Forecasting the Stock Market using News Sentiment Analysis

Marinus Teunis Bakker
STUDENT NUMBER: 2025113

A THESIS SUBMITTED FOR THE PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:
Supervisor: prof. dr. Max Louwerse
Second reader: dr. Gonzalo Nápoles

Tilburg University
School of Humanities & Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 2021

[This Page Intentionally Left Blank]

Preface

I am thrilled to present my thesis which is the culmination of my Master's in Data Science & Society at Tilburg University. As a result of conducting research during the spring semester, I devoted myself to the subject of predicting the stock market, especially the S&P 500 index movements based on news article sentiment. Furthermore, because of the novelty of the dataset, I was able to contribute to the academic literature. Of course, I encountered several difficulties while conducting this research. However, I could not overcome all these challenges by myself. Therefore, I would like to thank everyone who has helped, supported, and challenged me along the way. Specifically, Max Louwerse my supervisor during this research. Because he was always challenging me, it felt like I was a real professor conducting a worthful research instead of "just" a student writing his thesis for graduating from his masters. In addition, I would like to express my appreciation to my fellow student Renate, for proofreading my thesis. Finally, I would like to express my gratitude for the love and support of my roommates, family, and girlfriend during this research.

I wish you a lot of enjoyment while reading this study.

Martijn Bakker
Tilburg, May 2021

[This Page Intentionally Left Blank]

Forecasting the Stock Market using News Sentiment Analysis

Marinus Teunis Bakker

For decades long, investors are speculating when it is the best time to buy or sell a stock. To support investors, this study investigates whether the sentiment of news articles could predict the daily stock market movement directionalities. Many recent news articles, which were published in the United States of America in addition to the S&P 500 price movements were utilized in the current study. Combining both data sources, creates the final dataset which has a time span from 2013 to early 2020. The current study distinguishes itself from existing research by applying this unique dataset, comparing different feature sets constructed from a particular news sentiment analysis score and/or technical indicators, and examining different machine and deep learning models. The key finding, as well as the answer to the central research question, is that the stock market can be predicted based on the sentiment polarity of news articles. Furthermore, it was discovered that using news sentiment polarity in conjunction with technical analysis features resulted in even better results. The results of the best performing combination of news sentiment analysis method, technical analysis, and the proposed support vector machine model are very promising. Based on these outcomes and with the deployment of the final model, investors can make better decisions, to buy or sell a stock.

Keywords: *classification analysis, stock market prediction, news sentiment analysis, S&P500, support vector machines.*

Contents

1 Introduction

The year 2019 will go down in history as a tragic year due to the emergence of the coronavirus pandemic (COVID-19). The number of people infected by the virus has grown rapidly and almost every country in the world has been affected. After the virus spread among Europe and the United States of America, an enormous number of negative news articles concerning the virus was published in February and March 2020. Every day people were reading the news to follow the development of the deaths and new infections by the virus. While the coronavirus was world news the Standard & Poor's 500 (S&P 500), a market-capitalization-weighted index representing the 500 largest publicly traded companies in the U.S., fell from about 3,386 on February 19th, 2020, to about 2,237 on March 23rd, 2020, comparable to a 41% decrease (Yilmazkuday, 2020). Thus, one may assume a negative correlation between the published coronavirus pandemic news articles and the stock market prices. This suggests that the stock market can be predicted using news article sentiment.

If a significant correlation can be discovered between news articles sentiment and the stock market value in the future, investors can benefit from it and increase their trading profits by reacting (buy/sell) on recently published news articles. Therefore, this thesis will investigate whether news articles sentiment can be used to predict the daily S&P 500 movements. Therefore, several machine learning and deep learning algorithms will be used to develop models and assess the prediction power of news article sentiment for the S&P 500 movements. Accordingly, the following main research question will be answered:

"Can news sentiment data be applied to predict stock market movements?"

Ican and Celik (2017) state that forecasting the stock market movements in the competitive financial market has been addressed extensively, since it could enable people to increase their trading profits. Therefore, researchers who were able to predict the stock market prices with some certainty may not always publish their work, because they could take advantage of it by making much profit on the stock market exchanges. Thus, from a practical point of view, the outcome of this research, in particular the machine or deep learning models, can be applied to improve trading decisions and increase profits. To explain, the models will predict the market price movements based on news articles, so an investor can speculate to buy or sell stocks.

Furthermore, this research will address a gap in the scientific stock market literature, namely the prediction of the stock market movements using a wide variety of recent news articles from different sources in the United States of America from 2013 to early 2020. Also, various machine and deep learning algorithms will be used to build models that are able to predict the stock market movement. These models will be executed with different datasets, sentiment analysis methods and with or without a technical price indicator. Hence, the datasets as well as the sentiment analysis methods can be compared, and the prediction power of news article sentiment can be assessed. How this research distinguishes itself from existing studies is further described in the following section.

2 Related Work

This chapter will focus on the latest and most relevant research regarding the prediction of the stock market using news articles or similar sources sentiment. The following sections will discuss theories about stock market prediction, news sentiment analysis, whether the research will apply classification or regression, and state-of-the-art algorithms and their results which has been achieved in previous studies to predict the stock market.

2.1 Theories of Stock Market Prediction

Nowadays, forecasting the stock market is a daily discussed topic by investors. What is the best time to buy or sell a stock? For decades this has been a difficult and important decision. Economists and investors are trying to time the market as effectively as possible by using different forecasting philosophies. However, the financial market has long believed in theories that claim predicting the stock market is impossible. These “old” theories include the Efficient Market Hypothesis (EMH) introduced by Fama (1965) and the Random Walk Theory (RWT) established by Regnault in 1863 (Sewell, 2011).

2.1.1 Efficient Market Hypothesis

The EMH claims that at any given point, all available information about a stock is already incorporated into its price. Consequently, a stock is never overvalued or undervalued, but has a fair price on exchanges (platform to buy or sell a stock). Therefore, the theory states that investors are not able to beat the market by speculating. The EMH consists of three different categories: weak, semi-strong and strong (Schumaker & Chen, 2006).

The weak EMH suggest that the history of the price and information is incorporated in the current price. Thus, it claims that the prediction of the stock market based on price data (technical analysis) is not possible since the stocks are valued by following a random walk in which successive changes are non-correlated. The semi-strong EMH claims that the stock price does not only depend on historical price data but also currently available public information (Schumaker & Chen, 2006). This includes additional trading information such as volatility, and fundamental data such as profit prognoses and earning reports. The strong form goes another step further by extending the historical public information by private information, such as insider information in the stock price (Falinouss, 2007).

The weak and semi-strong approach of the EMH have been supported by several studies (Low & Webb, 1991; White, 1988). However, more recent studies proved that the EMH should be rejected since models were able to predict the stock market (Ican & Celik, 2017). Even Fama (1991) the introducer of the EMH, rejected his own theory in the early 90s.

2.1.2 Random Walk Theory (RWT)

The RWT has a different view on predicting the stock market, it argues that outwitting the stock market is infeasible since the stocks are valued by random. Contrary to the EMH, this theory claims that even with all publicly available-historical and real-time information it is impossible to confidently predict the stock values (Sewell, 2011).

2.2 Approaches to Stock Market Prediction

Even though the two previous discussed theories argue that predicting the stock market prices is not feasible, many researchers are taking on this challenge because of the increased availability and accessibility of data. Additionally, new technologies such as artificial intelligence and data mining have become available (Ican & Celik, 2017). The three main approaches or philosophies applied by researchers to predict the stock market movements are: (1) a technical approach, (2) a fundamental approach, and (3) a combination of these techniques. In the following sections the technical as well as the fundamental approach will be described in detail.

2.2.1 Technical Analysis

A technical approach contains a prediction of the stock price based on historical stock features such as the volatility and price. Data mining techniques are used to discover repetitive trends of the technical indicators in the financial time series data (Bohn, 2017). There are several well-established techniques for technical analysis. The main characteristics which form the bases of technical analyses are:

- *Stock market prices reflect everything.* According to Murphy (1999) and the EMH, technical analysis assumes that all information of a stock is always incorporated in its price.

- *Stock market prices have a repetitive pattern.* The key concept here is that if a trend is already underway, the probability of continuation is higher than of reversal. This presumption is crucial because technical analyst will not be profitable without it.

- *The patterns of stock price fluctuations do not change.* Murphy (1999) claims that technical analysis is primarily based on human psychology, which he believes never changes over time. Thus, technical analysts can use the same historical patterns to predict a new bear or bull market. Hence, the best times to buy or sell a stock can be forecasted based on historical data.

