


Bachelor's Thesis

Title of Bachelor's Thesis (English)	The Effect of Social Media Sentiment on Stock Price Prediction Models: A Comparative Study.
Title of Bachelor's Thesis (German)	Die Auswirkung der Stimmung in sozialen Medien auf Modelle zur Vorhersage von Aktienkursen: Eine vergleichende Studie.
Author (last name, first name):	Petschinger, Manuel
Student ID number:	12026023
Degree program:	Bachelor of Science (WU), BSc (WU) 
Examiner (degree, first name, last name):	Ph.D., Alexander, Mührmann

I hereby declare that:

1. I have written this Bachelor's thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.
2. This Bachelor's Thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.
3. This Bachelor's Thesis is identical with the thesis assessed by the examiner.
4. (Only applicable if the thesis was written by more than one author): this Bachelor's thesis was written together with

The individual contributions of each writer as well as the co-written passages have been indicated.

12.08.2024

Date



Signature

Bachelor Thesis

The Effect of Social Media Sentiment on Stock Price Prediction Models: A Comparative Study

Manuel Petschinger

Date of Birth: 04.01.2002

Student ID: 12026023

Subject Area: Finance

Studienkennzahl: 12026023

Supervisor: Univ.Prof. Alexander Mürmann, Ph.D.

Co-Supervisor: Yuan Chen, MSc

Date of Submission: 12.08.2024

Department of Finance, Accounting & Statistics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria



Contents

1	Introduction	8
1.1	Background and Motivation	8
1.2	Problem Statement	8
1.3	Research Objectives	9
1.4	Structure of the Thesis	9
2	Literature Review	10
2.1	Financial Market Prediction Models	10
2.2	Role of Sentiment Analysis in Finance	12
3	Economic Channels of Social Media Sentiment	14
3.1	Information Efficiency and Market Reactions	14
3.1.1	Efficient Market Hypothesis (EMH)	14
3.1.2	Information Asymmetry	14
3.1.3	Reduced Time of Market Reaction	14
3.2	Behavioral Finance	15
3.2.1	Herd Mentality	15
3.2.2	Cognitive Distortions	15
3.2.3	Media and Attention	15
4	Methodology	17
4.1	Correlation Analysis	17
4.1.1	Pearson's Coefficient	17
4.1.2	Spearman's Coefficient	18
4.2	Machine Learning Models	18
4.2.1	Linear Regression Model	18
4.2.2	Decision Tree Model	19
4.2.3	Random Forest Model	20
4.2.4	Artificial Neural Network	22
4.3	Sentiment Analysis	23
4.3.1	TextBlob Polarity	23
4.3.2	Fine-tuned DistilRoBERTa model	23
4.4	Performance Metrics	24
4.4.1	Coefficient of Determination (R^2)	24
4.4.2	Mean Squared Error (MSE)	25
4.4.3	Mean Absolute Error (MAE)	25

5	Experimental Setup	26
5.1	Data Sources	26
5.1.1	Reddit Data	26
5.1.2	Financial Data	27
5.2	Feature Engineering	27
5.2.1	Financial Metrics	27
5.2.2	Technical Indicators	29
5.2.3	Sentiment Scores	31
5.3	Feature Selection	33
5.4	Model Implementation	34
6	Results	35
6.1	Correlation Analysis Results	35
6.2	Comparative Analysis	37
6.2.1	Linear Regression	38
6.2.2	Decision Tree	39
6.2.3	Random Forest	40
6.2.4	Artificial Neural Network	41
7	Discussion	42
7.1	Interpretation of Results	42
7.2	Economic Implications	42
7.3	Limitations of the Study	43
7.4	Recommendations for Future Research	43
8	Conclusion	45

List of Figures

1	Decision tree model to predict the return for the upcoming two weeks of the Alphabet Inc. stock (sentiment scores included)	20
2	Graphical representation of the random forest model [1].	21
3	Graphical representation of a neural network [2].	22
4	TextBlob polarity (blue) vs. DistilRoBERTa score (red) for Amazon stock.	33
5	Correlation Analysis - Linear Regression	36
6	Scatterplots with Regression Lines	37
7	LR - R^2	38
8	LR - MSE	38
9	LR - MAE	38
10	DT - R^2	39
11	DT - MSE	39
12	DT - MAE	39
13	RF - R^2	40
14	RF - MSE	40
15	RF - MAE	40
16	ANN - R^2	41
17	ANN - MSE	41
18	ANN - MAE	41

List of Tables

1	TextBlob sentiment scores for sample text.	23
2	DistilRoBERTa sentiment scores for sample text.	24
3	Number of Reddit posts per stock	32
4	Pearson's correlation coefficients	35
5	Spearman's correlation coefficients	35
6	Comparison Table of the Linear Regression Analysis	38
7	Comparison Table of the Decision Tree Analysis	39
8	Comparison Table of the Random Forest Analysis	40
9	Comparison Table of the Artificial Neural Network Analysis	41

Abstract

The prediction of future stock prices with machine learning models is a key challenge in the modern financial industry, as it is influenced by a large number of complex and dynamic factors. This bachelor's thesis analyses to what extent the use of sentiment scores, derived from posts from the social media platform Reddit, can improve the predictive accuracy of traditional financial models for five US technology stocks: Alphabet Inc., Apple Inc., Amazon.com Inc., NVIDIA Corporation, and Microsoft Corporation. The correlation analysis, using Pearson's coefficient, Spearman's coefficient, and linear regression, shows a significant correlation between social media sentiment scores and the stock return of the subsequent 2 weeks. Various prediction models, including linear regression, decision tree, random forest, and artificial neural networks, were evaluated with and without inclusion of sentiment scores. The results show that sentiment scores have a certain predictive power, which varies depending on the model and stock. Especially for stocks such as Amazon and Microsoft, a significant improvement in model performance was observed by including sentiment scores. The economic implications of this research are promising. Investors can use sentiment scores to make more informed investment decisions and better manage risk. However, the study also shows that further research and additional data sources are needed to identify the conditions under which sentiment scores are most useful.

1 Introduction

1.1 Background and Motivation

In recent decades, the prediction of stock prices has become increasingly important, not only due to better technical systems, processing methods, and models, but also due to the increase in data streams. The prediction of future stock prices by machine learning techniques offers various possibilities for investors, which will be explained in more detail in the following study. Financial markets are complex systems that are influenced by a variety of factors, including macroeconomic indicators, corporate finances, and market sentiment. The rapid rise of social media platforms has created a huge new pool of data and with it the opportunity to capture and analyse public opinion in real time. By analysing posts on social media platforms such as Reddit and Twitter, investors can gain insight into the collective emotions and opinions of market participants, which can help them make more informed decisions.

1.2 Problem Statement

Although many studies recognise the importance of sentiment analysis for stock price prediction, most research focuses mainly on the use of historical price data. For example, Nguyen and his team made the following suggestion in the section "Conclusion & future work" of their study:

"One of the weaknesses of our method is that only the historical prices and sentiments derived from social media are considered. In future, we will try to find and integrate more factors which can affect the stock prices to develop a more accurate stock prediction model. For example, co-variance between stocks, macroeconomic indicators and the financial conditions of the company, which can be guessed from the income statement, balance sheet and cash flow, are important factors to be considered in the stock prediction model." [3]

This bachelor thesis will address exactly this problem and include balance sheets, income statements, cash flow statements, key interest rates from the central bank, and technical indicators in addition to historical prices. This allows us to analyse whether previous studies have only identified an improvement by including social media sentiment data because valuable financial data was missing or whether it can be shown that, despite traditional financial data, the sentiment of social media posts still improves price prediction models.

1.3 Research Objectives

The main objective of this research is to investigate the impact of incorporating sentiment scores from social media posts on the accuracy of stock prediction models. The following specific objectives are to be achieved:

1. Analysing the correlation between sentiment scores and future stock returns.
2. Evaluation of the effect of an inclusion of sentiment scores into machine learning models.
3. Evaluation of models and stocks for which the inclusion of sentiment scores significantly improves forecast accuracy.
4. Analysing the economic implications of the results for investors and analysts.

1.4 Structure of the Thesis

This thesis is divided into the following chapters:

2 Literature Review: This chapter provides an overview of existing research and models for predicting financial markets and the role of sentiment analysis in the financial industry.

3 Economic Channels of Social Media Sentiment: This chapter examines how social media sentiment influences economic decisions and market movements.

4 Methodology: The methods and models used to analyse the data and predict share prices are described in detail here.

5 Experiment Setup: This section describes the data collection, the processing tasks and the implementation of the models.

6 Results: The results of the correlation analysis and prediction models are presented.

7 Discussion: This chapter interprets the results, highlights the economic implications, and discusses the limitations of the study. Recommendations for future research are also given.

8 Conclusion: The main findings of the research are summarised and concluding remarks are given.

The Jupyter notebooks that contain the coding part of this thesis can be downloaded from the following github repository:

https://github.com/manuelpetsch/bachelor_thesis_petschinger.git

2 Literature Review

The Literature Review provides a comprehensive overview of the existing research and theoretical frameworks that underpin the study of financial market predictions and, in particular, the role of sentiment analysis in this domain. By synthesising the findings of various studies, this chapter sets the foundation for the subsequent analysis, offering insight into the strengths, limitations, and potential of integrating social media sentiment into financial models.

2.1 Financial Market Prediction Models

The prediction of financial markets is an area that attracts the interest of many researchers and practitioners in order to gain an advantage by predicting future share prices, market movements and overall financial trends. Although methods that attempt to predict stock prices have been around for several decades, this topic has experienced a considerable upswing in recent years. Thanks to continuous improvement in machine learning techniques, ever better hardware components and the use of artificial intelligence, these models have been improved and are constantly setting new benchmarks in terms of accuracy and reliability [4]. This chapter analyses various machine learning techniques that are used to predict financial markets and examines their individual areas of application, strengths, weaknesses and limitations.

Before describing different models and their characteristics, we will analyse the intended applications of machine learning techniques in the financial market. Zou and his team from the University of Adelaide [5] in Australia identify four key stock market prediction tasks:

1. Stock price prediction: The use of time-series data to anticipate future values for stocks and financial assets.
2. Stock movement prediction: Categorises stock trends into three categories - uptrend, downtrend and sideways.
3. Portfolio management: Strategic selection and oversight of a range of investments to achieve financial objectives.
4. Trading strategies: A pre-established set of guidelines is used to buy and sell stocks.

