

**T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL**

**INVESTIGATING THE IMPACT OF SOCIAL NETWORKS
ON STOCK MARKET WITH SENTIMENT ANALYSIS
AND MACHINE LEARNING METHODS:
AN APPLICATION ON BIST 30**

MUSTAFA KEMAL MAYUK

**A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF BUSINESS ADMINISTRATION**

ADVISOR: ASSOC. PROF. DR. FARİD HUSEYNOV

DECEMBER 2024

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**INVESTIGATING THE IMPACT OF SOCIAL NETWORKS ON STOCK
MARKET WITH SENTIMENT ANALYSIS AND MACHINE LEARNING
METHODS: AN APPLICATION ON BIST 30**

MUSTAFA KEMAL MAYUK

A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF BUSINESS ADMINISTRATION

ADVISOR: ASSOC. PROF. DR. FARİD HUSEYNOV

DECEMBER 2024

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SOSYAL AĞLARIN HİSSE SENEDİ BORSASINA ETKİSİNİN DUYGU
ANALİZİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE
ARAŞTIRILMASI: BIST 30 ÜZERİNE BİR UYGULAMA

MUSTAFA KEMAL MAYUK

YÜKSEK LİSANS TEZİ
İŞLETME ANABİLİM DALI

DANIŞMAN: DOÇ. DR. FARİD HUSEYNOV

ARALIK 2024

MASTER Of SCIENCE APPROVAL FORM

A thesis submitted by Mustafa Kemal MAYUK, defended on 19/12/2024 before the jury formed with the 16/12/2024 date and 2024/66 numbered decision of the GTU Graduate Administration Board, has been accepted as a MASTER of SCIENCE thesis in the Department of Business Administration.

JURY

MEMBER

(THESIS ADVISOR) : Assoc. Prof. Dr. Farid HUSEYNOV

MEMBER : Assoc. Prof. Dr. Mehtap ÖZŞAHİN

MEMBER : Assist. Prof. Dr. Beyhan Hilal YASLIDAĞ

APPROVAL

Gebze Technical University Graduate Administration Board
...../...../..... date and/..... numbered decision.

SIGNATURE/SEAL



To my dear spouse and daughter,

ABSTRACT

With the development and widespread use of portable devices such as mobile phones, the emergence of social networks operating on these devices has enabled people to share their feelings and thoughts instantly. Whether the opinions produced through social networks affect the behavior of individuals is the subject of many studies. Today financial behavior is a part of daily human behaviors.

Classical economic theory is based on consumer as a rational being. Today, there are many studies claiming that there are irrational factors affecting people's financial behavior. The fact that individuals influence each other's behavior through verbal, written or visual means, especially the shaping and formation of consumer behavior through these means has led to the formation of research areas such as behavioral economics. Today, social networks are the biggest channels through which people can share and spread their verbal, written or visual messages.

Just like a product or service, the stock market is a market with buyers and sellers, so the buying and selling decisions of investors in the stock market can be influenced by the thoughts, words and news spread by other investors in the market. The stock market differs from products or services only in that it functions as an investment instrument. Stocks are not bought or sold for function, convenience or social status, as in the case of goods or services, but only for future price expectations. In the stock market, fundamental and technical analysis methods have been used for many years to predict the direction of the price and the point it will reach. While technical analysis is based on various mathematical techniques on temporal price charts, fundamental analysis evaluates companies' balance sheets, micro and macro-economic factors together. Depending on whether the investment strategy is long or short term, investors may use these methods with different weights.

It would not be correct to say that all investors in the market invest only on the basis of technical and fundamental analysis methods. Emotions, influences, news flow, and the emotional states of individuals depending on the developments in the country and the world can also affect the trading decisions of stocks. This study aims to examine this effect that may occur through social networks by using data science methods such as sentiment analysis and regression.

Keywords: Machine Learning, Social Networks, Artificial Intelligence, Data Science, Behavioral Economics, Financial Technologies

ÖZET

Cep telefonları gibi taşınabilir cihazların ve bu cihazlar üzerinde çalışan sosyal ağların gelişmesi ve yaygınlaşması, insanların düşünce ve duygularını anlık olarak paylaşabilmelerine olanak tanıdı. Sosyal ağlar aracılığıyla üretilen kanaatlerin bireylerin davranışlarını etkileyip etkilemediği birçok araştırmanın konusudur. Günümüzde finansal davranış da bireylerin günlük davranışlarından biridir.

Klasik iktisat bireylerin rasyonel varlıklar olduğu düşüncesine dayanır. Günümüzde, iktisadi davranışın arkasında irrasyonel etkenlerin de olduğunu ortaya koyan birçok araştırma mevcuttur. Bireylerin birbirlerinin davranışlarını sözlü, yazılı ya da görsel yollarla etkilemesi, özellikle tüketici davranışlarının bu yollarla şekillenmesi ve oluşması, davranışsal ekonomi gibi araştırma alanlarının oluşmasına neden olmuştur. Günümüzde sosyal ağlar, bireylerin tüketim ya da finansal davranışlarını etkileyebilecek sözlü, yazılı ve görsel içeriklerin paylaşılıp yayılabildiği en büyük ortamlardır.

