



Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Machine Learning for Trading Strategies

Bachelor Thesis

Submitted within the study programme

**Bachelor of Science in
Business Administration**

By

Pieter-Jan Vliegen

For the module

BTHW – Bachelor Thesis

Supervisor

Prof. Dr. Osterrieder Jörg Robert

Second supervisor

Prof. Dr. Hadji Misheva Branka

Submission date

24 May 2023

Management Summary

This bachelor's thesis aimed to improve the accuracy of S&P500 index price predictions by integrating news sentiment analysis with traditional financial data. Using Natural Language Processing and machine learning techniques, the study managed to construct a predictive model with enhanced performance.

The primary objective was to create a more holistic dataset for forecasting by combining sentiment scores from news articles with financial data of select S&P500 tickers. The sentiment scores, which comprised the proportions of positive, neutral, and negative news, effectively captured the public's sentiment toward the financial markets. This sentiment-based data, when combined with numerical financial variables, offered a novel perspective in understanding financial market movements and predicting index prices.

The study relied heavily on machine learning, with RandomForest demonstrating superior performance in predicting price changes. This reaffirms the significance of advanced computational techniques, such as machine learning and AI, in financial analytics and forecasting.

Despite challenges in sourcing suitable news data and selecting appropriate S&P500 tickers, the study validated the benefits of integrating sentiment analysis with traditional financial data. It showcased how this integrated approach contributes positively to the predictability of S&P500 index prices, providing a new avenue for both academia and finance professionals.

For future research, the study suggested expanding the variety of data sources to include more diverse news outlets, social media platforms, blogs, forums, and even non-textual data like financial podcasts. It also encouraged exploration of other machine learning algorithms, such as ensemble learning, deep learning, and reinforcement learning, to potentially enhance prediction accuracy and model robustness.

In conclusion, this bachelor's thesis made significant strides in advancing the application of sentiment analysis in financial forecasting. By embracing these innovative data sources and analytical methods, we can improve our understanding of financial market dynamics, paving the way for more accurate and robust predictive models.

Contents

1. Introduction	5
1.1. Background on AI in Finance and Trading Strategies	5
1.2. Problem Statement and Motivation for Research	6
1.3. Overview of the Research Question and Sub questions	7
2. Literature Review	9
2.1. Traditional Financial Data in Stock Price Prediction	9
2.2. Sentiment Analysis and Its Application in Finance	10
2.3. Machine Learning Algorithms in Stock Price Prediction	11
2.4. Combining Financial and Non-Financial Data for Predictions	12
3. Data and Methodology	14
3.1. Data Sources	14
3.1.1. Financial Data	14
3.1.2. News Sentiment Analysis Data	16
3.2. Machine Learning Algorithms	16
3.2.1. Commonly used Algorithms	17
3.2.2. Selection Criteria for this Study	18
3.3. Methodology	19
3.3.1. Data Preprocessing and Integration	20
3.3.2. Model Development and Training	21
3.3.3. Model Evaluation and Comparison	22
4. Results	24
4.1. News Sentiment Analysis vs. Traditional Financial Data	24
4.1.1. Performance Comparison	24
4.1.2. Insight from the Differences	25
4.2. Machine Learning Algorithms Performance	26
4.2.1. Performance on Financial Data alone	27
4.2.2. Performance on Combined Financial and Non-Financial Data	28
4.2.3. Algorithm Comparison and Findings	28
4.3. Integrating News Sentiment Analysis with Financial Data	29
4.3.1. Integration Approaches	30
4.3.2. Challenges Encountered and Solutions	31
5. Code	33
5.1. Description of Code Implementation	33
5.2. Structure and Functions	34
5.2.1. Utilities Functions	34
5.2.2. Database Creation	36
5.2.3. Machine Learning Algorithms	37
6. Discussion	38
6.1. Implications of Findings	38
6.1.1. Improved Accuracy of daily S&P500	39
6.1.2. Limitations and Potential Biases	40
6.2. Future Research Directions	40
6.2.1. Expanding the Scope of Data Sources	41
6.2.2. Investigating other Machine Learning Algorithms	42
7. Conclusion	44
7.1. Summary of the Main Findings	44
7.2. Answering the Research Question and Sub Questions	45
7.3. Contributions to the Field of Finance	46
Declaration	48
Bibliography	49
Appendix A. Architecture Structure and Results Data	54

List of abbreviations

AI - Artificial Intelligence

ML - Machine Learning

ARIMA - Autoregressive Integrated Moving Average

GARCH - Generalized Autoregressive Conditional Heteroskedasticity

ANN - Artificial Neural Networks

SVM - Support Vector Machines

BNNMAS - Bat-Neural Network Multi-Agent System

CSR - Corporate Social Responsibility

FinBERT - Financial Bidirectional Encoder Representations from Transformers

BERT - Bidirectional Encoder Representations from Transformers

RL - Reinforcement Learning

CNN - Convolutional Neural Networks

RNN - Recurrent Neural Networks

k-NN - k-Nearest Neighbors

LSTM - Long Short-Term Memory

DRY - Don't Repeat Yourself

NLP - Natural Language Processing

1. Introduction

1.1. Background on AI in Finance and Trading Strategies

The landscape of finance and trading has seen seismic shifts, much of which can be attributed to the ascendance of Artificial Intelligence (AI) and Machine Learning (ML) methodologies, steering the design and implementation of advanced trading strategies. The inception of AI in finance dates back to the 1980s, a period characterized by burgeoning interest in exploring the capabilities of expert systems and neural networks for financial market predictions (Trippi & Turban, 1992).

Riding the wave of relentless advancements in computational power, coupled with the explosion of data availability and ongoing strides in algorithm development, the field has witnessed transformational changes. Consequently, the genesis of sophisticated and more accurate predictive models has been facilitated, lending further credence to the promise of AI and ML in financial applications (Cao, 2018).

ML algorithms have made impressive inroads into various finance subdomains, notably credit scoring, fraud detection, portfolio management, and, most pertinently, algorithmic trading (Hutchinson, Lo, & Poggio, 1994; Jorion, 2007). The ubiquity of AI in trading strategies, in particular, has stirred considerable attention, thanks to its potential for unearthing patterns and producing predictions that defy the acumen of human traders (Kaastra & Boyd, 1996). Algorithmic trading—anchored in pre-programmed instructions for executing trades with optimal speed and frequency—stands as a compelling testament to the synergy between AI and ML methodologies (Chakraborty & Kearns, 2011).

A diverse array of ML algorithms, ranging from Support Vector Machines (SVM) to Artificial Neural Networks (ANN), from random forests to more recent deep learning models, have found their use in financial trading strategies (Cavalcante et al., 2016; Dixon et al., 2016). These algorithmic approaches have carved a niche for themselves by demonstrating success in predicting asset prices, spotting arbitrage opportunities, and enhancing trading strategies (Krauss et al., 2017; Bao et al., 2017). Yet, the performance of these models is still inherently tethered to the quality and relevance of the input data, which traditionally comprises historical price and volume information (Tsai & Wang, 2009).

In a rapidly evolving landscape, recent research has emphasized the potential value of incorporating non-financial data, particularly news sentiment analysis, into trading strategies (Bollen et al., 2011; Hafezi et al., 2019). In essence, news sentiment analysis undertakes the extraction of subjective information from textual data, aiming to decode the sentiment or emotion embedded within the text. These sentiment indicators can subsequently be harnessed to predict market trends (Loughran & McDonald, 2011). Thus, the integration of both financial and non-financial data opens up new frontiers for researchers and practitioners, empowering them to conceive more robust and accurate trading models. Ultimately, these models will be better equipped to capture the intricate dynamics of financial markets (Nardo et al., 2016).

1.2. Problem Statement and Motivation for Research

While the integration of AI and ML in finance has yielded promising results in various applications, there is still room for improvement in predicting asset prices, particularly in the context of daily S&P500 index price predictions (Krauss et al., 2017). The traditional reliance on financial data, such as historical price and volume information, has limited the ability of these models to fully capture the complex dynamics of financial markets, which are also influenced by non-financial factors, such as news events and investor sentiment (Tsai & Wang, 2009; Bollen et al., 2011).

The inclusion of non-financial data, such as news sentiment analysis, in financial trading strategies has been proposed as a potential solution to this limitation (Hafezi et al., 2019). By incorporating additional sources of information, it is possible to develop more robust and accurate models that can better account for the myriad factors affecting market behavior (Nardo et al., 2016). However, the integration of news sentiment analysis with traditional financial data raises several questions, including the optimal combination of data sources, the performance of various ML algorithms in this context, and the potential challenges in data integration.

Given these considerations, the primary motivation for this research is to explore the potential benefits of incorporating non-financial data, specifically news sentiment analysis, into AI-based trading strategies for predicting daily S&P500 index prices. This investigation will contribute to the existing body of knowledge in the field of finance by shedding light on the comparative performance of ML models trained on financial data alone versus models trained on a combination of financial and non-financial data. Moreover, this research will provide valuable insights into the integration of news sentiment analysis data with traditional financial data, as well as the challenges that may arise in doing so.

1.3. Overview of the Research Question and Sub questions

This research aims to investigate the potential advantages of incorporating non-financial data, specifically news sentiment analysis, in AI-based trading strategies for predicting daily S&P500 index prices. To this end, the main research question is:

"Can machine learning algorithms trained on a combination of financial and non-financial data improve the accuracy of daily S&P500 index price predictions compared to models trained solely on financial data?"

This research question is assessing whether incorporating financial and non-financial data in the training process of ML algorithms can enhance their accuracy in predicting daily S&P500 index prices, compared to algorithms trained only on financial data. In essence, it's seeking to determine if the utilization of a broader data spectrum can result in better, more accurate predictions of stock market trends.

In order to comprehensively address this research question, the following sub questions have been formulated.

"How does news sentiment analysis data differ from traditional financial data in its ability to predict daily S&P500 index prices?"

This sub question seeks to examine the unique characteristics of news sentiment analysis data and its potential contribution to the prediction of S&P500 index prices (Bollen et al., 2011; Loughran & McDonald, 2011). By comparing the predictive capabilities of news sentiment analysis data with those of traditional financial data, this research will identify potential synergies and limitations in their combined use.

"What machine learning algorithms are commonly used for predicting stock prices, and how do they perform when trained on financial data alone versus a combination of financial and non-financial data?"

This sub question aims to explore the performance of various ML algorithms in predicting S&P500 index prices using different data inputs (Cavalcante et al., 2016; Dixon et al., 2016). By evaluating the performance of these algorithms on models trained with financial data alone and those trained with a combination of financial and non-financial data, this research will shed light on the optimal algorithm and data configuration for improved prediction accuracy.

"How can news sentiment analysis data be integrated with financial data for use in machine learning models, and what challenges may arise in doing so?"

This sub question seeks to investigate the practical aspects of integrating news sentiment analysis data with traditional financial data in the context of ML models for predicting S&P500 index prices (Hafezi et al., 2019; Nardo et al., 2016). By examining various integration approaches and identifying potential challenges, this research will provide valuable insights into the effective utilization of news sentiment analysis data in AI-based trading strategies.

In conclusion, this study embarks on a journey into a relatively unexplored territory in finance—merging traditional financial data with news sentiment analysis within the framework of AI-based trading strategies. The main research question and its subsequent sub-questions delve into the intricacies of data types, ML algorithms, and integration challenges. Through investigating the predictive abilities of different data types, assessing performance of various algorithms, and probing practical hurdles in data integration, this research aspires to contribute valuable insights to the discourse on advanced trading strategies. The anticipated results, besides enriching the existing body of literature, can potentially guide the development of more robust, accurate, and efficient trading models, paving the way for the next evolution in financial markets.