Starting with models such as Linear Regression, Support Vector Machines (SVM), Random Forests (RF), and K-Nearest Neighbours (KNN), we first review

those models that have long been used to predict market movements [6]. Linear regression attempts to establish a linear relationship between the independent and dependent variables. Although this approach is very easy to understand and interpret, in most cases it fails to capture the complexity of financial data [7]. Support Vector Machines (SVM) classify stock price movements by finding the optimal hyperplane that separates different classes in the feature space [8]. This model can handle high-dimensional data very well, but is quite susceptible to overfitting [9]. A random forest model uses randomisation to create many decision trees and aggregates their results to improve the accuracy of the overall prediction. Random forests are better at handling overfitting and processing large datasets, although this can be computationally intensive [10]. K-Nearest Neighbour models calculate their predictions based on similarities in past data points. Although KNNs are effective for small amounts of data, they struggle with large, high-dimensional data sets [11].

Deep learning models are characterised above all by the fact that they use several layers of abstraction to recognise complex patterns in large data sets. As a result, they have significantly improved price predictions in the financial sector [12]. Artificial Neural Networks (ANN) consist of interconnected neurons that process input features to predict stock prices. ANNs are very good at recognising hidden patterns, but require a very large amount of data and computational resources to provide good results [13]. Long Short-Term Memory (LSTM) networks, which belong to the supercategory Recurrent Neural Networks (RNN), are very useful for predicting time series in stock prices, as they are good at handling sequential data and long-term dependencies [6]. Another type, Convolutional Neural Networks (CNN), were originally designed for image recognition but were then adapted for financial market prediction. They treat stock data like temporary images and can therefore recognise local patterns effectively [14].

Hybrid models attempt to improve conventional models with a combination of different algorithms. An example of this would be a support vector machine that could be improved using genetic algorithms or sentiment analysis techniques. These models combine the strengths of several techniques and can therefore achieve better performance. However, these more comprehensive models are also more complex to implement and understand [15].

In addition to machine learning models, it is also very important to obtain the right data in sufficient quality and to prepare it correctly. Huang [12] considers the following types of data to be important for forecasting financial markets:

1. Historical prices (daily exchange rates, ...).

2. Technical indices (moving average, relative strength index, ...).
3. Financial news (financial messages, sentiment trend scores, ...).
4. Financial report data (balance sheets, income statements, ...).
5. Macroeconomic data (interest rates, gross foreign exchange reserves of central banks, ...)
6. Stochastic data (daily stock price fluctuations, ...).

It is essential that these data are processed and prepared properly for machine learning to perform well. Techniques such as Principal Component Analysis (PCA) or Wavelet Transforms help to reduce errors and ensure that the models can focus on the most relevant information. Several methods can be considered for the evaluation. In his work, Huang mentions error rates such as Mean Squared Error (MSE) or Mean Absolute Value (MAE) as well as accuracy indices such as Directional Predictive Accuracy (DPA). For the evaluation of practical performance, financial performance metrics are used, including the total return or Sharpe ratio, which provide insight into the practical advantages of the models in real-world trading scenarios [12].

2.2 Role of Sentiment Analysis in Finance

In this section, we examine the role of sentiment analysis in the financial market, how it has been used for predictions, and its impact on market behaviour. Sentiment analysis has become an additional tool for investors to predict stock prices and market movements by processing large amounts of microblogging data from social media platforms [16]. With the rise of social networks and easy access to their data via APIs, it is now possible for investors to process real-time data about the opinions of people that can significantly influence their decision-making processes.

Common data sources include Twitter, Reddit, and financial message boards. For example, Nguyen, Shirari and Velcin [3] use Yahoo Finance message boards to extract sentiment on specific companies, while Sprenger and his team [17] rely on Twitter data for their analysis. Although text data contains a large amount of information, it is very unstructured, which means that more preprocessing is required than with other data structures. One of the main tasks is therefore the conversion of text data into numbers without losing the relevant meaning of the posts [18]. Natural language processing (NLP) techniques, machine learning algorithms and sentiment analysis models such as TextBlob, VADER, and BERT

are used to calculate valuable sentiment information from raw social media posts [19, 16]. Sprenger [17] uses the Naive Bayesian text classification tool, which is a classification algorithm based on Bayes' theorem, to categorise text messages as buy, hold, or sell signals. However, not only the sentiment score of a post alone is important for stock price prediction. Palomo [16] observed that the popularity of a Twitter post is also crucial information, as his weighted-sentiment approach, considering the number of retweets, provides better performance than an equally weighted sentiment method.

Empirical evidence also shows that sentiment analysis can significantly enhance the predictive power of stock price prediction models. For example, Nguyen et al. [3] demonstrated that their method using sentiment analysis outperformed models using historical price data alone by 2.07 percent on average. The improvement was even greater for those stocks that are generally more difficult to predict, achieving 9.83 percent better accuracy than the historical price method. Sprenger et al. [17] have also found significant predictive power in the sentiments expressed in stock-related discussions on Twitter.

3 Economic Channels of Social Media Sentiment

Social Media offers a new channel through which market information is spread. Chapter 3 examines the various economic channels through which social media influences financial markets, analysing how the real-time character and the large range of social media platforms can influence investor behavior, market reactions, and stock prices.

3.1 Information Efficiency and Market Reactions

This section focusses on how social media contributes to information efficiency and how markets react to the rapid dissemination of information through these platforms.

3.1.1 Efficient Market Hypothesis (EMH)

The theory of efficient markets states that all available information is reflected in current share prices. In this scenario, social media can be seen as a new source of information and is therefore, according to the efficient market hypothesis, included in the prices of financial instruments [20]. Bollen, Mao, and Zeng [21] showed that the sentiment of Twitter posts has predictive power for market movements, suggesting that social media information is quickly integrated into stock prices. In reverse, price changes on the stock market also influence the sentiment of social media posts. Fekrazad, Harun, and Sardar [20] found a reverse causality on a daily and hourly basis. This results in a feedback loop: "social media sentiment about a particular stock affects its returns, and this, in turn, brings more attention to the stock, resulting in more tweets reflecting that sentiment" [20].

3.1.2 Information Asymmetry

Social media makes it possible to distribute information very quickly throughout the world. Jiao, Veiga, and Walther [22] found that social media has a similar and thus significant effect on stock prices as traditional news media. This finding emphasises the role of social networks in the dissemination of market information. As a result, similar to ordinary media, social networks can reduce the information asymmetry between market participants.

3.1.3 Reduced Time of Market Reaction

Social media not only creates a further development in terms of the scope of information dissemination, but also speed. As information is spread in real time via social media, it is possible to react to the market almost as quickly. As mentioned

above, the study by Bollen et al. [21] shows that Twitter sentiment can predict market movements within a short period of time, indicating a rapid integration of this information into stock prices.

3.2 Behavioral Finance

Section 3.2 examines the field of behavioural finance, which explores the psychological influences and biases that affect the decision-making process of investors. This section aims to provide a deeper understanding of the human elements that drive market movements by analysing the phenomena of herd instinct, cognitive biases, and media attention. In particular, it looks at how these cognitive factors are reinforced by social media and thus shape market behaviour.

3.2.1 Herd Mentality

Humans are social beings, which is why going with the crowd is part of our nature. This characteristic is not only noticeable in fashion trends but also has an impact on investor behaviour. When investors see that others are praising a stock, they are more inclined to buy it themselves [23]. Social networks play an important role in this aspect. It is possible to find out within seconds what other people think about a particular stock and then make an investment decision based on their opinion. This dynamic is discussed in more detail in the study by Jiao et al. [22]. Their work emphasises the importance of social media in the dissemination of investor sentiment and its impact on markets.

3.2.2 Cognitive Distortions

The usage of social media, in general, can lead to increased cognitive distortions. People tend to look for information that confirms their existing opinions, which is reinforced by the echo chamber effect in social media. This confirmation error results in distorted market decisions, as discussed in the studies by Neal and Wheatley [24] and Stambaugh et al. [25].

3.2.3 Media and Attention

According to the attention hypothesis, increased media attention leads to an increase in trading volume and price volatility. When attention is drawn to certain stocks or market areas in social networks, this can lead to market fluctuations. With regard to this phenomenon, Ballinari, Audrino, and Sigris [26] highlight the significant differences between the attention of retail investors and the attention of institutional investors. Although greater attention among unprofessional investors to news publications leads to greater volatility in the stock market, this effect is

rather low or even slightly negative for institutional investors. Institutional investors help stabilise the market by speeding up the price adjustment process, whereas retail investors slow this adjustment process due to noise-based trading and incorrect interpretations of new information. The effect of investor attention on volatility varies depending on the type and topic of news. News that is difficult to interpret or has significant implications for the market value of a company (e.g., mergers and acquisitions) tends to result in higher post-release volatility when retail investors pay attention. The study by Ballinari et al. [26] shows that these effects are not only statistically significant but also economically meaningful. Portfolio management strategies that incorporate these findings can lead to better performance, quantified as dozens of basis points.

In addition, the study by Jiao, Veiga and Walther [22] shows that there is a substantial difference between traditional news media and social media posts. Whereas coverage by traditional news media leads to decreases in market volatility and turnover, social media coverage predicts increases in subsequent volatility and turnover. This finding is particularly interesting in conjunction with the previous study by Ballinari et al. [26]. It can probably be assumed that this effect is due to the fact that retail investors tend to rely more on news from social media platforms, whereas institutional investors are more likely to focus on traditional or other forms of media. However, it is important to note that this is a quite vague interpretation of the findings of these two studies, which is not directly proven.

4 Methodology

This chapter outlines the research methods used to assess the impact of social media sentiment on stock price prediction. It covers correlation coefficients, the application of various machine learning models and the integration of sentiment scores. The chapter also explains the evaluation criteria for these models, ensuring a clear understanding of the approach taken to analyse and interpret the data.

4.1 Correlation Analysis

This section provides an overview of the methods that were used to analyse the correlation between the sentiment of social media posts and future stock returns. The correlation analysis will not only help us understand the degree to which changes in sentiment are associated with changes in stock prices, but also reveal for which time intervals of future returns the correlation is highest or lowest. It also aims to show whether social media sentiment can be a reliable predictor for stock price prediction models. For this study, both Pearson's and Spearman's coefficients were used to ensure that the result was not dependent on one method alone.

4.1.1 Pearson's Coefficient

Pearson's correlation coefficient measures the strength and direction of the linear relationship between two variables. The coefficient is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho(x, y) = \frac{E[xy]}{\sigma_x \sigma_y} \quad (1)$$

where x and y are two real-valued random variables with zero mean, $E[xy]$ is the cross-correlation between x and y , and $\sigma_x^2 = E[x^2]$ and $\sigma_y^2 = E[y^2]$ are the variances of x and y [27, 28, 29].