Hisse senedi piyasası da tıpkı bir ürün ya da hizmet gibi alıcısı ve satıcısı bulunan bir pazardır, bu nedenle hisse senedi piyasasındaki yatırımcılarının alım-satım kararlarında da piyasa içindeki diğer yatırımcıların düşünceleri, sözleri ve yaydıkları haberler etkili olabilir. Ürün ya da hizmetten farklı olarak hisse senetleri sadece yatırım aracı olma işlevine sahiptirler. Hisse senetleri, ürün ya da hizmetlerdeki gibi işlev, kolaylık, sosyal statü gibi beklentilerle değil sadece ve sadece gelecekteki fiyat beklentileri için alınır ya da satılırlar. Hisse senedi piyasasında fiyatın yönünü ve ulaşacağı noktayı öngörmek için temel ve teknik analiz yöntemleri uzun yıllardır kullanılmaktadır. Teknik analiz zamansal fiyat grafikleri üzerinde çeşitli matematiksel tekniklere dayanırken temel analiz ise şirketlerin bilançolarını, mikro ve makroekonomik etkenleri bir arada değerlendirir. Yatırım stratejisinin uzun ya da kısa vadeli olmasına bağlı olarak yatırımcılar bu yöntemleri farklı ağırlıklarda kullanabilirler.

Piyasadaki tüm yatırımcıların sadece teknik ve temel analiz yöntemlerine bağlı kalarak yatırım yaptıklarını söylemek doğru olmaz. Duyum, etkilenme, haber akışı, bireylerin ülkedeki ve dünyadaki gelişmelere bağlı olarak içinde bulundukları duygu durumları da hisselerin alım-satım kararlarını etkileyebilmektedir. Bu çalışma, sosyal ağlar aracılığıyla oluşabilecek bu etkiyi duygu analizi, regresyon gibi veri bilimi yöntemlerini kullanarak incelemeyi amaçlamaktadır.

Anahtar Kelimeler: Makine Öğrenmesi, Sosyal Ağlar, Yapay Zekâ, Veri Bilimi, Davranışsal Finans, Finansal Teknolojiler

ACKNOWLEDGEMENTS

I am grateful to my supervisor Assoc. Prof. Dr. Farid HUSEYNOV, who has always provided technical and academic guidance in this study. Thanks to his suggestions and encouragement, this study was designed and shaped.

I would like to express my gratitude to my beloved spouse, Assoc. Prof. Dr. Seher GÜZELÇOBAN MAYUK, who has always encouraged me with her understanding, guidance and inspiration to conduct academic studies.



TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	vi
ÖZET	vii
ACKNOWLEDGEMENTS	viii
TABLE OF CONTENTS	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xi
LIST OF FIGURES	xii
LIST OF TABLES	xiii
1. INTRODUCTION	1
1.1. Background Information	2
1.2. Purpose of Study	3
1.3. Importance of Study	3
1.4. Research Questions	4
1.5. Structure of Thesis	5
2. LITERATURE	6
2.1. Stock Market	6
2.1.1. Definitions and Basic Concepts	6
2.1.2. Stock Market Indexes	7
2.1.3. Borsa Istanbul (BIST)	9
2.1.4. BIST 30 Index	10
2.2. Common Stock Analysis Methods	11
2.2.1. Fundamental Analysis	12
2.2.2. Technical Analysis	13
2.3. Behavioral Finance	14
2.4. Machine Learning	16
2.4.1. Definitions and Basic Concepts	16
2.4.2. Machine Learning Types and Algorithms	18
2.4.3. Neural Networks and Deep Learning	19
2.5. Natural Language Processing (NLP)	23
2.5.1. Recurrent Neural Networks and Long Short-Term Memory	24
2.5.2. Transformers and Large Language Models	26
2.5.3. Sentiment Analysis	28
2.6. Use of AI and Machine Learning in Finance	29
3. METHODOLOGY	32
3.1. Data Types, Characteristics and Sources	32
3.2. Technologies and Tools	34
3.3. Obtaining the Data	35
3.3.1. Collecting Tweets about Stocks in the BIST 30 Index	35
3.3.2. Obtaining BIST 30 Index Closing Values	36
3.3.3. Obtaining Closing Prices of Stocks in BIST 30	36
3.4. Data Pre-processing	37

3.5. Sentiment Calculation of Tweet Data	39
3.6. Creation of Final Datasets	40
4. FINDINGS	43
4.1. Data Exploration	43
4.2. Regression Model	49
4.3. Correlation Coefficients and p-values	52
5. DISCUSSION and CONCLUSION	56
5.1. Theoretical Contributions	56
5.2. Practical Contributions	58
5.3. Limitations and Future Research Suggestions	59
REFERENCES	61
BIOGRAPHY	66
PUBLICATIONS AND PRESENTATIONS FROM THE THESIS	67



