regression, Support Vector Machines (SVMs), Decision Trees, Random Forests, Artificial Neural Networks (ANNs), and Recurrent Neural Networks (RNNs). The fundamental analysis employs the modern NLP approaches to perform sentiment analysis to capture the directional movement from textual information. Another technical contribution in their study is the proposed architecture of ST-GAN which stands for Stochastic Time-series GAN and is a Generative Adversarial Network adapted for time series data. By incorporating textual conventional data with numerical financial data, this model can learn from both data kinds for a better stock predicting model. Regarding the performance measurement of these models, Root Mean Square Error (RMSE) as well as Normalized RMSE (NRMSE) indicators are adopted for forecast horizons embraced in the study, including 1-day, 15-day, and 30-day horizons. Based on the best RMSE and NRMSE results from the current and prior studies, it could be con-

cluded that the proposed ST-GAN model has outperformed both traditional and deeper models. This can be explained by setting the model to predict stock prices better due to the integration of sentiment analysis with technical factors. Thus, the assessment of the hybrid models demonstrates their potential in presenting a more comprehensive perspective of learning the stock market prediction based on numerical and textual data. These gaps suggest that future studies can enhance the model by expanding the range of data to be implemented and investigating the stock level to forecast corporations. In conclusion, the study finds that there is enhanced precision and relevant information in the application of machine learning and sentiment analysis.

Afterward, in the paper by Pahwa and Agarwal (2019) [13], the authors wanted to know how stock market analysis could be improved through the use of supervised machine learning. They employed the data from Quandl and the stock of Google (GOOGL) and endeavored to explain the relationship between several stock features and the closing price of the stock within a day. Linear regression was their compass. They made sure that there was no exposure to other factors that come with raw data by carefully preparing the data and selecting features for the model. However, there were some challenges that they had to go through on their way. They were posed with the problem of feature space and were assisted by the high-low percentage and percentage change as prime indicators of the volatile nature of the stock exchange. Previous researchers have given them something to learn about technical analysis, and they attempted to combine business knowledge to compute it. The final stage was to create a Machine Learning model that had an accuracy of above 95% through the use of optimization algorithms and the aid of the SciPy, Scikit-Learn, and Matplotlib libraries. It is considered a story of how data and algorithms create a new type of investor who can use machine learning to control the stock market and launch the era of predictive finance.

In the field of the stock market, accuracy and precision is very important. To introduce a more precise model in the market, Panwar et al. [17] developed a novel way of optimizing the performance of SVM in the area of predictive modeling including the detection of media bias. In the study, a new SVM model is considered to solve the drawbacks of the linear regression model in the classification through the use of non-linear techniques. In several tests, it was demonstrated that the performance of this architecture is suitable for working with large, non-linear, multidimensional data. To this end, kernel tricks were used to transform the input data into higher dimensional space to enhance the separability of the classes of the given dataset. The media articles employed in this study were numerous and broad to enable the evaluation of the model's efficiency. To support the proposed SVM model, the researchers conducted a comprehensive analysis of the existing linear regression models. This comparison indicated that the SVM-based approach yielded better performance and generalization ability compared to the other methods, particularly in cases where the datasets feature complex or non-linear patterns. During the analysis of the results it can be noted that there were improvements in the classification accuracy, and the achieved SVM model has better results compared to typical approaches with macro F1 score. The study also aimed at identifying potential areas of application of the developed improved SVM model in areas such as sentiment analysis and

predictive analysis. These applications demonstrated the versatility and relevance of the proposed model in solving concrete practical problems in the field of data science. In addition, the authors described specific suggestions for the improvement of the SVM model that are as follows: a) use of better feature selection methods; b) use of more data sources. Thus, the research conducted by Panwar et al is useful for the evaluation of predictive modeling as it reveals the impact of enhanced SVM architecture. As such these results advocate for the increased and continued use of nonlinear models as a way of enhancing the overall results during the analysis of the given data. This work therefore not only serves to enhance the theoretical understanding of SVM but also to offer direction on how the predictive capability of models in numerous fields could be improved.

In recent years, Supsermpol et al. (2023) [24] introduced logistic regression and random forest algorithms in the field of stock market analysis. In their study, Supsermpol et al. [21] sought to evaluate the financial performance of SET newbie firms during their first year of listing using logistic regression and random forest models. In this research, the financial information of the SET firms was gathered for the years 2003 to 2018 with special reference to the pre-listing year and up to three years post-listing. Such financial measures as ROA, ROE, and ROS were used as the key performance indicators for the company's performance appraisal. This study also used articles by Javed et al. (2014) and Durrah et al. (2016), which supported the variables of liquidity, leverage, and operational efficiency as factors that influence financial performance. For purposes of testing the hypothesis related to high, medium, and low classes of financial performance, two models were analyzed, namely logistic regression and a random forest algorithm. Logistic regression was used to measure the performance of the model by comparing the accuracy rates and the area under the receiver operating characteristic (ROC) curve. Preliminary observation of the results as shaded in Table 4 reveals that the average accuracy of the logistic regression model stands at 61% on the average. 64% and the test accuracies ranged between 52. 38% and 57%. 14%. On the other hand, the random forest model indicated better performance with an overall accuracy of 7214% as well as test accuracies ranging from 61.90% to 69.43%. The AUC of the random forest model was between 6388% and 81.35%, with an average of 84.16%. It can be said that the random forest model was better than other models. Logistic regression became less reliable in terms of its predictive capability as compared to the random forest algorithm when applied to large financial datasets to enhance prediction accuracy. The authors concluded that the random forest algorithm yielded a better prediction of the performance of the companies during transitional periods.

2.2.2 Financial Insights

Financial data and resources are valuable when assessing the state of the business because they provide real figures that reveal its financial health and efficiency. Other financial performance indicators such as sales revenue, gross and net profit margins, and operating cash flows are essential in ascertaining the financial fitness of the business. Furthermore, assets – both fixed and current, and also intellectual property speak about the company's future earning capacity and its financial solidity. Understanding these elements enables the stakeholders to evaluate the performance of

the company, recognize threats and opportunities, and consequently come up with appropriate investment and management decisions.

In predicting financial performance, the application of approaches such as neural networks and incorporating both fundamental and technical analysis. The integration between these two concepts has been studied by Lam (2003) [3] with a focus on its effectiveness in improving the accuracy of financial market predictions. To achieve this, his study utilizes a neural network, which fuses the findings of fundamental analysis that uses economic and financial data with technical analysis that involves price and volume patterns on the share. As mentioned, fundamental analysis when it comes to stock investing aims at establishing the true value of an investment based on various accounting ratios, economic figures, and conditions within the respective industry of the invested company. Nevertheless, the more conventional approaches do not capture such high-order structures and the nonlinear, dynamic nature of stock return data. Whereas, by superimposing levels of neural networks, Lam [3] overcomes these limitations. On the side of technical analysis, according to Lam, a neural network is best suited as it identifies the historical data in the market with ease. This capability remains vital for determining future price fluctuations and behaviors in the marketplace. These findings are corroborated by Lam's study, where he clearly shows the potential of neural networks for handling large quantities of data and classification that other methods cannot detect easily. According to Lam [3], his hybrid model is more accurate than a more conventional approach, as the ability to use historical data and adjust to it and the efficiency of working with complex, nonlinear relationships make his methodology rather helpful for financial analysts and investors. Therefore, the expansion of basic and technical analysis by incorporating neural network techniques in Lam [3] reveals the following critical impacts on financial performance prediction: Still, his methodology not only improves the statistical quality of the predictions but also provides a coherent architecture for managing substantially all the challenges that markets pose to investors, which in turn helps to make better investment decisions.

In 2014, Exploring the complex field of financial performance assessment, Ghadikolaei, Esbouei, and Antucheviciene (2014) [6] presented the application of fuzzy multicriteria decision-making (MCDM) methods for Iranian companies. To achieve their goal, their report aims to propose a hybrid framework of FAHP and multiple ranking methods to assess the financial performance of automotive companies listed on the TSE based on accounting and economic value measures. Their approach combines the fuzzy analytic hierarchy process (FAHP) and the fuzzy technique for order performance by similarity to the ideal solution (FTOPSIS), creating a robust framework that transcends conventional evaluation limitations. By leveraging fuzzy logic, they effectively capture the ambiguity and vagueness associated with financial data, enhancing the precision and reliability of performance assessments. The authors demonstrate the applicability of their model to the dataset of Iranian firms to show its efficacy in real-world contexts. They show how the fuzzy MCDM approach can help in the decision-making process by providing a more comprehensive and rational way of making decisions compared to conventional approaches. This study stands as a testament to the potential of fuzzy MCDM techniques for refining financial performance evaluations, paving the way for more accurate and comprehensive

analyses. In Ghadikolaei's et al. methodology [6], the fuzzy analytical hierarchy process is used to determine the weights of the main criteria and sub-criteria. The soft computing approach of fuzzy VIKOR, ARAS-F, and fuzzy COPRAS is employed for ranking the automotive companies listed on the Tehran stock exchange in the period of 2002–2011. Finally, they used mean ranks to determine the final ranking of companies. Their findings underscore the critical importance of incorporating advanced computational techniques in financial evaluations, highlighting the synergistic benefits of combining fuzzy logic with MCDM. This research not only advances the field of financial analysis but also provides a valuable framework for practitioners seeking to enhance their decision-making processes amidst uncertainty.

Analyzing and improving financial performance proved to be essential for the ongoing competitive struggle in the context of London businesses. A variety of studies have assessed different aspects related to financial performance and identified important factors that affect a business's competitiveness. The paper by Simionescu Mihaela (2016) [9] is aimed at presenting a methodology in the literature to assess the financial performance indicators of London-based firms in 2014 and to ensure the support of managerial decisions. Based on the literature review, it can be concluded that Mihaela [9] pointed out the fact that one of the significant features of managing the financial performance process is the applicability of statistical methods. For instance, PCA and cluster analysis are applied in the dimension reduction step for data relating to financial performance as well as for extracting KPIs. According to Mihaela [9], PCA is beneficial in the sense that it aids in the reduction of dimensionality in datasets. It also shows better results for big data and does not have the issue of multicollinearity of variables. Another method that he employed in his classification was the cluster method by which businesses are grouped based on similar financial features or attributes. Furthermore, in his method, correlation analysis is employed, in a bid to determine the relationship between different pieces of financial information. To determine the relevancy of the generated clusters, his methodology employed statistical tests. Last, the KS test is used to verify whether the distributions of the data are normal. Further, Mihaela's literature [9] synthesizes the effects of cash flow and leverage on financial performance. In his study, he notes that cash flow discipline and an appropriate level of gearing are two key factors necessary for the sustainability of the financial environment and its efficient future development. Mihaela [9] claims that firms whose performance on these factors is well managed are likely to record better performance. In conclusion, according to Mihaela's (2016) literature, the measurement of financial performance has many dimensions. It is well understood that ratio analysis measures like cash flow and leverage are instrumental in assessing the financial health of a firm.

Deep learning is an effective method for analyzing financial statements. Thomas Fischer and Christopher Krauss [11], while writing a paper in 2018, have penned on Deep learning with long short-term memory networks for financial markets, and in this, we find elaboration on the elasticity of deep learning concerning financial markets forecasting. The authors' Restricted Boltzmann Machine, RBM: LSTMs are interesting because here, the latter is appropriate for time series data owing to the features that enable it to maintain distant statistics in the working memory. The use of forecast models, especially, in the framework of intricate financial envi-

ronments, for instance, the definition of the right kind of prognosis models is very helpful in trading as a tool that assists in decision-making, and/or recommendation of USE strategies on investments. As the authors suggest, the historical data on the particular stock or some index such as SP 500 common stock index might be used for that purpose. LSTMs have also been compared to other standard and conventional machine learning techniques such as Logistic regression, Random forest, etc As the authors have pointed out earlier in this document the system produces more accurate results than the other techniques. According to the literature stated in the above section, it can be hypothesized that by using the LSTM network, the pattern of the stock price can be modeled and the intricate features of the Financial time series data localized therefore, the utilization of the LSTM network will have a positive impact on the performance of predictions on stock prices. Therefore, the presented list is not exhaustive and it has been applied within the scope of this paper only, along with the comparison drawn between the LSTM models and comparative method. This brings about a proportional increase in contrast and it should provide a much better answer to a very persistent and fundamental question, that is, what role does deep learning provide in assisting the refinement of the industry's forecast of its financial figures? In addition, the authors are not very counter-cultural regarding certain principles concerning the data pre-processing or the training of the models which in a way increases the reliability of the results and which are usually less inclined to fall for bias. However, the paper has its limitations, as mentioned below. These historical stock prices have been extracted from a single index the SP 500 and hence may not be true for other markets or financial instruments. LSTM models also bring about complications and high computational demands, which are also considered implementation issues. However, issues such as overfitting of the model and the ability to pattern future prices based on past price data are mentioned as some of the drawbacks. Altogether, this study greatly benefits the field of financial market prediction as it demonstrates the potential of the LSTM networks. It paves way for the future work and practical implementation where we found that deep learning can further improve the performance of the forecast models in the financial domain. The results imply, therefore, that the attempt should be made to further advance and optimize these sophisticated approaches in the best interest of their application in dynamic and multifaceted market contexts.

In a paper by Yong'an Zhang, Binbin Yan, and Memon Aasma [15] in 2020, they presented a new deep-learning model that can be used to forecast a given financial time series. This model employs CEEMD, PCA, and LSTM networks to enhance the accuracy or two degrees and reliability of the prediction model. The method employed in CEEMD decomposes the time series into several properties based on the generation of IMFs. It then goes further with the analysis using the PCA to reduce the dimensions of the IMFs, in other words, eliminating any redundancy. Subsequently, LSTM networks are used for the prediction of future values for each IMF that can learn long-term dependencies. This is because the final forecast is derived from the predicted values of all IMFs. These conclusions are supported by six benchmark stock indices from different markets showing that the proposed CEEMD-PCA-LSTM model outperforms traditional benchmark models. This leads to a lower mean squared error (MSE) and a higher directional symmetry than the other, thereby showing a better prediction ability. Analyzing the real trading outcomes of

the strategy, it is possible to identify that the model corresponds to higher average earnings and risk-adjusted performance. Moreover, the profit and loss graphs do not show large oscillations that can be associated with a high level of market turmoil, which indicates the stability of the model. However, it is also important to point out some of the following weaknesses of the model. This is computationally expensive due to the use of CEEMD, PCA, and LSTM, which calls for a high number of hardware resources. It is relatively more sensitive to the hyperparameters to be used for each component; hence, it often needs much tuning. However, the model uses past performance and such a plan may fail to capture changes in the markets or any other factor affecting the markets. In conclusion, the CEEMD-PCA-LSTM is an enhanced model with potential benefits in the efficiency of financial time-series forecasting. Its decryption, reduction, and reliability of the financial data put this tool in a vantage position for stock market prediction. However, these aspects involving computational needs and a high degree of parameter dependence are major issues that still must be solved for real-world usage. Future research could concentrate on these aspects to help improve the model's usability and effectiveness in real-life scenarios.

Afterward, Pratyush Muthukumar Jie Zhong (2021) [16] who wrote a paper named "A Stochastic Time Series Model for Predicting Financial Trends using NLP," developed a sophisticated ST-GAN (Stochastic Time-series Generative Adversarial Network) deep learning model that considers both numerical data and financial text for the prediction of stock trends. Prior techniques of stock price prediction have mainly emphasized data of numerical form with the various machine learning algorithms used in the past including Linear Regression, Logistic Regression, Support Vector Machines, Decision Trees, and Random forests besides the deep learning techniques like Artificial Neural Networks (ANNs), Recurrent Neural Network (RNNs), and Convolutional Neural Networks (CNNs). This is because the naivety in combining and training ST-GAN from Naive Bayes sentiment analysis on financial text data and numerical indicators for time series series makes for novelty in this field. This integration ensures that the model can be able to incorporate changes in textual information for instance financial news articles and relate it to the changes in the stock market trends which some models do not capture since all the data inputs contain numerical figures. Specifically, it is composed of an LSTM generator and a CNN discriminator constructed for training on data that is temporally ordered and that may be sparse or not strongly textured. The accuracy of the model to forecast physical night demand is higher in all forecast horizons analyzed (1-day, 15-day, and 30-day) than other models implemented previously in this case study. The model performance was assessed using Root Mean Square Error (RMSE) and Normalised RMSE (NRMSE) measures for a 30-day prediction for Boeing stock price in comparison to existing baselines and traditional models, proving accurate predictions. This work finds out that with the ST-GAN model that combines sentiment analysis and numerical data the possibility of financial forecasting is well enhanced. Further research will revolve around the examination of how the accuracy of sentiments can be improved when using only textual understanding without sensitized analysis and the implementation of the model using a wider portfolio of stocks as well as the application of this model in the portfolio management and decision-making process.