The methods used by technical stock market analysts to identify patterns include technical factors such as confirmation and divergence, price patterns, resistance levels, trendlines, and moving averages. Through the years, many technical stock price prediction methods have been developed and are still established on the grounds of these fundamental principles (Hellmstrom & Holmstrom, 1998). To conclude, technical analysts strongly believe that market timing is crucial, and opportunities can be discovered by analyzing graphs. They consider that price fluctuation depends on recent historical (this is a relative term since there are different time windows at which trends occur) price movement patterns (Schumaker & Chen, 2006).

2.2.2 Fundamental Analysis

Another stock market prediction philosophy is fundamental analysis, this approach suggests that historical price data is insignificant for predicting the stock price. However, it argues that the stocks are valued as a reflection of the economic and political situation (Vargas, Lima, & Evsukoff, 2017). Fundamental analysts are trying to determine a company its true value. They include data such as the organizations’ debt, absenteeism, customers satisfaction, recent acquisitions, and the overall economic and political situation. Because news describes some of these fundamental factors, they believe that it can influence the supply and demand of stocks. Therefore, it is important for investors using fundamental analysis, to pay attention to news articles. Of course, they have the same goal as technical analysts, but they argue that they can act before information is reflected in the stock price. Thus, trading opportunities are created when there is a lag between an occurrence and its market reaction.

The use of news articles and other textual data for predicting the stock market already proved some success while forecasting the stock market (Ican & Celik, 2017). The share price of a stock may be significantly affected by information about a company's report or breaking news reports. Many researchers have examined the impact of news stories on the stock market and the stock market's response to press releases. In general, studies show that the stock market responds to news, and previous studies' findings suggest that news stories influence stock market movement (Ican & Celik, 2017). Studies that applied fundamental or technical analysis to predict the stock market are explained in the Section 2.4. Furthermore, a detailed overview about news sentiment analysis methods is described in Section 3.2.

2.3 Regression vs. Classification

When attempting to forecast the stock market, one of the first and most important decisions is to determine which form of machine or deep learning algorithm will be applied. One possible approach is to forecast the directionality of the price movement (up/down). Since there are two possible outcomes, this is known as a binary classification problem. When it is predicted that the stock market will go upwards or does not change, the positive label will be given, while the negative label represents a downwards movement prediction. The binary classification problem approach is commonly applied by many researchers (de Fortuny, de Smedt, Martens, & Daelemans, 2014; Hagenau, Liebmann, & Neumann, 2013; Zhai, Hsu, & Halgamuge, 2007). To face the classification problem in greater depth, more possible outcomes should be defined. These possible outcomes can be categorized by directionality and/or intensity such as small, normal, and big up or downwards movements (Gunduz & Cataltepe, 2015; Mittermayer, 2004; Rachlin, Last, Alberg, & Kandel, 2007).

Another perspective to face the prediction of the stock market is to use a regression learning algorithm which can predict exact values of the stock price. Schumaker and Chen (2009) developed a model which achieved a good performance for prediction the stock market with regression models. Predominantly, the direction (up/down) of the price movement is more valuable than the exact value, so the regression approach is rarely chosen (Serebryannikova, 2018). Additionally, assessing and evaluating the model's performance is more difficult, since it requires the use of multiple evaluation metrics at the same time (e.g., closeness and directional accuracy). Alternately, the coefficient of determination can be used. However, since this score rises evenly proportional with the number of explanatory variables, it is difficult to compare models with different numbers of independent variables (Bohn, 2017). Since this study is mainly focused on the impact of news articles on the stock market movements, instead of exact price prediction, binary classification has been chosen. In addition, the performance of the models can easily be compared with existing models developed in prior research.

2.4 Existing Studies: Models and Methods

This section is devoted to review existing studies which applied deep or machine learning to predict the stock market. Since each research made use of different datasets and are mostly trying to solve different stock market prediction tasks, it is nearly possible to compare the performance of the models directly among other studies. Nonetheless, one thing is certain: forecasting stock market patterns from news stories is a well-known machine learning challenge and it seems to be a difficult one. Most analysts are already satisfied when their developed model outperforms random guessing. However, even when the performance is slightly better than random guessing, the outcome should not be considered unsatisfactory since even minor improvements in stock market forecasting will result in growing portfolio value (Serebryannikova, 2018). Another important notable assumption is that researchers who achieved a low performance on their model may not always publish their work. Also, researchers who were able to predict the stock market prices with high certainty may not always publish their work, because they could take advantage of it by making much profit on the stock market exchanges.

In general, existing studies developed models that achieved an accuracy in the range of 60-70% (Serebryannikova, 2018). The dataset that will be used in the current research is never used for the prediction of the stock market before and is described in Chapter 3. However, two existing studies used a comparable dataset with relatively "old" news articles from Bloomberg and Reuters. Ding, Zhang, Liu, and Duan (2015) used this dataset and developed a convolutional neural network which achieved an accuracy of 64,21% when forecasting the daily directionality of the price movement of the S&P 500 index. Serebryannikova (2018) also tried to develop a well-performing model, unfortunately he achieved an accuracy of only 51.95%. Another recent comparable study has been done by Attanasio, Cagliero, Garza, and Baralis (2019), they applied the combination of technical and fundamental analyses. After crawling the news articles from Reuters, sentiment analysis methods were executed to quantify the positivity or negativity of a news article. Thereafter, the stock prices and volatility indices of the S&P 500 index were gathered at Yahoo Finance and merged to the news articles based on the common column date. This dataset has been used to compute technical features and train a Support Vector Machines (SVM), bayesian classifier, an Artificial Neural Network (ANN), an ensemble method, and a distance-based classifier on. The researchers claim that their SVM model outperforms all the existing models that combine fundamental and technical analysis.

Even though most other studies in the field of stock market forecasting are not directly comparable to this study, they are still being studied to determine which models and methods will be appropriate for the research to predict the S&P 500 index based on news sentiment analysis. Hu, Liu, Bian, Liu, and Liu (2018) have performed a news-oriented sentiment analysis to predict Chinese stock prices and claim that their framework results in a significant improvement of the accuracy of stock trend prediction. However, they recommend future researchers to improve the model performance by combining their fundamental analysis with a technical analysis.

As the researchers suggested in the previous paragraphs, another philosophy for predicting the stock market is a hybrid method of technical and fundamental analysis. Khedr and Yaseen (2017) applied a combination of technical and fundamental analysis where the fundamental part of the research was a news sentiment review, and the technical part contained an analysis of the historical features of the stock prices. They state that their model achieved an accuracy of 89.80% and that their research outperformed former studies. However, such outcomes are uncommon, and they are mostly dependent on the applied dataset. In this case it is probably due to only three companies that are included and a narrow timeframe of the dataset.

As from the previous paragraphs, it has been concluded that the combined approach of using both fundamental and technical analysis seems to perform best in predicting the stock market through news sentiment analyses and stock market features, extracted from historical prices. Hence, the current research will apply this method through the use of news sentiment polarity in combination with technical features, this is described in more detail in Chapter 4.

It seems that the SVM algorithm is mostly applied to predict the stock market movements (Luss & d'Aspremont, 2015; Schumaker & Chen, 2009). Choudhry and Garg (2008) and Reddy

(2018) also developed a SVM which was able to predict the stock market. Both studies argue that the SVM outperformed the baseline models. Decision Tree (DT) based algorithms are also commonly applied to face the classification of the stock market movements (Rachlin et al., 2007). In addition, an ANN is often used to solve classification task problems (Ican & Celik, 2017). To conclude, the current research will compare an SVM, DT-based models, and an ANN.