Pearson's coefficient [28] ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship, and
- 0 indicates that there is no linear relationship.

For our analysis, a high positive or negative correlation would suggest that sentiment scores could be used to predict future stock returns, whereas a value around zero would mean that there is no linear relationship between those two variables.

4.1.2 Spearman's Coefficient

Spearman's rank correlation coefficient measures the strength and direction of the monotonic relationship between two ranked variables. Unlike Pearson's correlation, Spearman's coefficient is calculated by the ranks of the variables and is therefore less sensitive to outliers. The coefficient is calculated by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2)$$

where d_i is the difference between the ranks of the corresponding values of x and y , and n is the number of observations [30].

Similar to Pearson's coefficient, Spearman's coefficient [30] also ranges from -1 to 1, where:

- 1 indicates a perfect positive monotonic relationship,
- -1 indicates a perfect negative monotonic relationship, and
- 0 indicates that there is no monotonic relationship.

The Spearman coefficient is particularly useful in our analysis if the relationship between sentiment scores and future returns is not linear, but still monotonic. The ranking of the data also reduces the influence of extreme values, which contributes to a more robust measurement of the correlation.

4.2 Machine Learning Models

This chapter describes the machine learning models that are used in this study, as well as the approaches used to fine-tune the models for optimal performance. Through various optimisation techniques, the model parameters are adjusted to improve the accuracy of stock price predictions.

4.2.1 Linear Regression Model

Linear regression is a basic statistical technique that models a linear relationship between a dependent variable (future stock returns) and one or more independent variables (financial data, sentiment scores, etc.) to predict the outcome of future events. The simplicity and interpretability of the model make it a popular choice for various purposes in statistics or data science [31].

The general formula of a linear regression model [31] is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (3)$$

where:

- Y is the dependent variable (future stock return),
- X_i are the independent variables (financial data, sentiment scores, etc.),
- β_i are the coefficients for each independent variable, and
- ϵ is the error rate.

According to Kanade's article [31], the linear regression model only works well if a few conditions are met:

1. Linear relationship: the dependent and independent variables should be linearly related.
2. Normal distribution of residuals: if not, it is possible that the estimation of the model becomes too wide or narrow.
3. Multicollinearity: if several independent variables correlate, it is difficult to find out which one predicts the target variable.
4. Autocorrelation: the dataset should not be autocorrelated.
5. Homoscedasticity: the residuals or error terms must have constant variance.

These points already show that linear regression is probably not the optimal model for predicting future stock returns. Some of the points can perhaps still be discussed, but others are clearly not fulfilled. For example, financial data, especially share prices, often exhibit autocorrelation since the prices of today are strongly influenced by the prices from yesterday. Nevertheless, it makes sense to use this model in the project, as it serves as a good basis. It is easy to implement and interpret and shows which independent variables have a linear relationship with future returns.

4.2.2 Decision Tree Model

A decision tree is also a widely used model in data analysis. This non-parametric supervised learning method can be used for both classification and regression problems. The model develops simple if-else decision rules to divide the data into smaller subsets and consequently predict the target variable [32]. The following example shows how the decision tree makes its decisions for the data set used in this project.

Similarly to the linear regression model, this model is also very simple to interpret, which means that it is easy to understand how the model makes its decisions.

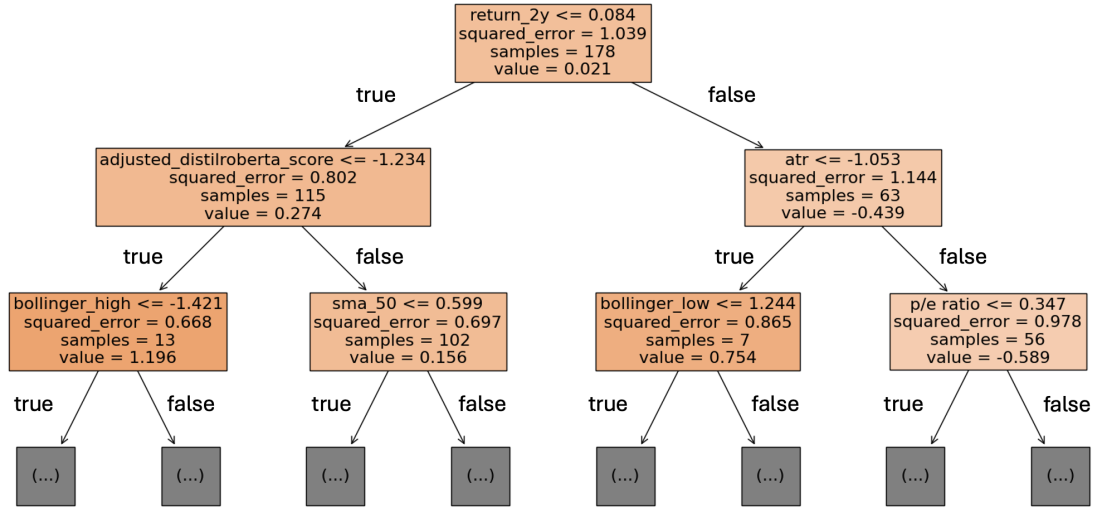


Figure 1: Decision tree model to predict the return for the upcoming two weeks of the Alphabet Inc. stock (sentiment scores included).

A decision tree therefore belongs to the category of white-box models. In contrast, the results of a black-box model (e.g., a neural network) are much more difficult to interpret. A disadvantage of decision trees, on the other hand, is that they tend to develop overly complex trees that do not generalise the data well. This phenomenon is also known as overfitting. However, to counteract this issue, there is a parameter that can be used to specify the maximum size of the tree. It is therefore essential to set this `max_depth` parameter optimally so that this model can perform at its best [32]. In order to ensure this 'optimal' choice of parameter, a loop was used in this project to test which parameter value results in the lowest mean squared error (on the test data set). The term mean-squared error and how it is calculated is explained in more detail in chapter "4.4 Performance metrics". In the loop just mentioned, values between 3 and 20 as well as a tree without a limit were tested for each subdata set on which a tree was trained.

4.2.3 Random Forest Model

The random forest model is an extended version of the decision tree developed by Leo Breiman and Adele Cutler. The model combines the output of several decision trees, which are trained on different samples of the data set, into a single result. This makes it possible to improve overall accuracy and reduce overfitting. The random forest algorithm uses both bagging and feature randomness to create an uncorrelated forest of decision trees [1, 33].

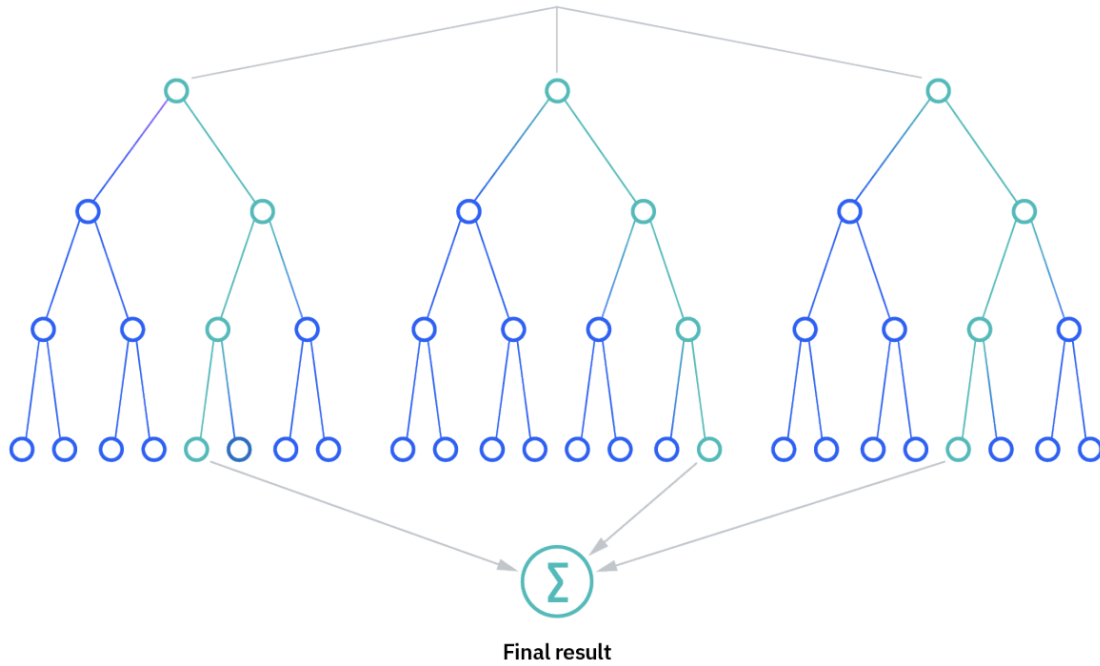


Figure 2: Graphical representation of the random forest model [1].

The biggest advantages of the Random Forest model include, as already mentioned, the reduced risk of overfitting, the wide range of applications, and the ease of determining the most important variables. However, there are also some challenges to this model. Since the data need to be processed for each individual decision tree, the model requires more computational resources for large datasets. The increased complexity compared to normal decision trees also means that it is more difficult to understand how the model makes its decisions. We therefore no longer talk about a white box model [1].

For this model, there are a few parameters to set. In this project, only the values for `max_depth` (maximum depth of the trees) and `n_estimators` (number of trees in the forest) are determined manually - for the other parameters, we use the standard settings of the Scikit-Learn library [34]. As before, we use a loop to determine the 'best' parameters. 'Best' in this case refers to the lowest mean-squared error on the test set. The values for the maximum depth of the trees are the same as for the decision tree, while the values for `n_estimators` are between 50 and 500 (in increments of 50).

4.2.4 Artificial Neural Network

Artificial neural networks (ANN) are deep learning models that are inspired by the way the human brain and nervous system work [35]. The networks consist of several interconnected neurones that are organised into three categories of layers [2]:

- Input layer: receives the input data.
- One or more hidden layers: process the data using weighted connections and activation functions.
- Output layer: returns the prediction.

The model runs through the network repeatedly and adjusts the weights of the connections in order to increase the accuracy of the prediction from run to run [35].