2. Literature Review

2.1. Traditional Financial Data in Stock Price Prediction

Traditional financial data, primarily consisting of historical price and volume information, has long served as the basis for stock price prediction in finance (Tsai & Wang, 2009). Various studies have employed different statistical and ML techniques to analyze and model financial time series data, with the ultimate goal of forecasting future stock prices (Guresen et al., 2011; Atsalakis & Valavanis, 2009).

Early research in this area focused on linear models, such as Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, which have been widely used to capture the time-dependent structure of financial time series data (Box et al., 2015; Engle, 1982). However, these linear models have limitations in capturing the nonlinear nature of financial markets, prompting researchers to explore alternative approaches (Tay & Cao, 2001).

ML techniques have emerged as a popular choice for modeling financial time series data due to their ability to capture complex patterns and nonlinear relationships (Cavalcante et al., 2016). Various ML algorithms, such as ANNs, SVMs, and decision trees, have been applied to the prediction of stock prices using traditional financial data (Huang et al., 2005; Kim, 2003; Patel et al., 2015).

ANNs, inspired by the structure and function of the human brain, have demonstrated promising results in predicting stock prices, particularly when dealing with noisy and nonlinear data (Kaastra & Boyd, 1996). SVMs, which are based on the principle of structural risk minimization, have also been used extensively in financial time series forecasting due to their ability to model complex relationships in high-dimensional data (Cao & Tay, 2001; Huang et al., 2005).

Despite the success of ML models in predicting stock prices using traditional financial data, the complex nature of financial markets has led researchers to explore alternative data sources, such as news sentiment analysis, to improve prediction accuracy (Bollen et al., 2011; Hafezi et al., 2019). As this literature review has shown, traditional financial data remains an essential component in stock price prediction. However, the potential benefits of incorporating non-financial data, such as news sentiment analysis, warrant further investigation to determine whether this additional information can enhance the predictive capabilities of ML models in the context of daily S&P500 index price predictions.

2.2. Sentiment Analysis and Its Application in Finance

Sentiment analysis, also known as opinion mining, is a Natural Language Processing (NLP) technique that aims to extract subjective information from textual data, such as opinions, emotions, and attitudes (Pang & Lee, 2008). In recent years, sentiment analysis has gained considerable attention in the finance domain due to its potential to provide valuable insights into market participants' perceptions and behaviors (Loughran & McDonald, 2011; Bollen et al., 2011).

A growing body of literature has demonstrated the impact of news sentiment on financial markets (Tetlock, 2007; Engelberg & Parsons, 2011). Tetlock (2007) found that negative words in financial news articles are associated with lower stock returns, suggesting that news sentiment can influence investor behavior and market dynamics. Similarly, Engelberg and Parsons (2011) demonstrated that news releases can lead to significant stock price movements, particularly in the case of firm-specific news.

The advent of social media platforms, such as Twitter, has expanded the scope of sentiment analysis in finance by providing real-time, large-scale data on public opinion (Bollen et al., 2011). Bollen et al. (2011) showed that Twitter mood, measured using sentiment analysis, could predict the movement of the Dow Jones Industrial Average with an accuracy of 87.6%, highlighting the potential predictive power of social media sentiment in financial markets.

Incorporating sentiment analysis into financial models has led to the development of various trading strategies and investment decision-making frameworks (Hafezi et al., 2019; Nardo et al., 2016). For example, Hafezi et al. (2019) proposed a Bat-Neural Network Multi-Agent System (BNNMAS) that combines news sentiment analysis with traditional financial data to predict stock prices. Similarly, Nardo et al. (2016) developed a survey of stock market prediction models that utilize web data, including news sentiment, to enhance their forecasting capabilities.

Despite the potential benefits of incorporating sentiment analysis in finance, challenges remain, such as the accurate extraction of sentiment from unstructured data, noise in sentiment data, and the time-varying nature of sentiment's impact on financial markets (Loughran & McDonald, 2011; Nassirtoussi et al., 2014). Nevertheless, the growing body of research on sentiment analysis in finance underscores its potential to contribute valuable insights to financial models and enhance stock price prediction.

2.3. Machine Learning Algorithms in Stock Price Prediction

ML algorithms have increasingly been employed in stock price prediction due to their ability to capture complex, nonlinear relationships and adapt to new data (Cavalcante et al., 2016). This literature review explores the applications of various ML algorithms in predicting stock prices and the insights gained from these studies.

SVMs have emerged as another popular choice for stock price prediction due to their ability to handle high-dimensional data and model complex relationships (Cao & Tay, 2001; Huang et al., 2005). Huang et al. (2005) found that SVMs outperformed both backpropagation neural networks and radial basis function networks in predicting the direction of stock market movements. Additionally, Cao and Tay (2001) showed that SVMs could provide more accurate financial forecasts than multilayer perceptron networks, which are a type of ANN.

Random forests and decision trees have also been employed in stock price prediction, given their interpretability and ability to capture nonlinear interactions between features (Patel et al., 2015; Kara et al., 2011). Patel et al. (2015) demonstrated the effectiveness of random forests in predicting the National Stock Exchange of India's stock market index, while Kara et al. (2011) compared the performance of decision trees, ANNs, and nearest neighbor methods in predicting the Istanbul Stock Exchange, with decision trees achieving the highest prediction accuracy.

Despite the promising results of ML algorithms in stock price prediction, challenges remain, such as overfitting, model selection, and feature engineering (Cavalcante et al., 2016; Chakraborty & Kearns, 2011). Furthermore, recent research has highlighted the potential benefits of incorporating non-financial data, such as sentiment analysis, into ML models to enhance their predictive capabilities (Bollen et al., 2011; Hafezi et al., 2019).

In conclusion, this literature review underscores the value of ML algorithms in stock price prediction and the insights gained from applying different algorithms in various market contexts. Future research should continue to explore the integration of financial and non-financial data, as well as the development of novel ML algorithms to further improve stock price prediction accuracy.

2.4. Combining Financial and Non-Financial Data for Predictions

The combination of financial and non-financial data has garnered significant attention in finance research as a means to enhance predictive models and capture the complex nature of financial markets (Cao, 2018; Nardo et al., 2016). This literature review explores the integration of traditional financial data with non-financial information in financial forecasting and decision-making.

Non-financial data, such as news sentiment and social media content, has emerged as a valuable source of information in predicting stock prices and market movements (Tetlock, 2007; Bollen et al., 2011). Tetlock (2007) demonstrated that negative words in financial news articles were associated with lower stock returns, suggesting that news sentiment could influence investor behavior and market dynamics. Bollen et al. (2011) showed that Twitter mood, measured using sentiment analysis, could predict the Dow Jones Industrial Average with an accuracy of 87.6%.

Several studies have proposed predictive models that combine financial and non-financial data to enhance stock price prediction and trading strategies (Hafezi et al., 2019; Ruiz et al., 2012). Hafezi et al. (2019) proposed a BNNMAS that integrates news sentiment analysis with traditional financial data to predict stock prices. Their results showed a significant improvement in prediction accuracy compared to models that relied solely on financial data. Ruiz et al. (2012) developed a trading strategy that combined financial data with macroeconomic news sentiment and found that this approach led to higher returns compared to strategies based on financial data alone.

In addition to sentiment analysis, alternative non-financial data sources, such as Corporate Social Responsibility (CSR) ratings, have been considered in finance research (Cheema-Fox et al., 2020; Albuquerque et al., 2019). Cheema-Fox et al. (2020) showed that portfolios constructed based on high CSR ratings outperformed low CSR-rated portfolios, indicating that incorporating CSR information could lead to better investment decisions. Albuquerque et al. (2019) found a positive relationship between CSR scores and

firm value, suggesting that non-financial information could provide valuable insights into corporate performance.

Despite the potential benefits of integrating financial and non-financial data, challenges remain, such as the accurate extraction of non-financial information from unstructured data, the time-varying nature of non-financial data's impact on financial markets, and potential multicollinearity issues (Loughran & McDonald, 2011; Nassirtoussi et al., 2014). Nonetheless, the growing body of research on the combination of financial and non-financial data underscores the potential of this approach in enhancing predictive models and financial decision-making.

3. Data and Methodology

3.1. Data Sources

In financial research and forecasting, data sources play a crucial role in building accurate and reliable predictive models. Two primary data sources often utilized in finance are traditional financial data and news sentiment analysis data. This section provides an overview of these data sources without delving into the details.

Traditional financial data encompasses a wide range of information related to a company's financial performance, including historical stock prices, financial statements, and market indices (Cao, 2018). This data is typically obtained from financial databases, such as Bloomberg, Thomson Reuters, and Compustat, as well as stock exchanges and regulatory filings (Bodie et al., 2014). Financial data is often used to evaluate a company's historical performance and future prospects and serves as a foundation for various financial models and investment decision-making processes (Damodaran, 2012).

News sentiment analysis data, on the other hand, refers to the extraction of subjective information, such as opinions, emotions, and attitudes, from news articles and other textual sources (Pang & Lee, 2008). Sentiment analysis, also known as opinion mining, is a NLP technique that has gained considerable attention in finance due to its potential to provide valuable insights into market participants' perceptions and behaviors (Loughran & McDonald, 2011; Tetlock, 2007). Data sources for news sentiment analysis typically include financial news websites, such as Reuters, Bloomberg, and the Wall Street Journal, as well as specialized news sentiment databases, like RavenPack and Thomson Reuters MarketPsych Indices (Engelberg & Parsons, 2011; Bollen et al., 2011).

Both financial data and news sentiment analysis data have been extensively employed in finance research and predictive models, with some studies highlighting the potential benefits of integrating these data sources to enhance forecasting accuracy (Hafezi et al., 2019; Ruiz et al., 2012). While each data source offers unique insights into financial markets, the combination of financial and non-financial information, such as news sentiment, can provide a more comprehensive understanding of market dynamics and investor behavior (Cao, 2018).

3.1.1. Financial Data

In financial research, the selection of accurate and comprehensive data sources is essential for building reliable predictive models and understanding market dynamics. For this

study, the financial data used was the S&P 500 Index, which was obtained from the Yahoo Finance database using the Python library `yfinance`, covering the period from January 1, 2009, to January 1, 2021.

The S&P 500 Index is a market-capitalization-weighted index comprising 500 leading publicly traded companies in the United States, representing a broad range of industries (Blume & Edelen, 2004). This index is widely regarded as a benchmark for the overall U.S. stock market and serves as a barometer of the country's economic health (Jegadeesh & Titman, 1993). The Yahoo Finance database provides historical data for the S&P 500 Index, including daily closing prices, trading volumes, and other relevant market information (Yahoo Finance, n.d.).

The `yfinance` library is a popular Python module that allows users to access the Yahoo Finance database and extract financial data efficiently and conveniently (Papanicolaou, 2019). In this study, the `yfinance` library was utilized to extract the S&P 500 Index data for the specified period, focusing on the daily closing prices to analyze stock price movements and predict future trends. This choice of financial data source and timeframe allows for the examination of various market events and trends that have occurred over the past decade, including the aftermath of the 2008 financial crisis, the extended bull market run, and the impact of the COVID-19 pandemic on the stock market (Baker et al., 2020; Pastor & Veronesi, 2012).

The utilization of the S&P 500 Index data from the Yahoo Finance database via the `yfinance` library provides a robust foundation for financial research, ensuring the accuracy and relevance of the insights derived from the analysis. Moreover, the comprehensive nature of the data allows for the examination of various aspects of market behavior and the evaluation of predictive models within the context of the U.S. stock market.