Furthermore, this study will compare different lexical-based approaches for sentiment analyses because of the following reasons: (1) according to previous studies, machine learning techniques have no significant contribution to news sentiment analysis compared with lexical approaches (Hutto & Gilbert, 2014), (2) these frameworks are state-of-the-art, well-performing and implementing them is computationally efficient (M. Silva, Giovanini, Fernandes, Oliveira &, C.S. Silva, 2020), and (3) previous studies which considered lexical-approaches mostly implemented various sorts and could not conclude which method fits the best in which case (Prajapati, 2020). There are several lexicons for sentiment analysis; previous studies mostly used one of the following lexicons: Valence Aware Dictionary and sEntiment Reasoner (VADER), TextBlob, Flair, SentiStrength, Loughran & McDonald (LM), and Hu & Liu (Canbaz, 2020). VADER was used as sentiment analysis in many researches such as studies by Kim et al. (2016), Steinert and Herff (2018), and Valencia, Gómez-Espinosa, and Valdés-Aguirre (2019). Prajapati (2020) also applied VADER while he studied the Bitcoin price prediction through social sentiments in news articles and Reddit messages. However, this study has compared VADER against Flair and Textblob. It has been concluded that in this case, there was no significant difference between these techniques. The LM lexicon was applied to meeting transcripts by Shapiro, Sudhof and Wilson (2017). Another study by Wang et al. (2014) also used LM to collect the sentiment of SeekingAlpha news articles and use it to predict the stock values. Matta, Lunese, and Marchesi (2015) used Twitter data in their study and gathered the sentiment scores while applying the SentiStrength model. During this study they were interest in opinions, SentiStrenght has been established for opinion mining. Hence, this sentiment analysis method is excluded because this study is only interest in subjective sentiment polarity. Lee, Hu and Lu (2018) uses the Hu & Liu lexicon approach for evaluating reviews of Tripadvisor a travel company. This sentiment method is also excluded because it tries to gather opinions instead of facts. As described above, there are many different sentiment analysis lexicon methods and there is no best practice. Therefore, the current research will compare three lexicons: VADER, TextBlob, and LM. The flair method is not included because the researcher's computer was not able to execute this algorithm on the dataset.

3 Methods

3.1 Algorithms

The previous section has explained which machine and deep learning models are selected to be applied to the current research to discover whether news article sentiment can be used to predict the daily S&P 500 index movements. This section will explain each of these chosen algorithms in more detail. The first subsection contains the description of a decision tree. Afterwards, other algorithms that are based on decision trees are explained. Finally, the SVM and an ANN will be illustrated.

3.1.1 Decision Tree

Decision trees can solve both regression and classification problems. The current research will focus on classification trees because the main purpose is to classify whether the stock price goes up or down the next day. Ahlin and Randby (2019) state that these models use the characteristics of an instance to classify to which binary class it belongs. A DT is an overview of features from an item that can be used to determine the target value in data mining and machine learning (Zhang, Zhou, Leung & Zheng, 2010). Decision tree-based models are commonly used techniques for machine learning classification problems because a decision tree is easy to explain and understand for people. Furthermore, the development of DTs is a manageable task, and these models perform well in general (Esmaily et al., 2018).

Classification trees are built in a top-down structure using the recursive binary-splitting method for the features. There are several methods used for binary splitting, one of these is the classification error rate. In order to calculate this error rate, the fraction of training entries in the under-represented class node is used. Equation 1 can be used to measure the classification error rate (E), where $\hat{p}_{mk}$ stands for the number of training observations in the m:th region from the k:th class.

$$E = 1 - max(\hat{p}_{mk}) \tag{1}$$

The classification error rate appears to be insufficiently sensitive for growing a decision tree. Therefore, two other calculations are preferred in practice; the Gini index and cross-entropy (Ahlin & Randby, 2019). The Gini Index, also known as Gini impurity, measures the likelihood of a certain prediction being wrongly labeled when chosen at random. Thus, it is a metric for determining the purity of a node. The index (G) would be high if the node contained an under-represented class. The formula for calculating the total variance for K classes is defined in Equation 2.

$$G = \sum_{K=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk}) \tag{2}$$

The Gini index can be used to evaluate the consistency of a class distribution by indicating when a group is underrepresented. The cross-entropy method, which evaluates the purity of the nodes, is another way to test the split quality of the classes. The cross-entropy (D) formula is defined in Equation 3.

$$D = -\sum_{K=1}^{K} \hat{p}_{mk}\, log(\hat{p}_{mk}) \tag{3}$$

3.1.2 Bagging and Boosting

A commonly problem with decision trees is high variance, this implies that if the training dataset is divided at random into a number of subsets, the algorithm will achieve quite different scores on each subset when assessing the model performance (Ahlin & Randby, 2019). Bagging is one of the several bootstrap-aggregation approaches which can solve the high-variance problem. This method will improve the ML models' stability and classification accuracy for both regression and classification models. It also helps avoid overfitting and reduces variance (DT problem) (Zhang et al., 2010). It can be used with various models, including the ANN (West, Dellana, & Qian, 2005), but it is most commonly used with DTs (Dietterich, 2000). By sampling observers randomly and with substitution, bagging divides the training collection into new sample subsets. As a result, certain observers may appear in several subsets. So, bagging uses the subsets to produce a number of new models. The final output will be obtained through voting in case the tasks is to solve a classification problem (Zhang et al., 2010).

Boosting is an approach which is quite similar to bagging, but it grows the trees in a sequential manner, with the trees learning from the trees that are generated before them. Hence, this algorithm is quite a slow learning one compared to the bagging method. Furthermore, instead of fitting to the target variable, the boosting approach is fitted to the model residuals. As a result, the predictions improve when the model performance is lacking, and the trees are relatively small. The Gradient Boosting (GB), Light Gradient Boosting (LGB), and Extreme Gradient Boosting (XGB) classifiers use this boosting strategy. Each of these algorithms will be discussed in the following sections.

3.1.3 Random Forest (RF)

Breiman (2001) developed a method for improving the accuracy of a DT known as an RF. A RF consists of several DTs. When compared to a DT, the benefits of an RF are improved accuracy and less overfitting. The RF is built up based on DTs and the bagging method, as discussed in the previous subsections. The difference between an RF and DT is that an RF uses subsets of the predictors to set up the splitting of the trees as efficiently as possible. The Gini index (Equation 2) or the cross-entropy (Equation 3) may be used for this.

3.1.4 Gradient Boosting

Gradient boosting is a machine learning strategy that employs both gradient descent and boosting. Gradient boosting, like the RF, makes use of a collection of DTs. These DTs are generated one at a time and fitted to correct the prediction errors produced by previous models. Boosting is a form of ensemble machine learning that is used in this learning technique. As the model is being fitted, the loss gradient is minimized using gradient descent. The most time-consuming aspect of designing a GB classifier is determining the best DT splitting points.

3.1.5 Extreme Gradient Boosting

Extreme gradient boosting uses GB combined with regularization, in which regularization prevents overfitting, and increases the algorithm's accuracy (Chen & Guestrin, 2016). Thus, XGB is a more advanced version of GB. To have a scalable, compact, and accurate library, this algorithm takes time and pushes the computation of the algorithms to their limits.

3.1.6 Light Gradient Boosting

Light gradient boosting is an open-source distributed gradient boosting platform created by Microsoft in 2016 (Ke et al., 2017). Because of the improvements in the training algorithm, this new architecture takes less time than GB and XGB. Furthermore, compared to GB or XGB algorithm, the gradient boosting approach has mostly achieved a better accuracy. Due to its high prediction accuracy, the LGB classifier has been commonly used in various fields (Fan et al., 2019).

3.1.7 Artificial Neural Network

An ANN consists of several neurons which are connected through several layers. These layers include an input layer, a collection of hidden layers, and an output layer (Wang, 2003). Figure 1 depicts this architecture, in which the lines connect the neurons and are associated with weights.

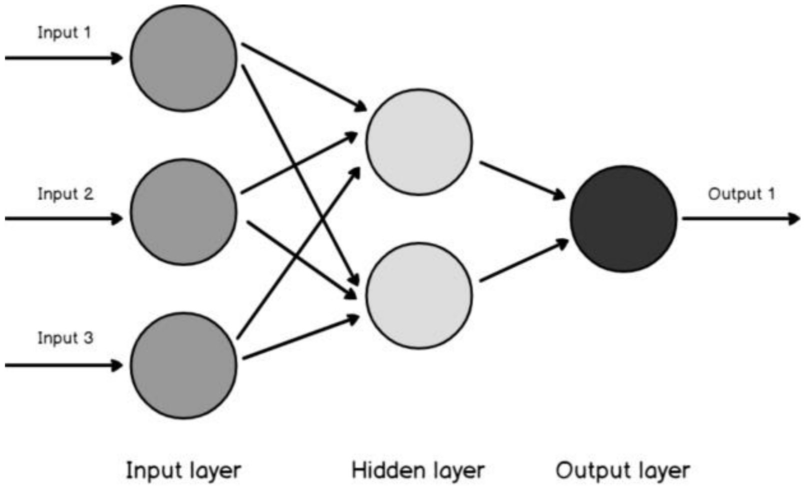

Figure 1 - Structure of an artificial neural network

In Equation 4 de formula for calculating the output (h_i) of a neuron (i) is defined.

$$h_i = \sigma\left(\sum_{j=1}^{n} V_{ij}\, x_j + T_i^{hid}\right) \tag{4}$$

where σ() represents activation function, N the number of input neurons, V_{ij} the weights, x_j the inputs to the input neurons, and T_i^{hid} the threshold terms of the hidden neurons. According to Wang (2003), the activation mechanism introduces non-linearity into the model and limits the importance of the neuron so that divergent neurons do not paralyze the neural network. There are several activation functions, a well-known activation function is the sigmoid (or logistic) function which is defined in Equation 5.

$$\sigma(u) = \frac{1}{1 + exp(-u)} \tag{5}$$

Any computable function can be approximated by an ANN, which is illustrated in Figure 1. The independent variables are the numbers given to the input neurons, while the dependent variables are the numbers predicted by the neural network. An ANN's inputs and outputs may be binary, for instance, 1 indicating an upwards movement and 0 indicating downwards movement. In addition, images may also be used as input neurons, for example, to determine if a human or an animal is portrayed. To conclude, an ANN can be used to solve a wide range of problems.