LIST OF SYMBOLS AND ABBREVIATIONS

BIST	: Borsa Istanbul
BIST 30	: Index of the 30 most active stocks in Borsa Istanbul
BIST 50	: Index of the 50 most active stocks in Borsa Istanbul
BIST 100	: Index of the 100 most active stocks in Borsa Istanbul
BIST 500	: Index of the 500 most active stocks in Borsa Istanbul
XU030	: International code of BIST 30
XU050	: International code of BIST 50
XU100	: International code of BIST 100
XU500	: International code of BIST 500
SPK	: Sermaye Piyasası Kurulu, stock market regulator in Türkiye
MACD	: Moving Average Convergence/Divergence
EMA	: Exponential Moving Average
AROON	: Aroon Oscillator
SO	: Stochastic Oscillator
RSI	: Relative Strength Index
OBV	: On-balance Volume
ADL	: Accumulation/Distribution Line
EMH	: Efficient Market Hypothesis
AI	: Artificial Intelligence
ML	: Machine Learning
ANN	: Artificial Neural Network
DNN	: Deep Neural Network
MLP	: Multi-Layer Perception
TLU	: Threshold Logic Unit
RNN	: Recurrent Neural Network
CNN	: Convolutional Neural Network
LSTM	: Long Short-Term Memory
LLM	: Large Language Model
BERT	: Bidirectional Encoder Representations from Transformers
SVM	: Support Vector Machine
SVR	: Support Vector Regression
R-Squared	: Variance of dependent variable explained by independent in regression
MAE	: Mean Absolute Error
MSE	: Mean Squared Error
RMSE	: Root Mean Squared Error

LIST OF FIGURES

	<u>Page</u>
Figure 2.1: MACD Applied BIST 30 Closing Graph	14
Figure 2.2: An Example of ML Overfitting	18
Figure 2.3: Basic Architecture of Deep Neural Network	20
Figure 2.4: Single Artificial Neural Network	20
Figure 2.5: Sigmoid Function	21
Figure 2.6: Multi-Variate Neural Network	21
Figure 2.7: Image Classification	22
Figure 2.8: DNN Extracts Complex from Simple Features	23
Figure 2.9: Architecture of RNN	25
Figure 2.10: Architecture of LSTM	25
Figure 2.11: Abstracted Architecture of Transformer	27
Figure 3.1: The Process of Obtaining Data Before Analysis	32
Figure 3.2: Data Cleaning and Pre-Processing Process	39
Figure 3.3: Testing Sentiment Analysis Model	40
Figure 3.4: Detailed Process of Creating Final Datasets	41
Figure 4.1: Stock Codes and Number of Tweets for Each Stock	44
Figure 4.2: Average Remaining Time to Close Distribution of Tweets	44
Figure 4.3: Trends of BIST 30 Components Between 2021 and 2022	45
Figure 4.4: BIST 30 Index Closing Trend Between 2021 and 2022	46
Figure 4.5: Average Closing of BIST 30 Stocks Between 2021 and 2022	46
Figure 4.6: Tweeting Activity over Months and Years	47
Figure 4.7: SASA Specific Tweeting Activity	47
Figure 4.8: Distribution of Tweet Sentiments	48
Figure 4.9: Clustering Model Between Sentiment and Average Closing	49
Figure 4.10: Linear Regression and SVR	50

LIST OF TABLES

	<u>Page</u>
Table 2.1: Indicator Basis Indexes in BIST	8
Table 2.2: Sector Basis Indexes in BIST	8
Table 2.3: BIST 30 Stock Components	10
Table 2.4: Most Commonly Used Ratios in Fundamental Analysis	12
Table 2.5: Most Widely Used Technical Indicators for Technical Analysis	14
Table 2.6: RNN and LSTM Comparison	16
Table 2.7: Comparison of The Most Commonly Used LLMs	28
Table 3.1: BIST 30 Index Closing Data for The First Five Days of 2021	36
Table 3.2: BIST 30 Stock Closings Sample	37
Table 3.3: Usage Of Fields for Each Dataset	41
Table 4.1: Stocks With a Number of Tweets Above the Average Number of Tweets	43
Table 4.2: Most Tweeted Stocks and Standard Deviation of Sentiment Scores	48
Table 4.3: SVR Metrics of BIST 30 Index	51
Table 4.4: SVR Metrics of Consolidated Data of BIST 30 Component Stocks	51
Table 4.5: R-Squared Metric of Individual Stocks	52
Table 4.6: Correlation Coefficients and P-Values for BIST 30 Index	53
Table 4.7: Correlation Coefficients and P-Values for BIST 30 Component Stocks	53
Table 4.8: Correlation Coefficients of Individual Stocks	54
Table 4.9: Distribution of Correlation Coefficient Values of Stocks	55
Table 4.10: P-Values for Sample Stocks	55

1. INTRODUCTION

Forecasting is one of the most fundamental activities that many professional traders in the stock market use in their decision-making process. Over the years, these techniques have evolved considerably and with the introduction of computers, both technical and fundamental analysis methods have come a long way. The rapid computability of statistical methods through computers has contributed to this. Although computers facilitate time-series analysis of prices and the accounting of macro- and microeconomic parameters, there are other challenges in stock market forecasting. Many people do not make purely rational decisions, and in the stock market, as in the commodity market, investors' emotions, beliefs and psychology can often be influential. The main difficulty in price forecasting arises from the difficulty in modeling human behavior (Schumaker et al., 2012).

Although predicting the future is difficult, it has been a topic that has attracted the attention of all researchers. Likewise, predicting financial markets and the impact of technology on finance have been among the most prominent research topics (Nasir et al., 2019). The difficulty in predicting financial markets is also expressed according to the "efficient market hypothesis". According to this hypothesis, opportunities in financial markets are seized when they arise, prices move randomly and are difficult to predict. Therefore, technical and fundamental forecasting methods used in financial markets focus on identifying buying and selling zones rather than predicting prices (Gidofalvi and Elkan, 2001).