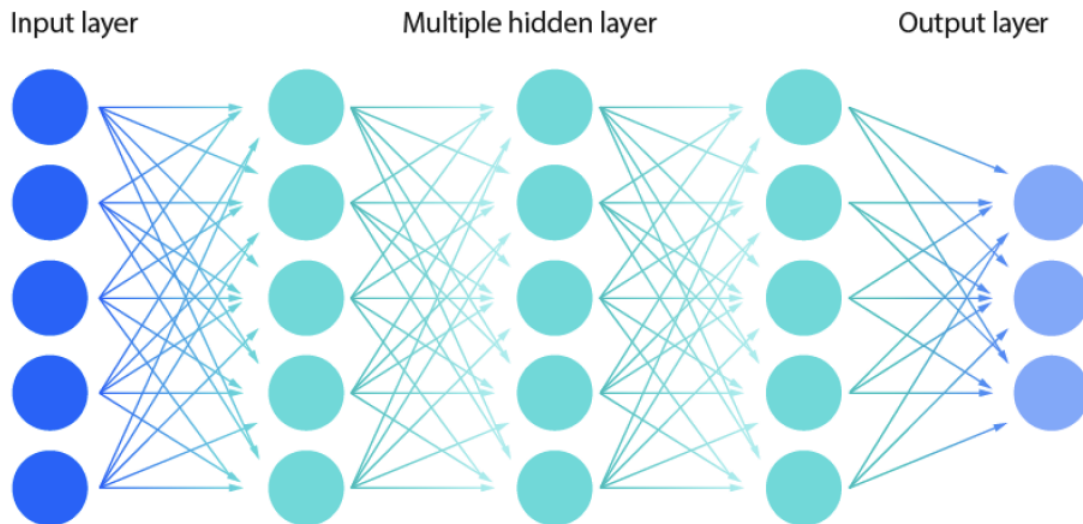


Figure 3: Graphical representation of a neural network [2].

For this project, a neural network was used from the Tensorflow and Keras libraries [36]. Before we continue with the setting of the parameters, it is important to have a look at the architecture of the neural network. In addition to the input and output layers, the model consists of six hidden layers that use the ReLU-activation function. The first hidden layer consists of 1,024 neurones, the second one contains half that number (512), the third half again (256), etc. The number of loops through the training data to adjust the weights (epochs) has been set

to 100 and the number of training examples used to update the weights once (batch_size) has been set to 32. 20 percent of the training data are used for validation (validation_split).

4.3 Sentiment Analysis

This chapter explains the methods used to convert text from social media posts into numerical sentiment scores, which represent the mood and emotions of the people who wrote them. The quality of these sentiment scores plays a crucial role in predicting future stock returns and is therefore very important for this project. The aim is for the sentiment scores to reflect public sentiment towards certain stocks or markets. In order to analyse both general sentiment (positive vs. negative) and finance-related sentiment (buy vs. sell), two different sentiment analysis models were used, which are described in more detail below.

4.3.1 TextBlob Polarity

TextBlob is a well-known Python library designed for simple natural language processing (NLP). TextBlob calculates the polarity of a text, which measures the degree of positivity, neutrality, or negativity. The output for the polarity of the text is a number between -1 and 1, whereby a negative value means a negative mood of the text, a value around 0 means a neutral mood and a positive value represents a positive mood. TextBlob can also be used to analyse the subjectivity of a text (between 0 and 1), but in this project only the polarity measurement is used [37].

The following example of the use of this model shows that it is very well suited to represent general mood but is not ideal for specific financial texts:

Text	Polarity	Subjectivity
The value of this stock will rise in the future!	0.0	0.125
This company is great.	0.8	0.750

Table 1: TextBlob sentiment scores for sample text.

4.3.2 Fine-tuned DistilRoBERTa model

To counteract this disadvantage, presented in Table 1, a more advanced natural language processing technique was integrated into the project in addition to a simple method such as TextBlob. The foundation for the upcoming model comes from a study by Christian Palomo [16], who used sentiment analysis of Twitter

data to predict the stock market. In his analysis, a BERT (Bidirectional Encoder Representations from Transformers) model that was refined with financial data outperformed all other models examined, with an accuracy and F1 score of 0,843.

DistilRoBERTa, the model used for this thesis, is a compressed version of the BERT model and is based on the Transformer architecture. A major advantage over TextBlob is that it is possible to refine this model for specific application purposes. However, since such a refinement would go beyond the scope of this bachelor thesis, a pre-trained and fine-tuned model from Hugging Face was used. Manuel Romero trained this model on a data set that contains 4.840 financial news sentences in English that are categorised by sentiment. The data set was manually divided by the agreement rate of five to eight annotators [38].

As visualised in Table 2, the model was applied to the same example sentences as before TextBlob. The label 'positive' stands for a buy signal, 'negative' for a sell signal, and 'neutral' for a hold signal. The value in the 'Score' column represents the certainty of the label [16]. To combine the two columns into a single value that can be interpreted in a way similar to the polarity of TextBlob, the labels were converted into numbers (positive to 1, neutral to 0, negative to -1) and then multiplied by the respective score.

Text	Label	Score
The value of this stock will rise in the future!	positive	0.999
This company is great.	positive	0.900

Table 2: DistilRoBERTa sentiment scores for sample text.

4.4 Performance Metrics

This section presents the key performance indicators used to evaluate the prediction models. Three different performance metrics, the coefficient of determination, the mean squared error, and the mean absolute error were used to evaluate and compare the accuracy and reliability of the models.

4.4.1 Coefficient of Determination (R^2)

The R^2 -coefficient is a number between 0 and 1 that indicates how well the independent variables predict the dependent variable. To be precise, the value indicates the proportion of the variance of the dependent variable that is explained by the independent variables. A higher value close to 1, indicates a better fit of the model to the data [39].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

where:

- y_i are the actual values,
- \hat{y}_i are the predicted values, and
- \bar{y} is the average of the actual values.

4.4.2 Mean Squared Error (MSE)

The MSE measures the mean square error between predicted and actual values. A lower value therefore indicates better model performance [40].

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

where:

- y_i are the actual values, and
- \hat{y}_i are the predicted values.

4.4.3 Mean Absolute Error (MAE)

In comparison to the MSE, the MAE does not measure the average squared error, but the average absolute error between predicted and actual values. The MAE value is highly correlated with the MSE value, wherefore a lower MAE also means better model performance. However, the MAE is easier to interpret as it indicates how far the predictions of the model are from the actual values on average [41].

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

where:

- y_i are the actual values, and
- \hat{y}_i are the predicted values.

5 Experimental Setup

This section describes in detail how the experiment was set up. It lays the foundation for conducting the experiments and subsequently analysing the results. After a detailed explanation of the data sources, it is shown how the data was prepared and implemented in the models.

5.1 Data Sources

As this project aims to combine traditional financial data with sentiment from social media, it requires two completely different types of data. Therefore, the chapter is divided into Reddit and financial data. While the first type comes from a single source, namely Reddit, the umbrella term financial includes information like historical share prices, FED fund rates, income statements, balance sheets, and cash flow statements, coming from three different data sources.

5.1.1 Reddit Data

Reddit is a popular social media platform where users can share posts and discuss various topics. For this study, the Reddit API was used to collect posts from five relevant subreddits relating to financial markets and equities: **r/wallstreetbets**, **r/stocks**, **r/investing**, **r/StockMarket**, and **r/technology**. The Python library PRAW (Python Reddit API Wrapper) helped to ensure a simple and stable interaction with the Reddit API. PRAW made it possible to implement a Python function that retrieves posts from the past year for a specific stock from a specific subreddit and saves them in a CSV file. The restriction on the past year was necessary due the limit of the API and ensured that the data was available in the desired density, e.g. enough posts per week. The function was used to query data from all five previously named subreddits for 21 stocks in total:

- | | | |
|--------------------------------------|------------------------------|---|
| – Apple Inc. | – Netflix, Inc. | – International Business Machines Corporation (IBM) |
| – Alphabet Inc. | – Adobe Inc. | |
| – Amazon.com, Inc. | – NVIDIA Corporation | – Oracle Corporation |
| – Advanced Micro Devices, Inc. (AMD) | – PayPal Holdings, Inc. | – Qualcomm Incorporated |
| – Microsoft Corporation | – Cisco Systems, Inc. | |
| – Tesla, Inc. | – CrowdStrike Holdings, Inc. | – Intel Corporation |
| – Meta Platforms, Inc. | – Uber Technologies, Inc. | – Salesforce, Inc. |

- The Trade Desk, Inc.
- Electronic Arts Inc.

5.1.2 Financial Data

Again it is noted that the aim of this thesis is not to compare a simple model with pure price data to a model with sentiment scores, but rather to have an already comprehensive model as a basis. Therefore, it is essential to retrieve comprehensive financial data.

However, the most important input for a stock price prediction model is **historical price data**. This data was obtained via the Yahoo Finance API [42] and includes daily closing prices, trading volumes, high prices, low prices, and adjusted closing prices that are adjusted for dividends and splits. These historical prices form the basis for calculating returns and technical indicators.

It is also necessary to obtain detailed **financial reports**. The Alpha Vantage API [43], although limited to 25 queries per day, made it possible to retrieve quarterly reports, including income statements, balance sheets, and cash flow statements. This data provides insights into the financial health and performance of companies. Income reports contain information on turnover, profit and operating results, balance sheets provide insight into the assets, liabilities, and equity of a company, and the cash flow statements show the cash flows from operating, investing, and financing activities. In addition, this data allows further financial metrics to be calculated, as shown in section 5.2.1.

In order to capture the macroeconomic environment of the companies, the historical federal funds rates of the US Federal Reserve were downloaded from the FRED website [44].

5.2 Feature Engineering

This chapter shows how the raw data was used to calculate valuable features that make a major contribution to the performance of the models.

5.2.1 Financial Metrics

Financial key performance indicators (KPIs) provide insight into the performance and health of a company. The following section outlines the key figures used and their calculation.