3.1.8 Support Vector Machines

In order to solve classification and regression tasks, an SVM can be used as well (Parrilla Gutiérrez, 2010). According to Vapnik (1998), who is also the creator of this method, the SVM is a more sophisticated version of the ANN since it focuses on fundamental mathematics. According to recent studies, the SVM is a non-linear binary classifier that is commonly used, and it can be represented geometrically (Noble, 2006). Meyer and Wien (2015) argue that SVM makes predictions based on separating a hyperplane by optimizing the margin of the two nearest data points of two separate groups. This optimal hyperplane maximizes the generalization of the model and tackles the problem of high dimensionality (Inoue & Abe, 2001). An SVM can be structured in several ways (Trafalis & Gilbert, 2006, Zhu & Hastie, 2005), but all the structures share that they maximize margin and applied the kernel method.

3.2 Sentiment Analyses Methods

As discussed in Section 2.4, there are various methods to determine the sentiment of a news article. It has been decided to focus on some of them, namely, VADER, TextBlob, and LM. Each method will be further elaborated below.

3.2.1 VADER

Hutto and Gilbert (2014) established the VADER method for sentiment analysis as an open-source tool. This widely used approach is mainly applied within the social media domain, where it assigns polarity to social media content (positive/negative). This polarity is calculated through a Lexicon and rule-based approach, which makes use of features such as punctuation, capitalization, and adverbs such as extremely, quite, slowly. VADER computes a positive, negative score, normalized, weighted, and compound score.

3.2.2 TextBlob

TextBlob is capable of a wide range of Natural Language Processing (NLP) tasks such as noun phrase extraction, part-of-speech tagging, sentiment analyses, tokenization, and spelling correction (Loria, 2018). During the current study, the sentiment analysis feature of TextBlob has been used. This analysis tool uses a sentiment lexicon and a pattern (English) sentiment analysis engine (Sohangir, Petty, & Wang, 2018). When TextBlob is executed on a news article, it assigns a polarity and subjectivity score.

3.2.3 Loughram & McDonald

Loughran and McDonald (2011) created a finance-specific word list based on many samples of SEC 10-K filings, arguing that although the existing lexicons considers certain terms to be negative, they are neutral or even positive when used in the finance domain. This study discovered that in the commonly used Harvard lexicon, almost 75% of the word identified as negative should not be considered negative in the financial domain. In finance, for example, the word "liability" is typically neutral but negative in everyday language. Therefore, the LM lexicon is now widely used in the financial sector for sentiment analysis (Canbaz, 2020).

4 Experimental Setup

This chapter will address the experimental study setup. First the dataset and its pre-processing will be described. Subsequently, the setup of the different machine and deep learning models, sentiment analysis methods, and datasets will be explained. Finally, the evaluation metrics to compare the models will be addressed.

4.1 Data

4.1.1 Dataset Descriptions

The dataset used for this study is a merger of two datasets from different sources. One dataset contains the news articles that will be used to assess the predictions power of daily news sentiment. The other dataset consists of the S&P 500 prices. Each of these datasets is described in detail in the following sections.

News articles
The news article dataset is created by Andrew Thompson on March 4, 2020[1]. He is part of Components, a research group that assembles, investigates, and editorializes large datasets. The news article dataset contains 2.7 million news articles and essays from 27 American publishers such as the Financial Times, Reuters and Cable News Network (CNN) from 2013 to early 2020. The dataset columns include the date, title, publication, article text, section, year, month, and URL (for some). As described in the previous chapters, this dataset is unique since it contains news articles from plenty of different news sources, and it contains recent news compared to existing studies. Therefore, this dataset has been chosen to be applied to the research to predict the daily directionality of the S&P 500 price movement. Furthermore, creating a new dataset by scraping the websites that contain news articles is only accessible for last year's publications. If more historical news articles were to be gathered, the proposed research would require a larger budget. Therefore, the recently created reliable dataset gathered by Andrew Thompson will be used during the research.

S&P 500 prices
The S&P 500 index price dataset is generated using the YahooFinancials package in Python to gather the S&P 500 prices on Yahoo Finance (finance.yahoo.nl). This dataset will be used during the research because the price data is reliable and the YahooFinancials package enables scraping the price data. The dataset consists of 1,782 rows and 5 columns. These columns include the high, low, open, and close price, volume, and formatted_date fields. The 1,782 rows are unique by the date column and has a time span 01/01/2013 till 01/31/2020.

4.1.2 Data Preprocessing

After collecting the news article and S&P 500 index datasets, each dataset will be pre-processed and afterwards the datasets will be merged. The pre-processing steps will be discussed in this subsection, also the merger of the two datasets will be addressed.

News articles
First, the columns are filtered so that only the meaningful variables will be in the dataset. These include the date, title, section, and publication. The researcher has chosen to filter out the article

[1] https://components.one/datasets/all-the-news-2-news-articles-dataset/

text because his computer was not able to execute any of the sentiment analysis methods on the whole article text. Therefore, the title of the article will be considered to represent the news article sentiment. Since the sentiment of the news articles depends on the title, all the rows without a title are filtered to assure that each entry has a filled title column.

Some publishers do not have a filled section column; these entries are replaced with "unknown". However, this study is mainly focused on financial and business news articles and with an "unknown" section column it is impossible to decide whether it belongs to a financial and business article or not. Therefore, this study uses two news article datasets, (1) with all the news articles with a financial or business-related section, and (2) all the news articles with unknown sections combined with dataset 1.

News articles sentiment scores

After preprocessing the news article dataset, some code should be executed to gather the different sentiment score features. As mentioned in Section 3.2 the following sentiment scores are calculated and added to the two news articles datasets: VADER, TextBlob, and LM. Hence, each row (news article) has several sentiment score columns. Since the news article and price datasets will be merged on the date columns, a group by date function has been executed. While grouping the rows by the date column, the sentiment scores will be averaged.

S&P 500 prices

The S&P 500 dataset is filtered so that only the following meaningful columns are included in the dataset: close price, volume, and date. Another column has been added with the movement, this is de binary dependent variable in this study. A 1 is assigned if the stock market value went up or remained the same compared to the day before and 0 if the stock market values went down. Since it has been decided to use fundamental as well as technical analysis, a column has been added with a technical indicator. Although there are many technical indicators which can be calculated with the features in the S&P 500 dataset, the most common one has been chosen, which is the simple moving average (SMA) (Nti, Adekoya, & Weyori, 2019). To calculate the SMA, all the closing prices are summed up over a given period and divided by the number of periods (Thirunavukarasu & Maheswari, 2014). So, if 100 days are considered as the time period, the SMA is the average of the last 100 closing prices. The formula for this calculation is defined in Equation 6 below, A_n stands for the price of an asset at period *n*, whereas n represents the number of total periods.

$$SMA = \frac{A_1 + A_2 + \ldots. + A_n}{n} \tag{6}$$

Since the prediction of a stock for the following day is considered as a short-term prediction, the period should is set on 10 days (Lauren & Harlili, 2014).

Final datasets

After the datasets are preprocessed, the price dataset has been left joined to the news article datasets to create two datasets that will be used as input and output for the machine and deep learning models. The datasets are merged on the column date. Thus, the two dataset which will be used for this research are fully preprocessed and both contains 1,096 rows. Because both datasets consist of the same features only one table has been created to give an overview of all the features in the datasets. Table 1 on the next page has a compact overview of all the features in the dataset.