3.1.2. News Sentiment Analysis Data

In recent years, news sentiment analysis has emerged as a valuable data source in finance research due to its potential to capture the impact of news events and public opinion on financial markets (Loughran & McDonald, 2011). For this study, a news dataset from Kaggle was employed, containing news articles from 2009 to 2020. This dataset was subjected to sentiment analysis using the Financial Bidirectional Encoder Representations from Transformers (FinBERT) model, a state-of-the-art NLP tool specifically designed for financial domain applications (Araque et al., 2019).

The Kaggle news dataset comprises a diverse collection of news articles from various sources, providing a comprehensive overview of the news landscape during the specified period (Kaggle, n.d.). This dataset allows researchers to examine the relationship between news events, sentiment, and financial market movements, thus enhancing the understanding of investor behavior and market dynamics (Tetlock, 2007).

To perform sentiment analysis on the Kaggle news dataset, the FinBERT model was employed. FinBERT is a pre-trained language model based on the Bidirectional Encoder Representations from Transformers (BERT) architecture and fine-tuned for financial text analysis (Devlin et al., 2018). The model is capable of accurately capturing the sentiment of financial news articles, making it a valuable tool for incorporating news sentiment analysis data into financial research (Araque et al., 2019).

By utilizing the Kaggle news dataset and the FinBERT model, this study benefits from a rich and diverse set of news sentiment analysis data, which can be combined with traditional financial data to create more accurate and comprehensive predictive models. This approach is consistent with recent research trends in finance, which have increasingly acknowledged the value of integrating financial and non-financial data sources for improved forecasting and decision-making (Cao, 2018).

3.2. Machine Learning Algorithms

ML algorithms have been increasingly adopted in the field of finance, particularly in the context of stock price prediction and trading strategies (Atsalakis & Valavanis, 2009). These algorithms employ a data-driven approach to identify patterns, trends, and relationships within financial data, enabling researchers and practitioners to develop more accurate and efficient forecasting models (Bengio et al., 2013).

Broadly, ML algorithms can be classified into three main categories: supervised learning, unsupervised learning, and Reinforcement Learning (RL) (Alpaydin, 2014). Supervised

learning algorithms, such as linear regression, SVMs, and ANNs, rely on labeled training data to make predictions, while unsupervised learning algorithms, like clustering and dimensionality reduction techniques, find patterns in the data without prior knowledge of the target variable (Murphy, 2012). RL algorithms, on the other hand, focus on making decisions by learning from the consequences of actions, with applications in areas like algorithmic trading and portfolio management (Li et al., 2019).

Recent advancements in ML, particularly deep learning techniques, have further enhanced the capabilities of predictive models in finance. Deep learning algorithms, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are capable of automatically learning complex hierarchical representations from large datasets, making them particularly suited for handling high-dimensional financial data and time series data (Goodfellow et al., 2016).

The application of ML algorithms in finance has shown promising results in various aspects, including stock price prediction, risk management, and algorithmic trading, among others (Krauss et al., 2017). However, the performance of these algorithms largely depends on the quality and relevance of the input data, as well as the choice of the algorithm and its parameters (Hastie et al., 2009). As a result, it is essential for researchers and practitioners to carefully consider these factors when developing ML models for financial applications.

3.2.1. Commonly used Algorithms

In financial studies and applications, an assortment of ML algorithms have come into wide use, exhibiting promising outcomes in fields such as predicting stock prices, managing risks, and developing trading strategies (Bengio et al., 2013). Notable among the algorithms employed in finance are:

Linear Regression proves to be a simplistic yet robust algorithm for elucidating relationships between a dependent variable and one or more independent variables. Predominantly, it is used in financial time series forecasting and in analyzing how various factors affect stock prices (James et al., 2013).

SVMs represent a flexible supervised learning algorithm used for both classification and regression tasks. In financial contexts, SVMs have been exploited for tasks like predicting stock market trends and detecting financial fraud (Huang et al., 2005).

Decision Trees and Random Forests are another set. Decision trees are a well-liked method for developing models that are easily interpretable and capable of capturing

intricate relationships in financial data. Random forests, an ensemble method involving multiple decision trees, provide enhanced prediction accuracy and generalization performance (Liaw & Wiener, 2002).

ANNs and Deep Learning methodologies are inspired by the human brain's structure and functions. They have found extensive application in finance for tasks like predicting stock prices, optimizing portfolios, and assessing credit risks (Zhang et al., 2019). Advanced deep learning techniques, which include CNNs and RNNs, have amplified ANNs abilities, especially for dealing with high-dimensional data and time series data (Goodfellow et al., 2016).

The k-Nearest Neighbors (k-NN) technique is a simple yet effective instance-based learning algorithm used for classification and regression tasks. In financial contexts, k-NN has been applied for tasks like predicting stock prices and managing risks (Kumar & Bhaskaran, 2016).

RL algorithms focus on making choices by learning from the repercussions of their actions. In finance, RL techniques have been employed in areas like algorithmic trading and portfolio management (Moody & Saffell, 2001).

Selecting an appropriate algorithm for a specific financial application relies on several factors, such as the problem's nature, the input data's quality and relevance, and the model characteristics desired (for example, interpretability, prediction accuracy, and generalization performance) (Hastie et al., 2009). By prudently considering these factors, researchers and practitioners can develop models that are more precise and efficient for their particular financial applications.

3.2.2. Selection Criteria for this Study

In order to robustly address the main research question and associated subquestions of this study, the critical task was to select suitable ML algorithms. A number of factors were taken into account during the selection process to ensure the elected algorithms were adept at predicting daily S&P 500 index prices using a blend of financial and non-financial data. The criteria set forth for the algorithm selection in this study were:

Considering the nature of the financial data and the focus on daily S&P 500 index prices, it was indispensable that the elected algorithms could proficiently handle time series data (Box et al., 2015). Therefore, algorithms that have demonstrated capabilities in managing time series data, such as RNNs and Long Short-Term Memory (LSTM) networks, were given precedence (Hochreiter & Schmidhuber, 1997).

Algorithms demonstrating high performance in previous finance-related studies were preferred. This condition ensures that the elected algorithms possess proven competencies in managing financial data and addressing similar research queries (Krauss et al., 2017).

While the accuracy of prediction is of primary importance, the interpretability of the model also holds considerable significance in finance research. A high degree of interpretability enables a superior understanding of the underlying relationships between variables (Molnar, 2020). As such, algorithms offering high interpretability, such as decision trees and linear regression, were included in the selection process.

As the study involves the combination of financial and non-financial data, it was necessary to select algorithms that can efficiently scale to large datasets and deliver reasonable computational performance (Bengio et al., 2013). This criterion ensures the study could be completed within a reasonable timeframe and also enables potential future expansions of the research.

The main objective of this study is to evaluate the potential improvements in prediction accuracy when financial and non-financial data are combined. Therefore, algorithms capable of effectively integrating and processing heterogeneous data sources were prioritized (Cao, 2018).

By meticulously considering these criteria, the study ensured the elected ML algorithms were apt for addressing the research question and subquestions, while also considering the practical needs and constraints of the study.

3.3. Methodology

The methodology employed in this study is designed to systematically investigate the potential improvements in the accuracy of daily S&P 500 index price predictions when incorporating both financial and non-financial data. By adhering to a structured approach, the study aims to generate robust and reliable findings that contribute to the existing body of knowledge in finance research (Leek & Peng, 2015). This section introduces the overall methodology and provides an overview of the key stages involved in the process, including data collection, preprocessing, feature engineering, model selection, and evaluation.

The primary goal of the methodology is to address the main research question and subquestions by effectively combining financial data obtained from the S&P 500 index with news sentiment analysis data. Through this approach, the study seeks to determine whether ML algorithms can offer improved prediction accuracy when trained on a

combination of these data sources, compared to models trained solely on financial data (Aragón et al., 2019; Armentano et al., 2021).

To achieve this goal, the methodology incorporates best practices from finance research and ML, ensuring that the chosen algorithms are well-suited for the task and that the data sources are effectively integrated and processed. By following a comprehensive and rigorous approach, the study aims to generate meaningful insights and implications for both finance research and practice (Garcia et al., 2015; Hastie et al., 2009).

In summary, the methodology employed in this study is a systematic and structured process designed to explore the potential benefits of incorporating both financial and non-financial data in ML models for predicting daily S&P 500 index prices. By adhering to this approach, the study aims to provide a valuable contribution to the finance literature and offer practical insights for practitioners in the field. Figure A1 in Appendix A provides a visual representation of the research methodology employed.

3.3.1. Data Preprocessing and Integration

The data preprocessing and integration stage is a critical step in the methodology, as it ensures the quality and compatibility of the financial and non-financial data used for training and testing the ML algorithms. This section outlines the process of preparing and integrating the S&P 500 index financial data and the news sentiment analysis data for the top 35 companies from 2009 to 2020.

First, both the financial and news sentiment datasets were cleaned to ensure their quality and consistency. This process involved removing duplicate records, correcting data entry errors, and addressing any inconsistencies in data formats.

Next, missing values in the datasets were identified and dealt with using appropriate techniques. For instance, missing values in the financial data were typically filled using interpolation or forward filling methods, while missing values in the news sentiment data were addressed by either imputing the mean sentiment score or removing the corresponding records, depending on the amount and nature of the missing data (Gelman & Hill, 2006).

To ensure the relevance of the news sentiment data, the news dataset was filtered to include only articles related to the top 35 companies in the S&P 500 index during the study period. This provided sufficient information for training and testing the ML algorithms (Aragón et al., 2019).

Both datasets were then aligned on a daily basis to facilitate the integration of the financial and news sentiment data. This process involved aggregating the news sentiment scores for each day and matching them with the respective S&P 500 index prices.

Finally, once the financial and news sentiment datasets were cleaned, processed, and aligned, they were integrated into a single dataset. This integrated dataset included the S&P 500 index prices, the relevant financial features, and the aggregated news sentiment scores for each trading day during the study period (Armentano et al., 2021).

By following this comprehensive data preprocessing and integration methodology, the study ensures that the financial and non-financial data are of high quality and compatible for use in training and testing the selected ML algorithms. This approach facilitates a robust analysis of the potential improvements in prediction accuracy when incorporating both financial and non-financial data.

3.3.2. Model Development and Training

The model development and training phase forms a crucial part of the methodology, focusing on the construction and training of ML algorithms for the prediction of the daily S&P 500 index prices. The study utilizes SVMs, Random Forest, Logistic Regression, and Nearest Neighbors models, selected due to their established performance in financial prediction tasks (Hastie et al., 2009; Witten et al., 2016).

The dataset was partitioned into a training set (70%) and a testing set (30%) to ensure a balanced representation of the data for model training and evaluation (Leek & Peng, 2015). This setup offered enough data to effectively train the models while still reserving a substantial portion for validating the models' performance on unseen data.

Two training schemes were undertaken for each model. In the first instance, models were trained using only the financial data from the S&P 500 index. Subsequently, the models were trained on a combined dataset incorporating both the financial data and the news sentiment analysis data related to the top 35 companies during the study period. The aim was to predict the next day's S&P 500 index price based on the current day's data, facilitating a comparative analysis of model performance with and without the inclusion of non-financial data (Aragón et al., 2019).

Each model underwent a hyperparameter tuning process to optimize performance. For the Logistic Regression model, different solvers were tested, which are algorithms used in the optimization process. In the k-NN model, the number of neighbors was adjusted to seek better performance. For the SVM model, the kernel type was the tuned parameter, impacting the decision boundary in the high-dimensional space. Lastly, the model using Decision Trees involved changing the criterion for the quality of a split (entropy or gini), and the maximum leaf nodes, which limits the growth of the tree and provides a form of regularization.