Gross Profit Margin

The gross profit margin shows the percentage of total revenue that remains as gross profit and is therefore a measure of efficiency in production or provision of services of a company [45].

$$\text{Gross Profit Margin} = \frac{\text{Gross Profit}}{\text{Total Revenue}} \quad (7)$$

Operating Margin

The operating profit margin shows what proportion of total revenue remains as operating profit after deducting operating costs, excluding interest and taxes. This key figure helps to assess the operational efficiency of a company [46].

$$\text{Operating Margin} = \frac{\text{Operating Income}}{\text{Total Revenue}} \quad (8)$$

Net Profit Margin

The net profit margin measures the proportion of total sales that remains as net profit after deducting all costs, including taxes and interest. The net profit margin provides a comprehensive overview of the profitability of a company [45].

$$\text{Net Profit Margin} = \frac{\text{Net Profit}}{\text{Total Revenue}} \quad (9)$$

Current Ratio

The current ratio shows the liquidity of the company by evaluating the ability of a company to cover its current liabilities with its current assets [45].

$$\text{Current Ratio} = \frac{\text{Total Current Assets}}{\text{Total Current Liabilities}} \quad (10)$$

Debt to Equity Ratio

The debt to equity ratio shows the ratio of the total liabilities of a company to its equity and therefore helps to assess its risk [45].

$$\text{Debt to Equity Ratio} = \frac{\text{Total Liabilities}}{\text{Total Shareholder Equity}} \quad (11)$$

Earnings per Share (EPS)

Earnings per share is an important indicator to assess the profitability of a company and is calculated by dividing the net profit by the number of outstanding shares [47].

$$\text{EPS} = \frac{\text{Net Income}}{\text{Common Stock Shares Outstanding}} \quad (12)$$

Price to Earnings Ratio (P/E-Ratio)

The price per earnings ratio compares the current share price of a company with its earnings per share. Therefore, it helps to determine whether a share is over- or under-valued [48].

$$\text{P/E Ratio} = \frac{\text{Adj Close}}{\text{EPS}} \quad (13)$$

Free Cash Flow

Free cash flow measures the amount of money that a company has left over after deducting capital expenditure from operating cash flow [49].

$$\text{Free Cash Flow} = \text{Operating Cash Flow} - \text{Capital Expenditures} \quad (14)$$

5.2.2 Technical Indicators

Technical indicators are mathematical calculations based entirely on historical prices that can help identify important trends. The following technical indicators were calculated using the Python library `ta`.

Simple Moving Average (SMA)

The simple moving average (SMA) calculates the average price of a stock over a certain period of time - in this case 50 days [50].

$$\text{SMA}_{50} = \frac{1}{50} \sum_{i=1}^{50} \text{Adj Close}_i \quad (15)$$

Exponential Moving Average (EMA)

The exponential moving average (EMA) weights recent prices more heavily than the SMA [50].

$$\text{EMA}_{50} = \text{Adj Close}_t \cdot \frac{2}{1 + 50} + \text{EMA}_{t-1} \cdot \left(1 - \frac{2}{1 + 50}\right) \quad (16)$$

Relative Strength Index (RSI)

The relative strength index (RSI) measures the speed and change in price movements. An RSI above 70 indicates an overbought situation, while an RSI below 30 indicates an oversold situation [50].

$$RSI_{14} = 100 - \frac{100}{1 + \frac{\text{Average Gain over 14 days}}{\text{Average Loss over 14 days}}} \quad (17)$$

Bollinger Bands

The upper Bollinger Band is 2 standard deviations above the SMA, while the lower Bollinger Band is 2 standard deviations below the SMA. They help measure the volatility of a stock price [50].

$$\text{Bollinger High} = \text{SMA}_{20} + 2 \cdot \text{Standard Deviation}_{20} \quad (18)$$

$$\text{Bollinger Low} = \text{SMA}_{20} - 2 \cdot \text{Standard Deviation}_{20} \quad (19)$$

Moving Average Convergence Divergence (MACD)

The MACD shows the relationship between two exponential moving averages (usually 12- and 26-day EMAs). A positive value indicates an upward trend and a negative MACD indicates a downward trend [50].

$$\text{MACD} = \text{EMA}_{12} - \text{EMA}_{26} \quad (20)$$

Stochastic Oscillator

The stochastic oscillator compares the current price of a share with its prices over a certain period of time. It helps identify over-bought and over-sold conditions [50].

$$\text{Stochastic \%K} = \frac{\text{Adj Close} - \text{Lowest Low}_{14}}{\text{Highest High}_{14} - \text{Lowest Low}_{14}} \times 100 \quad (21)$$

$$\text{Stochastic \%D} = \text{SMA}_3 \text{ of Stochastic \%K} \quad (22)$$

Average True Range (ATR)

The average true range (ATR) measures market volatility by calculating the average true range of a time period. A higher ATR value indicates a greater volatility [51].

$$\text{ATR}_{14} = \text{SMA}_{14} (\max(\text{High} - \text{Low}, |\text{High} - \text{Close}_{\text{prev}}|, |\text{Low} - \text{Close}_{\text{prev}}|)) \quad (23)$$

Ichimoku Cloud

The Ichimoku cloud is a comprehensive indicator that identifies trends and momentum, among other insights [50].

$$\text{Ichimoku A} = \frac{\text{Conversion Line} + \text{Base Line}}{2} \quad (24)$$

$$\text{Ichimoku B} = \frac{\text{Highest High}_{52} + \text{Lowest Low}_{52}}{2} \quad (25)$$

Commodity Channel Index (CCI)

Commodity Channel Index (CCI) measures the deviation of the share price from its statistical average. High positive values indicate an overbought situation, while high negative values indicate an oversold situation [50].

$$\text{CCI}_{20} = \frac{\text{Typical Price} - \text{SMA}_{20}(\text{Typical Price})}{0.015 \cdot \text{Mean Deviation}} \quad (26)$$

$$\text{Typical Price} = \frac{\text{High} + \text{Low} + \text{Close}}{3} \quad (27)$$

5.2.3 Sentiment Scores

In chapter 4.3, we have examined two models to convert text into sentiment scores: TextBlob and a fine-tuned DistilRoBERTa model. This section describes how the data was processed before and after using TextBlob and DistilRoBERTa to calculate the sentiment scores.

After checking the raw csv files for missing values, inconsistent data types or other errors, the single files were merged to one data frame per stock. This made it possible to analyse the number of Reddit posts that could be queried for each stock with the API (Table 3). To prepare the posts for analysis, URLs were stripped from the text and post-scores were normalised within each stock group to standardise the data for comparative analysis. The text data went through further preprocessing, where all content was converted to lowercase for uniformity. The Spacy English SM model [52] was used to remove stop words and perform lemmatisation. This ensures that the texts are reduced to their base forms, increasing the accuracy of TextBlob and DistilRoBERTa.

Figure 4 shows the sentiment scores of TextBlob and DistilRoBERTa over one year for Reddit posts about Amazon stock. Since the outputs of DistilRoBERTa are always fixed labels - negative, neutral, positive - which are then multiplied by

Stock	Number of Posts
Alphabet Inc.	687
Apple Inc.	629
Amazon.com, Inc.	597
NVIDIA Corporation	535
Microsoft Corporation	520
Tesla, Inc.	483
Meta Platforms, Inc.	410
Advanced Micro Devices, Inc. (AMD)	394
Intel Corporation	312
Netflix, Inc.	206
PayPal Holdings, Inc.	192
Uber Technologies, Inc.	150
Oracle Corporation	82
Adobe Inc.	81
International Business Machines Corporation (IBM)	76
Salesforce, Inc.	67
Cisco Systems, Inc.	62
CrowdStrike Holdings, Inc.	41
Qualcomm Incorporated	31
The Trade Desk, Inc.	25

Table 3: Number of Reddit posts per stock

the score of the model, the range of the output values is much higher than with TextBlob, whose values tend to be around zero.

To predict the future return of a stock, Palomo [16] found that “a weighted-sentiment approach based on the number of retweets provides better performance than equally weighted sentiments.” Therefore, to calculate the weighted mean of TextBlob and DistilRoBERTa scores, a Python function was defined that looks at the sentiment data for a specific stock within a seven-day window leading up to a given date. This function filters the relevant sentiment entries and computes weights based on post-scores that represent the popularity of a Reddit post. Using these weights, the function calculates the weighted mean for both the TextBlob polarity and the DistilRoBERTa scores.

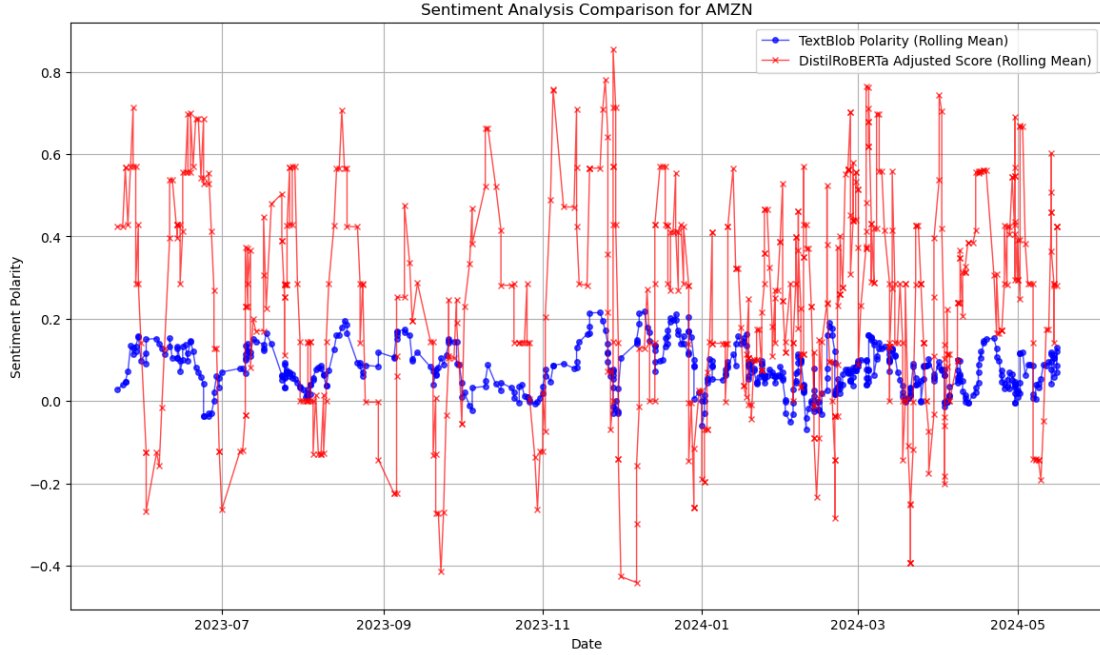


Figure 4: TextBlob polarity (blue) vs. DistilRoBERTa score (red) for Amazon stock.

5.3 Feature Selection

The large number of variables that were calculated during the feature engineering task makes it necessary to implement feature selection techniques to reduce them to the most important ones. This part of the project ensures that the models can focus on the most relevant input variables.