#	Features	Types	Input descriptions
1	Date	date	Date
IV1	Close	float 64	Close price of S&P 500
IV2	Volume	float 64	Volume of S&P 500
IV3	SMA	float 64	SMA (with $n = 7$) of S&P 500
IV4	LM score	float 64	Average LM score of news articles
IV5	VADER score	float 64	Average VADER score of news articles
IV6	TextBlob score	float 64	Average TextBlob score of news articles
DEP	Movement	integer	1 if price went up or remained the same compared to the previous day, and 0 if the price went down.

Table 1 - Features of the final dataset

The dependent variables have 589 entries that the stock market movement directionality went down, and 480 entries where the stock market price went up or remained the same. Hence, there is some imbalance in the dataset. This is not a problem because the evaluation metrics will be chosen based on this imbalance in the dataset, this is in detail described in Section 4.3.

One small adjustment has been made to the final datasets, which is known as standardization. This method may help to minimize dataset dissimilarities. Rescaling the features to give them the characteristics of a regular normal distribution is known as standardization (or Z-score normalization) (Mohamad & Usman, 2013). This technique is primarily used in the development of machine learning models and in statistics when comparing measurements with data dissimilarities. Equation 7 is defined below to standardize each feature ($X_{:,i}$):

$$z'_{:,i} = \frac{X_{:,i} - mean\ (X_{:,i})}{std\ (S_{:,i})} \tag{7}$$

4.2 Data Modeling

4.2.1 Training, Validation, and Test Set

For training and evaluating a machine or deep learning model, it is necessary to split the dataset. The dataset for this study has been divided into two parts, 80% for training and 20% for testing the models. Hence, the training data contains 855 observations, and the test data consists of 214 observations. The training data is used to optimize the model, and the test data is used to assess the performance independently. Xia, Broadhurst, Wilson, and Wishart (2013) state that only the test data can be used to determine a model's true effectiveness. Both the training and test data should be a good reflection of the entire data set. If the samples are not representative, the test prediction would be subject to sampling bias (Mendez, Reinke & Broadhurst, 2019). Real-world datasets should use a stratified 10-fold cross-validation (Kohavi, 1995). The cross-validation method ensures that the training and test sets include the same proportions of both target groups, which are in this study the upwards and downwards movement of the S&P 500 index price. Cross-validation guarantees the model's validity and reliability; see Section 4.3 for a more comprehensive explanation. The validation set was used as part of a grid search, covered in more detail in Subsection 4.2.3.

4.2.2 Data Models

Section 2.4 elaborates on several machine and deep learning models. It has been concluded that some algorithms fit the classification of the stock market movement best. These algorithms are already described in the previous chapter and include the RF, GB, LGB, XGB, ANN, and SVM. Each of these algorithms was used to create models, and a grid search was applied to find the best model settings, as described in the following subsection. Subsequently, the models were evaluated, and the best performing model for predicting daily S&P 500 movements was selected.

4.2.3 Hyperparameters Optimization

While running a supervised machine or deep learning algorithm, some hyperparameters should be set. These hyperparameters can be the default values in the model's library, manually configured or automatically configured. The latter two generally yield a boost in model performance (Feurer & Hutter, 2019). In order to search for the best hyperparameters, the manual or automated method should be used to set the values. Since it is very inefficient to manually tune hyperparameters (Probst, Bischl, & Boulesteix, 2018), automated hyperparameter optimization (HPO) has been used in this study. Grid search, also known as full factorial design, is a commonly used HPO tool which automatically finds the best hyperparameters for an algorithm (Montgomery, 2017). During the development phase of this study, the grid search has been deployed in a pipeline. A pipeline is an iterative method where all the models and different datasets are coded, and once it runs, it generates outputs with all the different models and dataset combinations and their performance. In Chapter 5, the best-performing model's and its best hyperparameters are shown.

4.3 Model Evaluation

While developing a model that classifies whether the price will go up or down the next day, two fundamental errors may occur. A False Positive (FP) occurs when it is predicted that the stock market goes up, but in fact went down. Conversely, if it is predicted that the stock market goes down, but it went up, it is called a False Negative (FN). Naturally, this research tried to optimize the models to predict as most as possible outcomes right, the so-called true predictions. A True Positive (TP) occurs when the stock market is predicted to go up and it indeed went up. Contrary, a True Negative (TN) if it is forecasted that the stock market goes down, and it went down. These classification results are often represented in a confusion matrix, such as the one shown in Appendix I (Jiao & Du, 2016).

While assessing the performance of the developed classification models, the accuracy or the F1-score are widely considered by researchers as evaluation metrics (Chicco & Jurman, 2020; Tharwat, 2020). Nevertheless, both metrics may yield unreliable scores on class imbalanced datasets because they neglect the ratio of the positive and negative in the dataset. An example may be if a dataset is non-balanced all the entries can be predicted as the over-balanced class in order to achieve a high accuracy. A more suitable metric for class-imbalanced datasets is the so-called area under the curve of the receiver operation characteristics (AUC-ROC) score (Haixiang et al., 2017). By taking both the TPs and FPs into account, the AUC-ROC score decreases the negative effect of a class-imbalanced dataset. This AUC-ROC score baseline is always 0.5, indicating that the TPs and FPs are distributed evenly.

Although the AUC-ROC score could handle a class-imbalanced dataset it has one problem, namely it pays little attention to FPs and mainly considers the positive (up) class (Tharwat, 2020). The area under the curve of precision and recall (AUC-PR) score is a metric which can solve this problem by taking all the TPs, FPs, and TNs into account and calculates the most optimal precision/recall trade-off (Siblini, Bruce & Patel, 2020; Tharwat, 2020). Even though this method is less popular than the AUC-ROC score, it has mostly been prioritized by data scientists. They argue that the AUC-PR score is a better metric, especially when the dataset is unbalanced among the different classes (Chicco, 2017; Ozenne, Subtil, & Maucort-Boulch, 2015). For testing binary imbalanced datasets, Chicco and Jurman (2020) recommend that data scientists should consider both the AUC-PR and the AUC-ROC score, with an emphasis on the former. To calculate the AUC-PR base score, all the TPs must be divided by the total number of entries. Thus, the AUC-PR baseline score in this study is, the number of positives, which is 480, divided by the number of total entries, which is 1069, that resulting in around 0.45.

In addition, to assess the models' AUC-PR and AUC-ROC score on the test set, a 10-fold stratified cross-validation will be integrated in the pipeline of the models. The main advantages of this cross-validation technique are that the generalization can be assessed, and it prevents overfitting (M. Santos, Soares, Abreu, Araujo & J. Santos, 2018). Stratified cross-validation assures that each validation subset consists of the same ratio of classes as the training subset it belongs to. According to Kohavi (1995), Stratified cross-validation decreases bias and variance and outperforms traditional cross-validation techniques. The AUC-ROC and AUC-PR metrics will be compared with the baseline scores in order to assess the models' performance. Finally, the accuracy of the best performing model will be gathered in order to compare the model to existing models from other studies which are described in Section 2.4. The formula to calculate the accuracy *(Acc)* is defined in Equation 8 below.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

4.4 Setup Overview

The pre-processing of the date as well as the development of the models has been done in Python. To provide a short summary of the experimental setup of this study, Table 2 below has been made. It shows the different variables that are used in several combinations to find the best performing one. Technical Analysis (TA) represents the volume, close price and the SMA.

Datasets	Models	Feature sets
Dataset 1 - Financial/business news articles	RF	LM score
Dataset 2 - All news articles	GB	TextBlob score
	LGB	VADER score
	XGB	TA
	ANN	LM score + TA
	SVM	TextBlob score + TA
		VADER score + TA

Table 2 - Variables to discover the best performing combination

In total, there are 84 (2 datasets x 6 models x 7 feature sets) combinations of these variables. Since each combination achieved the AUC-PR as well as the AUC-ROC score a total of 168 scores are gathered and discussed in the following chapter.

5 Results

This chapter will present and compare the results of the different machine and deep learning models, news sentiment analysis methods, and the two datasets. While comparing these combinations of methods, models, and datasets, many scores are collected. To provide an adequate overview of the results, this chapter will only present the most important scores. For a more detailed overview of all the results see Appendix II and III.

5.1 Algorithms

A total of seven different machine and deep learning algorithms have been used to develop models to predict the daily stock market movements. The SVM and ANN algorithm outperformed all the other algorithms in terms of the AUC-PR and AUC-ROC score. Therefore, the following sections will only present the scores of the SVM and ANN while comparing the datasets, news sentiment analysis methods, and models. Again, for more detail on the AUC-PR and AUC-ROC scores of the other models, see Appendix II and III.