Until the development of smartphones and social media, it was only possible to gain insight into human behavior through people's own statements. However, the rapid increase in the amount of data that people voluntarily create after recent developments makes new perspectives possible in the field of economics and social sciences (Rancho et al., 2015). Thanks to social networks such as Twitter, the last few decades have seen the most idea-generating data generation period in human history. This data contains a lot of information about people's behavior. Many companies have been established and many studies have been initiated in institutions to mine ideas on this data (Liu, 2012).

Using this abundance of data and capabilities of disruptive technologies have opened the door to inherently complex and unpredictable research topics.

1.1. Background Information

Other disciplines besides data science are often needed to study sense of data. As the volume, variety, velocity, veracity and value of the data increases, the data becomes "big data". To work with big data, it is imperative that disciplines such as software engineering, data engineering, data science and domain knowledge work together. Especially in recent years, additional disciplines such as artificial intelligence and high-performance computing are also needed, and data science is evolving with the development of technology (Saltz and Grady, 2017).

The domain or business expertise for this study is financial markets, a branch of finance. More specifically, the stock markets and behavioral finance should be examined. To conduct this study, the basic mechanisms of the stock market and existing methods used in forecasting are examined.

Software and data engineering skills are used to extract data from digital data sources, store, process and prepare the data. This knowledge is a necessity in data science studies in order to research, build and test the necessary software libraries and configure them according to the needs. The creation of development platforms may also require not only software and data engineering knowledge but also operating system knowledge.

Data science is key to this kind of studies. It is the responsibility of the data scientist to prepare the data, find appropriate algorithms and develop the model. However, since the main purpose of such studies is to do data science and make sense out of data, the role of data science in organizations is not limited to these functions. The data scientist should follow the entire process from end to end, understand the data equivalent of the questions in question for data science, and support engineering roles where necessary (Saltz and Grady, 2017).

As mentioned before, the size of the studies and the volume of data used can affect the level of information needed. Within the scope of this study, a sufficient level of domain knowledge, software and data engineering and data science knowledge is sufficient.

1.2. Purpose of Study

Human behavior is one of the most important parameters that cause the difficulty in researching and studying phenomena in social sciences. The fact that human psychology can also affect prices, especially when it comes to consumer or investor behavior, is also a subject of behavioral finance, which has been extensively studied in recent years (Almansour et al., 2023). However, the availability of large amounts of voluntarily-generated opinion data, and the technology and capabilities to make sense of it, offer great opportunities for both forecasting in financial markets and behavioral finance to capture patterns of investor behavior.

This study aims to contribute to the studies on financial markets by utilizing emerging technologies. In addition to the use of classical data science methods, it is aimed to present an example of the use of artificial intelligence topics such as natural language processing, which has recently been the subject of more research, in the field of behavioral finance. The main focus of this study is to contribute to both theoretical and practical areas, and to develop insights into both the economic and scientific value of data. This study aims to do all this by examining the relationship between social network data containing opinions and the pricing of a financial market instrument, the stock market.

1.3. Importance of Study

Today, there is a lot of talk about data as a new economic resource. Almost all companies, institutions and governments are engaged in many projects and studies on making sense of data. Also in social sciences, as in all fields, data science can provide answers to many research questions. Surveys have been a long-standing way of gathering people's opinions. Today, data is generated naturally and in large quantities as a result of people's digital activity.

This study aims to contribute to the existing social science studies which use data science by presenting an application example. Such studies, which provide examples of the economic and scientific contributions of data, are very important. However, this study will not only focus on the contributions of making sense of data, but also on ways to learn the general opinion of the society by using technology. Although this

study specifically examines the relationship between sentiment analysis from investors' social network data and the stock market, emphasizes more general goals, such as identifying the general opinion on a topic.

Opinions shape our choices, and understanding them is crucial for both individuals and organizations. Traditionally, gathering opinions involved personal networks and focused research, but it has become a substantial business itself. Studying on opinions can provide great opportunities in science, finance, marketing, business and management (Liu, 2012).

1.4. Research Questions

This study aims to investigate the relationship between Twitter data, a social network example, and the Turkish stock market. The effects of quantity, sentiment and time variables in Twitter data on the closing prices of both stocks and the index will be analyzed.

Daily basis, tweets about stocks between the opening and closing hours of the stock exchange are included in the scope of this study. The number of daily tweets about a stock is a variable that can provide insight into social media activity about that stock (Liu et al., 2023). Recent research supports the hypothesis that information in twitter data may be statistically related to financial indicators (Souza et al., 2015). The timing of tweets about stocks can also have an impact on the closing value. There are studies indicate that a certain period of time must pass for social media data to have an impact on the stock market (Yu et al., 2012).

In line with previous research, it is possible to examine the impact of the number of tweets, the sentiments in the content of the tweets, and the time between the time stamps of the tweets and the closing time of the stock market on the closing values of the index and stocks.

1. Does the number of tweets have an effect on the closing value of the BIST 30 index, average closing prices of stocks and closing prices of individual stocks?
2. Does the average sentiment score have an effect on the closing value of the BIST 30 index, average closing prices of stocks and closing prices of individual stocks?