According to Gatawa in 2021 [25], financial metrics play the highest role in evaluating a company's performance, which help investors and stakeholders assess a company's overall health and market position. Gatawa explored the value relevance of financial metrics using data from 300 publicly listed companies on the NYSE and NASDAQ. His study found that a company's state is heavily dependent on financial metrics (Example: Revenue Growth, Profitability). His research shows that stock performance gives mixed results. On the other hand, changes in market price per share are heavily linked by financial metrics the most. His study highlights that stock performance alone will not fully picture a company's value. Instead, financial metrics are very important in evaluating a company's health, along with stock performance, sentiment analysis and other metrics.

2.2.3 Text Analysis

A closer look at the specific comments as well as the different newspaper articles makes it possible to evaluate the general market sentiment about a particular company and the state of the company in question. This makes social media analytics useful in determining the degree to which consumers are happy with their purchases, key markets, and potential reputational risks. It builds on the concept of traditional financial analysis and offers even more information about the conditions in which a particular company is operating and its position within the industry. Thus, by increasing efficiency, becomes possible to make decisions and change strategies regarding positive and negative feedback quickly.

The use of sentiment analysis on news articles to forecast stock price movement has been a topic of interest in the recent past. Initially, Bollen, Mao, and Zeng (2011) showed that the mood of feeds that are posted on Twitter can predict the movement of the stock market thereby creating a link between the sentiment of the public and the stock markets. Similarly, Tetlock (2007) investigated the effect of media content on investors' sentiment and market returns and determined that high levels of media pessimism feed negative pressure on market prices. In the domain of textual analysis, Schumaker and Chen (2009) proposed the AZFin text system for forecasting stock prices using breaking financial news, which confirms that textual information is as useful as quantitative information for forecasting the stock market. It has proven its work as having laid a foundation in the use of natural language processing when making financial forecasts. The Efficient Market Hypothesis as expounded by Fama referred to stock prices as representing all available data. This postulate underpins most investigations of market efficiency, asserting that any new information or sentiment, for example, about the news, should prompt an immediate adjustment of the actual prices of the shares. Loughran and McDonald (2011) built upon this by developing domain-specific sentiment dictionaries for financial texts that enhanced the role of sentiment analysis in forecasting market trends. It focuses on the development of context-dependent dictionaries in the analysis of financial texts. There were improvements made by Nassirtoussi et al. (2014), who conducted a systematic review of text mining techniques for market prediction. They emphasized the effectiveness of using news sentiment alongside other factors that determine the market. Further, Ranco et al. (2015) improved this by combining news sentiment with web traffic data, thus improving the forecast of intraday price changes. Li et al., (2014)

looked at the immediate effect of news sentiment on stock return supporting the evidence that positive news sentiment is positively related to stock price. Similarly, Zhai, Hsu, and Tan (2007) used both news and technical indicators and concluded that the combined approach had a higher predictive accuracy. Engelberg and Gao (2011) investigated how investor attention which can be defined by media coverage affects the movements in the market and thus the relationship between the media, the attention of the public, and the markets. In total, these studies underscore the significance of sentiment analysis in aspects of financial predictions and the advancement of such methodologies to improve the prediction of stock prices.

Leveraging these insights, researchers have applied deep learning to financial disclosures. Such as Kraus and Feuerriegel (2017) [10] attempted to extend the analysis of financial disclosures in making stock trading decisions to use a deep learning approach as compared to text mining. The authors' main interest was to recognize the difficulties stemming from natural language in financial reports and texts in an attempt to enhance the efficiency of the prediction of stock prices. To undertake this study, the data was collected from the financial reports of the firms that are listed on the stock exchange. These documents are important in financial decision-making processes and automated trading as they contain information that, in one way or another, affects stock prices. The authors of this paper found these articles and aimed to determine useful patterns and signals that could be of help for stock price movements. To achieve this, the authors employed several computational models, which include the basic text mining model, such as the bag of words model, a more advanced form of TfidfVectorizer, which uses the term frequency-inverse document frequency, and deep learning models, particularly recurrent neural networks with LSTMs.The efficiency of these models was assessed based on the extent to which they were able to mimic the tendency of stock price movements. The most effective model was the LSTM model, which had an accuracy of 87 percent. The model has an average accuracy of an average accuracy of 5%, thus making it suitable for the management of sequential and complex data in financial texts. Therefore, based on the findings of this study, it was ascertained that the TF-IDF Model is 80% accurate, and although it is higher than the BoW, it is still not effective at capturing textual data. The Bag-of-Words (BoW) Model also had the lowest accuracy of 75%, which states that a word-frequency-based approach is not sufficient when it comes to processing financial text documents. The last part of the study discussed a comparison of the proposed LSTM networks and conventional text mining techniques, and it was seen that LSTM networks provide better and more accurate results when it comes to the analysis of financial disclosures. This enhancement of the computational models is very essential to financial investors and automated trading systems, as it enhances the functionality of the decision support models.

Extending on the prior research on text-mining and deep learning in Specific contexts, the latest research has incorporated these methodologies into both strategic management and financial decisions. Later on, Prollochs and Feuerriegel (2018) [14], therefore, aim to extend the literature on strategic management by developing a text-mining framework that can automatically detect firms' problems from their textual data, particularly financial disclosures. The authors tried to determine the most significant problems that are relevant to the companies and analyze the language

in the context of threats and possibilities. The empirical data for the study was collected from the U.S. Securities and Exchange Commission's EDGAR database, which contains 249,481 filings, of which 29,874 were from the energy sector for the period from 2004 to 2017. They applied Latent Dirichlet Allocation (LDA) for topical analysis of these filings into twenty topics, including earnings releases and FY guidance, production guidance, and MA activity. Their approach involved evaluating the level of risk and the level of optimism for every topic, and the results revealed that topics like earnings results were rather optimistic, with the mean score being 60%, and low risk with a mean score of - 80% conversely. Prollochs and Feuerriegel performed a performance assessment of the proposed model and compared the automated process with the manual process. Based on the LDA-based topic modeling, they also revealed that their approach was more accurate because it offered a more detailed and automated way of identifying strategic issues. Traditional techniques such as the SWOT analysis could only offer analysis of key observable factors at the business unit level but not with the depth, detail, and frequent real-time refresh rate of the risk-optimism framework proposed. The paper did not present specific percentages regarding the performance of the computational models, but it did refer to the application of computer applications and their ability to present data that is up to date. Using LDA has been particularly useful in identifying the primary topics and quantifying the valence and volume of SC, as well as out-competing the manual method in terms of performance and time efficiency.

Consequently, The study by Dogra et al. (2022) [18] contributes to the literature by exploring the differential response of private and public sector banks to news events. The authors gathered the news articles belonging to the seven major news events from the period of 2017-2020 from the prominent financial news portals. These events were grouped into sentiments, positive and negative, using a Random Forest classifier with DistilBERT embeddings, achieving an accuracy level of 94%. The research on bad news, also reveals that public banks are worse and permanently affected than private banks. For instance, while bad news impacts private bank stocks for one month, it has a similar impact on public bank stocks for the subsequent two months. On the one hand, positive information affected public banks for 5 days but private banks for almost a month. This research contributes to the existing literature by investigating how various types of banks respond to news events while highlighting the significance of sentiment analysis in the context of finance. The authors emphasize that the analysis of the sentiment behind the news events is important for investors and policymakers to make proper decisions. These findings indicate that negative news has a greater impact on public banks rather than private banks and can be attributed to several reasons such as market sentiment and regulation.

Similarly, Sinjanka and Yusupha [23] (2023) aim to discuss the use of text mining and natural language processing in the achievement of business intelligence. The primary goal was to address the issue of finding the necessary information within the vast flows of text that were rather unstructured. The authors employ a data set of customer reviews, news articles, financial data, and social media updates made by the company. Several forms of prediction models were studied, including the neural network model with backpropagation and transform models such as the BERT,

GPT-3, GPT-4, and so on. The authors also found out that the effectiveness of BERT and GPT-4 surpassed other models when it comes to sentiment analysis and stock price prediction, with figures standing at 85% and 90%, respectively. Of the developed model, the backpropagation neural network that has incorporated the financial statement variables with macroeconomic factors achieved 75 percent accuracy. The research also emphasized that transformer models, especially GPT-4, achieved the highest accuracy in analyzing context and sentiment from various inputs. On the other hand, the traditional neural network models produced the lowest accuracy, implying that there is a need for smarter ways of handling unstructured data. The studies reveal that the use of sophisticated NLP techniques can be instrumental in achieving greater accuracy in financial forecasting and the assessment of market sentiment, thus improving the decision-making process in organizations. The study also focuses on the integration of quantitative and qualitative data to measure the well-being of a company and provides insight into the limitations of traditional financial analysis tools.

Additionally, In a research work by Joshi et al. [8], the authors have proposed the application of stock trend prediction using news sentiment analysis to predict future stock trends with the help of non-quantitative data obtained from financial news articles. This research hypothesis is that news articles are useful in explaining stock market movements and therefore seeks to determine the relationship between news sentiment and stock trends. The authors chose three models for the classification of news articles, namely Random Forest, Support Vector Machines, and Naive Bayes to predict positive or negative news and stock movements. The data for the study included articles published in the news and the share price of Apple Inc. between February 2013 and April 2016. In this study, the sentiment scores of the news articles were identified, and the text was transformed into the TF-IDF feature vector for classification. The classifiers' performances were measured in terms of accuracy, precision, recall, and ROC area under the curve. Validation methods like 5-fold, 10-fold, 15-fold cross-validation, 70%, and 80% split of the data set were implemented to evaluate the performance of the models. According to the analysis, Random Forest outperformed the other tested models with an accuracy of 88%-90% in all test conditions. The naive Bayes model's accuracy was 83% and the SVM model's accuracy was 86%. The study encompasses an unknown test set of news articles from January to April 2016 for validation purposes, where the RF classifier outperformed. According to the study, using the news sentiment analysis with the Random Forest model which achieved the best performance, the stock trends could be predicted. Joshi et al. in [8] also proposed many ways for future research such as increasing the sample size of the companies used for analysis, utilizing other type of data like Twitter sentiments.

In recent years, Pornpawee Supsermpo et al. (2023) [24] in their large-scale experimental work on the application of NLP for stock market prediction, the authors discuss the analysis of financial news sentiment along with historical stock prices for better prediction. The proposed research is mainly oriented towards the use of the FinBERT model, which is a BERT-based model fine-tuned for performing a financial sentiment analysis. The authors applied neural network models, including ARIMA, GRU, CNN, and LSTM, and compared them with the FinBERT-based

models. In their study, they found out that the FinBERT-based model performed better than the other models with a least RMSE of 370. 155 on a test set of 656 days. This research employed both Dow Jones Industrial Average (DJIA) prices and the 20 most popular daily headlines of the Wall Street Journal from January 2008 to December 2020. This combination allowed the researchers to link the sentiment of the news to stock price fluctuations adequately. By feeding sentiment scores extracted from FinBERT and historical prices into a GRU cell with a linear layer, the model demonstrated improved prediction accuracy compared to traditional and other modern NLP models. An interesting feature of the study is the high scientific relevance and methodological work of the researchers. The authors used several configurations and concluded that FinBERT, including the GRU unit, is the most effective. This approach also validates the possibility of utilizing finetuned DELPRC like FinBERT for predicting financial data. Additionally, it discusses the effects of utilizing different types of information to improve prediction models, including current news and past prices. However, the authors also identified certain constraints of the study which include – employing only one news source and the DJIA index which comprises quite several firms. They suggest further this work by researching new related to certain firms and internal factors and by enlarging the data set and including addendums of, for example, the comments of users and reactions of users to the articles. The combined approach could further sharpen theory and broaden the usefulness of those methods in different areas of finance.

Smith and O'Hare in 2022 [19] analyzed the comparative impact of social media sentiment and financial news on the movements of stock prices. The study was about whether CEO Twitter sentiment or news sentiment could predict the prices of stocks or could influence stock prices. Data for this research was collected from 23 major SP 500 companies between 2019 and 2020. This also included stock performance, CEO tweets, and sentiment scores sourced from Forbes and AP News. In order to assess relationships, Pearson correlation and time-lagged cross-correlation were used. The results showed that financial news sentiments had a stronger correlation with stock price movements, as compared to social media sentiments. CEO Twitter sentiments showed inconsistent and weak relationships, that were also often negatively correlated, but in contrast, the financial news sentiments were significantly more correlated in 20% of the cases. The study also found out that during stable periods like 2019, a little correlation could be observed between stock prices and sentiments. But on the other hand, during the economic turbulence of 2020, financial news sentiments became more predictive. Time-lagged analysis showed that it was stock prices that were driving the social media sentiments rather than the other way around. But it was financial news that had some predictive potential.

Chapter 3

Workplan

The work plan for the thesis project is structured into three main phases: pre-thesis 1, pre-thesis 2, and final thesis or defence.

Pre-Thesis 1 was a three-month phase that started in January 2024 and ended in May 2024. Firstly, there was team formation and domain selection. Using the above domain, we selected our supervisor. Consequently, after familiarising ourselves with the use of NLP in the financial and business sectors, we defined our particular area of research interest and presented it to our supervisor. Next, we wrote an abstract for the paper. We then aimed at compiling and studying papers belonging to the area. After identifying the existing research gaps, we defined our research statement and research objective, and then wrote an introduction. It ended with the writing and submission of the report.

The second phase, Pre-Thesis 2 was a three-month from June 2024 and ended in August 2024. First of all, we searched for the data necessary for the analysis, paying special attention to its relevance and provenance. In the next move, we pre-processed the selected data aiming at preparing it for model training. From the literature identified in Pre-Thesis 1, we chose the appropriate models to apply in our research. We then created and assessed these models, taking our time to study their performance. As a conclusion of our work, we wrote up a Pre-Thesis 2 report.

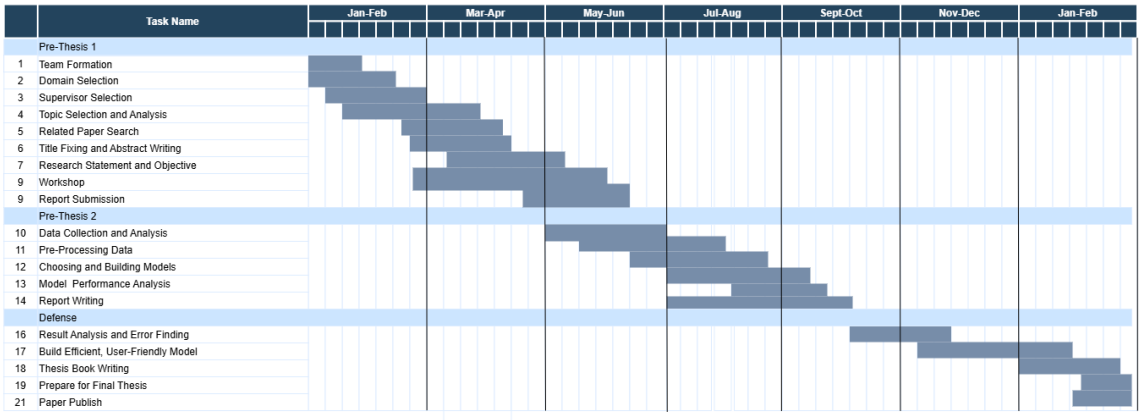

Figure 3.1: Workflow of Comprehensive Thesis Research

The Final Thesis phase started in September 2024 and ended in January 2025. First, we reviewed the results and found possible mistakes. Thus, resolving these issues improved the efficiency and practicality of our model. With reference to normal measures of assessment, the models were assessed and efforts were made to increase their performance ratios. After this, preparations for the final defense will be made, then the thesis book will be written after confirmation. Last but not least, it will be our intention to disseminate the results of the study in relevant peer-reviewed academic and professional journals.

Beyond these existing phases there are strategic post-defense goals to enhance research results from this thesis. Our team will concentrate on improving the thesis document after passing the defense and organizing additional research material including datasets code and model specification details for use as references and future replications.

A workshop seminar will be planned at our institution for presenting our research outcomes together with the approaches we used so that participants can provide feedback and faculty insights. We will participate in an interactive forum to discover upcoming research applications which might enhance our study's ongoing development.