5.2 Dataset

This study used two news article datasets, (1) with all the news articles with a financial or business-related section, and (2) all the news articles with unknown sections combined with dataset 1. In order to discover which dataset is most suitable for forecasting the S&P 500 index movements, the performance has been assessed. A detailed overview of the AUC-PR score for the SVM and ANN classifier on each dataset is depicted in Table 3 below.

Independent variables	**Dataset 1 - Financial/business news articles**		**Dataset 2 - All news articles**	
	ANN	SVM	ANN	SVM
LM	0.54	**0.56**	0.54	0.54
TextBlob	0.59	**0.63**	0.55	0.38
VADER	0.52	**0.61**	0.49	**0.61**
TA	0.63	**0.71**	0.61	0.70
LM + TA	0.65	**0.73**	0.66	**0.73**
TextBlob + TA	0.62	**0.75**	0.64	0.74
VADER + TA	0.56	**0.74**	0.61	**0.74**

Table 3 - The AUC-PR and AUC-ROC score of the models on the two datasets

When observing Table 3, the bold depicted scores are the highest of its row. With these results it can be concluded that dataset 1, containing only the news articles that have a business or financial related section, performed better than dataset 2. Moreover, dataset 1 outperformed dataset 2 on four of the seven different feature sets, and on three occasions, both datasets achieved the same score. Therefore, the following sections will only focus on the performance of the sentiment analysis method and models on dataset 1.

5.3 News Sentiment Analysis Method

As from the previous two sections, only one dataset and two algorithms remained. In order to discover which feature set or independent variables are the best input for the ANN and SVM, an overview of all the AUC-PR and AUC-ROC scores are shown in Table 4 below.

	AUC-PR		**AUC-ROC**	
Feature sets	ANN	SVM	ANN	SVM
LM	0.54	**0.56**	0.56	**0.59**
TextBlob	0.59	**0.63**	0.63	**0.67**
VADER	0.52	**0.61**	0.48	**0.60**
TA	0.63	**0.71**	0.70	**0.72**
LM + TA	0.65	**0.73**	0.69	**0.75**
TextBlob + TA	0.62	**0.75**	0.71	**0.76**
VADER + TA	0.56	**0.74**	0.60	**0.75**

Table 4 - The AUC-PR and AUC-ROC scores of the models on each feature set

Both the AUC-PR and AUC-ROC scores are highest when using the TextBlob score and TA (SMA, volume, closing price) as shown on the bold and highlight depicted cells in Table 4. In general, the TextBlob news sentiment analysis method also performed better on its own than when only the LM or VADER method was used. The next section zooms in on the performance of the models on the TextBlob and TA features only.

5.4 Model

As shown in Table 4, the SVM outperformed the ANN in terms of AUC-ROC and AUC-PR (bold depicted scores), no matter what independent variables were considered. Nevertheless, the SVM achieved the highest scores when the TextBlob and TA feature set was used as shown in the bold and highlight depicted cells. It achieved a performance of 0.75 AUC-PR and 0.76 AUC-ROC score. The leading metric during this study was the AUC-PR score, this trade-off between the precision and recall made by the SVM is depicted in Figure 2 below.

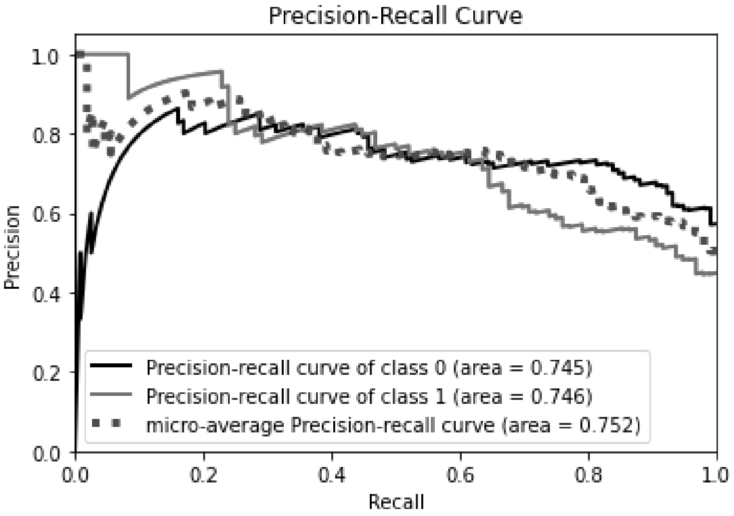

Figure 2 - ROC curve of the precision and recall of the SVM

The best performing model that made use of the SVM algorithm was optimized by a grid search to discover the best model settings. These best settings/hyperparameters that are revealed through a grid search are presented in Table 5 below.

Hyperparameter	Optimal value	Description
C	100	The C parameter trades off correct classification of training examples against maximization of the decision function's margin.
Gamma	0.01	The gamma parameter determines how far an entry could reach while it still influences the model.
Kernel	rbf	The kernel maps the data points into several groups.

Table 5 - Optimal hyperparameters for SVM and its description

In order to compare our model with existing studies the accuracy has been calculated. The corresponding confusion matrix of the SVM is depicted in Figure 3 below. The values inside the confusion matrix represents the TPs, FPs, TNs, and FNs, as explained in Appendix I. These values can be used to calculate the accuracy. The accuracy can be calculated by the sum of TPs and TNs, divided by the total number of observations. Thus, the accuracy of the SVM is (37 + 110) / (37 + 110 + 8 + 59) = 0.69.

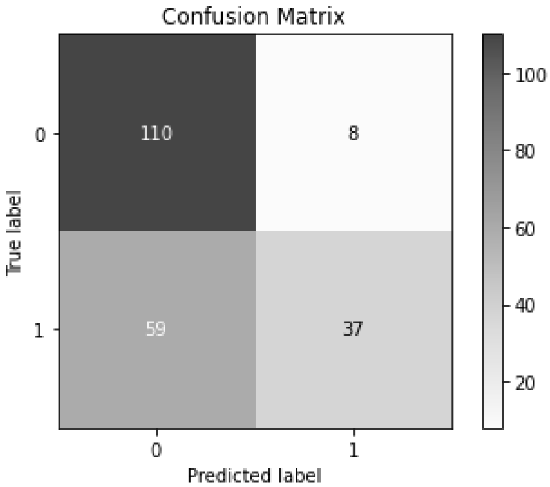

Figure 3 - Confusion matrix of the SVM

6 Discussion

The purpose of this study was to examine whether news articles sentiment could predict the S&P 500 index movement directionalities. The findings of the best dataset, news sentiment analysis method, feature set, and model will be discussed in this chapter. Furthermore, the results will be compared with the literature. Finally, the impact on the field of the stock market will be explained and the limitations and future research will be addressed.

6.1 Findings

The key findings of this research represent the best performing combination of the three variables for the classification task of the daily stock market movements. These variables include the dataset, news sentiment analysis method (feature set), and the model and its best hyperparameters. After choosing the best variables while evaluating the results of Chapter 5, the main research question will be answered. Lastly, the academic impact in the field of predicting the stock market using machine and deep learning will be discussed.

6.1.1 Dataset

During this study different machine and deep learning models were built and executed on two different datasets, where dataset (1) contained all the news articles with a financial or business-related section, and (2) all the news articles with unknown sections combined with dataset 1. Afterwards, the leading AUC-PR metric scores are compared, and it has shown that dataset 1 is most suitable for predicting the stock market movements. This conclusion is based on the discussed results in Section 5.2. Considering these results, it turned out news articles expressing positive or negative sentiment about a financial or business-related subject, for example the S&P500, have a greater impact on stock market movements than a news story about, for instance local entrepreneur. This may one assumed that investors are more interest in world business news than that local entrepreneur news article. Based on this conclusion, investors value financial or business world news more than a local news article about entrepreneurs.

6.1.2 News Sentiment Analysis Method

The previous two sections have enlightened which dataset and two algorithms are achieved the highest performance. Therefore, this section will only consider the dataset containing all the financial and business-related section, and the chosen ANN and SVM. In order to discover the most suitable feature set, that function as input for the ANN and SVM, an overview of all the AUC-PR and AUC-ROC scores is shown in Table 4 in Section 5.3. Both AUC-PR and AUC-ROC score are the highest while using the TextBlob score and TA (SMA, volume, close price). However, the TextBlob news sentiment analysis method also performed better on its own than when only the LM or VADER method was used. Comparing these results with the academic literature, turns out that Hu et al. (2018) their recommendation to use not only fundamental analysis but combine it with technical analysis to achieve a better performance has been proven. To conclude, the TextBlob score combined with the TA features seems to perform the best. Therefore, the following section zooms in on the performance of the models on these features.