3. Does the time remaining until the closing time of stock market have an effect on the closing value of the BIST 30 index, average closing prices of stocks and closing prices of individual stocks?

1.5. Structure of Thesis

The introduction briefly describes the background of the study. In this section, it is explained that the study will seek answers to research questions related to both financial asset forecasting and behavioral economics by using data science methods. The following sections are titled as literature, data and methodology, findings, discussion and conclusion.

In the literature section, the theoretical background of the areas subject to the study is explained. In this section, the main topics are the stock market, behavioral finance, and machine learning and artificial intelligence, where the technical information used in the study is explained. Basic concepts and definitions related to the stock market and current analysis methods on the stock market are given. The current research on behavioral finance and the importance of the subject today are also discussed. In the literature chapters on the technical knowledge of the application, natural language processing, which is the main topic of the sentiment analysis method used in this study, and the classical machine learning methods used afterwards are mentioned.

The data and methodology section examines how tweet data is collected and processed, how sentiment analysis is performed, and how the data is combined with price information into final datasets.

The findings section presents both statistical results and regression performance metrics on the final data sets, correlation between variables and p-value data.

In the discussion and conclusion section, the theoretical and practical contributions of the study, limitations, challenges and suggestions for future studies are explained.

2. LITERATURE

In this section, the basic concepts of the headings, which are the components of this study, and other related studies in the literature are mentioned. This study has two main components: the stock market and machine learning. Within the stock market topic, it is possible to talk about basic concepts and analysis methods.

Machine learning techniques, which are the main methodology of this study, are another important topic. The commonalities and differences of concepts such as artificial intelligence, machine learning and deep learning and their relevance to this study will be discussed.

This study can be considered as a behavioral finance study as it aims to examine the impact of human behavior on financial markets by measuring it in some way. For this reason, a brief behavioral finance literature is also mentioned.

2.1. Stock Market

The stock market is one of the main metrics of a country's economy, widely mentioned in our daily lives and followed by the media. The stock market is one of the leading indicators for the real economy and one of the most important factors affecting individuals' participation in the economy. Since investors' participation in the stock market is important, confidence in stock markets is also important (Constantinescu, 2012).

2.1.1. Definitions and Basic Concepts

Stock market also can be called as equity market or share market is the aggregation of buyers and sellers of stocks. In financial markets, a stock (or a share) is a unit of equity ownership within a company's capital stock and may refer to units of a mutual fund, limited partnership and real estate investment trust. A share or share certificate is one of the equivalent parts of a company's capital stock. Buyers of a company's shares become shareholders of that company. A share certificate expresses the relationship between the company and the shareholder.

Companies may distribute dividends per share to their shareholders in proportion to their profits in certain periods. This is called a dividend, and people who buy shares in companies that pay dividends specifically to collect income in this way are called dividend investors. Dividend investing is another way to earn returns without buying and selling stocks.

Confidence is an important topic for participation of investors to stock market (Constantinescu, 2012). There are institutions in charge of monitoring and enforcing these regulations. In Türkiye, this task belongs to the Capital Markets Board (Sermaye Piyasası Kurulu or SPK).

Investors can invest in stocks of different companies in any proportion according to various strategies. They can determine how much to invest in which company's stock according to their expectations based on macro or micro economic factors. The information on which stocks an investor owns and in what proportion is called the investment portfolio.

This paper is based on stocks themselves, but there are also assets that take stocks as their underlying financial asset and are themselves different financial instruments. Funds and derivatives are the most well-known examples. Many investment firms may market their own portfolios as funds, which can be exchanged in the same way as stocks. Derivatives, on the other hand, are financial assets that can be traded on a highly leveraged basis through contracts with forward time or price. They can use underlying such as foreign exchange, precious metals, and, quite commonly, stocks as underlying assets. The spot market nomenclature for the underlying asset is often used to distinguish between the derivatives market and the actual underlying asset market, which is the subject of this study.

2.1.2. Stock Market Indexes

Indexes are created to measure the performance of stocks as a whole or by sector. In Türkiye, there are indicator indexes such as XU100 (or BIST 100) for the 100 most traded stocks or sector indexes such as XBANK for the bank stocks. There are also more specific indices such as the dividend stock index and the participation finance index.

The stocks that make up the index components are reviewed periodically and updates are made if necessary. The index makes it easy to track the trend of the stock market in general or on the basis of specific characteristics such as sectors. Table 2.1 shows the indicator indices in the Turkish stock exchange together with the dates of their creation (BIST-1, 2024).

Table 2.1: Indicator Basis Indexes in BIST

Index Name	Index Code	Creation Date
BIST 100	XU100	01.01.1986
BIST 50	XU050	28.12.1999
BIST 30	XU030	27.12.1996
BIST 500	XU500	01.11.2023

Table 2.2 shows sector basis indexes with their detailed names, codes and creation dates.