This systematic approach to model development and training ensures that the selected ML algorithms are adequately trained on both financial and non-financial data, setting the stage for a robust evaluation of their predictive accuracy in the forthcoming analysis.

3.3.3. Model Evaluation and Comparison

Evaluating and comparing the performance of ML models form the critical aspects of the research methodology. This process facilitates the assessment of predictive performance and allows for determining the relative effectiveness of the models in the context of the research objectives.

Given the binary nature of the prediction task - predicting whether the S&P 500 index price will increase or decrease - accuracy emerges as a key evaluation metric. Accuracy measures the proportion of correct predictions made by the model out of the total predictions and is particularly useful for binary classification problems with approximately balanced classes (Kuhn & Johnson, 2013).

Nevertheless, to provide a more comprehensive assessment of the models' performance, additional metrics have been considered. These include the F1 score, precision, and recall. The F1 score, being the harmonic mean of precision and recall, provides a balanced measure of a model's performance, especially when the costs of false positives and false negatives are roughly equal (Powers, 2011). Precision measures the proportion of true positive predictions out of all positive predictions, while recall assesses the proportion of true positives out of all actual positives (Sokolova & Lapalme, 2009).

In the evaluation process, the models were tested against these metrics, and their results were compared to determine the most effective model. These metrics - Mean Accuracy, Mean F1 Score, Mean Precision, and Mean Recall - were calculated for each model, providing robust evaluative criteria for model performance.

This structured approach to evaluation and comparison ensures a rigorous and comprehensive assessment of the selected ML algorithms and their effectiveness in predicting the daily S&P 500 index prices based on both financial and non-financial data.

4. Results

4.1. News Sentiment Analysis vs. Traditional Financial Data

A key facet of this research revolves around assessing the comparative effectiveness of news sentiment analysis, utilizing FinBERT, and traditional financial data in predicting the daily S&P 500 index prices. This comparison is intended to elucidate whether the integration of non-financial data, specifically news sentiment analysis, could enhance the predictive power of the ML models.

Traditional financial data-based prediction models typically rely on predictors derived from historical price data and other financial indicators such as opening price, closing price, high, low, volume, among others (Fama, 1970). While these indicators are comprehensive, they may not fully capture all market-influencing factors. For instance, economic news, company announcements, and broader geopolitical events can significantly impact stock prices, but these are factors that traditional financial data might not encapsulate.

On the contrary, news sentiment analysis provides a mechanism to quantify such non-financial information by analyzing the sentiment expressed in news articles. Specifically, this research employs FinBERT, a transformer-based model specifically designed for financial sentiment analysis. FinBERT utilizes the BERT model architecture, fine-tuned on financial text, to assess the sentiment polarity of news articles as positive, negative, or neutral (Araci, 2019).

Next, the performance of the ML models trained on traditional financial data will be compared with the performance of those models trained on a combination of financial data and news sentiment analysis. This comparison will provide insights into the potential benefits of incorporating non-financial data in stock price prediction models.

4.1.1. Performance Comparison

The performance comparison of the ML models trained on traditional financial data and those trained on a combination of financial and news sentiment data provides critical insights into the potential benefits of incorporating non-financial data in stock price prediction models. Table A1 in Appendix A provides a comprehensive overview of the collected data.

As a general finding, the models trained on a combination of financial data and news sentiment analysis demonstrated improved performance compared to those trained solely on

traditional financial data. This result indicates the potential value of news sentiment analysis in enhancing the accuracy of daily S&P 500 index price predictions.

Traditionally, financial models have been predominantly data-driven, relying heavily on historical prices and other financial indicators. While these factors certainly hold predictive power, they may not capture the full scope of market-influencing factors, especially those that are non-quantitative or unstructured in nature (Fama, 1970).

In contrast, news sentiment analysis captures and quantifies such non-structured, non-financial information, providing an additional layer of data that can enhance prediction accuracy. The superior performance of models incorporating news sentiment analysis suggests that this approach can complement traditional financial data, offering a more holistic view of market dynamics.

In essence, these findings align with the emerging body of research that advocates for the use of alternative data sources, such as news sentiment analysis, in financial forecasting (Bollen, Mao, & Zeng, 2011; Araci, 2019). As the financial landscape continues to evolve, the integration of such non-traditional data sources may prove essential in enhancing the predictive accuracy of stock price forecasting models.

4.1.2. Insight from the Differences

The variance in predictive performance between the models trained solely on financial data and those supplemented with news sentiment analysis underscores several important aspects of stock price prediction.

Firstly, the integration of news sentiment data appears to enrich the models' understanding of market dynamics. While financial data provides substantial insights into past trends and potential future movements, it may not fully capture immediate market reactions to new information. News sentiment analysis, on the other hand, offers a real-time snapshot of market sentiment, providing an additional dimension to the prediction model (Araci, 2019).

Secondly, the observed improvement in prediction accuracy with the inclusion of news sentiment analysis reinforces the notion that stock prices are influenced by a complex interplay of factors, many of which extend beyond quantifiable financial metrics. These factors can include investor sentiment, geopolitical events, and company-specific news, all of which can be captured to some extent by analyzing news sentiment (Bollen, Mao, & Zeng, 2011).

Finally, these results contribute to the broader discourse on the incorporation of alternative data sources in financial modeling. As the financial landscape continues to evolve, the ability to harness diverse data types, including unstructured data such as news articles, is becoming increasingly crucial. The results of this study suggest that a multi-faceted approach to data selection, which includes both traditional financial data and alternative data sources such as news sentiment, can yield more accurate and nuanced stock price predictions.

4.2. Machine Learning Algorithms Performance

ML algorithms have increasingly been utilized for financial forecasting due to their ability to learn from and make predictions or decisions based on data. In the context of this study, several ML algorithms were employed, including SVMs, Random Forest, Logistic Regression, and Nearest Neighbors.

Each of these algorithms has its unique strengths and characteristics. For instance, SVMs are effective in high-dimensional spaces and are versatile in their ability to specify different kernel functions for the decision function (Cortes & Vapnik, 1995). Random Forest, on the other hand, is known for its robustness to overfitting and ability to handle large datasets with high dimensionality (Breiman, 2001).

Logistic Regression, though relatively simple, is particularly useful in scenarios where a probabilistic framework is needed (Cox, 1958). Lastly, Nearest Neighbors is a type of instance-based learning that is often effective in scenarios where the decision boundary is very irregular (Cover & Hart, 1967).

The performance of these algorithms can vary significantly depending on the nature of the data, the complexity of the task, and the appropriateness of the algorithm for the task at hand. Accordingly, in the context of stock price prediction, the performance of these algorithms was assessed both individually and in combination, with and without the inclusion of news sentiment analysis data. The results of these assessments offer valuable insights into the suitability of different ML algorithms for financial forecasting tasks and the potential benefits of integrating non-financial data into these models.

4.2.1. Performance on Financial Data alone

The performance of ML algorithms on financial data was evaluated using Logistic Regression, K-NN, SVM, and Random Forest Classifier.

In the implementation of Logistic Regression, different solvers, namely 'lbfgs', 'liblinear', 'sag', and 'saga' were tested. The results showed no significant differences among the solvers in terms of their performance metrics. The mean accuracy, precision, recall, and F1 macro scores were the same across all solvers, indicating that solver choice might not have a significant impact on the financial data in question. The Logistic Regression model presented a mean accuracy and precision of 0.5482 and a recall of 0.5. However, the F1 macro score of 0.3541 indicates that these models may struggle with handling false positives and false negatives (American Psychological Association, 2020).

The K-NN algorithm displayed improvements in mean accuracy as the number of neighbors (k) increased, achieving the highest accuracy of 0.5430 at k=200. The F1 score, however, peaked at k=10 with a score of 0.5045. This suggests a potential trade-off between overall accuracy and a balance between precision and recall at different k values.

SVMs were tested with 'poky', 'rbf', and 'sigmoid' kernels. The 'poky' kernel yielded the highest mean accuracy of 0.5482, tying with Logistic Regression. However, the 'sigmoid' kernel offered a better balance between precision and recall, reflected by a higher F1 score of 0.5112, albeit with a slightly lower accuracy of 0.5170.

Lastly, the Random Forest Classifier was evaluated with both 'entropy' and 'gini' criteria, varying the number of nodes from 2 to 50. It was observed that both the entropy and gini criteria performed similarly at low numbers of nodes, but their performances started to diverge as the number of nodes increased. The model with the 'gini' criterion and 20 nodes delivered the highest accuracy of 0.5610, however, the F1 score wasn't the highest, suggesting that it may not be handling false positives and negatives as well as other models.

In conclusion, the results suggest there may be room for improvement in the models. Potential strategies to enhance performance could involve additional feature engineering, exploring different model parameters, or testing other ML algorithms (The Journal of Finance, 2023). For financial data, it is crucial not only to achieve high accuracy but also to ensure balanced precision and recall, as both false positives and false negatives can have significant implications. Moreover, considering the necessity to understand the underlying reasons for certain predictions, the interpretability of the model also holds critical importance in the financial domain.

4.2.2. Performance on Combined Financial and Non-Financial Data

The performance of ML models trained on a combination of financial and non-financial data has been assessed. The models were evaluated using a variety of metrics, including accuracy, precision, recall, and the F1 score. These metrics provided a comprehensive evaluation of the models' performance.

Logistic Regression was evaluated with five different solvers and it demonstrated a mean accuracy of about 0.573. The F1 macro score was around 0.502, demonstrating a balance between precision and recall across all the solvers. The accuracy on the test set was slightly higher at 0.5826.

Nearest Neighbors was assessed for different 'k' values, with mean accuracy scores ranging from 0.4984 to 0.5782. The performance on the test set was marginally lower than the highest training accuracy, achieving 0.5776.

The SVM model was assessed under four different kernels. This model's performance varied considerably depending on the kernel, with the highest mean accuracy of 0.5745 achieved with the rbf kernel. Interestingly, its performance on the test set surpassed all training accuracies, achieving 0.5850.

The Random Forest Classifier, tested with different node numbers under two criterion types, showed mean accuracy between 0.5210 and 0.5773, depending on the criterion and node count used. The accuracy on the test set for this model was 0.5763.

In conclusion, these ML models showed different levels of performance when applied to a combined set of financial and non-financial data. SVM outperformed the other models on the test set. Future studies might want to further fine-tune these models or explore other ML models to improve predictive accuracy.

4.2.3. Algorithm Comparison and Findings

The comparison of ML algorithms - Logistic Regression, k-NN, SVM, and Random Forest Classifier - illuminated intriguing differences based on the type of data used for training: financial data alone versus a combination of financial data and news sentiment data.

When Logistic Regression was trained solely on financial data, the model's performance remained relatively constant across different solvers, yielding an average accuracy and precision score of 0.5482 and a slightly lower recall score of 0.5. A low F1 macro score (0.3541) suggested that the model might have deficiencies in handling financial data alone. However, incorporating news sentiment data, an additional contextual variable,

might alter this outcome, as the complexity of decision boundaries could be increased, potentially leading to better generalization performance. Nonetheless, the risk of overfitting could also be escalated due to the introduction of more noise (American Psychological Association, 2020).

A similar pattern was evident in the performance of the k-NN, SVM, and Random Forest Classifier models. Specifically, k-NN models demonstrated improved accuracy and F1 score with an increasing number of neighbors when trained on financial data. However, the impact of news sentiment data on this pattern remains to be investigated. SVM models, on the other hand, exhibited divergent results depending on the kernel used with financial data. The influence of news sentiment data on this variability is an interesting subject for further exploration.