6.1.3 Hyperparameters and Model

As described in Section 5.4, the SVM model achieved the best AUC-PR and AUC-ROC score. The hyperparameters for this SVM are found after deploying a grid search and these include: C - 100, gamma - 0.01, and the kernel - rbf. The SVM obtained an AUC-PR score of 0.75, whereas the baseline 0.45 was increased by 30%. Furthermore, it obtained an AUC-ROC score of 0.76, this means that the baseline of 0.50 is improved by 26%.

6.1.4 Findings for main research question

The previous three sections have described the best dataset, models, and methods. For predicting the daily directionality of the S&P500 index, the best performing combination considers only the news articles with a financial or business-related section, the TextBlob sentiment analysis method in combination with TA features (SMA, volume, close price), and the SVM model. Although this combination increased the baseline models around 30%, it cannot be concluded if news sentiment analysis can predict the stock market movements yet, because TA features are also considered, so it can be the case that only the TA features have predictive power. Therefore, it is necessary to look in Table 4, for the scores achieved while only the TextBlob score served as input feature for the machine and deep learning models. The table shows in that in case of only the TextBlob classifier is considered, the SVM achieved an AUC-PR score of 0.63, which improved the baseline score by 18% and 0.67 AUC-ROC score, where the baseline is increased by 17%. Even though, news sentiment in combination with TA achieved a better performance, it can be assumed that news sentiment on its own hold some predictive power regarding the S&P 500 index movements. Based on this conclusion the findings of Ican and Celik (2017) that the stock market can be predicted based on news articles sentiment should be confirmed.

6.1.5 Impact on this field

This research has shown that news sentiment analysis in combination with TA can predict the S&P 500 index movements with around 30% more certainty than random guessing. Because the S&P 500 index represents the 500 largest publicly traded companies in the U.S, it may be assumed that the whole stock market can be predicted with the methods used during this study. Although many researchers already argued that the stock market can be predicted with news sentiment with an accuracy in range of 60-70%, this study has approved this theory with an accuracy of 69% (see Section 5.4) on a different and newer dataset. In addition, the research found that the combination of fundamental and technical analysis performed the best. Furthermore, based on the findings of this research, the EMH and RWT theories, which are described in Chapter 2, should be rejected.

6.2 Limitations and Further Research

The current research has a valuable academic contribution towards the predicting of the stock market based on fundamental analysis (news article sentiment) and technical analysis (SMA, volume, close price). However, this research had some limitations that are relevant to address in this section.

The dataset had its limitations because it was unknown how the dataset was exactly gathered by Andrew Thompson, who is working for Components, a research group that assembles, investigates, and editorializes large datasets. Although the researcher has found that this is a reliable and well-performing organization, the exact pre-processing steps, such as several filters, that may have been used, are unknown. Therefore, it is almost impossible to recreate a whole new dataset to use the suggested methods in order to increase trading profits. However, it is possible to reach out to this organization and ask them how this dataset was created.

The purpose of this study was to discover whether news article sentiment can be used to predict the stock market. This has been researched while classifying whether the S&P 500 index price is going up, or down the next day. Hence, a classification problem was considered. Although it was not in the interest of this research; as described in Section 2.3, exact price prediction can give more context to the predictive power of news article sentiment towards the stock market. Thus, while applying a regression model, exact prices can be predicted, so that speculators may be able to take less risks. To explain, if it is predicted that the price will go up with more than 50%, than it is less risky for an investor to buy a share rather than predicting that the stock market will only go up with 1%. To summarize, future research can be conducted with this dataset while applying a regression model.

The main research question of the study was formulated in such a way that the analysis of news sentiment, which also belongs to fundamental analysis, was central. Besides the best performing news sentiment analysis, TextBlob, it has been found that it best performed in combination with TA features. However, this research was mainly focused on fundamental analysis, so there was only a small research done into technical analysis factors. Therefore, a recommendation for further research is to compare the used TA features of this study with others state-of-the-art TA metrics. Afterwards, the best performing TA features should be combined with the TextBlob sentiment score on news articles which may improve the model.

Furthermore, it would be very interesting if the proposed SVM and feature set can be applied on contemporary data. When a simulation can be executed on this experiment with a certain amount of dollars, where the simulation buys a stock if the price is predicted to go up the next day, and contrary sells if the price is predicted to go down. After this experiment has been conducted for some time, the return on investment can be calculated. This can make the outcomes of the current research even more explainable for traders.

To summarize, even though future research is recommended, the results of the best performing combination of dataset, news sentiment method, TA, and the SVM model are very promising. Investors can use these results to make better decisions whether to buy, sell or hold a stock.

7 Conclusion

The purpose of this study was to investigate whether news articles sentiment could predict the stock market movement directionalities. Accordingly, the following main research question was formulated: "Can news sentiment data be applied to predict stock market movements?". The current study discovered that the best prediction model contained only financial or business-related news articles rather than all the news articles when classifying the stock market movements. In order to gather the news article sentiment polarity, different sentiment analysis methods were applied. Subsequently, several machine and deep learning algorithms were used to develop models that classified the stock market movements. While evaluating the results, it has been found that the scores of the TextBlob news sentiment analysis method are most suitable as input for the models. Furthermore, the best results were obtained by combining this approach with an SVM model whose hyperparameters were optimized using a grid search. The SVM was capable to predict the stock market based on the sentiment polarity of news articles. Additionally, it was discovered that using the TextBlob feature combined with technical analysis resulted in even better model performance.

The results of this study confirm the findings of Ican and Celik (2017) that the stock market can be predicted based on news articles sentiment. Furthermore, Hu et al. (2018) their recommendation to use not only fundamental analysis but combine it with technical analysis to achieve a better performance has been proven.

To conclude, the current study provides evidence that the stock market can be predicted based on news articles. However, combining those news articles with technical analysis is recommended. Furthermore, investors can use the proposed dataset, news sentiment analysis method, technical analysis features, and SVM model to enhance their stock buying and selling decisions in order to increase the value of their portfolio.

Reference List

Ahlin, M., & Ranby, F. (2019). *Predicting marketing churn using machine Learning models*. (Master's thesis, Umeå University) Retrieved from http://www.diva-portal.org/smash/get/diva2:1335397/FULLTEXT01.pdf

Attanasio, G., Cagliero, L., Garza, P., & Baralis, E. (2019, November). Combining news sentiment and technical analysis to predict stock trend reversal. *2019 International Conference on Data Mining Workshops*, 514–521.

Bohn, T. A. (2017). *Improving long term stock market prediction with text analysis.* (Master's thesis, Western University) Retrieved from https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=6267&context=etd

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Canbaz, A. G. (2020). *Can investor's sentiment from forum posts predict bitcoin return?* (Doctoral dissertation). Retrieved from http://research.sabanciuniv.edu/41148/1/10237272_Canbaz_Ay%C5%9Fe__Gul.pdf

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*, 35.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 6.

Choudhry, R., & Garg, K. (2008). A hybrid machine learning system for stock market forecasting. *World Academy of Science, Engineering and Technology*, *39*, 315-318.

De Fortuny, E. J., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, *50*, 426-441.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*, 139–157.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., & Amirabadizadeh, A. (2018). A Comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal of Research in Health Sciences*, *18*, 412.

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance, 46*, 1575-1617.

Falinouss, P. (2007). *Stock trend prediction using news articles: A text mining approach*. (Master's thesis, Luleå University of Technology) Retrieved from http://epubl.ltu.se/1653-0187/2007/071/index-en.html

Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business*, *38,* 34-105.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management*, *225*, 105758.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning* (pp. 3-33). Springer, Cham.

Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications*, *42*, 9001-9011.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*, 685-697.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73,* 220-239.

Hellström, T., & Holmström, K. (1998). Predicting the stock market.

Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 261-269).

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media, 8.*

Ican, O., & Celik, T. B. (2017). Stock market prediction performance of neural networks: A literature review. *International Journal of Economics and Finance*, *9*, 100-108.

Inoue, T., & Abe, S. (2001, July). Fuzzy support vector machines for pattern classification. *International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*.

Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, *4*, 320-330.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems,* 3146-3154.

Khedr, A. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, *9*, 22.

Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, *11*, e0161197.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence, 13,* 1137-1145.

Lauren, S., & Harlili, S. D. (2014, August). Stock trend prediction using simple moving average supported by news classification. In *2014 International Conference of Advanced Informatics: Concept, Theory and Application* (pp. 135-139). IEEE.

Lee, P. J., Hu, Y. H., & Lu, K. T. (2018). Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics*, *35*, 436-445.