Table 2.2: Sector Basis Indexes in BIST

Index Name	Index Code	Creation Date
BIST INDUSTRY	XUSIN	28.12.1990
BIST SERVICES	XUHIZ	27.12.1996
BIST FINANCIAL	XUMAL	28.12.1990
BIST TECH.	XUTEK	30.06.2000
BIST TECH. WEIGHT LIMITED	XTKJS	01.12.2016
BIST BANKING	XBANK	27.12.1996
BIST ARACI KURUMLAR	XAKUR	04.11.2019
BIST BILISIM	XBLSM	30.06.2000
BIST ELEKTRIK	XELKT	27.12.1996
BIST FIN. KIR. FAKTORING	XFINK	27.12.1996
BIST GAYRIMENKUL Y.O.	XGMYO	28.12.1999
BIST FOOD	XGIDA	27.12.1996
BIST HOLDING VE YATIRIM	XHOLD	27.12.1996
BIST TELECOMMUNICATION	XILTM	24.07.2000
BIST INSAAT	XINSA	01.02.2013
BIST WOOD PAPER PRINT	XKAGT	27.12.1996
BIST CHEM. PETROL PLASTIC	XKMYA	27.12.1996
BIST MINING	XMADN	01.02.2013
BIST MENKUL KIYM. Y.O.	XYORT	27.12.1996
BIST METAL MAIN	XMANA	27.12.1996
BIST METAL ASSET MACHINES	XMESY	27.12.1996
BIST INSURANCE	XSGRT	27.12.1996
BIST SPORT	XSPOR	31.03.2004
BIST STONE and SOIL	XTAST	27.12.1996
BIST TEXTILE and LEATHERR	XTEKS	27.12.1996
BIST TICARET	XTCRT	27.12.1996
BIST TOURISM	XTRZM	27.12.1996
BIST TRANSPORTATION	XULAS	27.12.1996

BIST indices are calculated as free float market capitalization weighted by using several parameters (WEB-2, 2020).

$$E_t = \frac{\sum_{i=1}^n (F_{it}/D_t) * N_{it} * H_{it} * K_{it}}{B_t} \quad (2.1)$$

E_t	Value of the index at time t
n	Number of constituent stocks
F_{it}	Price of the stock “i” at time t
N_{it}	Total number of shares of stock “i” at time t
H_{it}	Free float ratio used in index calculation of the stock “i” at time t
K_{it}	Weighting factor of the stock “i” at time t
D_t	Exchange rate at time t
B_t	Divisor of the index at time t

Similar stock indices are available in all countries or trading regions. In the United States, for example, the NASDAQ-100 and the S&P 500 are the two most important main indices followed. The NASDAQ-100 includes mostly stocks of technology companies, while the S&P 500 is the index of the 500 largest US companies. Other important indices include the CSI 1000 in China, FTSE 100 in the UK, the DAX in Germany, the CAC 40 in France and the Nikkei 225 in Japan.

2.1.3. Borsa İstanbul (BIST)

Borsa Istanbul, or BIST, was established in 1985 under the short name IMKB to provide custody and settlement services to investment firms and brokerage houses. Stock trading activities within Borsa Istanbul are monitored by the state's SPK institution (capital markets institution of Türkiye, abbreviation for Sermaye Piyasası Kurulu) and regulations are applied, when necessary, by this institution. The SPK is responsible for protecting investor rights, ensuring the enforcement of capital market laws, and ensuring the fair and efficient functioning of the securities market. In addition to its duties of protection, monitoring, supervision and regulation, this institution also has the authority to impose penal sanctions on those who do not comply with the rules.

Today, there are many digital opportunities for investors to trade on BIST. They can use the applications of banks as well as the platforms provided by various brokerage houses to trade stocks. Brokerage houses are required to hold various certificates provided by the SPK and are subject to the SPK's regulations. Companies traded on the stock market are obliged to announce their balance sheets and activities to the

public at regular intervals. Similarly, investment firms that do not directly offer stocks but market fund-like products must also publish fund allocations and detailed information about the fund. All this information is published on a platform called KAP (public disclosure platform, abbreviation for Kamu Aydınlatma Platformu).

2.1.3. BIST 30 Index

The BIST 30 Index, which is the subject of this study, is an index formed by measuring the performance of the stocks of the 30 companies traded in Borsa Istanbul with the highest trading volume and market capitalization. The stocks included in the BIST 30 index are also included in the BIST 50 and BIST 100 indices.

The components of the BIST 30 index are reviewed quarterly and may be updated if necessary. The subject of this study is the BIST 30 constituents in 2021 and 2022. Table 2.3 shows stock components of BIST 30 (WEB-1, 2024).

Table 2.3: BIST 30 Stock Components

Stock Code	Company Name
AKBNK	Akbank
AKSEN	Aksa Enerji
ALARK	Alarko Holding
ARCLK	Arçelik
ASELS	Aselsan Elektronik Sanayi ve Ticaret
BIMAS	BİM Birleşik Mağazalar
EKGYO	Emlak Konut GYO
EREGL	Ereğli Demir ve Çelik Fabrikaları
FROTO	Ford Otomotiv Sanayi
GARAN	Türkiye Garanti Bankası
GUBRF	Gibre Fabrikaları
HEKTS	Hektaş Ticaret
ISCTR	Türkiye İş Bankası
KCHOL	Koç Holding
KOZAA	Koza Anadolu Metal Madencilik İşletmeleri
KOZAL	Koza Altın İşletmeleri
KRDMD	Kardemir Karabük Demir Çelik Sanayi ve Ticaret
ODAS	Odaş Elektrik Üretim Sanayi Ticaret
PETKM	Petkim Petrokimya Holding
PGSUS	Pegasus Hava Yolları
SAHOL	Sabancı Holding
SASA	SASA Polyester Sanayi
SISE	Türkiye Şişe ve Cam Fabrikaları
TAVHL	TAV Havalimanları Holding
TCELL	Türkcell İletişim Hizmetleri
THYAO	Türk Hava Yolları
TKFEN	Tekfen Holding
TOASO	TOFAŞ Türk Otomobil Fabrikası
TUPRS	TÜPRAŞ Türkiye Petrol Rafinerileri
YKBNK	Yapı Kredi Bankası