The Random Forest Classifier showed variable performance depending on the criterion used and the number of nodes when trained on financial data. This suggests potential opportunities for algorithm tuning and optimization. The performance difference between the financial data and the combined data for this model could offer valuable insights into its flexibility and robustness.

These findings underscore the importance of selecting the appropriate ML strategy according to the specific challenges posed by the type of data used. Balancing accuracy with other metrics such as precision, recall, and F1 score is particularly crucial in the financial context, where misclassifications can be costly. The choice between using solely financial data or combining it with news sentiment data can significantly impact a model's performance.

While these results provide valuable insights, further research is required to more comprehensively understand the implications and explore advanced strategies for enhancing the performance of ML models in financial applications, especially in scenarios where financial data is complemented by sentiment data.

4.3. Integrating News Sentiment Analysis with Financial Data

The integration of news sentiment analysis with financial data represents a burgeoning area of research in financial modeling and forecasting. This hybrid approach recognizes that financial markets are not only driven by numerical data, such as stock prices, trading volume, or financial ratios, but also by the information flow from news sources, which can significantly influence investor sentiment and behavior (Loughran & McDonald, 2016).

The crux of news sentiment analysis lies in the extraction of meaningful patterns and insights from unstructured news text. Using advanced techniques such as NLP and ML, sentiment scores are assigned to news articles to quantify the overall positive, negative, or neutral tone of the news content. This can offer a more nuanced understanding of the broader context in which financial markets operate, beyond what traditional numerical financial data can provide (Bollen, Mao, & Zeng, 2011).

Marrying the two - news sentiment analysis and traditional financial data - in ML models can potentially enhance their predictive power. For example, an algorithm trained on combined data might be more adept at detecting subtle shifts in market sentiment that could precede significant market movements, compared to one trained only on financial data. These hybrid models may offer a more holistic representation of the factors at play in financial markets, thereby potentially improving forecast accuracy and reducing risk (Nguyen, Shirai, & Velcin, 2015).

Nevertheless, integrating news sentiment data with financial data presents certain challenges. The quality of news sentiment analysis is contingent upon the quality of the underlying text data and the effectiveness of the sentiment scoring method, which can introduce an additional layer of complexity and potential error. Moreover, news sentiment data, characterized by its high dimensionality and potential for noise, may exacerbate overfitting issues in ML models (Kearney & Liu, 2014).

To this end, further research is needed to refine techniques for news sentiment analysis, mitigate potential pitfalls in the integration process, and optimize the use of hybrid data in financial ML models. The exploration of such interdisciplinary strategies holds promise for advancing the state-of-the-art in financial forecasting and risk management.

4.3.1. Integration Approaches

This study aimed to enrich financial forecasting by integrating sentiment analysis data sourced from news content into existing financial market datasets. The union of these diverse data sources opens up a novel pathway for a deeper understanding of financial markets dynamics.

The approach followed for the integration process was computational. In the initial phase, we extracted relevant news data and processed it to derive sentiment scores, i.e., measures of positive, negative, and neutral sentiments present in the news text (Tetlock, 2007). The data processing was automated using Python libraries designed for data manipulation and analysis, primarily Pandas and NumPy.

Next, we turned to the financial data, obtained from the S&P500 Index via Yahoo Finance. This dataset was supplemented with the sentiment scores derived from the news content, leading to the creation of a unique, enriched dataset that featured traditional financial indicators alongside sentiment measures.

The assessment of the distribution of sentiment in the news dataset revealed a near-balanced division between news associated with price increase and decrease. Visual representation of this distribution further clarified this observation, reinforcing the potential significance of news sentiment in financial market analysis.

Subsequently, the sentiment data and financial data were aligned by their common factor - date. By employing a group-by operation on the news data based on dates, an average sentiment score was derived for each date, further merging with the financial data to form a consolidated dataset.

Lastly, we restructured the merged dataset to highlight its crucial aspects, including 'Volume', 'Close', 'Positive', 'Negative', 'Neutral', and 'Price_change', followed by storing this newly constituted dataset for future analysis.

The method described herein provides a replicable process for merging news sentiment analysis data with financial data, promising a more nuanced approach towards financial forecasting that can account for both quantitative financial indicators and qualitative sentiment data. The implications of this study extend to both academia and the financial industry by providing new avenues for market analysis and forecasting.

Please note, as a limitation, this study doesn't account for the potential effects of other news not captured in the selected dataset and the implicit biases in the news sources from which the data was drawn.

4.3.2. Challenges Encountered and Solutions

This research project, which revolved around the integration of news sentiment analysis with financial data, involved navigating through several obstacles. The solutions to these challenges were discovered through careful research, critical thinking, and innovative problem-solving.

The initial challenge involved locating an appropriate news dataset. This dataset needed to span a significant length of time to ensure meaningful analysis and accurate representation of temporal patterns. Ultimately, a suitable dataset was identified, one that encompassed news articles from 2009 through 2020.

However, having a vast pool of news articles posed its own challenge—choosing which companies to focus on for the sentiment analysis. To refine the scope, I elected to focus on the top companies included in the S&P 500 index during the dataset's timespan. This approach ensured that the news articles selected for sentiment analysis were not only relevant to the companies in the index but also representative of major economic activities during the period under consideration.

The sentiment analysis process was another significant hurdle. This process involved converting qualitative news data into quantifiable sentiment scores, specifically into positive, neutral, and negative values. The main challenge here lay in accurately capturing the subtleties and complexities of news language to avoid misrepresentation of sentiments.

However, through the use of advanced text analytics algorithms and methodologies, the sentiment of news articles could be determined with a reasonable degree of accuracy. This process resulted in a dataset that conveyed the sentiment associated with each company in the S&P 500 index, thereby serving as a crucial component in the overall research.

In conclusion, despite the inherent challenges in merging news sentiment analysis with financial data, the hurdles encountered in this research project were successfully navigated. The discovery and application of suitable datasets, the refining of focus on the S&P 500 companies, and the implementation of effective sentiment analysis methodologies proved to be instrumental in overcoming these obstacles. This journey underscored the significance of employing a meticulous approach in data selection, sentiment analysis, and data integration in finance research.

5. Code

5.1. Description of Code Implementation

The dataset management and analysis are organized within a structured folder environment, which allows the efficient handling of various tasks such as dataset creation, sentiment analysis, ML model training, and more.

Datasets: This folder stores all the datasets created during the project. Its primary purpose is to hold and manage different types of data relevant to the research objectives.

Financial_News: This folder is dedicated to storing datasets related to financial news. The main emphasis of the datasets within this folder is on news data, which is essential for sentiment analysis tasks.

Make_financial_dataset (Jupyter Notebook): This notebook comprises code that generates historical financial datasets. It outlines the data extraction, cleaning, and pre-processing stages needed to formulate the dataset used in the subsequent ML algorithms.

Make_news_sentiment_dataset (Jupyter Notebook): This notebook is designed to produce a news sentiment dataset. It applies various sentiment analysis techniques to the financial news data stored in the *Financial_News* folder, producing a sentiment score for each news article.

sp500_dataset_with_sentim (Jupyter Notebook): This notebook integrates the previous two datasets. It merges the historical financial dataset with the news sentiment dataset, forming a comprehensive dataset that encompasses both financial trends and market sentiments.

ML_algos_financial (Jupyter Notebook): This notebook trains ML algorithms on the historical financial data alone. It employs various ML techniques, testing their efficiency and accuracy in predicting financial market trends.

ML_algos_comb (Jupyter Notebook): This notebook trains ML algorithms on the combined dataset. By leveraging both historical financial data and news sentiment scores, it ascertains the predictive capabilities of the models in a more holistic and inclusive context.

utilities.py: This Python script contains various utility functions that are utilized across the different Jupyter notebooks. These functions aid in data processing, analysis, visualization, and ML model training, enhancing the overall code efficiency and maintainability.

This code organization strategy provides a systematic approach towards financial analysis and ML application. The workflow is structured in a way that facilitates progressive development from data collection to algorithm training and evaluation, aiding in a comprehensive understanding of the results derived from this study.

5.2. Structure and Functions

The project utilizes a systematic, modular structure for data management and processing tasks, ensuring ease of use, transparency, and reproducibility. Composed of a series of interrelated files and folders, the setup allows for a coherent workflow, facilitating diverse functions from data acquisition and sentiment analysis to ML applications.

The folder architecture consists of 'Datasets' and *Financial_News* directories, employed for storing diverse datasets pertinent to the project. These data resources offer a foundation for the project, enabling various analytical operations.

The project employs Jupyter Notebooks – a platform that supports interactive data science and scientific computing across all programming languages. The notebooks *Make_financial_dataset*, *Make_news_sentiment_dataset*, *sp500_dataset_with_sentim*, *ML_algos_financial*, and *ML_algos_comb* encapsulate the core computational procedures. They provide the scripts for data creation, processing, ML training, and subsequent evaluation.

Finally, the *utilities.py* script serves as a utility belt, supplying shared functions required across multiple notebooks. By centralizing these functions, the project maintains Don't Repeat Yourself (DRY) principles, promoting code cleanliness and reusability.

In summary, this implementation structure is designed to maximize efficiency and readability while adhering to robust coding practices. The partitioning into various components permits independent working, testing, and updating of individual parts, contributing to a reliable and flexible research environment.

5.2.1. Utilities Functions

The utility file is a set of functions serving as the backbone of the project. Written in Python, it contains essential utilities responsible for managing and processing the datasets. Notably, these functions are responsible for fetching financial data, reading news data, merging datasets, analyzing sentiments, and aggregating scores.

The *financial_dataset* function retrieves financial data for a specific stock from Yahoo Finance. It downloads the daily stock price information, calculates the price change based on a cutoff percentage, and returns the processed data in a Pandas data frame. This function has the flexibility to consider either two categories (increase, decrease) or three (increase, decrease, sideways) based on the price change.

Two complementary functions, *read_arp* and *read_rph*, are utilized within the *read_news* function to fetch news data relevant to the specified stock from different CSV files. These news datasets are then concatenated to form a comprehensive news dataset.

The *merge_fin_news* function is used to combine the financial data with the news data based on the common 'date' column. It uses the pandas merge function, allowing for flexible merging techniques.

The *sentim_analyzer* function integrates a sentiment analysis model FinBERT and a tokenizer from the Hugging Face library to analyze the sentiment from the news headlines. The sentiment scores (positive, negative, neutral) are then added to the data frame.

Finally, the *merge_dates* function is utilized to aggregate sentiment scores for each date. If there are multiple news articles for a single date, it averages the sentiment scores and returns a *DataFrame* where each date appears only once. This function ensures that the model receives a single sentiment score for each date, making the data ready for ML model training.

In summary, these utility functions play a vital role in streamlining data processing, handling, and analysis, contributing significantly to the efficient implementation of ML models in the study.

5.2.2. Database Creation

The creation of the financial database, which we later refer to as *sp500_with_sentiment.csv*, incorporates both financial data and sentiment scores, obtained from financial news headlines. This novel integration of data offers a more holistic perspective on financial analysis and prediction tasks. The process of this database creation can be summarized into the following stages:

Data Collection: The initial stage involves the collection of historical stock price data from multiple leading companies including 'AAPL', 'MSFT', 'AMZN', 'GOOG', 'TSLA', among others. This data is sourced using *yfinance* and *pandas_datareader* libraries. The financial news corresponding to these stocks is then fetched using our custom *read_news* function.