Loria, S. (2018). textblob Documentation. *Release 0.15*, *2*.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*, 35-65.

Lowe, D., & Webb, A. R. (1991, February). Time series prediction by adaptive networks: A dynamical systems perspective. *In IEE Proceedings F (Radar and Signal Processing), 138,* 17-24. IET Digital Library.

Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. *Quantitative Finance*, *15*, 999-1012.

Matta, M., Lunesu, I., & Marchesi, M. (2015). Bitcoin spread prediction using social and web search media. In UMAP workshops, 1–10.

Mendez, K.M., Reinke, S.N. & Broadhurst, D.I. (2010.) A comparative evaluation of the generalized predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, *15***,** 150.

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*, *28*.

Mittermayer, M. A. (2004, January). Forecasting intraday stock price trends with text mining techniques. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (pp. 10-pp). IEEE.

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*, 3299-3303.

Montgomery, D. C. (2017). *Design and analysis of experiments.* John wiley & sons.

Murphy, J. J., & Murphy, J. (1999). *Technical Analysis of the Financial Markets* (2de ed.). Pearson Professional Education.

Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology*, *24*, 1565- 1567.

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2019). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, *53*, 3007–3057.

Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, *68*, 855-859

Parrilla Gutiérrez, J. M. (2010). *Support vector machines: Similarity functions to work with heterogeneous data and classifying documents*. (Master's thesis, Univesrity of Southern Denmark) Retrieved from https://upcommons.upc.edu/bitstream/handle/2099.1/11809/622 80.pdf?sequence=1&isAllowed=y

Prajapati, P. (2020). Predictive analysis of Bitcoin price considering social sentiments. *arXiv preprint arXiv:2001.10343*.

Probst, P., Bischl, B., & Boulesteix, A. L. (2018). Tunability: Importance of hyperparameters of machine learning algorithms. *arXiv preprint arXiv:1802.09596*.

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007, March). Admiral: A data mining based financial trading system. In *2007 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 720-725). IEEE.

Reddy, V. K. S. (2018). Stock market prediction using machine learning. *International Research Journal of Engineering and Technology*, *5*, 1033-1035.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine*, *13*, 59-76.

Schumaker, R., & Chen, H. (2006). Textual analysis of stock market prediction using financial news articles. *AMCIS 2006 Proceedings*, 185.

Serebryannikova, A. (2018). *Predicting stock market rrends from news rticles*. (Master's thesis, Charles University) Retrieved from https://lct-master.org/getfile.php?id=3770&n=1 &dt=TH&ft=pdf&type=TH

Sewell, M. (2011). History of the efficient market hypothesis. London: *Research Note*, *11*, 04.

Shapiro, A. H., Sudhof, M., & Wilson, D. (2020). Measuring news sentiment. *Federal Reserve Bank of San Francisco.*

Siblini, W., Fréry, J., He-Guelton, L., Oblé, F., & Wang, Y. Q. (2020, April). Master Your Metrics with Calibration. In *International Symposium on Intelligent Data Analysis* (pp. 457-469). Springer, Cham.

Silva, M., Giovanini, L., Fernandes, J., Oliveira, D., & Silva, C. S. (2020). Facebook ad engagement in the Russian active measures campaign of 2016. *arXiv preprint arXiv:2012.11690*.

Sohangir, S., Petty, N., & Wang, D. (2018, January). Financial sentiment lexicon analysis. In *2018 IEEE 12th International Conference on Semantic Computing* (pp. 286-289). IEEE.

Steinert, L. & Herff, C. (2018). Predicting altcoin returns using social media. *PloS one*, *13*, e0208119.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics, 17*.

Thirunavukarasu, A., & Maheswari, U. (2013). Technical analysis of Fuzzy Metagraph based decision Support system for capital market. *Journal of Computer Science*, *9*, 1146.

Trafalis, T. B., & Gilbert, R. C. (2006). Robust classification and regression using support vector machines. *European Journal of Operational Research*, *173*, 893-909.

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, *21,* 589.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Vargas, M. R., de Lima, B. S. L. P., & Evsukoff, A. G. (2017, juni). Deep learning for stock market prediction from financial news articles. In *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, 1–6.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., & Zhao, B. Y. (2014). Crowds on wall street: Extracting value from social investing platforms. *arXiv preprint arXiv:1406.1137.*

Wang, S.C. (2003). Artificial neural network. *Interdisciplinary Computing in Java Programming,* 81–100.

West, D., Dellana, S., & Qian, J. X. (2005). Neural network ensemble strategies for financial decision applications. *Computers and Operations Research, 32*, 2543–2559

White, H. (1988, July). Economic prediction using neural networks: The case of IBM daily stock returns. In *International Conference on Neural Networks*, *2*, 451-458.

Xia, J., Broadhurst, D. I., Wilson, M., & Wishart, D. S. (2013). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics, 9*, 280–299.

Yilmazkuday, H. (2020). COVID-19 effects on the S&P 500 Index. *SSRN Electronic Journal*, 2–3.

Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007, June). Combining news and technical indicators in daily stock price trends prediction. In *International Symposium on Neural Networks* (pp. 1087-1096). Springer, Berlin, Heidelberg.

Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, *37*, 7838–7843.

Zhu, J., & Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, *14*, 185-205.

Appendix I – Confusion Matrix

	Predicted true	**Predicted false**
Actual true	True Positives (TPs)	False Negatives (FNs)
Actual false	False Positives (FPs)	True Negatives (TNs)

Appendix II – AUC-PR Scores

This appendix depicts the AUC-PR scores for each machine learning model. Because this study used two news article datasets, (1) with all the news articles with a financial or business-related section, and (2) all the news articles with unknown sections combined with dataset 1, the scores are divided by the dataset.

Dataset 1 - Financial/business news articles

Feature sets	RF	GB	LGB	XGB	ANN	SVM
LM	0.51	0.47	0.47	0.49	0.54	0.56
TextBlob	0.59	0.45	0.51	0.50	0.59	0.63
VADER	0.52	0.50	0.51	0.50	0.52	0.61
TA	0.52	0.44	0.45	0.45	0.63	0.71
LM + TA	0.55	0.55	0.44	0.46	0.65	0.73
TextBlob + TA	0.54	0.56	0.46	0.46	0.62	0.75
VADER + TA	0.55	0.53	0.51	0.47	0.56	0.74

Dataset 2 - All news articles

Feature sets	RF	GB	LGB	XGB	ANN	SVM
LM	0.54	0.51	0.53	0.50	0.54	0.54
TextBlob	0.55	0.48	0.49	0.48	0.55	0.38
VADER	0.50	0.42	0.49	0.45	0.49	0.61
TA	0.52	0.44	0.45	0.45	0.69	0.70
LM + TA	0.54	0.54	0.48	0.47	0.66	0.73
TextBlob + TA	0.57	0.58	0.48	0.45	0.64	0.74
VADER + TA	0.53	0.53	0.49	0.46	0.61	0.74

Appendix III – AUC-ROC Scores

This appendix depicts the AUC-ROC scores for each machine learning model. Because this study used two news article datasets, (1) with all the news articles with a financial or business-related section, and (2) all the news articles with unknown sections combined with dataset 1, the scores are divided by the dataset.

Dataset 1 - Financial/business news articles

Independent variables	**RF**	**GB**	**LGB**	**XGB**	**ANN**	**SVM**
LM	0.54	0.52	0.49	0.53	0.56	0.59
TextBlob	0.64	0.48	0.56	0.54	0.63	0.67
VADER	0.52	0.51	0.50	0.52	0.48	0.60
TA	0.55	0.49	0.48	0.51	0.70	0.72
LM + TA	0.57	0.55	0.49	0.52	0.69	0.75
TextBlob + TA	0.59	0.61	0.52	0.51	0.71	0.76
VADER + TA	0.59	0.60	0.51	0.51	0.60	0.75

Dataset 2 - All news articles

Independent variables	**RF**	**GB**	**LGB**	**XGB**	**ANN**	**SVM**
LM	0.54	0.55	0.55	0.56	0.55	0.46
TextBlob	0.61	0.54	0.54	0.54	0.62	0.63
VADER	0.53	0.46	0.51	0.50	0.55	0.54
TA	0.55	0.49	0.48	0.51	0.63	0.75
LM + TA	0.58	0.56	0.53	0.54	0.70	0.74
TextBlob + TA	0.60	0.57	0.51	0.49	0.70	0.74
VADER + TA	0.58	0.57	0.49	0.47	0.67	0.74