2.2. Common Stock Analysis Methods

Stock analysis, whether on an individual stock basis or on an index basis, aims to predict trends, find the stock of the company that will earn the most going forward and identify buying and selling zones. According to the efficient market hypothesis, the market price at any given moment contains all information, so it is not possible to generate above-market returns by forecasting in financial markets. According to all three variants of the efficient market hypothesis, it is not possible to achieve high returns through forecasting (Shah et al., 2019).

- i) There is no relationship between price changes, price changes are random. Therefore, technical analysis cannot beat the market.
- ii) Prices change rapidly according to information such as political developments, economic news, balance sheet announcements. Therefore, very large returns cannot be achieved with fundamental analysis.
- iii) Price contains all information, so information is public, this is a barrier to high returns through forecasting.

Many researchers and investors disagree with these predictions of the efficient market hypothesis and think that forecasting can be done using financial market data using many data (Shah et al., 2019). From a methodological point of view, two main methods are used in stock analysis; fundamental analysis and technical analysis (Bentes and Navas, 2013).

Both fundamental analysis and technical analysis have their advantages, disadvantages and complementarities. It is common for these two methods to be used together. Many fundamental analysts use technical analysis to identify buy and sell points. Many technical analysts also use fundamental analysis to identify companies that can be really profitable in portfolio construction (Bentes and Navas, 2013).

Artificial intelligence and machine learning can make new contributions to both of these types of analysis. Machine learning is widely used in technical analysis to detect patterns or anomalies in time-series price charts, and in fundamental analysis to establish regression between company data and price. Artificial intelligence and machine learning techniques propose new methods to augment and combine these existing analysis methods.

2.2.1. Fundamental Analysis

The fundamental analysis examines companies' financial and economic reports, considers all quantitative and qualitative data to identify companies' value. More specifically it is an examination process of reports such as profit and loss accounts, balance sheets and also indicators such revenues, earnings, liabilities, expenses and assets. It should not be assumed that only companies' information is taken into account in the fundamental analysis process. International and national indicators such as inflation, prices, GDP growth rates, interest rates, unemployment data are also significant parameters (Bentes and Navas, 2013).

An important step in fundamental analysis is to calculate some ratios using the company's financial statements. Table 2.4 shows most commonly used ratios in fundamental analysis (Bentes and Navas, 2013).

Table 2.4: Most Commonly Used Ratios in Fundamental Analysis

Prices	Shares	Profitability	Solvability	Efficiency	Market
Price Earnings Ratio	Earnings per Share	Return on Asset	Current Ratio	Accounts Receivable Turnover	Free Float
Dividend Yield	Price Book Value	Return on Equity	Leverage Ratio	-	Index Trading
Price Cash Flow	-	Return of Investment	Long Term Debt	-	Frequency Index
-	-	Capital Turnover	Inventory Turnover	-	-
-	-	Earnings Margin	-	-	-

Dozens more of these ratios are used in fundamental analysis. It is also part of the fundamental analysis process to consider macro events such as news flow, sectoral and political developments, economic crisis, pandemic, natural disasters. Fundamental analysis is based on periodic reports of the companies, and these reports are not instantaneous or frequent. In Türkiye, for example, these reports are published at the end of each quarter. Fundamental analysis is therefore commonly done for long-term investment purposes (Bentes and Navas, 2013).

2.2.2. Technical Analysis

Technical analysis assumes that all data about the company, such as national or international economic, investor psychology, etc., is already reflected in the price (Bentes and Navas, 2013). It is therefore sufficient to analyze the price historically. Nowadays, volume trading is done automatically by programs that follow certain algorithms, especially by investment companies. These programs also use methods enabled by technical analysis. This is called algorithmic trading (Gold, 2015).

In technical analysis, the parameters considered are mathematically or statistically calculated indicators. There are hundreds of technical indicators but some of them are used commonly. Also, technical indicators have different categories due to calculation methods or used data. The literature review shows that technical indicators can be categorized under three main categories: trend, momentum and volume. Table 2.5 shows some of the most commonly used technical indicators, with two examples from each category (Gold, 2015).

Table 2.5: Most Widely Used Technical Indicators for Technical Analysis

Category	Indicator	Abbreviation
Trend	Moving Average	MACD
	Convergence/Divergence	
Trend	Aroon Oscillator	AROON
Momentum	Stochastic Oscillator	SO
Momentum	Relative Strength Index	RSI
Volume	On-balance Volume	OBV
Volume	Accumulation/Distribution Line	ADL

For example, if the MACD indicator examined, it is seen that it is the difference of two EMAs (Exponential moving averages).

$$\text{MACD}_t = 12 \text{ days EMA}_t \text{ of closing} - 26 \text{ days EMA}_t \text{ of closing} \quad (2.2)$$

$$\text{Threshold}_t = 9 \text{ days EMA}_t \text{ of MACD} \quad (2.3)$$

Identifying buy and sell thresholds is an important aim in the technical analysis. When indicators are applied to the chart, they create an additional indicator-applied chart and a threshold calculation is also used to create trading zones. For the MACD, 9 days EMA of MACD is a threshold. If MACD is above threshold it is interpreted as a “bullish” (buy) signal, if it is below threshold means “bearish” (sell) signal (Gold, 2015).