Poletaev and his team [53] evaluated various feature selection techniques, of which either Granger Causality or Feature Correlation has been proven to be the best, depending on the task, model, and data. To get the best out of both methods, the best ten features were calculated with Granger Causality and the best ten features were chosen with Feature Correlation. Subsequently, the union of both ten feature sets was used to form the following set of 16 features in total:

- ichimoku_a – return_2y – atr – macd
- ema_50 – eps – return_1m – return_1d
- return_1y – ichimoku_b – return_2d – currentratio
- p/e ratio – sma_50 – bollinger_low – bollinger_high

5.4 Model Implementation

When implementing the models, it turned out that the performance increases rapidly if a separate model is trained for each stock rather than training a large model with all the data. For this reason, it was necessary to split the whole data frame into smaller data frames, each representing the data for one stock. However, if the analysis were to be carried out for all stocks (20 stocks as in Table 3), the project would lose clarity. Therefore, the five stocks with the highest number of Reddit posts were selected for analysis: Alphabet, Apple, Amazon, Nvidia, and Microsoft. All five data frames were converted to numpy arrays to improve the computational efficiency of machine learning models for stock price prediction. The data was scaled using the Scikit-Learn `StandardScaler()` function. The Standard Scaler calculates the mean and standard deviation during fitting and applies the transformation to standardise the data during transformation. This preprocessing step is crucial to improve the performance and stability of machine learning algorithms [54]. In addition, the data needed to be split into training and test data. After the models are fitted to the training data, the test data is used to evaluate and compare their results. With the help of the Scikit-Learn `train_test_split` function, the numpy arrays for each stock were randomly split into 90 percent training and 10 percent test data [55].

6 Results

6.1 Correlation Analysis Results

This section analyses the relationship between sentiment scores from social media posts and future stock returns. This correlation analysis helps to understand whether changes in sentiment are related to changes in stock returns. Pearson’s and Spearman’s correlation coefficients were used to ensure a robust assessment of the relationship. In addition, linear regression analyses were performed to assess the predictive power of the TextBlob and DistilRoBERTa scores.

	textblob_polarity	adjusted_distilroberta
return_next_1w	-0.085775	0.020140
return_next_2w	-0.114930	-0.106598
return_next_3w	0.018289	0.021373
return_next_1m	-0.068772	-0.012273

Table 4: Pearson’s correlation coefficients

	textblob_polarity	adjusted_distilroberta
return_next_1w	-0.084748	-0.012394
return_next_2w	-0.139872	-0.124290
return_next_3w	-0.030775	-0.007897
return_next_1m	-0.108654	-0.017794

Table 5: Spearman’s correlation coefficients

Both Pearson and Spearman correlation results show that the relationship between sentiment scores and future stock returns is mostly weak. The strongest correlation is observed between the TextBlob polarity and the returns of the next two weeks with values of -0.114930 (Pearson) and -0.139872 (Spearman), indicating that a poorer sentiment score leads to a higher stock return for the subsequent two weeks. For both sentiment analysis models and both correlation coefficients, the correlation between sentiment and the return of the next two weeks is the highest. It therefore makes the most sense to select the return of the future two weeks as the target variable for the predictions of the models.

Linear regression analyses were performed to further investigate the predictive power of sentiment scores. In contrast to the previous analysis part, only the

OLS Regression Results						
Dep. Variable:	return_next_2w	R-squared:		0.013		
Model:	OLS	Adj. R-squared:		0.012		
Method:	Least Squares	F-statistic:		13.17		
Date:	Sat, 03 Aug 2024	Prob (F-statistic):		0.000299		
Time:	17:05:52	Log-Likelihood:		1782.6		
No. Observations:	986	AIC:		-3561.		
Df Residuals:	984	BIC:		-3551.		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.0127	0.002	7.547	0.000	0.009	0.016
textblob_polarity	-0.0561	0.015	-3.629	0.000	-0.086	-0.026
Omnibus:	41.329	Durbin-Watson:		0.412		
Prob(Omnibus):	0.000	Jarque-Bera (JB):		65.546		
Skew:	0.347	Prob(JB):		5.85e-15		
Kurtosis:	4.055	Cond. No.		12.3		

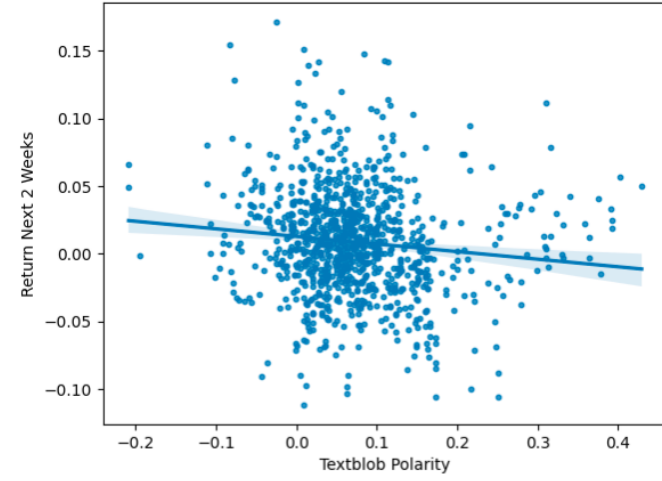
(a) textblob_polarity vs. return_next_2w

OLS Regression Results						
Dep. Variable:	return_next_2w	R-squared:		0.011		
Model:	OLS	Adj. R-squared:		0.010		
Method:	Least Squares	F-statistic:		11.31		
Date:	Sat, 03 Aug 2024	Prob (F-statistic):		0.000801		
Time:	17:05:52	Log-Likelihood:		1781.7		
No. Observations:	986	AIC:		-3559.		
Df Residuals:	984	BIC:		-3550.		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.0106	0.001	7.634	0.000	0.008	0.013
adjusted_distilroberta_score	-0.0101	0.003	-3.363	0.001	-0.016	-0.004
Omnibus:	44.121	Durbin-Watson:		0.407		
Prob(Omnibus):	0.000	Jarque-Bera (JB):		77.761		
Skew:	0.333	Prob(JB):		1.30e-17		
Kurtosis:	4.204	Cond. No.		2.48		

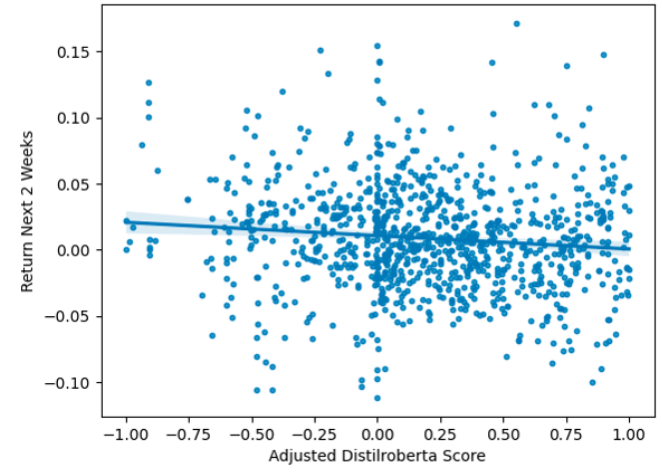
(b) adjusted_distilroberta vs. return_next_2w

Figure 5: Correlation Analysis - Linear Regression

relationship between sentiment scores and future two-week returns was analysed. The results show that both the linear regression between `textblob_polarity` and `return_next_2w` and between `adjusted_distilroberta` and `return_next_2w` have a low but significant predictive power.



(a) `textblob_polarity` vs. `return_next_2w`



(b) `adjusted_distilroberta` vs. `return_next_2w`

Figure 6: Scatterplots with Regression Lines

6.2 Comparative Analysis

This section analyses the differences between machine learning model performance with and without additional sentiment scores as independent variables.

6.2.1 Linear Regression

The following plots and table 6, representing the analysis of the linear regression model, show that the inclusion of sentiment scores improves the performance of the model for most stocks, as evidenced by higher R^2 values and lower MSE and MAE values, which means smaller deviations between predicted and actual values. A particularly strong improvement can be seen in the predictions for Amazon. On the other hand, the example of Nvidia shows that it is also possible that the inclusion of sentiment scores worsens model performance.

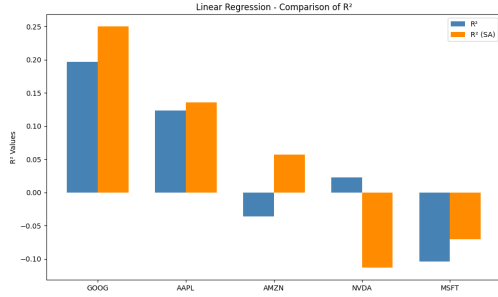


Figure 7: LR - R^2

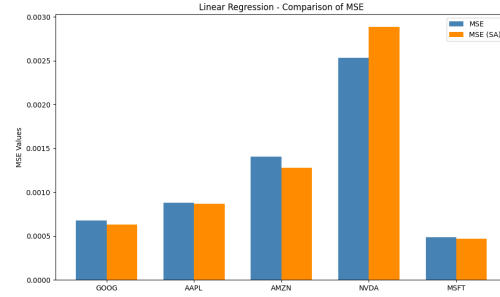


Figure 8: LR - MSE

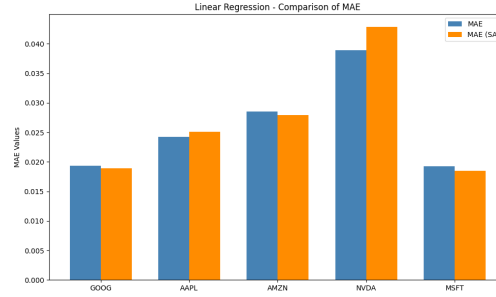


Figure 9: LR - MAE

Stock	R^2	R^2 (SA)	MSE	MSE (SA)	MAE	MAE (SA)
GOOG	0.196586	0.250002	0.000679	0.000634	0.019314	0.018942
AAPL	0.123083	0.135898	0.000880	0.000868	0.024243	0.025089
AMZN	-0.035865	0.056605	0.001405	0.001279	0.028522	0.027928
NVDA	0.022606	-0.113713	0.002532	0.002885	0.038886	0.042832
MSFT	-0.103925	-0.070515	0.000486	0.000471	0.019285	0.018476

Table 6: Comparison Table of the Linear Regression Analysis

6.2.2 Decision Tree

The results for the decision tree in table 7 show a very similar picture. For most stocks, the sentiment scores as additional variables are associated with an improvement in predictive performance, although there is again one example, namely Apple, where this is not the case. The predictions with sentiment analysis for the Apple stock result in a lower R^2 -Score and higher MSE and MAE values than with financial data alone. For most of the cases, the decision tree is much better at predicting future stock prices than the linear regression model.