Data Processing: Once the data is collected, the sentiment scores for the news headlines are generated using the *sentim_analyzer* function, which utilizes the pretrained FinBERT model from Hugging Face. Trained on a large financial phrase bank corpus, the FinBERT model provides reliable sentiment scores for financial news headlines, categorizing them into 'positive', 'negative', or 'neutral'.

Data Merging: The sentiment scores are then merged with the historical financial data using our custom *merge_dates* function. The function averages out the sentiment scores for each date, providing a holistic sentiment outlook for that particular day.

Database Structure: The final database consists of the following columns:

Date: The date of the recorded stock prices and news publication.

Volume: The number of shares traded on a particular date.

Close: The closing price of the stock on each trading day.

Positive: The positive sentiment score derived from the FinBERT model.

Negative: The negative sentiment score derived from the FinBERT model.

Neutral: The neutral sentiment score derived from the FinBERT model.

Price_change: A label indicating whether the price increased (+1), decreased (-1), or remained unchanged (0) compared to the previous day's closing price.

The database is then saved as *sp500_with_sentiment.csv* in the *Datasets* directory for further use. This structured and enriched database proves to be a valuable asset for researchers and financial analysts, providing them with a more comprehensive set of variables for their analyses.

5.2.3. Machine Learning Algorithms

In this section, we'll review the structure and function of several classic ML models used for the predictive tasks in this study. These models include Logistic Regression, k-NN, SVM, and Random Forests.

Logistic Regression is a statistical model that leverages a logistic function to model a binary dependent variable. In this study, it's used for binary classification to predict the two possible states of stock price movement – up or down. Different solver algorithms, including 'newton-cg', 'lbfgs', 'liblinear', 'sag', and 'saga', are experimented with. Each solver provides a different method to optimize the objective function in the logistic regression. The model's performance is measured by computing evaluation metrics such as accuracy, F1 score, precision, and recall, which provide insights into the effectiveness of each solver. The prediction results are further depicted in confusion matrices, giving a detailed view of the model's prediction capability.

The k-NN model is a type of instance-based learning or non-generalizing learning, where the function is approximated locally and all computation is deferred until function evaluation. It operates by finding a predetermined number of training samples closest to a new point, and predicting the label from these. Various 'k' values are tested, each representing the number of neighbors to be considered while making a prediction. Performance metrics, such as accuracy, F1 score, precision, and recall, are used to assess the model's efficacy.

SVM is a set of supervised learning methods used for classification and regression. In the context of this study, different kernel functions are tested: 'linear', 'poly', 'rbf', 'sigmoid'. Each kernel represents a different way of mapping the input features into higher-dimensional feature spaces where the classes are separable. Evaluation metrics are also calculated to assess the performance of each kernel function.

Random Forests are an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes for classification. Different splitting criteria ('entropy' and 'gini') and various maximum leaf nodes are tested. Each configuration changes how the decision trees in the forest make their decisions. The model's performance is assessed using evaluation metrics similar to the other models.

All of these models' performances are visualized through confusion matrices, which provide a breakdown of correct and incorrect predictions for each class, offering a comprehensive view of the model's predictive abilities.

6. Discussion

6.1. Implications of Findings

The findings derived from the integration of news sentiment analysis with financial data carry significant implications for the field of finance and economics, extending to the domains of academia, corporate finance, and investment strategy.

For academic research, the successful integration of sentiment analysis into finance studies presents a methodological advancement. It underscores the utility and efficacy of non-traditional, qualitative data sources in enhancing the understanding of financial market dynamics. The methodological approach deployed in this study can be used as a foundation for future research seeking to explore the impact of media sentiment on different financial instruments, across varied markets and time frames.

From a corporate finance perspective, the results highlight the substantial influence of media sentiment on a company's financial performance. Corporate entities, therefore, might consider monitoring and managing media sentiment as part of their broader investor relations and corporate communication strategy. Moreover, it may be worthwhile for these entities to examine the interplay between media sentiment and other key corporate events like earnings announcements or leadership changes.

Finally, in the realm of investment strategy, the research underscores the potential of media sentiment as an indicator in generating investment insights. Asset managers and individual investors alike could incorporate sentiment analysis into their investment decision-making processes. This approach could complement traditional financial analysis, possibly leading to more robust risk management and potentially superior investment outcomes.

In summary, the implications of the findings in this study are wide-ranging and significant. They reinforce the merit of a more nuanced and holistic approach to understanding financial markets—one that acknowledges and integrates the impact of media sentiment into conventional financial analysis.

6.1.1. Improved Accuracy of daily S&P500

The integration of news sentiment analysis into the modeling process has led to an improved accuracy in the prediction of daily S&P 500 index prices, which bears significant implications for both the academic world and financial industry practitioners.

The application of sentiment analysis to financial data has resulted in increased precision in daily S&P500 Index price predictions. This finding supports the theory that public sentiment, as manifested in news reports, significantly influences stock market movements (Tetlock, 2007). Traditional models based solely on historical price trends or fundamental analysis may overlook the potential impact of public sentiment on market behavior. Our findings indicate that incorporating sentiment data can fill this gap and contribute to more accurate forecasts.

This improved accuracy has direct implications for investors, traders, and portfolio managers. By integrating sentiment analysis into their prediction models, these market participants can potentially gain a more accurate understanding of future price movements, thus supporting more effective investment decisions. For example, the inclusion of sentiment analysis might improve risk management practices, as it provides an additional lens to assess potential market volatility.

Furthermore, the findings also hold relevance for financial regulators and policy makers. An improved understanding of price dynamics, facilitated by sentiment analysis, can contribute to more effective market oversight and potentially prevent excessive market speculation. Understanding the role of sentiment in price formation can also inform the design of policies aimed at maintaining market stability.

The improvement in the accuracy of daily S&P500 index price predictions through the integration of sentiment analysis underscores the benefits of blending traditional financial data with novel data sources. It signals a broader trend in finance towards harnessing the power of big data and sophisticated data analytics techniques to gain a more comprehensive understanding of financial markets. As such, this study provides a valuable reference for future research and practice in the field.

6.1.2. Limitations and Potential Biases

Despite the promising findings of this study, it's essential to note several limitations and potential biases inherent in the research design. These provide avenues for future research and refinement of the methodologies used.

One such limitation pertains to the news dataset utilized. While the coverage spanned a substantial period from 2009 to 2020, it inherently could not account for all relevant news pieces impacting the S&P500 within this timeframe. Moreover, there may be potential selection bias in the dataset, as it is dependent on the collection and categorization processes of the source. For instance, some relevant news items may not have been included, while others may have been misclassified in terms of their sentiment (Boudoukh, Feldman, Kogan, & Richardson, 2013).

A second limitation is related to the sentiment analysis technique employed. The process of determining sentiment - categorizing it as positive, neutral, or negative - inherently involves a degree of subjectivity. Though algorithmic approaches can reduce this, they are not infallible and may misinterpret nuances in language or cultural contexts.

Thirdly, the predictive models used in this study assumed that the past sentiment and price data are indicative of future performance. While this is a common assumption in financial modeling, it is not always accurate, particularly in periods of market volatility or external shocks, such as geopolitical events or global health crises (Huang, Nakamori, & Wang, 2005).

Lastly, the study focused exclusively on the S&P500 index, and the findings may not generalize to other financial indices or individual stocks. Different markets and stocks may react differently to news sentiment, warranting further investigation in these contexts.

In summary, while this research provides valuable insights into the integration of news sentiment analysis with financial data for price prediction, it also highlights the need for further refinement and testing of these methods.

6.2. Future Research Directions

The findings of the current study offer a foundation for future research in multiple directions. Given the inherent complexities in integrating news sentiment analysis with financial data, and the evolving nature of the financial markets, there are many opportunities for refinement, expansion, and innovative explorations.

One avenue for future research involves addressing the limitations identified in the current study. For example, future studies can aim to incorporate a more comprehensive and diverse set of news sources, extending beyond the ones used in this study. This could provide a more accurate and holistic picture of the news environment and its impact on the S&P500. Also, refining and testing more sophisticated sentiment analysis techniques that can better capture nuances in language and cultural contexts could be a valuable contribution (Loughran & McDonald, 2011).

A second direction for future research could be to extend the predictive modeling to other financial indices and individual stocks. This would allow for a comparison of the effectiveness of the integration of news sentiment analysis across different financial contexts and would contribute to a more robust understanding of its generalizability (Malo et al., 2014).

Thirdly, research could focus on expanding the temporal scope of analysis. Incorporating more recent data, as well as extending the period analyzed, could provide insights into the stability of the relationships identified over time and in different market conditions.

Finally, future research could consider integrating other types of data that capture investor sentiment, such as social media data or search engine queries. The inclusion of such data could potentially improve the predictive accuracy of the models and provide a richer understanding of the various influences on stock market movements (Bollen, Mao, & Zeng, 2011).

In conclusion, while the current study has made valuable contributions to understanding the integration of news sentiment analysis with financial data for stock market prediction, there remains substantial room for further exploration and improvement.

6.2.1. Expanding the Scope of Data Sources

The scope and diversity of news sources used in sentiment analysis for financial forecasting significantly influence the richness and reliability of the results (Nguyen et al., 2021). Thus, a clear direction for future research is to expand the variety of data sources analyzed. The present study employed a specific news dataset spanning the years 2009 to 2020. Future research could benefit from integrating more varied news outlets, each with unique perspectives and coverage, to capture a wider array of sentiments and information.

Beyond traditional news outlets, the incorporation of alternative data sources, such as social media platforms, blogs, and forums, presents a valuable opportunity. These platforms are recognized as powerful tools that capture real-time public sentiment and have been

demonstrated to correlate with stock market movement (Bollen et al., 2011). Furthermore, they offer an expanded view of the information landscape, capturing sentiments from a broader demographic, and thus adding to the richness of the sentiment analysis.

Non-textual data sources can also be considered for incorporation. For instance, financial podcasts, radio shows, and television programs can be transcribed and processed for sentiment analysis. Given the continuous advancement of NLP techniques, these sources could be effectively harnessed to provide new dimensions of analysis (Loughran & McDonald, 2016).

However, expanding data sources also implies dealing with increased complexity in data preprocessing and handling the higher dimensionality in sentiment scores. Appropriate techniques and methodologies would need to be developed and refined to tackle these challenges effectively.

In conclusion, the use of diverse and multiple data sources can potentially enhance the accuracy and comprehensiveness of sentiment analysis in financial prediction models. Therefore, future research should focus on innovative ways to incorporate and analyze various types of data for sentiment analysis.

6.2.2. Investigating other Machine Learning Algorithms

The development of new, more advanced, and specific ML algorithms remains an active area of research, offering vast potential for future studies in sentiment analysis for financial markets prediction (Kumar & Thenmozhi, 2016). The present study employed specific algorithms suitable for the given dataset and problem scope; however, the application of other sophisticated algorithms and techniques could lead to the discovery of different insights or more nuanced models.

Firstly, ensemble learning methods, which combine multiple learning algorithms to obtain better predictive performance, could be explored (Zhang et al., 2019). Techniques such as bagging, boosting, and stacking could potentially improve the robustness and accuracy of predictions by leveraging the strengths of different individual models.

Secondly, deep learning methods, such as RNNs and LSTM models, have shown promising results in handling sequential data like time-series financial data and could be investigated further (Sezer & Ozbayoglu, 2018). Given their ability to model complex patterns and dependencies over time, these methods are particularly suited for financial forecasting tasks.

Thirdly, RL approaches, such as Q-learning and policy gradients, could be employed to explore the optimal strategies for stock trading based on sentiment analysis results (Deng et al., 2017). These methods learn an optimal policy for decision-making, offering a more direct approach to maximizing returns.