Our team seeks industry specialist collaboration alongside academic experts in NLP and finance to develop real-world implementations of our study findings. These academic connections have the potential to generate multiple research pursuits as well as learning opportunities both for students and professionals seeking employment.

We will build an extensive knowledge base that benefits students and researchers who want to explore this field in the future. Our foundation will compile annotated research articles together with method documentation while offering advice for similar research examinations. Our aim is to boost academic and professional advancement across our research domain through these planned initiatives.

Chapter 4

Background Study

4.1 News & Compamy Review Analysis

For news sentiment analysis, RoBERTa is used as our models.

RoBERTa

The BERT architecture has been adapted in RoBERTa (Robustly optimized BERT approach) which is a transformer model that uses a self-attention model. In contrast to typical models, RoBERTa reads text bi-directionally, taking into account the context of a word in terms of both previous and subsequent occurrences. This capability helps the model understand nuances meanings and word relations that are crucial for understanding sentiment in complex sentences.

Transformer architecture has many layers; the layers involving self-attention are similar to feed-forward neural networks. Self-attention mechanism assesses the relationship between words, weighing it out based wholly on relevance. This allows RoBERTa to pay more attention to relevant text segments when deciding sentiment, intensifying its contextual understanding.

RoBERTa undergoes two primary training phases: **pre-training** and **fine-tuning**. In the pre-training phase, it learns language structure using masked prediction of subsequent sentences and words from large text corpora, which allows the model to understand the language very well. For our model, RoBERTa was trained and fine-tuned to fit specific features with labeled datasets, improving the sentiment classification of news titles.

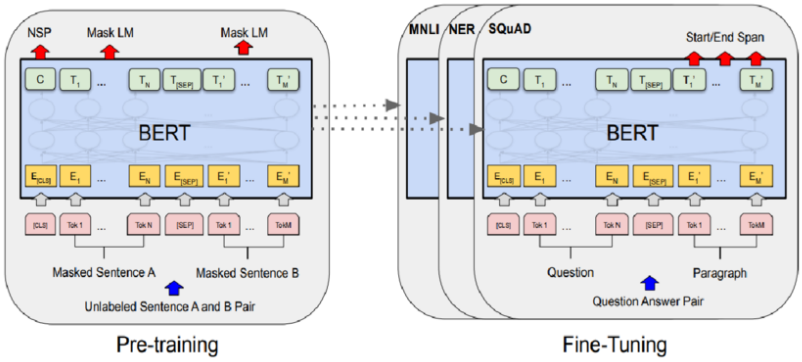

Figure 4.1: Pre-Training and Fine-Tuning BERT model [12]

4.2 Stock Analysis

In our Stock Market Analysis, we used a calculation-based method instead of using machine learning models to analyze the stock market dataset. The nature of the analysis we wanted to do was a stock performance evaluation using predefined financial indicators and calculations.

We loaded and preprocessed the stock market dataset we had using the **preprocess_dataset** module we built. The **preprocess_dataset** module handles data cleaning by removing rows that contained NaN values and sorting the dataset based on date to evaluate better. Another module we built was **processed_group**, which calculates Daily Returns, Gap, Daily Range, Moving Averages, Volatility, and RSI, to evaluate the stock state more efficiently. There is another module, **evaluate_group**, where we take different stock data like Moving Average, Relative Strength Index (RSI), Volatility, Daily Return, Gap, and Daily Range and calculate a "Score."

Finally, results are summarized by the most common score on a monthly basis within the **stock_analysis** method. Instead of storing these monthly scores, we store them in their corresponding year, assigning higher weights to recent months and lower weights to older months. It ensures that recent stock data gets more priority, which makes the model more accurate. For this, our module produced weight arithmetic progression for months and normalized them such that their sum is one. This permits for a calibrated rating where scores from the most recent months exert a far more important impression on the overall marks for the year. We then calculated the weighted yearly scores in the same manner. Our module in the end got the final score based on those monthly and finally yearly data. Upon defining thresholds, the final scores for each stock are placed into qualitative labels, for example, **Favourable** or **Challenging**.

Here, the primary aim was not to predict the future price of the stock or its trend but to assess the current performance of the stock. For this, Machine Learning models might be a way, but building functions to calculate the state is more precise. This provides an interpretable, clear output that is very crucial for Stock Market Analysis.

4.3 Finance

For financial analysis, we used various models to train our finance dataset and considered the best three working models for our methodology.

4.3.1 XGBoost

Extreme Gradient Boosting (XGBoost) is one of the most powerful machine learning algorithms and a black box regressor based on an ensemble of decision trees. It makes models in sequence: each trying to correct errors from previous models. XGBoost is efficient by design, very accurate, and quite flexible.

First, we carefully followed a multi-step procedure to easily incorporate XGBoost into our model and improve finance score regression performance. Therefore, at first, we fine-tuned the XGBoost model using a train-validate split of 85:15 of the dataset. It was configured with 100 estimators and a learning rate of 0.1.

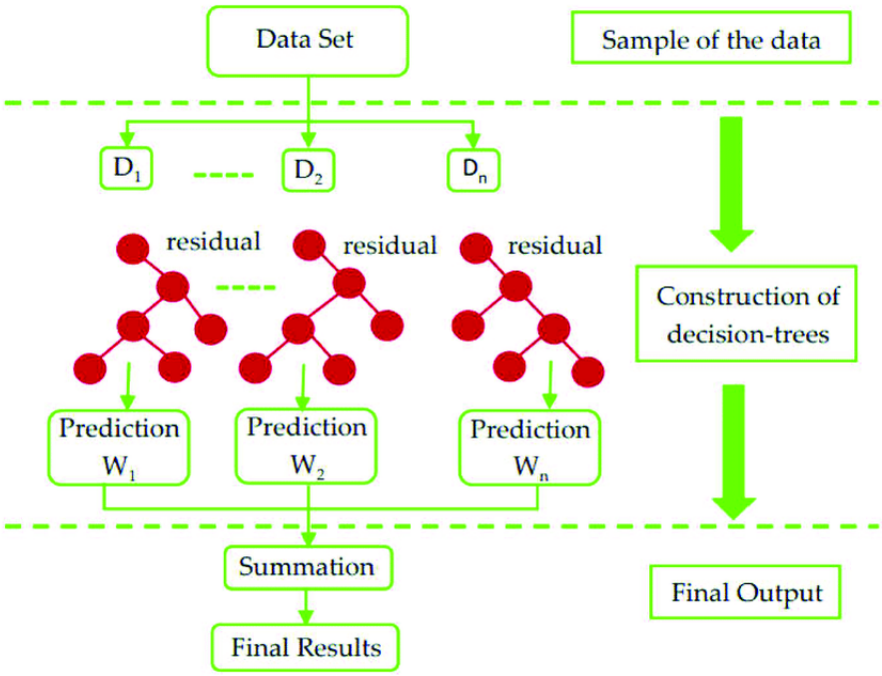

Figure 4.2: XGBoost Model [20]

4.3.2 Random Forest Model

Random Forest is an ensemble learning method. An ensemble learning method is an algorithm that combines the predictions of multiple individual sets of machine learning algorithms to give a more accurate and less variable prediction.

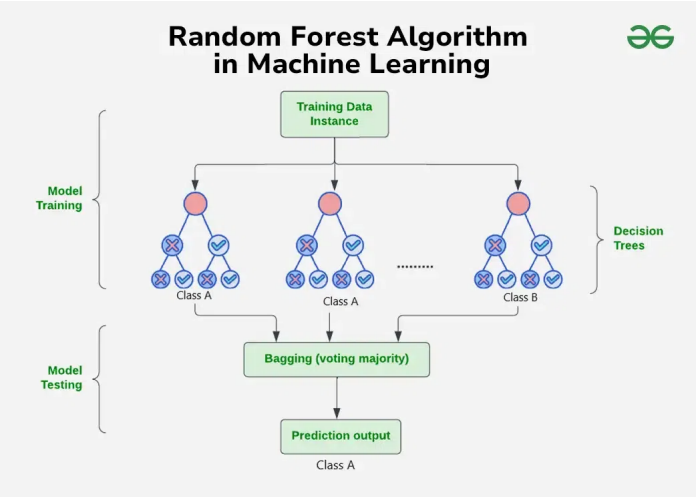

Figure 4.3: Random Forest Model [26]

We applied the Random Forest algorithm using multiple financial metrics to predict financial scores in our model. We selected the Random Forest algorithm because it performs well with non-linear datasets.

4.3.3 Decision Tree Model

Decision trees build regression models in the form of a tree structure. At the same time, it reduces the dataset to smaller and smaller subsets while creating an incremental decision tree.

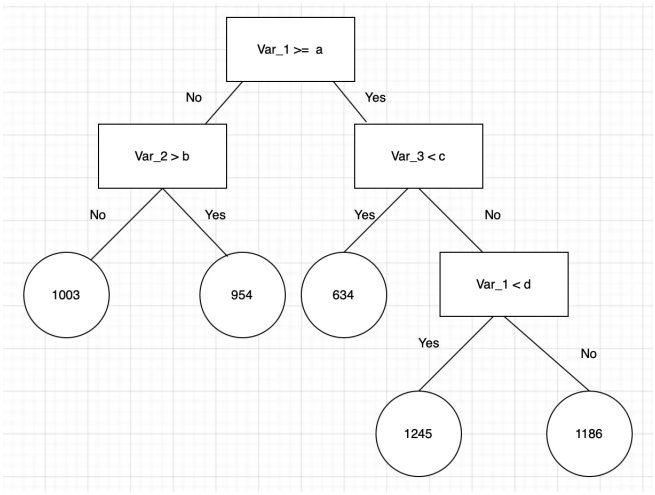

Figure 4.4: Decision Tree Model [21]

Decision Tree Regressor was our choice because it was a simple, and yet powerful method for capturing complex, non-linear relationships found in financial data.

4.3.4 KNN Model

KNN regression is a nonparametric method that fits the continuous outcome as a function of its independent variables through a simple averaging over observations in

the same neighborhood, attending both to the magnitude of the differences between variables and to that between output.

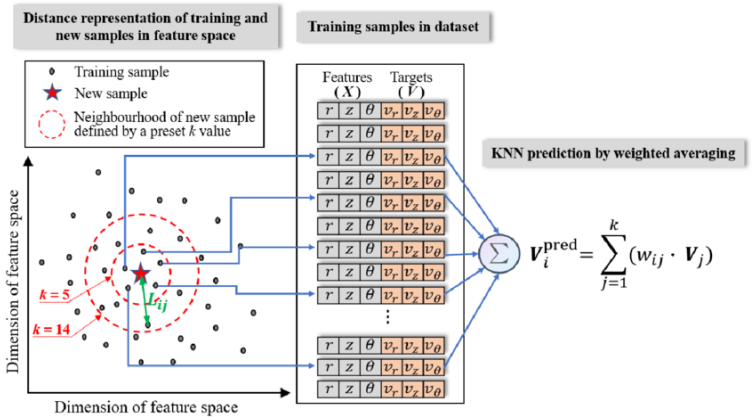

Figure 4.5: KNN Model [28]

For this model, we used KNN with 3 neighbors to predict financial scores, leveraging the relationships between financial metrics.

4.3.5 SVR Model

Support vector Regression (SVR) is a type of support vector machine (SVM) that is used for regression. It is a regression model which takes both linear and non-linear kernels.

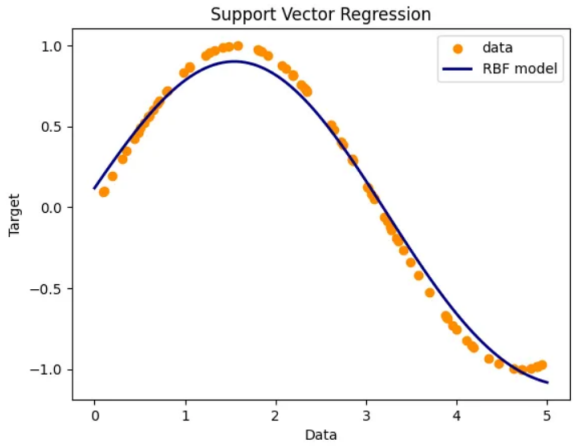

Figure 4.6: SVR Model [27]

SVR can use both linear and non-linear kernels. Our dataset is non-linear, so we trained the SVR using the Radial Basis Function (RBF) kernel to capture non-linear relationships in the financial dataset.

Chapter 5

Dataset

5.1 Analysis of Dataset

We combined data from four primary datasets for our model, each representing various aspects of company and market performance to provide a comprehensive overview. The datasets include:

- Company Review Dataset
- News Dataset
- Financial Dataset
- Stock Market Dataset

These datasets contain data from the following companies: AstraZeneca PLC, Broadcom Inc, Biogen Inc, Copart Inc, CoStar Group Inc, CrowdStrike Holdings Inc, Cintas Corp, Cisco Systems Inc, Costco Wholesale Corp, Cognizant Technology Solutions Corp, Datadog Inc, Dexcom Inc, Dollar Tree Inc, Electronic Arts, Exelon Corp, Fastenal Co, Meta Platforms Inc, Fiserv Inc, Fortinet Inc, Gilead Sciences Inc, Alphabet Class C, Alphabet Class A, Honeywell International Inc, Intel Corp, IDEXX Laboratories Inc, KLA Corp, Kraft Heinz Co, Lululemon Athletica Inc, MercadoLibre Inc, Marriott International Inc, Microchip Technology Inc, Mondelez International Inc, Microsoft Corp, NVIDIA Corp, NXP Semiconductors NV, Old Dominion Freight Line Inc, Paccar Inc, Paychex Inc, Qualcomm Inc, Ross Stores Inc, Starbucks Corp, Synopsys Inc, Tesla Inc, Texas Instruments Inc, T-Mobile US Inc, Verisk Analytics Inc, Vertex Pharmaceuticals Inc, Warner Bros Discovery Inc, Workday Inc, Xcel Energy Inc, Zscaler Inc, Airbnb Inc, Constellation Energy Corp, Amazon.com Inc, Amgen Inc, Coca-Cola Europacific Partners PLC, MongoDB Inc, DoorDash Inc, Apple Inc, Autodesk Inc etc. We excluded Microsoft because our company review source is Trustpilot which is a website of Microsoft. So, reviews can be biased towards Microsoft. Similarly, we also eliminated Google because news dataset was taken from Google News API. Despite the fact that Google has one-star ratings and unfavourable reviews, this approach was chosen to reduce any potential biases in the data. By excluding Google and Microsoft, we can guarantee the neutrality and reliability of our assessment system while prioritizing the safety and precision of its outcomes.

We have used **RoBERTa** for sentiment analysis and labelled the news dataset and review dataset into positive, negative, and neutral columns, as RoBERTa proved itself good for financial sentiment analysis. Although it typically mislabels texts like 'Company fired me for no reason' as neutral, it should be negative and vice versa for other examples, because the text can be positive or negative, but it doesn't fall in the finance category like 'Stock is rising.'. To fix this, we used the **VADER** model to recheck the texts that are categorised as neutral labels, as VADER is a more generalised model that can capture general sentiments.

5.1.1 Company Review Dataset

We utilized data scraping techniques to extract company reviews from **Trustpilot**, focusing on feedback provided by general users. For each company, up to the first 100 pages of reviews were collected, provided that this amount of data was available. Our company review dataset comprises **27262 rows** and **4 columns**. The sentiments of the posts are written in one of the columns.

This dataset reflects customers' posts of different companies over time:

Date	Post	Company	Sentiment
11/3/2024	AstraZenica prof	AZN	Negative
10/6/2021	There genetic m	AZN	Negative
7/30/2024	Don???t go ther	AZN	Negative
9/8/2022	Bye Bye Atra Ze	AZN	Negative
12/13/2023	Have you heard	AZN	Positive
5/11/2024	out of business s	AZN	Positive
3/27/2024	This company m	AZN	Negative
5/28/2021	Brilliant work in c	AZN	Positive
1/15/2022	LIBERTA LIBER	AZN	Positive
4/24/2024	Wonderful medic	AZN	Positive
11/9/2024	a forced subscip	AVGO	Negative
11/6/2024	Just bad..Broadc	AVGO	Negative
10/29/2024	Why can???t I g	AVGO	Negative
10/17/2024	They do not hon	AVGO	Negative

Figure 5.1: Sample of Company review dataset

Our dataset includes several key columns. The human-readable date and time of the post is stored in the **Date** column, **Post** column which contains the detailed review of a company. Finally, a **Company** column notes the related company that has been shared on and the **Sentiment** column contains the sentiment of each post. Each sentiment is then classified as **Positive**, **Negative** and **Neutral**.