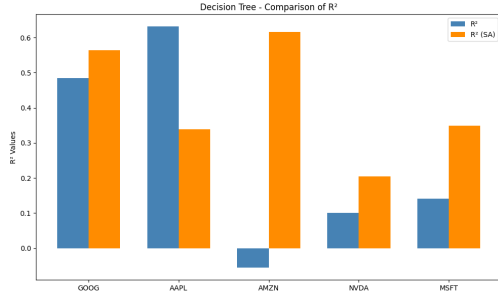


Figure 10: DT - R^2

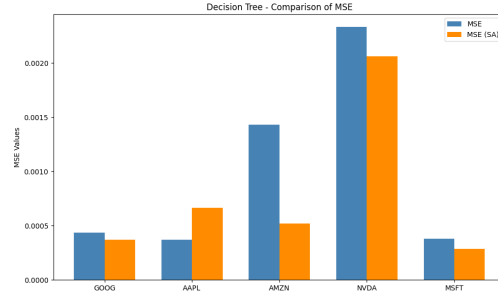


Figure 11: DT - MSE

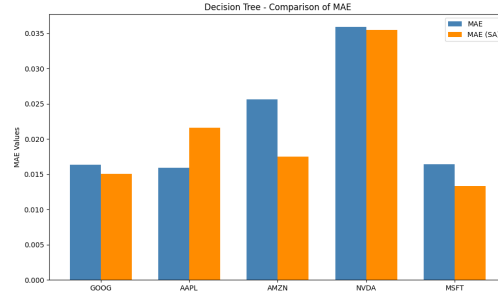


Figure 12: DT - MAE

Stock	R^2	R^2 (SA)	MSE	MSE (SA)	MAE	MAE (SA)
GOOG	0.484800	0.564745	0.000436	0.000368	0.016382	0.015072
AAPL	0.631964	0.338426	0.000370	0.000664	0.015907	0.021624
AMZN	-0.056238	0.616154	0.001432	0.000521	0.025596	0.017532
NVDA	0.100369	0.205030	0.002331	0.002060	0.035894	0.035487
MSFT	0.140253	0.348346	0.000379	0.000287	0.016421	0.013343

Table 7: Comparison Table of the Decision Tree Analysis

6.2.3 Random Forest

The Random Forest model, represented in table 8, also shows an improvement in model performance for the majority of the cases due to additional sentiment scores. Similar to the decision tree, Apple is again the outlier, showing a better performance without sentiment scores. In general, the Random Forest shows the best performance of all the models analysed and even achieves an R^2 value of 0.730194 and an MSE value of 0.000228 with the inclusion of sentiment analysis for the Alphabet stock. That means that the predictions for two weeks into the future only deviate from the true value by an average return of 1.298 percent.

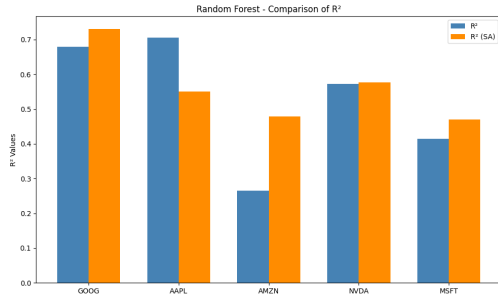


Figure 13: RF - R^2

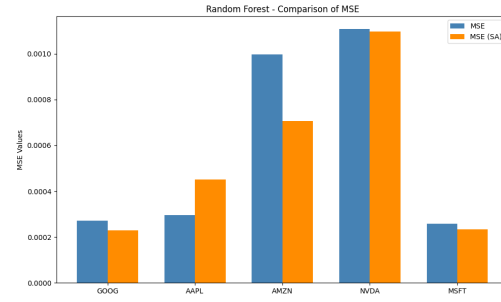


Figure 14: RF - MSE

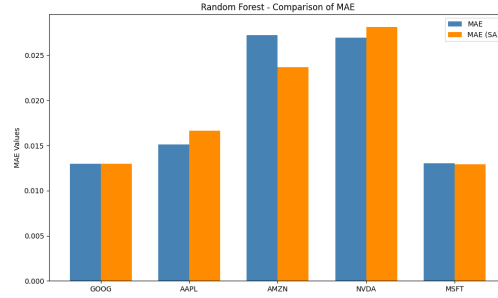


Figure 15: RF - MAE

Stock	R^2	R^2 (SA)	MSE	MSE (SA)	MAE	MAE (SA)
GOOG	0.679978	0.730194	0.000271	0.000228	0.012970	0.012980
AAPL	0.705421	0.551263	0.000296	0.000451	0.015120	0.016619
AMZN	0.264737	0.479519	0.000997	0.000706	0.027238	0.023659
NVDA	0.572552	0.576247	0.001107	0.001098	0.026934	0.028116
MSFT	0.414846	0.469629	0.000258	0.000234	0.013025	0.012927

Table 8: Comparison Table of the Random Forest Analysis

6.2.4 Artificial Neural Network

Table 9 presents the performance metrics for the analysis of artificial neural networks, which show mixed results when sentiment scores are included. While some stocks, such as Amazon and Microsoft, show a significant improvement in performance metrics, the results for other stocks such as Alphabet and Nvidia deteriorate. This suggests that the inclusion of sentiment scores in neural networks does not always lead to better model performance.

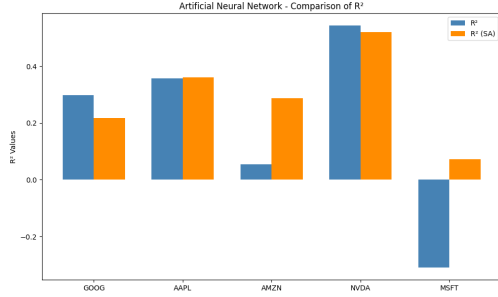


Figure 16: ANN - R^2

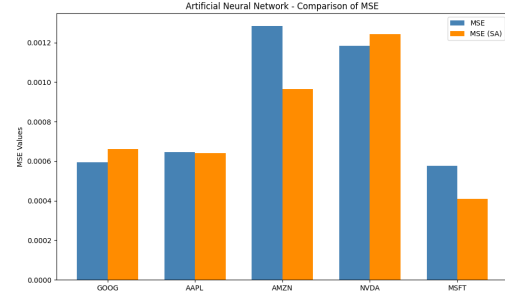


Figure 17: ANN - MSE

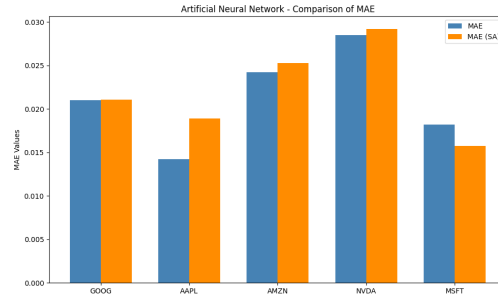


Figure 18: ANN - MAE

Stock	R^2	R^2 (SA)	MSE	MSE (SA)	MAE	MAE (SA)
GOOG	0.297374	0.217495	0.000594	0.000662	0.020986	0.021063
AAPL	0.357657	0.361241	0.000645	0.000641	0.014228	0.018887
AMZN	0.053867	0.288033	0.001283	0.000965	0.024194	0.025295
NVDA	0.543050	0.520068	0.001184	0.001243	0.028497	0.029173
MSFT	-0.309950	0.071645	0.000577	0.000409	0.018201	0.015763

Table 9: Comparison Table of the Artificial Neural Network Analysis

7 Discussion

This chapter interprets the results of the analyses from the previous section and discusses their economic implications. It also highlights the limitations of the study and makes recommendations for future research.

7.1 Interpretation of Results

The results of this study show that the inclusion of sentiment scores from social media posts has some predictive power for future stock returns, but this effect varies significantly depending on the model and stock used. For each of the four models, there is at least one test case in which the inclusion of social media sentiment leads to a deterioration in the prediction. However, in general, most of the test cases show that social media posts can be very valuable as an additional source of information that enhances stock price prediction models. Notably, for some cases, such as the decision tree with the Amazon stock, R^2 is negative with financial data alone but becomes reasonably positive when sentiment analysis is included. Especially for the prediction of Amazon and Microsoft stock returns, the additional sentiment scores appear to have a consistent and, in some cases, very strong effect. When comparing the models with each other, the Random Forest clearly shows the best and most stable results. With sentiment analysis included, its R^2 -score is above 0.46 for every test case conducted in this study.

7.2 Economic Implications

The results of this study have several economic implications.

Investment strategies: Investors could use the sentiment of social media posts to improve their investment strategies. Investors could benefit from this additional information, particularly in the case of stocks such as Amazon and Microsoft, where the models have been improved by sentiment scores.

Risk management: The inclusion of sentiment scores can also help to better manage risks. By analysing market sentiment, investors can react to positive or negative sentiment at an early stage and take appropriate action. It is also useful to understand the relationship between social media sentiment and stock market movements in order to assess how other, possibly large, market participants might react to changes in sentiment.

Market analysis: Financial analysts could integrate sentiment analysis techniques into their market analyses to create more comprehensive reports. This could lead to more precise forecasts and better recommendations for action.

7.3 Limitations of the Study

The following limitations of this study should be taken into account.

Data quality and availability: The restriction in the free availability of social media data in sufficient quality has led to a severe limitation of the study. The low density and short time horizon of the social media posts make a more precise analysis or implementation of more complex models impossible. This also leads to an instability of the results that vary greatly between the individual stocks and models. The inclusion of sentiment scores does not always lead to better predictions and the results depend heavily on the context. It can be assumed that this instability could be significantly reduced with a larger amount of data.

Model complexity: Models such as artificial neural networks require very large amounts of data in order to perform well. However, it was still useful to include such a model in the analysis to see how a neural network deals with the additional social media data, even if the general performance of the models suffers from the small amount of data.

Time restriction: Due to limitations of the Reddit API, the analysis is limited to a specific time period. Long-term trends and seasonal effects could not be taken into account.

7.4 Recommendations for Future Research

Based on the results and limitations of this study, the following recommendations are given for future research.

Expansion of data sources: Future studies should incorporate a broader range of data sources, including more social media platforms and traditional news sources, to refine the analysis. Most importantly, it would be interesting to see how the models would behave with data over a longer period of time and with higher density, e.g. more posts per week. A longer analysis period could help to better understand long-term trends and seasonal effects and to check the robustness of the models.