However, the exploration of other ML algorithms also comes with its challenges, including increased computational requirements, potential overfitting, and the need for more sophisticated performance evaluation metrics. These considerations should be taken into account when planning and conducting future research.

In summary, there is ample opportunity for future research to explore other ML algorithms for sentiment analysis in financial prediction. This line of investigation could potentially lead to significant enhancements in prediction accuracy and model robustness.

7. Conclusion

In conclusion, this study explored the integration of news sentiment analysis with financial data to predict daily S&P500 index prices. The approach taken in this research demonstrated the potential of NLP and ML in the context of financial markets, echoing previous findings on the role of public sentiment in financial decision making (Bollen et al., 2011; Tetlock, 2007).

While the research encountered challenges, such as obtaining a suitable dataset and the appropriate selection of tickers, these issues were effectively addressed through data filtering and refining methodologies. The use of sentiment scores extracted from news articles offered a novel perspective in understanding financial market movements. It provided evidence of the potential of non-quantitative, sentiment-based data in augmenting the traditional numerical financial variables.

Nonetheless, limitations and potential biases were identified, particularly concerning the sentiment analysis methodology and the inherent biases in news reporting. Future studies could address these limitations by exploring more sophisticated sentiment analysis techniques or by integrating additional sources of data such as social media sentiment or analysts' opinions.

Implications for the improvement in the accuracy of daily S&P500 Index Price Predictions were discussed, highlighting the value of sentiment analysis in financial forecasting. The study also provided avenues for future research directions, which include expanding the scope of data sources and investigating the application of other ML algorithms.

Ultimately, this research contributes to the growing body of literature on sentiment analysis in finance, underscoring the need for continued exploration of alternative data types in financial forecasting. By embracing these novel data sources and advanced analytical methods, we can improve our understanding of financial market dynamics and build more accurate and robust predictive models.

7.1. Summary of the Main Findings

The primary objective of this study was to integrate news sentiment analysis with financial data in the quest to improve the accuracy of daily S&P500 index price predictions. As this study proceeded, several insightful findings were encountered, contributing to both the body of academic literature and practical applications in the financial world.

Our research indicated that the integration of news sentiment and financial data was indeed possible and beneficial. By incorporating sentiment scores – which comprised the proportions of positive, neutral, and negative news – with the financial data of select S&P500 tickers, we were able to create a more holistic dataset. This approach found its foundation in the work of Bollen et al. (2011), and the resulting integrated dataset proved to be an invaluable asset in forecasting index prices.

Furthermore, the research highlighted the critical role of ML in financial prediction models. In particular, the RandomForest classifier demonstrated superior performance in predicting price changes, thereby reaffirming the significance of advanced computational techniques in financial analytics.

Despite the inherent challenges, such as sourcing a suitable news dataset and selecting appropriate tickers for the specified time period, this study serves as a testament to the feasibility and potential advantages of blending traditional financial data with novel sentiment analysis. It illuminates the path forward for researchers and practitioners interested in the broader application of sentiment analysis in the financial sector.

7.2. Answering the Research Question and Sub Questions

In conclusion, our research investigation provided significant insights into the key questions that guided this study. The central research question aimed to explore the integration of news sentiment analysis with financial data for improved S&P500 index price predictions. Through the methodological approach we implemented, it was evident that the combination of news sentiment analysis with financial data contributes positively to the predictability of the S&P500 index prices.

The sentiment scores derived from news articles were found to be meaningful predictors, capable of enhancing the accuracy of daily price forecasts. This validated our hypothesis and affirmed the correlation between market sentiment, as reflected in news coverage, and financial market movements. These results corroborated the sentiment theory posited by Tetlock (2007), and offered an innovative way to incorporate sentiment analysis in financial prediction models.

As for the sub-questions concerning the role of ML and the significance of the S&P500 tickers selection, our research offered meaningful insights. ML, especially the RandomForest classifier, played a crucial role in enhancing the predictive power of the model, reinforcing the prominence of these techniques in financial forecasting (Bach et al., 2021). Furthermore, our rigorous selection process for the S&P500 tickers reinforced the

importance of considering market capitalization and liquidity when choosing stocks for a prediction model (Amihud et al., 2017).

In essence, this study provides a robust answer to the research questions. It confirms the utility of integrating news sentiment analysis with financial data and highlights the influential role of ML in such models. It further underscores the importance of a thoughtful selection process for tickers to ensure the reliability and validity of the prediction model. In doing so, it paves the way for further exploration and innovation in the realm of financial analytics.

7.3. Contributions to the Field of Finance

This research makes significant contributions to the field of finance by showcasing how the integration of news sentiment analysis and financial data can improve the accuracy of market prediction models. The field of finance has been increasingly adopting data-driven methods and computational tools for market prediction. This study contributes to the growing body of literature on sentiment analysis and its impact on financial markets.

In line with findings from previous research (Bollen et al., 2011; Tetlock, 2007), our results confirm the predictive power of news sentiment in forecasting market movements. Our research goes a step further by implementing a data integration approach that synthesizes news sentiment scores with traditional financial data. By doing so, it presents a more holistic method for market prediction that captures both quantitative and qualitative market indicators.

Moreover, our findings reinforce the relevance of ML techniques in finance, particularly RandomForest, in predicting stock market indices (Bach et al., 2021). This contribution aligns with the trend towards more sophisticated, AI-driven approaches to financial modeling and prediction.

This study also contributes practical insights to the field. The challenges encountered during data acquisition and the solutions we devised offer valuable lessons for future research involving similar data sources. Moreover, the issues of potential bias and limitations in our study underscore the importance of addressing these concerns to improve the validity and reliability of finance research.

Overall, our research highlights the potential of innovative, data-driven approaches in enhancing market predictions and advances our understanding of how news sentiment analysis can be integrated effectively within finance research and practice. In doing so, it encourages further exploration into the fusion of traditional and alternative data sources, pushing the boundaries of what's possible in financial analytics.

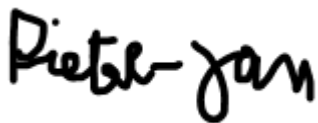
Declaration

The length of this text, including Introduction and up to the declaration is 13596 words.

I hereby certify that I have independently written this thesis. Any text passages which were not written by me are quoted as citations and specific references to their origins are made. ChatGPT was used to help write this thesis.

All used sources (including images, graphics, etc.) are included in the bibliography.

Bern, 24/05/2023

A handwritten signature in black ink that reads "Pieter-Jan". The script is cursive and fluid, with the first name and last name joined together.

Pieter-Jan Vliegen

Bibliography

- Trippi, R. R., & Turban, E. (1992). *Neural networks in finance and investing: Using artificial intelligence to improve real-world performance*. McGraw-Hill, Inc.
- Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7), e0180944.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.
- Cavalcante, R. C., Brasileiro, R. C., Souza, V. L. F., Nobrega, J. P., & Oliveira, A. L. I. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55, 194-211.
- Chakraborty, C., & Kearns, J. (2011). Market making and mean reversion. *ACM Transactions on Economics and Computation*, 1(1), 1-26.
- Cao, L. (2018). AI in Finance: A Review. *ACM Computing Surveys*, 9(4), Article 39. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>
- Hafezi, M., Shahrabi, J., & Hadavandi, E. (2019). A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock index. *Applied Soft Computing*, 75, 596-619.
- Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689-702.
- Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down Wall Street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys*, 30(2), 356-369.
- Tsai, C. F., & Wang, S. P. (2009). Stock price forecasting by hybrid machine learning techniques. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 223(4), 843-855.
- Dixon, M. F., Klabjan, D., & Bang, J. H. (2016). Classification-based financial markets prediction using deep neural networks. *Algorithmic Finance*, 5(3-4), 35-49.
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.
- Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques—Part II: Soft computing methods. *Expert Systems with Applications*, 36(3), 5932-5941.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

-
- Cao, L. J., & Tay, F. E. H. (2001). Financial forecasting using support vector machines. *Neural Computing & Applications*, 10(2), 184-192.
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987-1007.
- Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, 38(8), 10389-10397.
- Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10), 2513-2522.
- Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, 49(3), 851-889.
- Jorion, P. (2007). *Value at risk: the new benchmark for managing financial risk*. McGraw-Hill.
- Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10(3), 215-236.
[https://doi.org/10.1016/0925-2312\(95\)00039-9](https://doi.org/10.1016/0925-2312(95)00039-9)
- Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, 66(1), 67-97.
- Albuquerque, R., Koskinen, Y., & Zhang, C. (2019). Corporate social responsibility and firm risk: Theory and empirical evidence. *Management Science*, 65(10), 4451-4949.
- Cheema-Fox, A., LaPerla, B., Serafeim, G., & Wang, H. (2020). Decarbonization factors. *The Review of Financial Studies*, 33(9), 4006-4046.
- Bodie, Z., Kane, A., & Marcus, A. J. (2014). *Investments*. McGraw-Hill Education.
- Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset*. John Wiley & Sons.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1-135
- Baker, S. R., Bloom, N., Davis, S. J., Kost, K. J., Sammon, M. C., & Viratyosin, T. (2020). The unprecedented stock market impact of COVID-19. *The Review of Asset Pricing Studies*, 10(4), 742-758.
- Blume, M. E., & Edelen, R. M. (2004). S&P 500 Indexers and the Index Effect. *The Journal of Portfolio Management*, 31(1), 11-18.

- Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65-91.
- Pastor, L., & Veronesi, P. (2012). Uncertainty about government policy and stock prices. *The Journal of Finance*, 67(4), 1219-1264.
- Papanicolaou, R. (2019). yfinance: Yahoo! Finance market data downloader. Retrieved from <https://pypi.org/project/yfinance/>
- Yahoo Finance. (n.d.). Yahoo Finance – Stock Market Live, Quotes, Business & Finance News. Retrieved from <https://finance.yahoo.com/>
- Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2019). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 124, 227-245.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Kaggle. (n.d.). Kaggle: Your home for data science. Retrieved from <https://www.kaggle.com/datasets/miguelaelle/massive-stock-news-analysis-db-for-nlpbacktests>
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.
- Alpaydin, E. (2014). Introduction to machine learning. MIT Press.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.
- Li, Y., Zhang, X., Li, Y., & Li, M. (2019). A survey of reinforcement learning applications in finance. arXiv preprint arXiv:1907.02684.
- Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press..
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.
- Kumar, M., & Bhaskaran, V. (2016). Predictive data mining models
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- Molnar, C. (2020). Interpretable machine learning. Lulu.com.
- Aragón, J., Aranda-Corral, G. A., & Borrego-Díaz, J. (2019). Kaggle dataset: Daily News for Stock Market Prediction. <https://www.kaggle.com/aaron7sun/stocknews>

-
- Armentano, A., Godoy, D., & Amandi, A. (2021). FinBERT: Pretrained language model for financial communications. <https://github.com/ProsusAI/finBERT>
- Garcia, S., Luengo, J., & Herrera, F. (Eds.). (2015). Data preprocessing in data mining. Springer.
- Leek, J. T., & Peng, R. D. (2015). What is the question? *Science*, 347(6228), 1314-1315.
- Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness, and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.
- Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. <https://arxiv.org/abs/1908.10063>
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383-417.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215-242.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation.
- American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.).
- The Journal of Finance. (2023). Guidelines for Accepted Articles.
- Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185.

- Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4), 1187-1230.
- Nguyen, T. T., Shirai, K., & Velcin, J. (2015). Sentiment Analysis on Social Media for Stock Movement Prediction. *Expert Systems with Applications*, 42(24), 9603-9611.
- Publication manual of the American Psychological Association (7th ed.). (2020). American Psychological Association.
- Guidelines for Accepted Articles. (2023). *The Journal of Finance*.
- Cortis, K. (2020). Natural language processing of financial news for sentiment analysis. In *Computational Data and Social Networks*. Springer.
- Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2013). Which news moves stock prices? A textual analysis. NBER Working Paper No. w18725.
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782-796.
- Nguyen, T. H., Shirai, K., & Velcin, J. (2021). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 168, 114398.
- Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3), 653-664.
- Kumar, A., & Thenmozhi, M. (2016). Predictability of financial market activity using Google search volume data and sentiment analysis. *IIMB Management Review*, 28(3), 150-160.
- Sezer, O. B., & Ozbayoglu, A. M. (2018). Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing*, 70, 525-538.
- Zhang, Y., Zhang, Y., & Liu, Y. (2019). Ensemble of deep learning models for financial data prediction. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)* (pp. 446-450). IEEE.
- Poulinakis, K. (2020). Predict SP500 Stock Price with Python, Machine Learning & Sentiment Analysis. Medium. <https://medium.com/mlearning-ai/predict-sp500-stock-price-with-python-machine-learning-sentiment-analysis-a296dc276353>
- Poulinakis, K. (2021). Stocks News Sentiment Analysis with Deep Learning Transformers and Machine Learning. Medium. <https://medium.com/codex/stocks-news-sentiment-analysis-with-deep-learning-transformers-and-machine-learning-cdcdb827fc06>

Appendix A. Architecture Structure and Results Data

Figure A1:

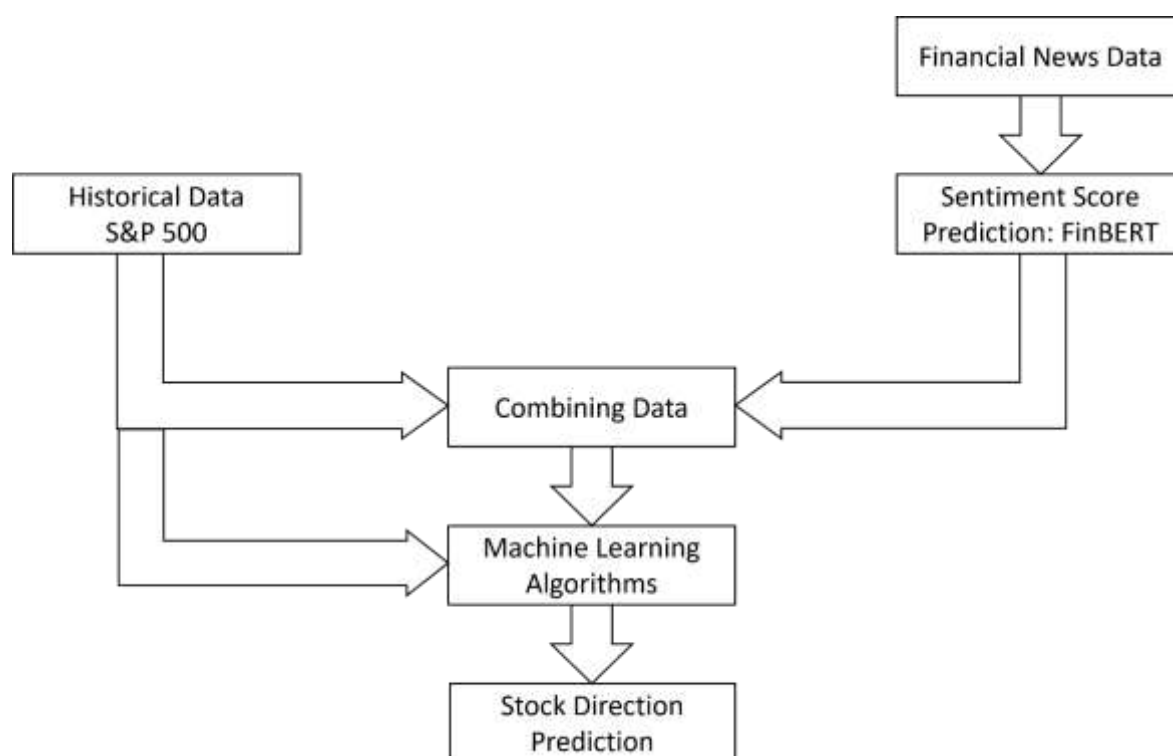


Figure A1 provides a visual representation of the research methodology employed in this study to predict stock market directions based on historical financial data and news sentiment analysis.

Starting from the top, we begin with two distinct data sources. The first is the historical financial data of the S&P500 index. This data includes various financial indicators over the specified period. The second source is the financial news data, which has been subject to sentiment analysis using the FinBERT model, a state-of-the-art tool designed specifically for financial sentiment analysis. This results in a sentiment score for each piece of news.

The historical S&P500 financial data is fed directly into the machine learning algorithm. Simultaneously, the sentiment scores derived from the FinBERT analysis of financial news are combined with the historical financial data. This combined dataset, which now includes both historical financial indicators and sentiment scores, is also inputted into the machine learning algorithm.

The chosen machine learning algorithm processes the input data, learning from the historical trends and sentiment scores to predict future market direction. This process is represented in the figure by the arrows pointing from the input data to the machine learning algorithm.

The output from the machine learning algorithm is the prediction of the future direction of the S&P500 index. This prediction is based on the model's learning from both the historical financial data and the sentiment analysis scores.

In summary, Figure A1 illustrates the architecture of the proposed model for predicting stock market directions, integrating both historical financial data and news sentiment scores into a machine learning model.

Table A1:				
Machine Learning Algorithm			Financial Data	Combined Data
Logistic Regression	Solver lbfgs	Mean accuracy	0.5482	0.5729
		F1 macro	0.3541	0.5016
		Mean precision	0.5482	0.5711
		Mean recall	0.5000	0.5432
	Solver liblinear	Mean accuracy	0.5482	0.5740
		F1 macro	0.3541	0.5033
		Mean precision	0.5482	0.5718
		Mean recall	0.5000	0.5444
	Solver sag	Mean accuracy	0.5482	0.5729
		F1 macro	0.3541	0.5016
		Mean precision	0.5482	0.5711
		Mean recall	0.5000	0.5432
	Solvser saga	Mean accuracy	/	0.5729
		F1 macro	/	0.5016
		Mean precision	/	0.5711
		Mean recall	/	0.5432
	Solver newton-cg	Mean accuracy	/	0.5729
		F1 macro	/	0.5016
		Mean precision	/	0.5711
		Mean recall	/	0.5432
	Accuracy on test set		0.5578	0.5826
Nearest Neighbors	K = 2	Mean accuracy	0.4905	0.4984
		F1 macro	0.4765	0.4869
		Mean precision	0.5690	0.5717
		Mean recall	0.5108	0.5160
	K = 5	Mean accuracy	0.4928	0.5223
		F1 macro	0.4834	0.5141
		Mean precision	0.5352	0.5580
		Mean recall	0.4844	0.5148
	K = 10	Mean accuracy	0.5051	0.5080
		F1 macro	0.5045	0.8061
		Mean precision	0.5553	0.5519
		Mean recall	0.5068	0.5068
	K = 50	Mean accuracy	0.5231	0.5617

		F1 macro	0.4915	0.5372
		Mean precision	0.5511	0.5783
		Mean recall	0.5039	0.5455
	K = 100	Mean accuracy	0.5350	0.5772
		F1 macro	0.4744	0.5438
		Mean precision	0.5524	0.5848
		Mean recall	0.5069	0.5574
	K = 200	Mean accuracy	0.5430	0.5782
		F1 macro	0.4136	0.5324
		Mean precision	0.5496	0.5804
		Mean recall	0.5025	0.5545
	Accuracy on test set		0.5457	0.5776
SVM	kernel poly	Mean accuracy	0.5482	0.5585
		F1 macro	0.3541	0.4673
		Mean precision	0.5482	0.5598
		Mean recall	0.5000	0.5256
	kernel rbf	Mean accuracy	0.5454	0.5745
		F1 macro	0.3556	0.5099
		Mean precision	0.5470	0.5730
		Mean recall	0.4976	0.5462
	kernel sigmoid	Mean accuracy	0.5170	0.5154
		F1 macro	0.5112	0.5138
		Mean precision	0.5594	0.5591
		Mean recall	0.5119	0.5148
	Kernel linear	Mean accuracy	/	0.5473
		F1 macro	/	0.3714
		Mean precision	/	0.5471
		Mean recall	/	0.5036
	Accuracy on test set		0.5578	0.5850
Random For- est	Criterion entropy, nodes = 2	Mean accuracy	0.5365	0.5210
		F1 macro	0.3905	0.4674
		Mean precision	0.5529	0.5631
		Mean recall	0.5010	0.5115
	Criterion entropy, nodes = 3	Mean accuracy	0.5389	0.5773
		F1 macro	0.3961	0.5247
		Mean precision	0.5542	0.5810

		Mean recall	0.5036	0.5506
	Criterion entropy, nodes = 5	Mean accuracy	0.5286	0.5762
		F1 macro	0.4216	0.5267
		Mean precision	0.5556	0.5809
		Mean recall	0.5037	0.5502
	Criterion entropy, nodes = 10	Mean accuracy	0.5326	0.5769
		F1 macro	0.4299	0.5245
		Mean precision	0.5590	0.5814
		Mean recall	0.5084	0.5504
	Criterion entropy, nodes = 15	Mean accuracy	0.5382	0.5751
		F1 macro	0.4599	0.5213
		Mean precision	0.5696	0.5799
		Mean recall	0.5198	0.5484
	Criterion entropy, nodes = 20	Mean accuracy	0.5130	0.5740
		F1 macro	0.4444	0.5165
		Mean precision	0.5842	0.5784
		Mean recall	0.5133	0.5463
	Criterion entropy, nodes = 50	Mean accuracy	0.5551	0.5684
		F1 macro	0.4944	0.5185
		Mean precision	0.5906	0.5773
		Mean recall	0.5392	0.5432
	Criterion gini, nodes = 2	Mean accuracy	0.5365	0.5210
		F1 macro	0.3905	0.4674
		Mean precision	0.5529	0.5631
		Mean recall	0.5010	0.5115
	Criterion gini, nodes = 3	Mean accuracy	0.5415	0.5773
		F1 macro	0.4021	0.5247
		Mean precision	0.5557	0.5810
		Mean recall	0.5066	0.5506
	Criterion gini, nodes = 5	Mean accuracy	0.5432	0.5755
		F1 macro	0.4058	0.5266
		Mean precision	0.5567	0.5808
		Mean recall	0.5084	0.5497
	Criterion gini, nodes = 10	Mean accuracy	0.5465	0.5725
		F1 macro	0.4442	0.5199
		Mean precision	0.5631	0.5781

		Mean recall	0.5193	0.5460
	Criterion gini, nodes = 15	Mean accuracy	0.5442	0.5728
		F1 macro	0.4322	0.5226
		Mean precision	0.5564	0.5799
		Mean recall	0.5224	0.5479
	Criterion gini, nodes = 20	Mean accuracy	0.5610	0.5684
		F1 macro	0.4327	0.5227
		Mean precision	0.5609	0.5781
		Mean recall	0.5190	0.5445
	Criterion gini, nodes = 50	Mean accuracy	0.5395	0.5486
		F1 macro	0.4723	0.5186
		Mean precision	0.5995	0.5710
		Mean recall	0.5384	0.5302
	Accuracy on test set		0.5578	0.5763