Post	Neutral	Positive	Negative
AstraZenica profiteering at the expense of cancer.....	0	0	1
Have you heard AstraZeneca?Good.....	0	1	0
I spend lots of money 3x a week	1	0	0

Table 5.1: Sample Table with Sentiment Classification

Table [5.1] is an example of **Positive**, **Negative** and **Neutral** sentiment of rows. Sentiment analysis is applied on text that can give us value. The values in the other column are **0** and the label of the Positive column is **1**, if the title and the body hold positive sentiment. And that is also the case with the negative and neutral columns.

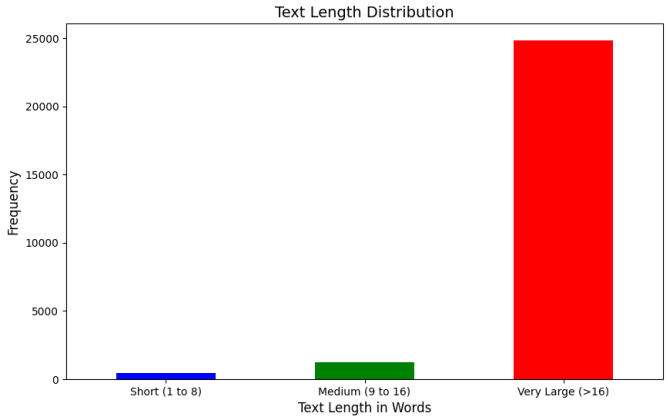

Figure 5.2: Text Length Distribution of company review dataset

Fig. [5.2] illustrates the distribution of text lengths across three categories: Short is Total Number of words in which illustrates occurrence is between **1** and **8** and total instances of a variable are > 9 but < 16. Medium is the total number of words in which word occurrence is more than 9, but not more than 16. Very Large is Total number of words in which the occurrence is more than 16. Over **95%** fall into the Very Large category and have over 24000 occurrences, so by definition most of the texts used in the dataset are longer than 16 words.

5.1.2 News Dataset

The news dataset contains news headlines from Google News that shows the company's performance and their activities in the market from **2021 to 2024**. This article provides some insights on the media coverage based on the company's strategies and consequences. The news dataset has **6 columns** and **9005 rows**.

This dataset reflects News Titles of different companies over time:

PublishedAt	Title	Description	URL	Company	Sentiment
11/26/2024 15:1	Head of Santanc	<a href="https://i	https://news.goo	ARM	Positive
11/25/2024 22:1	Construction wo	<a href="https://i	https://news.goo	ARM	Negative
11/25/2024 18:4	Elon Musk?s Ne	<a href="https://i	https://news.goo	ARM	Neutral
11/26/2024 12:2	Kremlin, Medvec	<a href="https://i	https://news.goo	ARM	Negative
11/27/2024 23:3	Ravens TE Char	<a href="https://i	https://news.goo	ARM	Negative
11/26/2024 23:0	Lawrence woma	<a href="https://i	https://news.goo	ARM	Negative
11/24/2024 23:2	Man in Pawtucke	<a href="https://i	https://news.goo	ARM	Negative
11/27/2024 18:4	Tether-Backed N	<a href="https://i	https://news.goo	ARM	Neutral
11/28/2024 0:45	Victim speaks ou	<a href="https://i	https://news.goo	ARM	Negative
11/28/2024 0:00	Charlie Kolar Su	<a href="https://i	https://news.goo	ARM	Negative
11/27/2024 19:0	Elon Musk's Neu	<a href="https://i	https://news.goo	ARM	Neutral
11/28/2024 2:33	Ravens TE Char	<a href="https://i	https://news.goo	ARM	Negative
11/27/2024 16:1	Teen farmer surv	<a href="https://i	https://news.goo	ARM	Negative
11/27/2024 17:2	Baltimore Raven	<a href="https://i	https://news.goo	ARM	Negative

Figure 5.3: Sample of the dataset

Fig. [5.3] contains numerous key columns that include important data for analysis. The Published at column indicates the date each news item was published, whereas the Title column gives the article's headline. The URL column contains links to the full articles for further reference. Additionally, the Company column shows which company is featured in each headline. Finally, the Sentiment column assigns each headline's sentiment to one of four categories: positive, negative and neutral.

Title	**Neutral**	**Positive**	**Negative**
UAE state oil group ADNOC sets up international investment arm XRG - Reuters	1	0	0
Head of Santander's Portuguese arm says Novo Banco could be of interest - Reuters	0	1	0
Ravens TE Charlie Kolar Out With Broken Arm - pro-footballrumors.com	0	0	1

Table 5.2: Sample table with Sentiment Classification

The table [5.2] is an example of positive, negative and neutral sentiment rows. If the title hold a positive sentiment, the positive column's label is 1, and the other column's values are 0. Same goes for the negative and neutral columns.

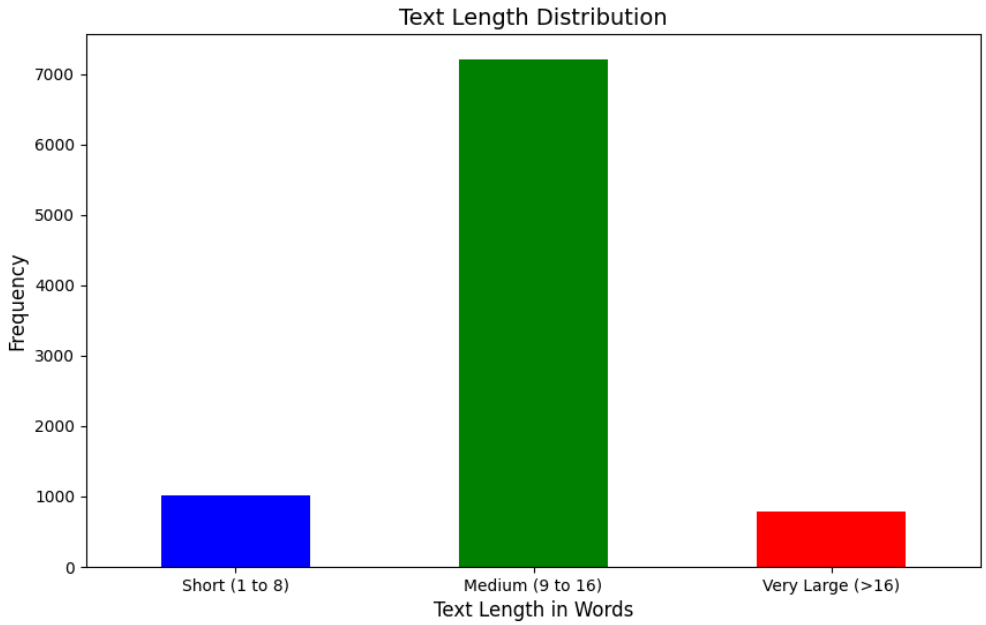

Figure 5.4: Text Length Distribution of News Dataset

Fig. [5.4] demonstrates the title sentence length of each data row over the whole dataset. The **most common length is 9-16**, categorized as medium length, where almost 7000 titles belong to. Almost 1000 titles belong to the short category, which has a 1-8 word range. More than 500 titles ended up in the very large category, which have more than sixteen words.

5.1.3 Finance Dataset

To ensure the reliability of the dataset, we extracted financial data using Alpha Vantage's API. In the Alpha Vantage dataset, We have collected financial data from 2022 to 2025 of several companies from Alpha Vantage. Our dataset consisted of 720 rows with 9 columns. All of those companies have **Ticker Symbol**, **Company Name**, **GICS sector**, **Gross margin**, **Net profit margin**, **Debt to equity ratio**, **ROA**, **After Tax ROE** and **years** for each.
This dataset reflects the financial state of different companies over time:

Ticker Symbol	Company Name	GICS Sector	Gross Margin	Net Profit Margir	Debt-to-Equity R	ROA	ROE	Years
ARM	Arm Holdings pl	MANUFACTURI	0.9573459716	0.00473933649	0.6948901506	0.000436935624	0.00074055788	2023-03
ARM	Arm Holdings pl	MANUFACTURI	0.9540740741	0.1555555556	0.5873015873	0.01567164179	0.02487562189	2023-06
ARM	Arm Holdings pl	MANUFACTURI	0.9429280397	-0.136476427	0.4267756128	-0.016152717	-0.023046302	2023-09
ARM	Arm Holdings pl	MANUFACTURI	0.9563106796	0.1055825243	0.42186251	0.01222768798	0.01738609113	2023-12
LIN	Linde plc Ordina	LIFE SCIENCES	0.190841554	0.1429789307	0.8932569886	0.01418439716	0.02732583851	2022-03
LIN	Linde plc Ordina	LIFE SCIENCES	0.08506862281	0.0440132513	0.9286938549	0.00477627271	0.009376417805	2022-06
LIN	Linde plc Ordina	LIFE SCIENCES	0.1948969131	0.1450051259	0.9403635591	0.01712794155	0.03383118954	2022-09
LIN	Linde plc Ordina	LIFE SCIENCES	0.221910828	0.1691719745	0.956105726	0.01667126968	0.03317677626	2022-12
LIN	Linde plc Ordina	LIFE SCIENCES	0.24301128	0.1858754291	0.9750312735	0.01887732231	0.03792844633	2023-03
LIN	Linde plc Ordina	LIFE SCIENCES	0.2534347399	0.1932041217	0.9388389166	0.02000813029	0.03946280474	2023-06
LIN	Linde plc Ordina	LIFE SCIENCES	0.2532347505	0.1928527418	0.9663478842	0.02010870264	0.04023343102	2023-09
LIN	Linde plc Ordina	LIFE SCIENCES	0.2474340176	0.1885386119	0.9998992951	0.01909393523	0.0388469285	2023-12
AZN	AstraZeneca PL	LIFE SCIENCES	0.6917471466	0.03388937665	1.758778206	0.00384949089	0.01062190424	2022-03
AZN	AstraZeneca PL	LIFE SCIENCES	0.721660013	0.03342308049	1.68729823	0.003727518405	0.01001892464	2022-06

Figure 5.5: Sample of the Finance dataset

Correlation Heatmap Analysis: The heatmap in Fig. [5.6] shows the correlation matrix of five financial metrics: Gross Margin, Net Profit Margin, Debt-to-Equity Ratio, ROA, and After-Tax ROE. A high positive correlation (0.97) is observed between Debt-to-Equity Ratio and After-Tax ROE, indicating that as one increases, the other tends to increase as well. Net Profit Margin shows a moderate positive correlation (0.26) with ROA, suggesting that improvements in ROA are associated with higher Net Profit Margins. Gross Margin has slight positive correlations with both ROA (0.17) and Net Profit Margin (0.18), reflecting a weaker but consistent relationship. Meanwhile, the correlation between Debt-to-Equity Ratio and ROA is nearly neutral (-0.0046), indicating little to no interaction between them. Overall, the analysis highlights the interconnectedness of key financial metrics, offering insights into a company's financial dynamics.

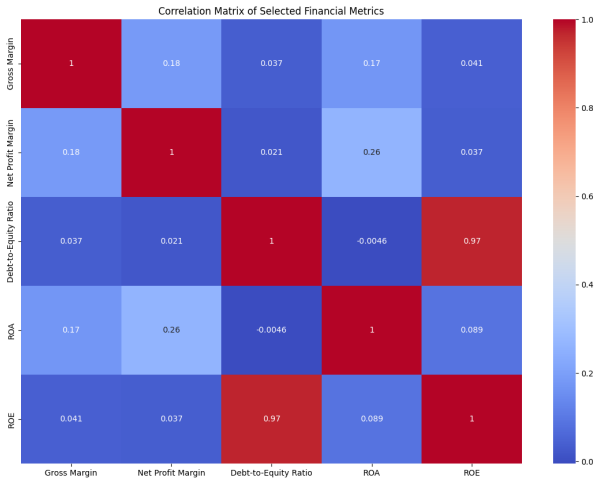

Figure 5.6: Correlation Heatmap of Finance Metrics

Analyze Relation Between Features: The pair plot in Fig. [5.7] illustrates the relationships between financial metrics: Gross Margin, Net Profit Margin, Debt-to-Equity Ratio, ROA, and After-Tax ROE. The diagonal density plots display the distribution of each metric, revealing that Gross Margin and Debt-to-Equity Ratio have skewed distributions. Positive correlations are evident between metrics like Gross Margin and Net Profit Margin, as well as between ROA and Net Profit Margin, indicating that these metrics tend to increase together. Conversely, the Debt-to-Equity Ratio shows little to no correlation with other variables, as seen in its scattered data points. This analysis highlights both the relationships and independence among key financial metrics.

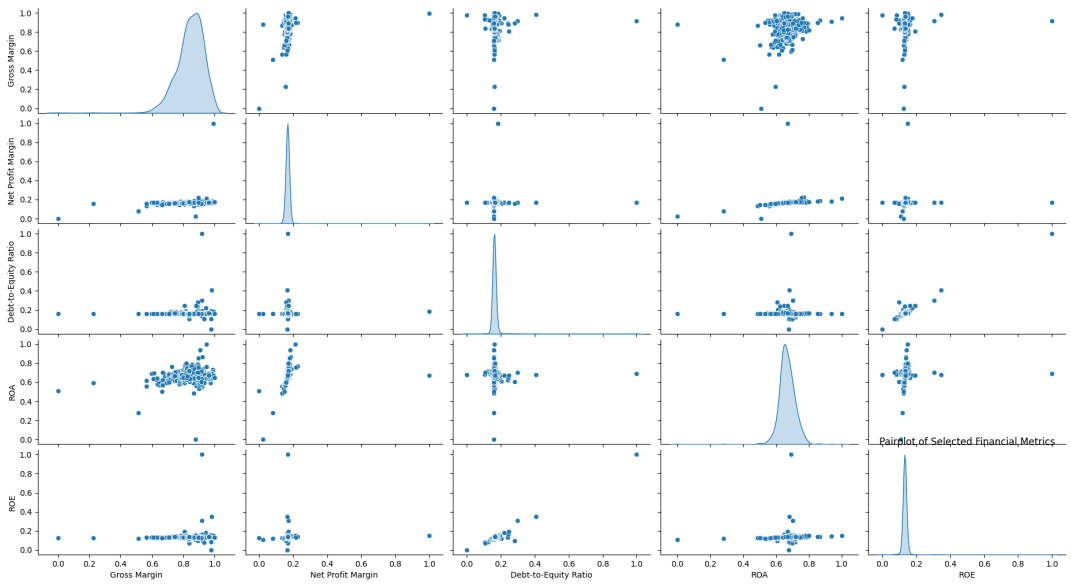

Figure 5.7: Relationships between Financial Metrics

5.1.4 Stock Dataset

The stock dataset sourced from Yahoo Finance includes data from 2022 to 2024 of 14 companies. It consists of **66,390** rows and **8** columns. The key columns for the stock market analysis are **Date, Open**[1], **High**[2], **Low**[3], **Close**[4], **Adj Close**[5], **Volume**[6], **Symbol.**

Date	Open	High	Low	Close	Adj Close	Volume	Symbol
2023-09-14	56.09999847	66.27999878	55.54000092	63.59000015	63.59000015	130534500	ARM
2023-09-15	68.62999725	69	60.75	60.75	60.75	74568900	ARM
2023-09-18	57.95000076	58.74100113	55.02000046	58	58	34571900	ARM
2023-09-19	56.25	56.77999878	53.88000107	55.16999817	55.16999817	18308600	ARM
2023-09-20	55.16999817	55.40000153	51.52000046	52.90999985	52.90999985	16369500	ARM
2023-09-21	51.77999878	52.79999924	49.84999847	52.15999985	52.15999985	14980700	ARM
2023-09-22	52.90000153	52.90000153	50.35499954	51.31999969	51.31999969	8430200	ARM
2023-09-25	51.11999893	54.5	50.02000046	54.43999863	54.43999863	12896500	ARM
2023-09-26	53.93000031	54.52999878	52.70000076	53.52000046	53.52000046	6283500	ARM
2023-09-27	54.40000153	54.40999985	51.78900146	52.99000168	52.99000168	6866700	ARM
2023-09-28	53.5	56.5	52.86000061	55.49000168	55.49000168	8765800	ARM
2023-09-29	56.20999908	56.79000092	53.08000183	53.52000046	53.52000046	7797500	ARM
2023-10-02	53.47499847	54.18999863	51.90999985	52.25999832	52.25999832	8857100	ARM
2023-10-03	51.95999908	52.23500061	50.79999924	51.56999969	51.56999969	4687700	ARM

Figure 5.8: Sample of the Stock Market dataset

This Stock Market dataset reflects the financial performance of the company over time.