Optimisation of data processing and models: Further research could focus on further preprocessing, optimising the model parameters and integrating

additional variables to improve the accuracy of predictions. Depending on the amount of data and the area of application, it would be very interesting to see how other types of models, e.g. Support Vector Machines or Long Short-Term Memory models, would deal with sentiment scores.

Investigation of further stocks: The analysis should be extended to a larger number of stocks, for example in other sectors or countries, in order to be able to draw more generally valid conclusions.

Consideration of macroeconomic factors: Future research could incorporate macroeconomic factors to further refine the models and obtain a more comprehensive picture of market dynamics.

8 Conclusion

This bachelor thesis analysed the extent to which sentiment scores from social media posts can contribute to the prediction of future stock returns. The results show that there is a significant correlation between TextBlob or DistilRoBERTa sentiment scores and the stock return for the subsequent two weeks. Further analysis of four machine learning models demonstrated that sentiment scores, calculated from social media posts, can be a valuable addition to stock prediction models. However, it is clear that further research is necessary to identify the conditions under which these scores are most useful and to further improve the machine learning models included. The integration of sentiment scores offers a promising field for future studies and practical applications in the area of financial market prediction.

References

- [1] IBM, “What is random forest?” <https://www.ibm.com/topics/random-forest>, 2024, accessed: 2024-07-23.
- [2] —, “What is a neural network?” <https://www.ibm.com/topics/neural-networks>, 2024, accessed: 2024-07-28.
- [3] T. Nguyen, K. Shirai, and J. Velcin, “Sentiment analysis on social media for stock movement prediction,” *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417415005126>
- [4] C. W. R. University, “Advancements in artificial intelligence and machine learning,” <https://online-engineering.case.edu/blog/advancements-in-artificial-intelligence-and-machine-learning>, 2024, accessed: 2024-08-12.
- [5] J. Zou, Q. Zhao, and Y. Jiao, “Stock market prediction via deep learning techniques: A survey,” *arXiv preprint arXiv:2212.12717*, 2023.
- [6] P. Soni, Y. Tewari, and D. Krishnan, “Machine learning approaches in stock price prediction: A systematic review,” *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012065, 2022.
- [7] J. M. Sangeetha and K. J. Alfia, “Financial stock market forecast using evaluated linear regression based machine learning technique,” *Measurement: Sensors*, vol. 31, 2023.
- [8] A. Mammone, M. Turchi, and N. Cristianini, “Support vector machines,” *WIREs Computational Statistics*, vol. 1, no. 3, pp. 283–289, 2009. [Online]. Available: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.49>
- [9] Scikit-learn, “1.4. Support Vector Machines — scikit-learn 1.5.1 documentation,” 2023, accessed: 2024-07-28. [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html>
- [10] S. Rigatti, “Random Forest,” *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017. [Online]. Available: <https://doi.org/10.17849/insm-47-01-31-39.1>
- [11] L. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, revision #137311.

- [12] J. Huang, J. Chai, and S. Cho, “Deep learning in finance and banking: A literature review and classification,” *Frontiers of Business Research in China*, vol. 14, no. 1, p. 13, 2020. [Online]. Available: <https://doi.org/10.1186/s11782-020-00082-6>
- [13] M. Vijn, D. Chandola, and V. A. Tikkiwal, “Stock closing price prediction using machine learning techniques,” in *Procedia Computer Science*, vol. 167. Elsevier, 2020, pp. 599–606.
- [14] Y. L. e. a. W. Lu, J. Li, “A cnn-lstm-based model to forecast stock prices,” *Complexity*, vol. 2020, no. 1, p. 6622927, 2020. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/6622927>
- [15] A. Khan, A. Shah, and A. Ali, “A performance comparison of machine learning models for stock market prediction with novel investment strategy,” *PLOS ONE*, 2024.
- [16] C. Palomo, “Tweet sentiment analysis to predict stock market,” Stanford University, Stanford, CA, Tech. Rep., 2020, stanford CS224N Custom Project. [Online]. Available: <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-170049613.pdf>
- [17] T. Sprenger, A. Tumasjan, and P. Sandner, “Tweets and trades: the information content of stock microblogs,” *European Financial Management*, vol. 20, no. 5, pp. 926–957, 2014. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-036X.2013.12007.x>
- [18] S. Jansen, *Machine Learning for Algorithmic Trading - Second Edition*. Packt Publishing, 2020.
- [19] N. Das, S. Gupta, and S. Das, “A comparative study of sentiment analysis tools,” in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)*, 2021, pp. 1–7.
- [20] A. Fekrazad, S. Harun, and N. Sardar, “Social media sentiment and the stock market,” *Journal of Economics and Finance*, vol. 46, no. 2, pp. 397–419, 2022. [Online]. Available: <https://doi.org/10.1007/s12197-022-09575-x>
- [21] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S187775031100007X>

- [22] P. Jiao, A. Veiga, and A. Walther, “Social media, news media and the stock market,” *SSRN Electronic Journal*, 2020, accessed: 2024-07-21. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755933
- [23] A. Hayes, “Herd instinct: Definition, stock market examples, and how to avoid,” <https://www.investopedia.com/terms/h/herdinstinct.asp>, 2022, accessed: 2024-07-23.
- [24] R. Neal and S. M. Wheatley, “Do measures of investor sentiment predict returns?” *The Journal of Financial and Quantitative Analysis*, vol. 33, no. 4, pp. 523–547, 1998. [Online]. Available: <http://www.jstor.org/stable/2331130>
- [25] R. Stambaugh, J. Yu, and Y. Yuan, “The short of it: Investor sentiment and anomalies,” *Journal of Financial Economics*, vol. 104, no. 2, pp. 288–302, 2012, special Issue on Investor Sentiment. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0304405X11002649>
- [26] D. Ballinari, F. Audrino, and F. Sigrist, “When does attention matter? the effect of investor attention on stock market volatility around news releases,” *International Review of Financial Analysis*, vol. 82, p. 102185, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1057521922001466>
- [27] O. Dunn and V. Clark, *Applied Statistics: Analysis of Variance and Regression*. New York: Wiley, 1974.
- [28] K. Pearson, “Mathematical contributions to the theory of evolution.â€”iii. regression heredity and panmixia,” *Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896.
- [29] J. Rodgers and W. Nicewander, “Thirteen ways to look at the correlation coefficient,” *Amer. Statistician*, vol. 42, pp. 59–66, Feb. 1988.
- [30] C. Spearman, “The proof and measurement of association between two things.” *American Psychological Association*, 1961.
- [31] V. Kanade, “What is linear regression? types, equation, examples, and best practices for 2022,” <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/>, 2023, accessed: 2024-07-23.
- [32] Scikit-learn, “1.10. decision trees â€” scikit-learn 1.2.2 documentation,” <https://scikit-learn.org/stable/modules/tree.html>, 2024, accessed: 2024-07-23.
- [33] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>

- [34] Scikit-learn, “sklearn.ensemble.randomforestregressor,” <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>, 2024, accessed: 2024-07-28.
- [35] S. Walczak and N. Cerpa, “Artificial neural networks,” in *Encyclopedia of Physical Science and Technology (Third Edition)*, third edition ed., R. Meyers, Ed. New York: Academic Press, 2003, pp. 631–645. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B0122274105008371>
- [36] TensorFlow, “Tensorflow quickstart: Beginner,” <https://www.tensorflow.org/tutorials/quickstart/beginner>, 2024, accessed: 2024-07-28.
- [37] TextBlob, “Textblob quickstart,” <https://textblob.readthedocs.io/en/dev/quickstart.html>, 2024, accessed: 2024-07-28.
- [38] M. Romero, “distilroberta-finetuned-financial-news-sentiment-analysis,” <https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis>, 2024, accessed: 2024-07-23.
- [39] N. University, “Coefficient of determination r squared,” <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html>, 2024, accessed: 2024-07-29.
- [40] K. Stewart, “Mean squared error,” <https://www.britannica.com/science/mean-squared-error>, 2024, accessed: 2024-07-23.
- [41] M. Ahmed, “Understanding mean absolute error (mae) in regression: A practical guide,” <https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide\ -26e80ebb97df>, 2023, accessed: 2024-07-29.
- [42] Yahoo, “Yahoo finance api,” <https://developer.yahoo.com/api/>, 2024, accessed: 2024-08-06.
- [43] AlphaVantage, “Alpha vantage api documentation,” <https://www.alphavantage.co/documentation/>, 2024, accessed: 2024-08-06.
- [44] F. R. B. of St. Louis, “Effective federal funds rate (fedfunds),” <https://fred.stlouisfed.org/series/FEDFUNDS>, 2024, accessed: 2024-08-06.
- [45] T. Stobierski, “13 financial performance measures managers should monitor,” <https://online.hbs.edu/blog/post/financial-performance-measures>, 2020, accessed: 2024-08-03.

- [46] A. Hayes, “Operating margin: What it is and the formula for calculating it, with examples,” <https://www.investopedia.com/terms/o/operatingmargin.asp>, 2024, accessed: 2024-08-03.
- [47] J. Schmidt, “Earnings per share formula (eps),” <https://corporatefinanceinstitute.com/resources/valuation/earnings-per-share-eps-formula/>, 2024, accessed: 2024-08-03.
- [48] R. Berger, “How to understand the p/e ratio,” <https://www.forbes.com/advisor/investing/what-is-pe-price-earnings-ratio/>, 2023, accessed: 2024-08-03.
- [49] Handelszeitung, “Free cash flow,” <https://www.handelszeitung.ch/finanzlexikon/free-cash-flow#>, 2024, accessed: 2024-08-03.
- [50] M. Cutkovic, “Best trading indicators: A list of the 17 most used technical indicators,” <https://www.axi.com/eu/blog/education/trading-indicators>, 2024, accessed: 2024-08-04.
- [51] C. Markets, “Average true range (atr),” <https://www.cmcmarkets.com/de-at/hilfe/glossar/a/average-true-range>, 2024, accessed: 2024-08-04.
- [52] spaCy, “spacy models: en_core_web_sm-3.7.1,” https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.1, 2023, accessed: 2024-08-03.
- [53] A. Poletaev, B. Liu, and L. Li, “Improve the prediction in the digital era: Causal feature selection with minimum redundancy,” *Journal of Digital Economy*, vol. 3, pp. 14–36, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2773067024000086>
- [54] Scikit-learn, “sklearn.preprocessing.standardscaler,” <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>, 2024, accessed: 2024-08-01.
- [55] —, “sklearn.model_selection.train_test_split,” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html, 2024, accessed: 2024-08-03.