[1]**Open:** The stock price at the start of the trading day.
[2]**High:** The highest stock price during the trading day.
[3]**Low:** The lowest stock price during the trading day.
[4]**Close:** The stock price at the end of the trading day.
[5]**Adj Close:** The adjusted closing price (accounting for factors like dividends).
[6]**Volume:** The number of shares traded during the day.

Chapter 6

Methodology

6.1 Dataset Extraction

6.1.1 Model Training Dataset

NASDAQ's one of the top 56 companies' financial, stock market, news, and company review data was collected. All the data was spanned between **2022 and 2025**.

Financial Model Training Dataset

The **Alpha Vantage**[1] API, which offers historical and current financial data for several companies, was used to gather the financial dataset for training the Finance model. **Alpha Vantage** was selected because of the accuracy and reliability of its data. The extracted dataset contains Ticker Symbol, Company Name, GICS Sector, Gross Margin, Net Profit Margin, Debt-to-Equity Ratio, Return on Assets, Return on Equity, and Years.

News Model Training Dataset

The **Google** [2] API, which offers up-to-date news, was used to gather news datasets for training the News model. The extracted dataset contains PublishedAt, Title, Description, URL, and Company.

Company Review Model Training Dataset

To extract public reviews of companies from **Trustpilot**[3], data scraping was used for the first 100 pages of reviews for each company. **Trustpilot** was chosen because of its large number of users and reviews. The extracted dataset contains Date, Post, and Company.

Stock Market Dataset

The Stock Market dataset of those 56 companies was extracted from **Yahoo Finance**[4] with their API. **Yahoo Finance** was selected because of the accuracy and

[1] **Alpha Vantage:** https://www.alphavantage.co/
[2] **Google News:**https://news.google.com/home
[3] **Trustpilot:**https://www.trustpilot.com/
[4] **Yahoo Finance:**https://finance.yahoo.com/

reliability of its data. The extracted dataset contains Date, Open, High, Low, Close, Adjusted Close, Volume, and Symbol.

News and Review Sentiment Labelling for NLP Training

RoBERTa was used for sentiment analysis, classifying the Company Review Dataset and News Datasets as neutral, negative, and positive. **RoBERTa** was chosen because of its shown effectiveness in financial sentiment analysis. However, it is not the most effective for non-financial texts and frequently classifies non-financial negative texts as neutral. For instance, "I got fired from the company for no reason!" will be categorized as neutral by **RoBERTa**. As a result, the **VADER model** was applied to recheck neutral texts, as it is effective at general sentiment analysis.

6.1.2 User-Specific Company Dataset

The training dataset extraction method and the data extraction method are identical. Users will provide a **Ticker Symbol** of the company they wish to analyze, and the method will use the Ticker Symbol to retrieve datasets containing that company's news, stock, and finance data automatically. Unfortunately, because **TrustPilot** recently blocked web scraping on their website, data will not be retrieved from **TrustPilot** for company review for this section. Since some companies may have similar identical names, Ticker Symbol will be used to retrieve data, because, each company's Ticker Symbol is unique. Also, the Ticker Symbol is very well-known and used throughout the world.

6.2 Dataset Preprocessing

6.2.1 Company Review Dataset Preprocessing

In company review dataset, we preprocessed the data to prepare it for our model to run, ensuring that the outputs are not irrelevant or noisy. The following steps were taken:

- **Date Formatting:** We converted the "Date" column to datetime format, and generated the "YearMonth" column which is the combination of year and month.
- **Handling Missing Data:** We removed rows with missing "Post" values to make sure each post had texts. We replaced all NaN values in only "Post" columns with empty strings.
- **Lowercasing:** We converted all text in the "Post" column to lowercase to avoid any variability.
- **Cleaning Column Names:** Extra white space was removed where it existed in column names and the column names were all made lowercase for consistency.

- **Sentiment Standardization:** For sentiment standardization, we removed extra whitespaces from the sentiment column entries while ensuring that all sentiment labels remained in lowercase.
- **Filtering Sentiments:** We dropped rows that had unrelated sentiment labels. To make sure the dataset was meaningful for analysis, we only kept 'positive', 'negative', and 'neutral' labels.
- **Text Cleaning:** We preprocessed up the 'Post' column further by getting rid of any extra text, punctuation, formatting that were not useful to our analysis.
- **One-Hot Encoding:** We one-hot encoded the "Sentiment" column and converted it into labels to use it as input for model training.
- **Dataset Splitting for Training:** The data was splitted into 80% for training, 10% into testing, and 10% into validation.

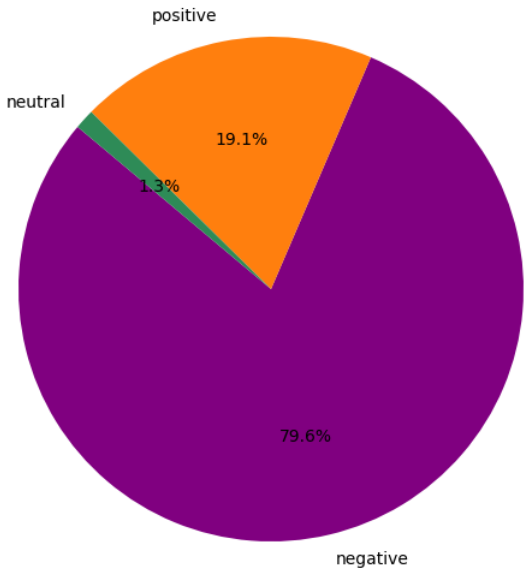

Figure 6.1: Piechart of Sentiments in Company Review dataset

We can find from Fig. [6.1] that the distribution of sentiment in social media posts is divided into 3 main sections. Negative sentiment gets the largest share, making up 79.6%. Instead, only 1.3% of the posts are neutral. However, 19.1% of the posts are positive.

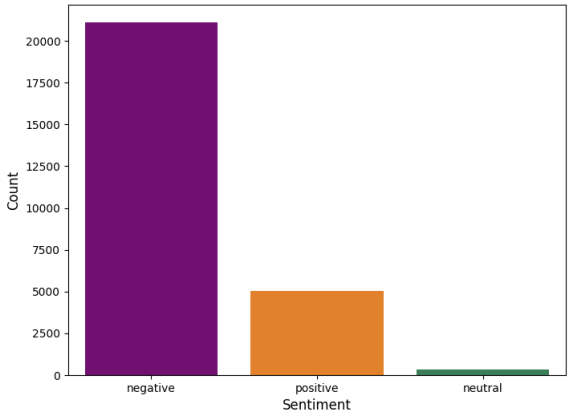

Figure 6.2: Distribution of Company Review Sentiments

From Fig. [6.2] we can observe the sentiment distribution in social media posts based on their counts. The most prominent category is negative, with over 20,000 posts, which strongly suggests that a significant portion of users are engaging in discussions highlighting dissatisfaction, criticism, or complaints. This is followed by positive sentiment, with around 5,000 posts, indicating a smaller yet notable number of users posting optimistic or encouraging content. The least represented category is neutral, with fewer than 1,000 posts, implying that users are less inclined to engage in purely informational or non-opinionated discussions.

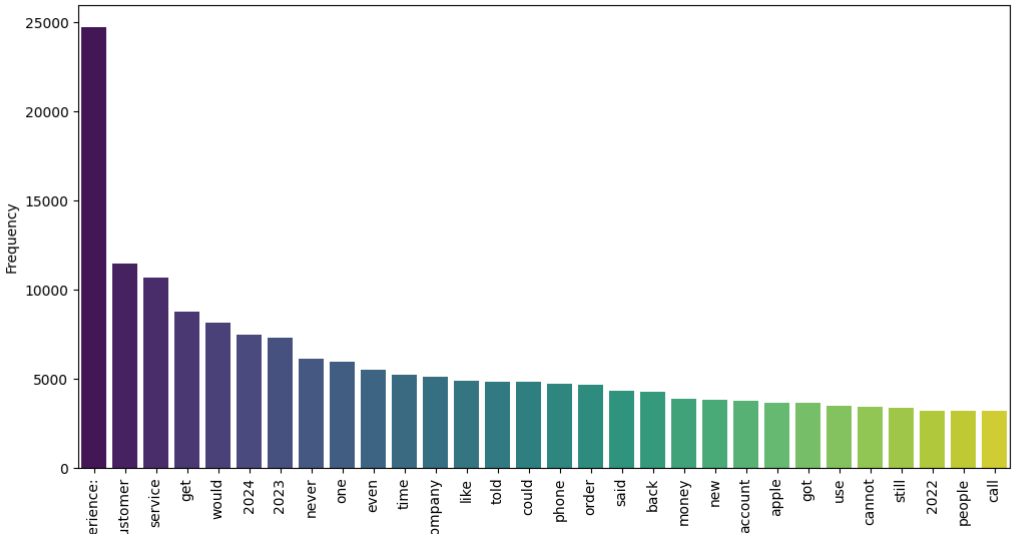

Figure 6.3: Most Frequently used Words in Company Review

As shown in Fig. [6.3], we can observe the top 30 most frequent words in Company Reviews. The most common word, **experience**, occurs with a frequency of over 25,000. Following closely are **customer** and **service**, each appearing over 10,000 times. Other frequently used words include **get**, **2024**, and **would**, with frequencies exceeding 5,000 each. Words such as **company**, **like**, and **phone** also feature prominently in the dataset, demonstrating a focus on discussions about services,

companies, and user experiences.

Figure 6.4: Word Cloud for Company Review Posts

The word cloud visualization from Fig. [6.4] illustrates the most frequently mentioned terms in Company Reviews. Prominent words such as **experience**, **customer**, **service**, **company**, and **phone** highlight key topics of discussion centered around user experiences, customer service interactions, and products or services. These words appear larger due to their high frequency, reflecting the dominant themes in public discourse across Company Reviews platforms.

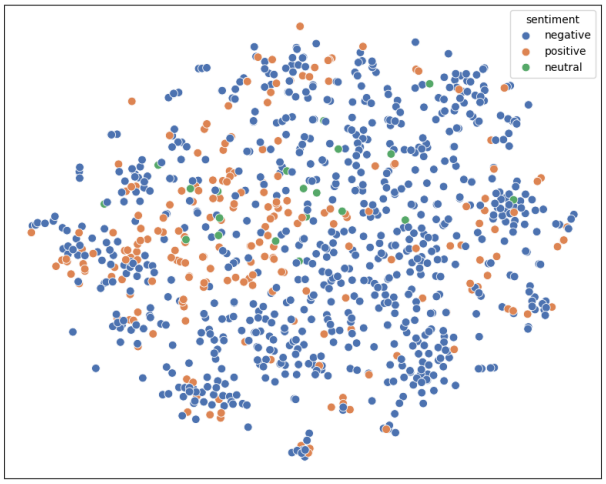

Figure 6.5: Scatter Points of Different Sentiments

The scatter plot in Fig. [6.5], which utilizes t-SNE for dimensionality reduction, visualizes the sentiment distribution (positive, neutral, and negative) within the dataset. Each point in the plot represents a data entry, with its color corresponding to a specific sentiment: blue dots signify negative sentiments, orange dots represent positive sentiments, and green dots denote neutral sentiments. The distribution demonstrates no clear clustering based on sentiment, suggesting that the sentiment categories are dispersed uniformly throughout the sample.

6.2.2 News Dataset Preprocessing

In the News dataset, we preprocess the data to prepare it for our model in almost the same way as the Company review dataset. The following steps were taken:

- **Dropping Unnecessary Columns:** We dropped unnecessary columns such as those not required for further analysis.
- **Text Normalization:** Text was normalized to improve consistency and prepare the data for processing.
- **Standardizing Sentiment Labels:** Sentiment labels were standardized for uniformity across the dataset.
- **Removing Invalid Rows:** Rows where the *Sentiment* column had the value "not applicable" for the company were removed.
- **Preprocessing Titles:** The *Title* column was preprocessed, and the cleaned text was stored in a new column named *Cleaned_Title*.
- **Text Cleaning:**
  - Removed extra whitespace: Trimmed multiple spaces to single spaces.
  - Removed URLs: Deleted web links starting with `http://` or `www.`.
  - Removed characters: Stripped `@mentions` and `#hashtags`.
  - Expanded contractions: Used a custom `fix()` function to expand short forms.
  - Kept specific punctuation: Removed non-alphanumeric characters except for allowed punctuation (`.,!?;:'"()/-`).
  - Reduced punctuation repetition: Compressed repeated punctuation marks into a single occurrence.

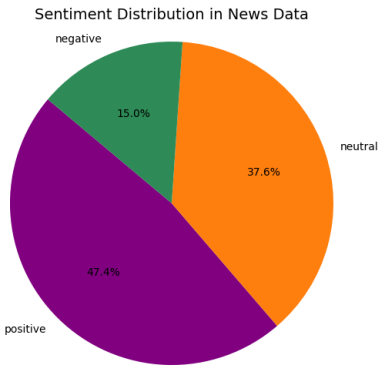

Figure 6.6: Piechart of Sentiments in News dataset

Refer to Fig. [6.6] and we can see that 37.6% of all titles in this dataset were neutral to the companyies in question. Further, 47.4% of the news titles were positive. However, 15.0% of the news titles showed negative sentiment, possibly indicating that news was disseminated about the companies unfavorable. It may include loss of selling prices due to news of financial losses, legal challenges, poor trends in markets, or some other setback.

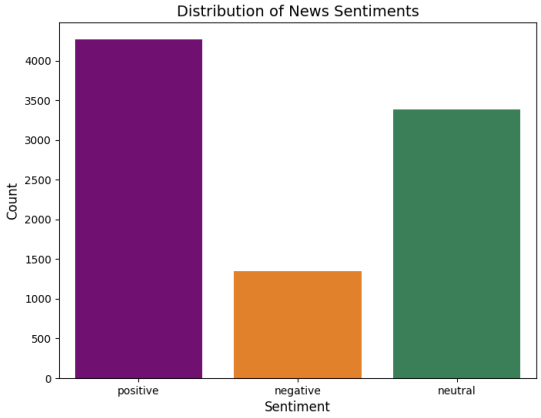

Figure 6.7: Distribution of News sentiment

Fig. [6.7] depicts the distribution of three sentiment types: positive, negative, and neutral. Among them, the positive sentiment has the highest count, exceeding 4,000, followed by the neutral sentiment at around 3,500, and the negative sentiment with a count slightly above 1,000. The bars are color-coded: purple represents positive sentiment, orange denotes negative sentiment, and green indicates neutral sentiment. The x-axis displays the sentiment categories, while the y-axis represents their respective counts.

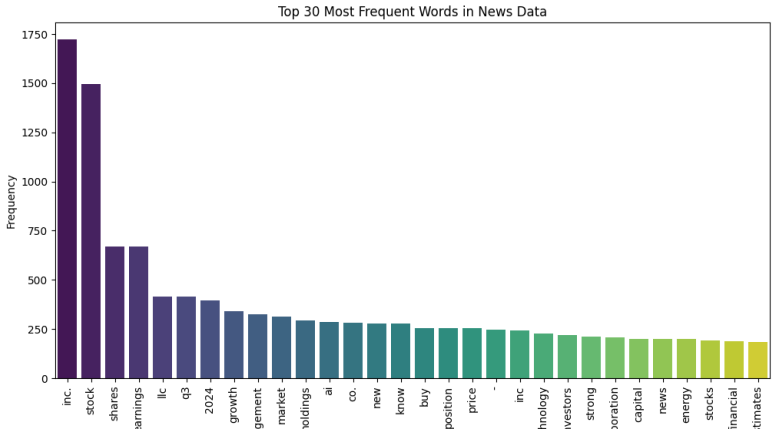

Figure 6.8: Most frequently used words in News

The dataset's analysis reveals that the word **inc.** is the most frequent, with over **400 occurrences**. This high frequency signifies its central importance in the dataset. Other terms like **stock** (350+ occurrences), **international** (almost 200 times), **American** (180+ times), and **group** (150+ times) also appear frequently, indicating a focus on business-related entities, companies, and financial terminology.

Figure 6.9: Word Cloud for News

The word cloud visualization from Fig. [6.9] helps us identify the most frequently mentioned terms in the review posts. The prominent words, such as **Group**, **Inc**, and **Apple**, highlight the focus on technology companies, automotive discussions, and often **product**-related conversations. These words appear larger due to their high frequency, indicating public interest or discourse about these topics.

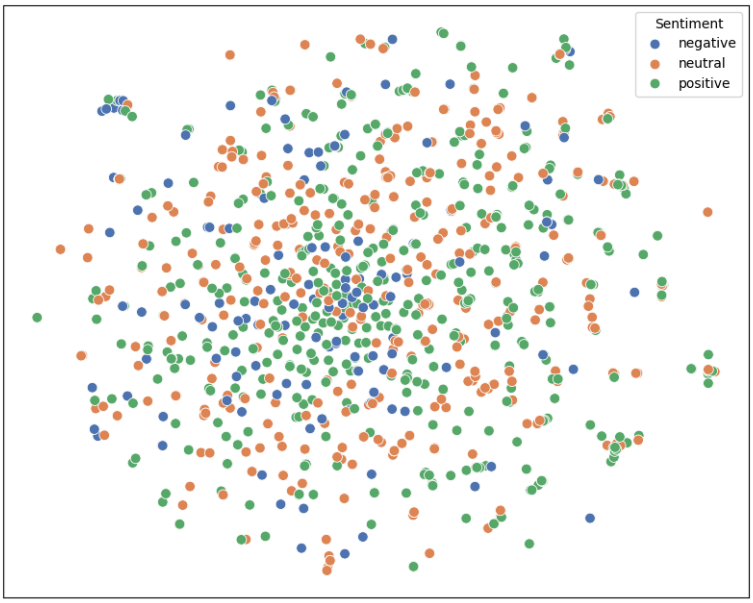

Figure 6.10: Scatter Points in Different Sentiments

Figure [6.10] shows a scatter plot with each point representing a news article and color-coded by sentiment: blue for neutral, orange for positive, and green for negative. The distribution displays clusters of articles across different emotion categories, with neutral sentiments appearing to dominate and spread out across the plot. **Positive and negative attitudes are more scattered**; however, there are

some distinguishable groups for each sentiment. This demonstrates that the news pieces analyzed had a variety of emotional tones.

6.2.3 Finance Dataset Preprocessing

In our financial dataset analysis, we did various key preprocess and analytical steps so that our data is ready to train our models. Here's a breakdown of the process:

- **Normalization of Feature Column Using MinMax Scaler:** To standardize the dataset, the *Feature* column was normalized using a MinMax scaler. The MinMax scaler adjusts the values in the column to a specific range, typically between 0 and 1. This process ensures that all features are scaled proportionally and eliminates any discrepancies caused by varying scales of measurement. By confining the values to a consistent range, this step prepares the data for meaningful comparison and ensures that no single feature disproportionately influences subsequent calculations.
- **Score Calculation Using Average Formula:** After normalizing the features, a composite *Score* was generated by applying an average formula that combines multiple key metrics:

  $$\text{Score} = \frac{\text{ROE} + \text{Debt-to-Equity Ratio} + \text{Gross Margin} + \text{Net Profit Margin} + \text{ROA}}{5}$$

  All metrics are treated equally by averaging them. This approach simplifies the calculation while still incorporating key financial indicators.
- **Normalization of Score Column Using MinMax Scaler:** The generated *Score* column was also normalized using a MinMax scaler to confine its values within the range of 0 to 1. This step ensures consistency across the dataset and makes the scores easier to interpret. Normalizing the scores enhances their comparability and prevents any bias caused by large variations in numerical values.

6.2.4 Stock Dataset Preprocessing

The stock dataset's preprocessing step includes a number of essential operations to get the data prepared for analysis. The steps are:

- Load the dataset from a predefined source.
- Handle missing values, typically by forward filling.
- Transform the **Date** column to a datetime format.
- Sort the dataset chronologically by **Symbol** and **Date**.
- Compute financial indicators such as Daily Return, Daily Range, Moving Average (MA30), Rolling Volatility (VT30),Relative Strength Index (RSI30).

- Group data by stock symbols for individual stock analysis.
- Eliminate any resulting NaN values to maintain data integrity.

Finally, the processed dataset is ready for visualization, allowing for comprehensive analysis of stock trends and performance.

Several graphs are generated during the analysis to indicate how each company's stock has performed. Here, Apple Inc. (AAPL) is used as an illustration example for visualization.

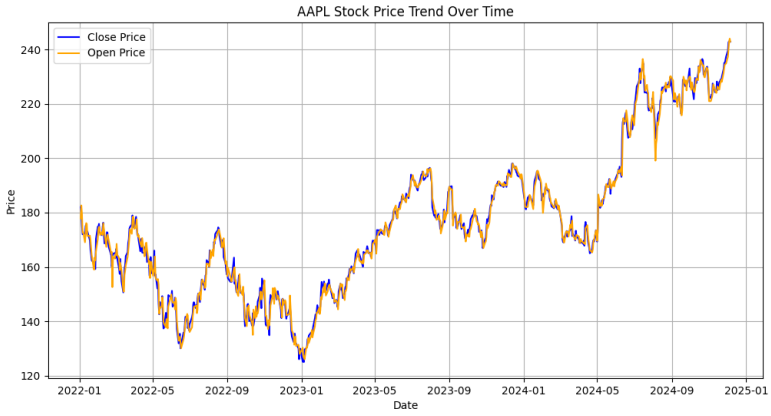

Figure 6.11: Stock Price Trend Over Time

The above graph of Fig. [6.11], with the x-axis representing dates and the y-axis representing stock prices in USD, illustrates the trend of AAPL's stock price from early 2021 to mid-2024. The blue line represents daily closing prices, while the orange line represents daily opening prices. Both lines exhibit similar patterns in terms of volatility and an overall upward trend over time. The stock experiences periods of decline and recovery, with a notable surge in price during mid-2024, indicating a significant increase in value. This representation captures the overall trends and fluctuations in AAPL's stock price during the specified timeframe.

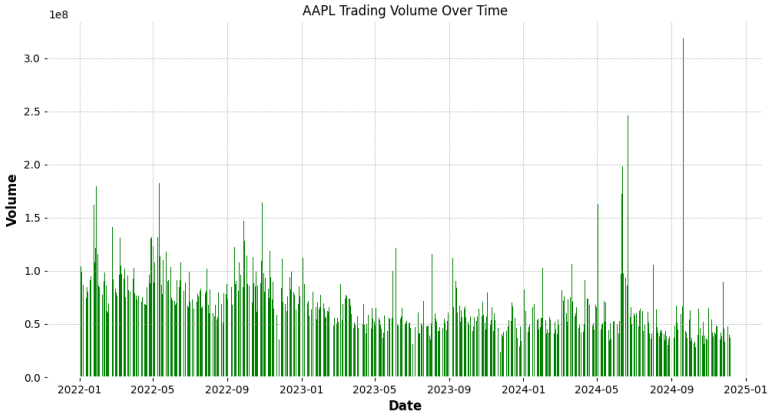

Figure 6.12: Trading Volume Over Time

The Fig. [6.12] displays the trading volume of AAPL stock transactions from early 2021 to mid-2024. The x-axis represents the date, while the y-axis represents the trading volume in units. Vertical green bars illustrate the daily trading volume, highlighting fluctuations in market activity. Noticeable peaks occur in 2021 and 2022, signifying days with significantly higher trading activity. However, these spikes become less frequent over time, with trading volumes appearing relatively lower and more stable during 2023 and the first half of 2024. This visualization emphasizes periods of heightened interest and trading activity in AAPL shares.

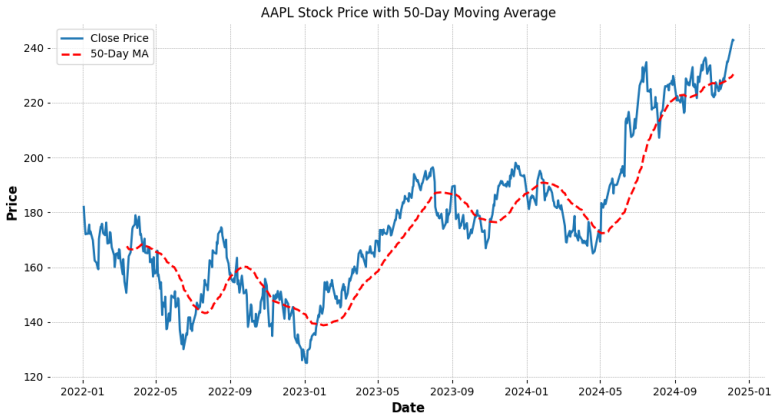

Figure 6.13: Stock Price with 50-Day Moving Average

The Fig. [6.13] shows the price of AAPL's stock from the beginning of 2022 to the start of 2025. The 50-day moving average is shown by the red dashed line, while the closing price is represented by the blue line. The stock starts near 180 in early 2022, drops to about 140 by mid-2022, and remains relatively flat until the end of 2022. In early 2023, it rises steadily, surpassing 180 by mid-year, and continues to climb, reaching over 240 by the beginning of 2025. The stock's overall upward trend is closely tracked by the 50-day moving average, which smooths out short-term fluctuations.

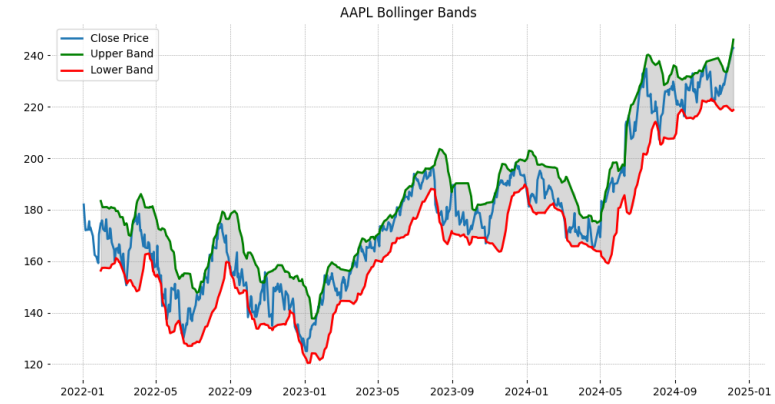

Figure 6.14: Bollinger Bands

Bollinger Bands in Fig. [6.14], include an upper band (green), a lower band (red), and the stock's closing price (blue) from early 2022 to early 2025. The bands expand

and contract to indicate periods of high and low volatility. The stock price often fluctuates within the bands, with peaks near 180 in late 2022, surpassing 200 in mid-2024, and nearing 240 by early 2025. During negative trends, such as mid-2022 and late 2023, the price approaches the lower band, while sharp upward movements cause the price to touch or cross the upper band.

6.3 Model Development

6.3.1 Financial Model Development

We started by preparing the data by grouping financial data by **GICS Sector**[5] to do sector-wise model training, extracting the target variable (Score) for regression, and dropping unnecessary columns. After that, we split the data into an **85% training set** and a **15% validation set**. Next, we trained multiple models.

Decision Tree Regressor: We trained the model so that we could capture non-linear relationships in financial data. We evaluated it using matrices like **MSE**, $\boldsymbol{R^2}$, and **MAE** that measured its performance. It is effective for financial datasets where relationships are not linear.

K-Nearest Neighbors (KNN): Our model was trained with **3 neighbors**. KNN makes no assumption about data distribution, which makes this ideal for diverse financial metrics that may not be similar to standard distributions. We also calculated performance metrics so that we could evaluate its accuracy.

Support Vector Regressor (SVR): We used an **RBF(Radial Basis Function)** kernel to train the model. RBF kernel enabled it to capture non-linear relationships in financial data. The SVR model is able to reduce the impact of outliers in the data using margins. This can model complex, non-linear patterns, which makes this ideal to do financial analysis.

Random Forest Regressor: Our model was trained using multiple decision trees and it focused on measuring its accuracy. Random Forest is well-suited for handling complex, non-linear datasets, which makes it perfect for financial data with varying patterns and relationships.

XGBoost Regressor: We trained our model with **100 estimators** and the learning rate was **0.1**. XGBoost is able to capture nonlinear relationships. This can prevent overfitting and also help us understand key financial metrics; eg. ROE and Debt-to-Equity Ratio. XGBoost Regressor gives priority to accuracy and scalability. This makes it one of the best-performing algorithms when it comes to structured financial data.

After training, all models were saved as **.pkl** files using **Joblib**. After that, we stored the performance metrics in a **JSON** file for further analysis.

6.3.2 Stock Analysis Model

In our Stock Market Analysis, we used a calculation-based method instead of using machine learning models to analyze the stock market dataset. The nature of the

[5]**GICS (Global Industry Classification Standard) Sector:** It is a classification system developed by MSCI and SP Dow Jones Indices to categorize companies based on their primary business activities.

analysis we wanted to do was a stock performance evaluation using predefined financial indicators and calculations.

We loaded and preprocessed the stock market dataset we had using the **preprocess_dataset** module we built. The **preprocess_dataset** module handles data cleaning by removing rows that contained NaN values and sorting the dataset based on date to evaluate better. Another module we built was **processed_group**, which calculates Daily Returns, Gap, Daily Range, Moving Averages, Volatility, and RSI, to evaluate the stock state more efficiently. There is another module, **evaluate_group**, where we take different stock data like Moving Average, Relative Strength Index (RSI), Volatility, Daily Return, Gap, and Daily Range and calculate a "Score."

Finally, results are summarized by the most common score on a monthly basis within the **stock_analysis** method. Instead of storing these monthly scores, we store them in their corresponding year, assigning higher weights to recent months and lower weights to older months. It ensures that recent stock data gets more priority, which makes the model more accurate. For this, our module produced weight arithmetic progression for months and normalized them such that their sum is one. This permits for a calibrated rating where scores from previous months exert a far more important impression on the overall marks for the year. We then calculated the weighted yearly scores in the same manner. Our module in the end got the final score based on those monthly and finally yearly data. Upon defining thresholds by median score value, the final scores for each stock are placed into qualitative labels, for example, **Challenging** or **Favourable**.

Here, the primary aim was not to predict the future price of the stock or its trend but to assess the current performance of the stock. For this, Machine Learning models might be a way, but building functions to calculate the state is more precise. This provides an interpretable, clear output that is very crucial for Stock Market Analysis.

6.3.3 News Model

For the News Model Implementation, our approach is as follows:

- **One-Hot Encoding for Sentiment:** We applied One-hot encoding in order to turn the 'Sentiment' column into numerical representations that match the requirement of the model and are compatible with it.
- **Removing Stopwords**: We processed a column named 'Cleaned_Title'. This was done to filter out common stopwords, resulting in a new feature column, named 'Cleaned_Title_No_Stopwords'. This improved the quality of the input data by removing the redundant words.
- **Preprocessing News Data:** The news data was thoroughly preprocessed, which included cleaning the text and splitting it into training and evaluation datasets for model training and validation.

- **Model Training with RoBERTa:** RobertaForSequenceClassification model was used here which is a transformer-based model. We configured it to work with the following three sentiments: positive, negative, and neutral.
- **Tokenization of Text Data:** RoBERTa tokenizer was used to tokenize the text data. Here, padding and truncation were applied in order to maintain consistency in the size of the input. The maximum length of the sequence was set up to 512 tokens.
- **Training Configuration:** A well-defined set of training arguments was used to train the model. The epoch count was set to 30. This was done in order to ensure sufficient training iterations. Batch size of 15 was used in order to balance memory efficiency and performance of training. Weight decay of 0.01 was applied to reduce overfitting and to improve generalization. An evaluation was conducted at the end of each epoch, monitoring the model's performance while training.
- **Saving the Model and the Tokenizer:** We kept a directory named Model-Save where we saved the model after training. This enabled future use of the model without the requirement of re-training.
- **Performance Evaluation:** To assess the performance of the model in depth, matrices like accuracy, precision, recall and F1 score were used.

6.3.4 Company Review Model

This is how we implemented the Review Model:

- **One-Hot Encoding for Sentiment:** One hot encoding was applied to the Sentiment column in order to transform it into a numerical feature which is compatible with the model.
- **Dropping Unnecessary Columns:** We dropped columns containing data that are not relevant and not required for our analysis. We only kept required and relevant columns. redundant words.
- **Tokenization of Training Data:** We applied truncation and padding in order to tokenize the data. This made input data be of consistent input length. The maximum length of a sequence was set to be 512 tokens.
- **Defining Training Arguments:** In order to get an optimized performance from the model, we configured the training arguments. The configuration included 25 epochs, 15 as batch size and 0.01 as weight decay. We also added epoch-based evaluation and logging settings.
- **Saving the Model and Tokenizer:** The model that we have trained and the tokenizer were saved in a directory to make sure that they can be reused without the necessity of retraining.
- **Performance Evaluation:** The performance of the review model was evaluated through matrices like accuracy, precision, recall and F1-score in order to provide an in depth evaluation of the effectiveness of the model.

6.4 Score Calculation Process

6.4.1 Financial State Score

The best pre-trained models we saved after training our finance models based on various **GICS Sectors** are used for predicting the score after pre-processing the user's company's financial dataset. By evaluating their **error rates**, the best pre-trained models are selected. Using the best pre-trained model in the sector, the model predicts financial scores for companies in that industry.

The methodology creates a weighted score for that company over numerous years, with the most recent year receiving the highest priority, and so on. The weighted scores are then standardized by comparing them to the sector's **median score**. Scores below the median are assigned to a **0-50** scale. Scores over the median are assigned to the **50-100** range. The median is used because some companies' scores can be extremely high while others can be extremely low, affecting the total mean value, and comparing the mean to the output score may not yield an appropriate result.

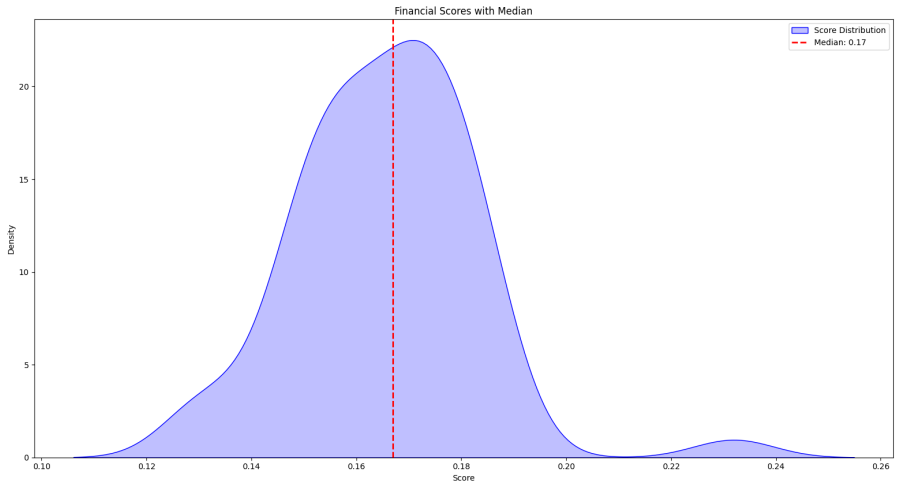

Figure 6.15: Standardized Financial Scores

Based on the standardized score, companies are characterized as having a good or bad financial state. If the company falls between **0% and 50%**, it is considered bad. If it falls between **50% and 100%**, it is considered good. In addition, the final standardized score is given, which ranges from **0% to 100%**.

6.4.2 Stock State Score

The model first groups the dataset by year. The score of the company is based on the stock market and depends on **moving average trends**, the **relative strength index of 30 days**, **volatility of 30 days**, **daily returns**, **price gaps**, and **daily**

price range stability. For every favorable condition met (for example, the 10-day moving average outperforms the 30-day moving average), that day's stock is awarded a point, and vice versa for other metrics. The daily score is calculated based on all the point averages of that day. Finally, the yearly score is calculated by the average daily scores of that year. Then the model calculates a weighted average score of all the years. The immediate years get the most priority. The weighted average score percentage is the final score of the company.

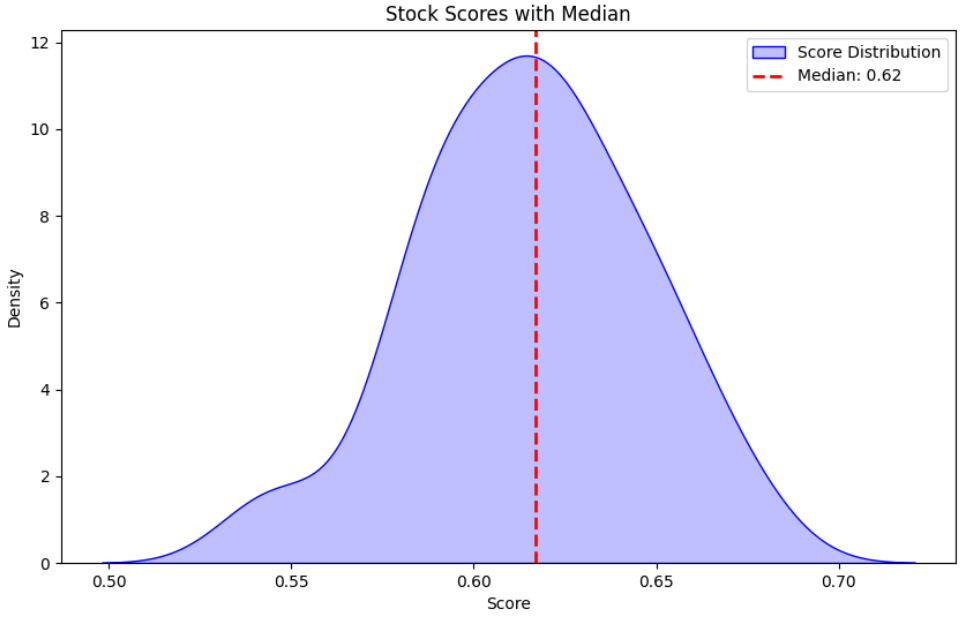

Figure 6.16: Standardized Stock Scores

Then, the weighted scores are standardized by comparing the scores to the sector's **median score**. The score which is below the median score is assigned to the **0-50** range. The score which is above the median score is assigned to the **50-100** range. The median score is used because the scores of some companies can be extremely high while others can be extremely low, affecting the total mean value, and comparing the mean with the output score may not produce an appropriate result. A score between **0% and 50%** suggests challenging performance, while a score between **50% and 100%** indicates favorable performance. In addition, the final standardized score is given, which ranges from **0% to 100%**.

6.4.3 Sentiment Score Calculation for Company News

After pre-processing the user's company's news dataset, we use the **RoBERTa** model we saved from training our company models to predict the sentiment of each news headline. Each news headline is classified as positive, negative, or neutral.

The model then calculates yearly sentiment scores using the following formula:

$$\text{Score} = \frac{\text{Positive News Count} - \text{Negative News Count}}{\text{Total News Count}}$$

The immediate years are given the highest priority when the model applies a weighted average of the scores. The final score for that company's news dataset is the weighted average. The final score ranges from **0%** to **100%**.

The company is then classified according to the final score. The company has more good news than bad news if the score is **higher than 0**. The company has more bad news if the score is **less than zero**. The company has equal amounts of good and negative sentiment, or all neutral feelings if the score is **exactly zero**.

6.4.4 Sentiment Score Calculation for Company Reviews

Due to recent limitations on retrieving data from **Trustpilot**, the model verifies whether the user-provided company exists in the core data set. If it does, the model first extracts the provided company rows from the dataset, then the model groups the dataset by year, and then the model calculates the score based on the count of positive and negative posts, similar to the sentiment score calculation for company news.

The company is then classified according to the final score. The company has more positive reviews than negative reviews if the score is **higher than 0**. The company has more negative reviews if the score is **less than zero**. The company has equal amounts of positive and negative reviews, or all neutral reviews if the score is **exactly zero**.

6.4.5 Cumulative Company Score

Gatawa (2021) [25] provided solid evidence suggesting that financial metrics play the most important role when it comes to evaluating a company's performance, making this the fundamental factor in evaluating the health of a corporation. Moreover, his research suggested that even though stock performance is relevant, it can not alone reflect the true value of a company. Instead, market price fluctuations are heavily influenced by financial metrics, meaning that stock prices act as a secondary indicator rather than a standalone measure of company health.

Smith and O'Hare (2022) [19] provide evidence that while news and social media sentiment can influence stock prices, their impact is limited compared to fundamental financial metrics. Additionally, their time-lagged analysis revealed that stock prices were often driving social media sentiment, rather than the other way around. This indicates that while news sentiment can have some predictive potential, social media sentiment (including CEO tweets and public reviews) tends to be inconsistent and often weakly correlated.

This supports the idea that news sentiment should have a higher weight than reviews, but neither should be prioritized over financial or stock performance. While financial news can sometimes provide valuable insights, public reviews and social media sentiment tend to be more volatile and should be given the lowest weight in company evaluation.

For the overall score of a company, the model will first check if the company exists in the trained company review dataset. If it does, then the overall score of a company is determined by prioritizing four key metrics in a structured manner:

- Priority of Finance Score, $\boldsymbol{P_F}$ (Highest priority)
- Priority of Stock Score, $\boldsymbol{P_S}$
- Priority of News Score, $\boldsymbol{P_N}$
- Priority of Review Score, $\boldsymbol{P_R}$ (Lowest priority)

Each metric is assigned a gradual weight based on its priority. Less important factors are given lower weights, while more critical factors receive higher weights. Weighted sum:

$$\boldsymbol{W_F} = \frac{\boldsymbol{P_F}}{\boldsymbol{W}_{\text{total}}}, \quad \boldsymbol{W_S} = \frac{\boldsymbol{P_S}}{\boldsymbol{W}_{\text{total}}}, \quad \boldsymbol{W_N} = \frac{\boldsymbol{P_N}}{\boldsymbol{W}_{\text{total}}}, \quad \boldsymbol{W_R} = \frac{\boldsymbol{P_R}}{\boldsymbol{W}_{\text{total}}}$$

After calculating the weighted sum for each sector, the model calculates the overall score of the company. The overall score of the company will be:

$$\begin{aligned} \textbf{Overall Score} = \text{Finance Score} \cdot \boldsymbol{W_F} + \text{Stock Score} \cdot \boldsymbol{W_S} \\ + \text{News Score} \cdot \boldsymbol{W_N} + \text{Review Score} \cdot \boldsymbol{W_R} \end{aligned}$$

If the company does not exist in the core company review dataset, then the overall score of a company is determined by prioritizing three key metrics in a structured manner:

- Priority of Finance Score, $\boldsymbol{P_F}$ (Highest priority)
- Priority of Stock Score, $\boldsymbol{P_S}$
- Priority of News Score, $\boldsymbol{P_N}$ (Lowest priority)

Each metric is assigned a gradual weight based on its priority. Less important factors are given lower weights, while more critical factors receive higher weights. Weighted sum:

$$\boldsymbol{W_F} = \frac{\boldsymbol{P_F}}{\boldsymbol{W}_{\text{total}}}, \quad \boldsymbol{W_S} = \frac{\boldsymbol{P_S}}{\boldsymbol{W}_{\text{total}}}, \quad \boldsymbol{W_N} = \frac{\boldsymbol{P_N}}{\boldsymbol{W}_{\text{total}}}$$

After calculating the weighted sum for each sector, the model calculates the overall score of the company. The overall score of the company will be:

$$\textbf{Overall Score} = \text{Finance Score} \cdot \boldsymbol{W_F} + \text{Stock Score} \cdot \boldsymbol{W_S} + \text{News Score} \cdot \boldsymbol{W_N}$$

After calculating the overall scores for all companies, we determine the **median score** across the dataset. Each company's score is then compared to this median to evaluate it's state:

- **Above Median:** The company is in a *favorable position* relative to its peers.
- **Below Median:** The company is in a *challenging position* relative to its peers.
- **Equal to Median:** The company is in a *balanced position*, reflecting alignment with the broader industry average.

6.5 Model Architecture

The model architecture illustrated below includes the design for data extraction, data processing, feature extraction, model training, model saving, output prediction, and score generation.

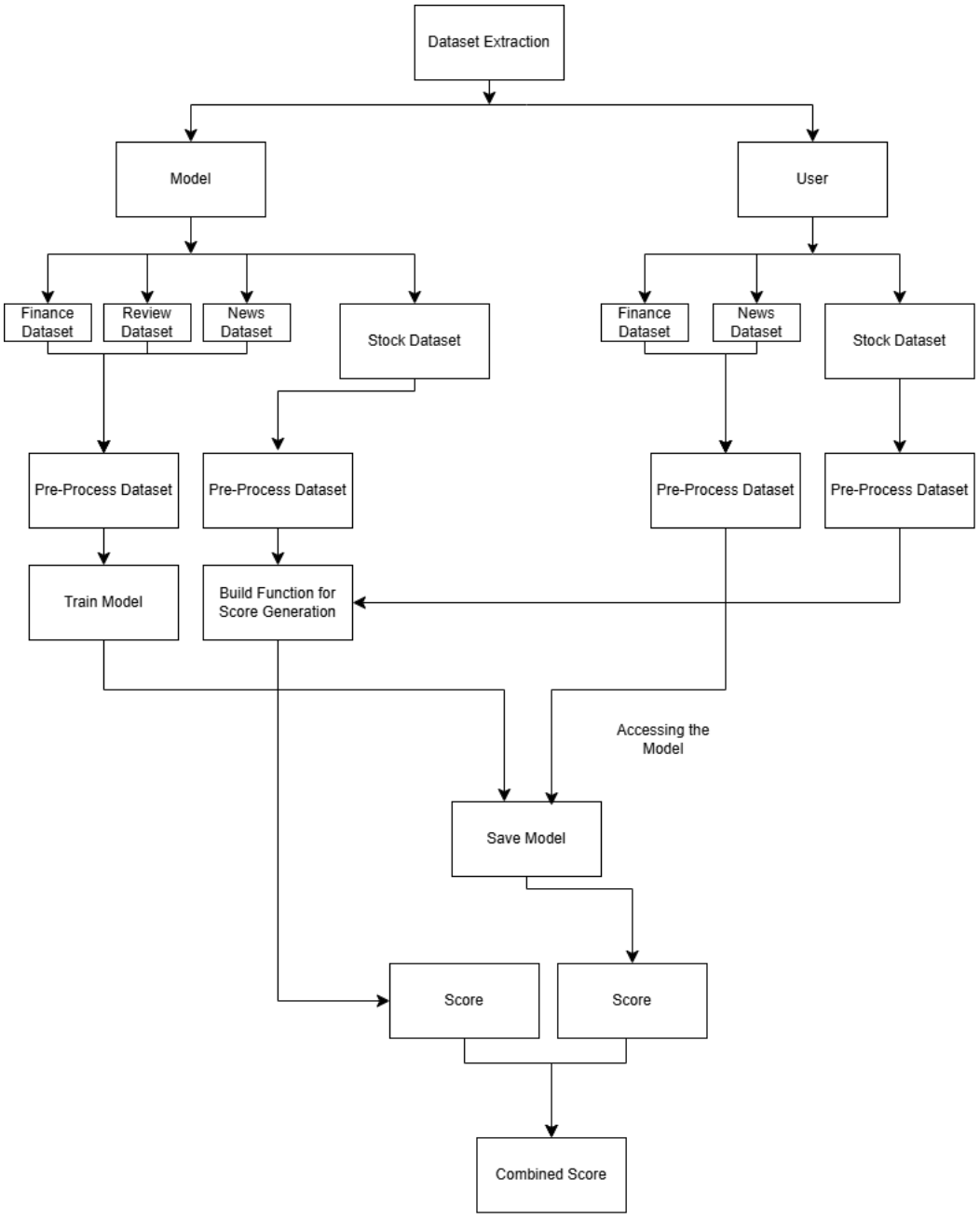

Figure 6.17: Model Architecture

Chapter 7

Result Analysis

7.1 News Result Analysis

Metric	Eval Loss	Eval Accuracy	Eval F1 Score	Eval Precision	Eval Recall
Value	0.4613	89.80%	89.78%	89.80%	89.78%

Table 7.1: Evaluation Results for RoBERTa Model

We trained this model for **30** epochs. The evaluation loss is **0.4613**, as indicated in Table 7.2, indicating that the RoBERTa model reaches a relatively low error in its predictions throughout the training.

The Eval Accuracy is **89.80%**, which means the model is correctly classifying most of the instances. The Eval Precision is **89.80%** and the Eval Recall is **89.78%**, indicating that the model is identifying positive cases while maintaining a low rate of false positives and false negatives.

The Eval F1 Score is **89.78%**, which means the model's performance is balanced in terms of precision and recall.

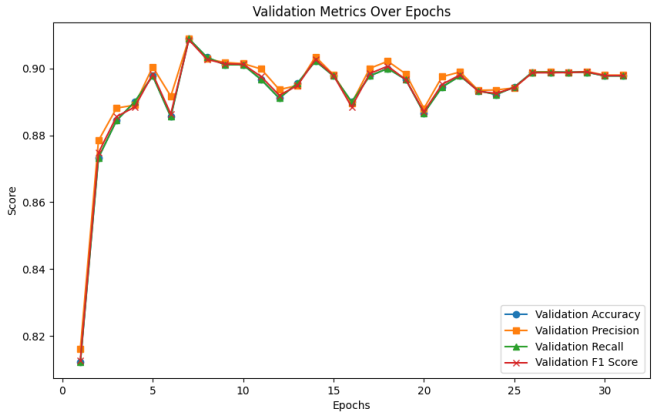

Figure 7.1: RoBERTa's Performance - News Dataset

All of the metrics rise during the first few epochs, with rare variations, as Fig 7.1

demonstrates. After a few epochs, it stabilized and had high values for all metrics, exceeding **89.00%**.

7.2 Company Review Result Analysis

Metric	Eval Loss	Eval Accuracy	Eval F1 Score	Eval Precision	Eval Recall
Value	**0.165**	**94.42%**	**94.48%**	**94.61%**	**94.42%**

Table 7.2: Evaluation Results for RoBERTa Model

We trained this model for **25** epochs. The Eval Loss is **0.165**, as indicated in Table 7.2, indicating that the RoBERTa model is achieving comparatively low error in its predictions throughout training.

The Eval Accuracy is **94.42%**, which means the model is correctly classifying most of the instances. The Eval Precision is **94.61%** and the Eval Recall is **94.42%**, indicating that the model is identifying positive cases while maintaining a low rate of false positives and false negatives.

The Eval F1 Score is **94.48%**, which means the model's performance is balanced in terms of precision and recall.

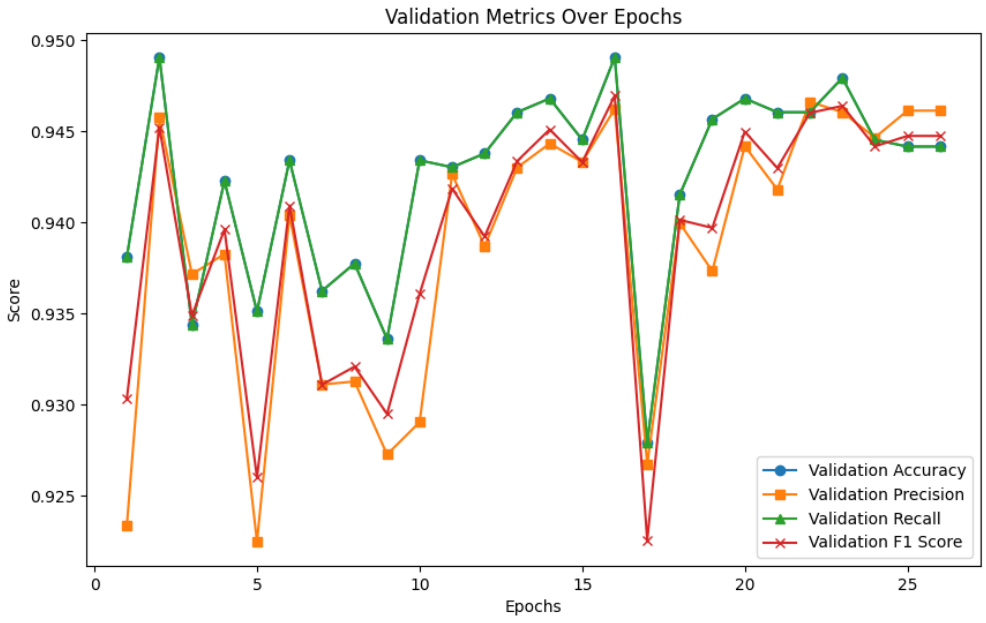

Figure 7.2: RoBERTa's Performance - Company Review

All of the metrics rise during the first few epochs, with rare variations, as Fig 7.2 demonstrates. After 18th epochs, it stabilized and had high values for all metrics, exceeding **0.95%**.

7.3 Finance Result Analysis

Class	Metric	XGBoost	Decision Tree	Random Forest	SVR	KNN
TECHNOLOGY	MSE	0.0207	**0.0150**	0.0175	0.0237	0.0191
	MAE	0.0293	**0.0251**	0.0268	0.0737	0.0285
	R2	0.0791	**0.3329**	0.2228	0.0528	0.1521
REAL ESTATE & CONSTRUCTION	MSE	0.0086	**0.0042**	0.0086	0.0150	0.0058
	MAE	0.0886	**0.0503**	0.0700	0.1072	0.0594
	R2	-0.4162	**0.3033**	-0.4454	-1.5097	0.0241
TRADE & SERVICES	MSE	0.0000	0.0000	0.0000	0.0004	0.0000
	MAE	0.0017	0.0016	**0.0012**	0.0166	0.0013
	R2	0.9574	0.9774	0.9836	-1.1212	**0.9872**
MANUFACTURING	MSE	0.0000	0.0000	0.0000	0.0002	0.0000
	MAE	0.0014	0.0013	0.0108	**0.0011**	0.0022
	R2	0.9756	0.9697	-0.0190	**0.9883**	0.8972
LIFE SCIENCES	MSE	0.0005	0.0002	**0.0001**	0.1359	0.0003
	MAE	0.0079	0.0065	**0.0044**	0.0169	0.0059
	R2	0.5007	0.8542	**0.9184**	0.0169	0.0059

Table 7.3: Industry-Wise Performance of Finance Models

According to Table [7.3], KNN, Random Forest, and Decision Tree are doing very well in the Technology sector with low MSE and MAE scores which means high accuracy in predictions. The top choices for this sector are Decision Tree and Random Forest, which deliver nearly identical results. However, R^2 is low, indicating that predictions are not well accurate for this sector (below **0.5**).

Decision Tree is the unique identifier for the Real Estate & Construction sector because it had minimum MSE (**0.0042**) and maximum R^2 (**0.3033**). However, R^2 is low, indicating that predictions are not well accurate, like the Technology sector.

For the Trade & Services sector, Decision Tree, XGBoost, KNN, and Random Forest achieved the lowest MSE of **0.0000** and also, they achieved the highest R^2 (over **0.97**), which represents its ability to understand and capture the complex relationships in this sector.

In the Manufacturing sector, Decision Tree, XGBoost, and KNN performed very well with pretty low error rates, but on the other hand, Random Forest performed poorly, particularly when it comes to its R^2 value, which drops to **-0.0190**.

Decision Tree acts very well in the Life Sciences sector in terms of MSE (**0.0002**) and R^2 (**0.8542**). This means that the Decision Tree can capture such complexities in Life Sciences data as well as any other model in this sector.

Chapter 8

Conclusion

8.1 Conclusion

Our research mainly focuses on a method for understanding what state a company is in. The aim of our research is to do this task by using Natural Language Processing and Machine Learning. Our research discusses a new method of analysis, which combines Financial Data, Stock Market Information, and News Titles with their sentiments and company's review from users with their sentiments. This explores beyond the usual focus, such as revenue, profit margins, and stock prices data, and tries to gain a deeper understanding of a company's state by analyzing textual data sources.

Our research targets to show in the future that sentiments expressed by News Titles and reviews can play a vital role in analyzing a company's state by providing important insights by merging all of these analyses. Sentiment analysis using NLP models can be significantly useful in understanding the public sentiment about a company, which can impact the overall image of that company. Moreover, a company's performance is also analyzed by using financial and stock market data, which focuses on important financial indicators and changes in stock prices. In our research, multiple machine learning models such as RoBERTa, XGBoost, Random Forest, SVR etc. were used to analyze all of the datasets. The target of our research is to show that by including all of these four types of sources while analyzing the state of a company, we can extract a better understanding of a company's present state. This analysis can assist investors, financial advisors, and corporate strategists in gaining a better understanding of the company's state.

According to our research and knowledge, no such research has been conducted previously that combined these four sectors of a company to produce a result that would help gain a greater understanding of a company's state. Our goal is to make it easier for everyone to gain an understanding of a company's state more accurately and easily.

8.2 Limitations

We have some research limitations because our system cannot extract user reviews from Trustpilot when we provide a company name. Currently, there are no free accessible APIs or publicly available mechanisms from such platforms to retrieve review data seamlessly. This limits the range of data sources we can use, which may result in the exclusion of important qualitative insights.

Our model depends on bringing together many different metrics to evaluate how well a company is performing. Our analysis faces a challenge because many platforms restrict access to detailed data metrics and we could not include enough metrics. The accuracy of the computed health score may be impacted by gaps in the analysis caused by the absence of direct access to some important data sources.

8.3 Future Work

Our future work will include:

- **Expanding Data Accessibility:** We want to find websites or program interfaces that give public access to user feedback with sentiment rating information. Our model needs recent and varied feedback from multiple online sources to replace Trustpilot data feeds.
- **Incorporating Additional Metrics:** Our model will analyze both reviews plus other useful information about employee happiness levels and how well companies manage ESG standards plus their supply chains. These metrics add new dimensions to corporate health evaluations and create a complete assessment perspective.
- **Developing Data Collection Workarounds:** We will research customized methods to gather needed data when direct access is absent including permitted web scraping and use of third-party aggregated datasets.
- **Model Optimization:** To enhance the scoring model we will test new ML approaches with bigger and varied data sets to produce more reliable performance metrics.

  Our extended research seeks to create a more inclusive corporate health assessment tool that investors and financial advisors can use reliably.

Bibliography

[1] R. Golan and W. Ziarko, "A methodology for stock market analysis utilizing rough set theory," in *Proceedings of 1995 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, 1995, pp. 32–40. DOI: 10.1109/CIFER.1995.495230.

[2] R. Lawrence, "Using neural networks to forecast stock market prices," Jan. 1998.

[3] M. Lam, "Neural network techniques for financial performance prediction: Integrating fundamental and technical analysis," *Decision Support Systems*, vol. 37, no. 4, pp. 567–581, 2004, Data mining for financial decision making, ISSN: 0167-9236. DOI: 10.1016/S0167-9236(03)00088-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923603000885.

[4] K. S. Vaisla and A. K. Bhatt, "An analysis of the performance of artificial neural network technique for stock market forecasting," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:1061505.

[5] S. G. Chowdhury, S. Routh, and S. Chakrabarti, "News analytics and sentiment analysis to predict stock price trends," 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:17231714.

[6] A. S. Ghadikolaei, S. K. Esbouei, and J. Antucheviciene, "Applying fuzzy mcdm for financial performance evaluation of iranian companies," *Technological and Economic Development of Economy*, vol. 20, no. 2, pp. 274–291, 2014. DOI: 10.3846/20294913.2014.913274. eprint: https://doi.org/10.3846/20294913.2014.913274. [Online]. Available: https://doi.org/10.3846/20294913.2014.913274.

[7] G. V. Attigeri, M. P. M. M, R. M. Pai, and A. Nayak, "Stock market prediction: A big data approach," in *TENCON 2015 - 2015 IEEE Region 10 Conference*, 2015, pp. 1–5. DOI: 10.1109/TENCON.2015.7373006.

[8] K. Joshi, N. Bharathi, and J. Rao, "Stock trend prediction using news sentiment analysis," *International Journal of Computer Science and Information Technology*, vol. 8, pp. 67–76, Jun. 2016. DOI: 10.5121/ijcsit.2016.8306.

[9] M. Simionescu, *The competition between london companies regarding their financial performance*, Jan. 2016. [Online]. Available: https://doi.org/10.7441/joc.2016.02.01.

[10] M. Kraus and S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," *Decision Support Systems*, vol. 104, pp. 38–48, 2017, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2017.10.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923617301793.

[11] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018, ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2017.11.054. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221717310652.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423/.

[13] K. Pahwa and N. Agarwal, "Stock market analysis using supervised machine learning," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 197–200. DOI: 10.1109/COMITCon.2019.8862225.

[14] N. Pröllochs and S. Feuerriegel, "Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling," *Information Management*, vol. 57, no. 1, p. 103 070, 2020, Big data and business analytics: A research agenda for realizing business value, ISSN: 0378-7206. DOI: https://doi.org/10.1016/j.im.2018.05.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378720617309254.

[15] Y. Zhang, B. Yan, and M. Aasma, "A novel deep learning framework: Prediction and analysis of financial time series using ceemd and lstm," *Expert Systems with Applications*, vol. 159, p. 113 609, 2020, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113609. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420304334.

[16] P. Muthukumar and J. Zhong, *A stochastic time series model for predicting financial trends using nlp*, Feb. 2021.

[17] B. Panwar, G. Dhuriya, P. Johri, S. S. Yadav, and N. Gaur, "Stock market prediction using linear regression and svm," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 629–631. DOI: 10.1109/ICACITE51222.2021.9404733.

[18] V. Dogra, F. S. Alharithi, R. M. Álvarez, A. Singh, and A. M. Qahtani, "Nlp-based application for analyzing private and public banks stocks reaction to news events in the indian stock exchange," *Systems*, vol. 10, no. 6, 2022, ISSN: 2079-8954. DOI: 10.3390/systems10060233. [Online]. Available: https://www.mdpi.com/2079-8954/10/6/233.

[19] S. Smith and A. O'Hare, "Comparing traditional news and social media with stock price movements; which comes first, the news or the price change?" *Journal of Big Data*, vol. 9, no. 1, 2022. DOI: 10.1186/s40537-022-00591-6.

[20] M. Jan, S. Hussain, R. Zahra, *et al.*, "Appraisal of different artificial intelligence techniques for the prediction of marble strength," *Sustainability*, vol. 15, p. 8835, May 2023. DOI: 10.3390/su15118835.

[21] K. Li, C. Savari, H. Sheikh, and M. Barigou, "A data-driven machine learning framework for modelling of turbulent mixing flows," *Physics of Fluids*, vol. 35, Jan. 2023. DOI: 10.1063/5.0136830.

[22] K. Puh and M. B. Babac, "Predicting stock market using natural language processing," *American Journal of Business*, vol. 38, no. 2, pp. 41–61, Apr. 2023. [Online]. Available: https://ideas.repec.org/a/eme/ajbpps/ajb-08-2022-0124.html.

[23] Y. Sinjanka, "Text analytics and natural language processing for business insights: A comprehensive review," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 9, pp. 1626–1651, Sep. 2023.

[24] P. Supsermpol, V. N. Huynh, S. Thajchayapong, and N. Chiadamrong, "Predicting financial performance for listed companies in thailand during the transition period: A class-based approach using logistic regression and random forest algorithm," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 9, no. 3, p. 100 130, 2023, ISSN: 2199-8531. DOI: 10.1016/j.joitmc.2023.100130. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2199853123002329.

[25] G. GATAWA, *The Value Relevance of Financial Metrics to Publicly Listed Firms: Evidence from Large-, Mid-, and Small-Cap Firms Listed in NYSE and NSDAQ*, https://www.researchgate.net/publication/353896528_The_Value_Relevance_of_Financial_Metrics_to_Publicly_Listed_Firms_Evidence_from_Large-_Mid-_and_Small-Cap_Firms_Listed_in_NYSE_and_NSDAQ?fbclid=IwZXh0bgNhZW0CMTEAAR29ryuLzh-oefHn4wzhebAZrigMEOK
u2IwdXYvUqvReyM_aem_i_vL9y_Ou6u5yhBCEIcEZQ, [Accessed 29-01-2025].

[26] *Random Forest Algorithm in Machine Learning - GeeksforGeeks — geeksforgeeks.org*, https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/?fbclid=IwY2xjawICuRlleHRuA2FlbQIxMAABHfXNKugEBVAVea
aem_1rrOxNG3wl7VddgbSiLGRQ, [Accessed 26-01-2025].

[27] N. VERMA, *An Introduction to Support Vector Regression (SVR) in Machine Learning — nandiniverma78988*, https://medium.com/@nandiniverma78988/an-introduction-to-support-vector-regression-svr-in-machine-learning-681d541a829a, [Accessed 26-01-2025].

[28] *Xoriant Blog on Decision Trees for Classification: A Machine Learning Algorithm — xoriant.com*, https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm, [Accessed 26-01-2025].

[1] [2] [4] [7] [8] [13] [17] [22] [24]

[3] [6] [9] [11] [15] [16] [5] [10] [14] [23] [18]

[12] [20] [26] [21] [28] [27]

[19] [25]