Several online or offline technical analysis tools provide automated indicator calculations over trend. Figure 2.1 shows the chart of the BIST 100 index for the first month of 2021 with the MACD indicator added (WEB-3, 2024).

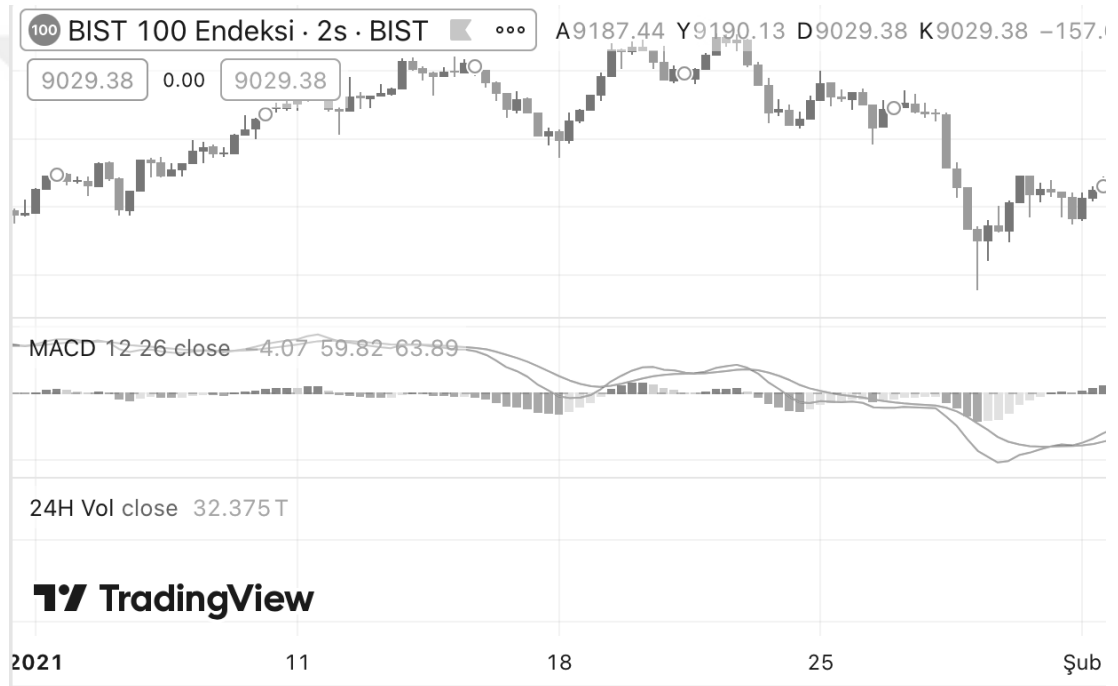


Figure 2.1: BIST 30 and MACD Added Chart (WEB-3, 2024)

2.3. Behavioral Finance

Keynes (1936) pointed out for the first time that economics cannot be explained entirely on the basis of mathematical models and that human behavior can also have an impact. In 2002, psychologist David Kahneman, who was awarded the Nobel Prize in Economics, demonstrated with experiments that people make risky choices in the face of offers that emphasize gains, while they make guaranteed choices in the face of offers that emphasize losses (Bildik, 2020).

The efficient market hypothesis is one of the most important theories of the 21st century in the field of financial markets. EMH assumes that all information about the market is quickly incorporated into the price, that the price is all-inclusive.

The EMH is based on three main arguments (Naseer and Tariq, 2015):

- i) Investors are rational and purely utility motivated.
- ii) If investors do not behave rationally, trades are assumed to be random and offset any effect on price.
- iii) Rational arbitrageurs eliminate the influence of irrational investors on the market.

Despite the impact of EMH, there have been many critical studies against EMH. Research shows that financial anomalies, which can be categorized as fundamental, calendar or technical, allow for high returns, contrary to EMH's predictions. In particular, studies in the field of behavioral economics and its extension, behavioral finance, reveal that the psychological and sociological conditions of investors will be effective in making irrational decisions and that the market should not be seen as an area consisting of completely rational players (Naseer and Tariq, 2015).

Psychological empirical research shows that when decision-making is difficult and feedback is delayed or noisy, people tend to rely more on intuition and practice. Research shows that this is also true in financial trading. Investors seem to act on behavioral biases when stock valuation is difficult. Price movements in the market may cause different behaviors due to people's behavioral biases. Valuation uncertainty is observed to have different behavioral effects. It has been observed that it increases the disposition effect that motivates some investors to buy on declines and sell on rises. The disposition effect generates behaviors such as frequently tracking prices to look for trading opportunities. In cases of valuation uncertainty, some investors also experience a gambling effect. Investors with high-risk winning behaviors are found to consider more losses. On the other hand, the disposition effect is found to be amplified for overconfident investors. Overconfident investors are less likely to accept their losses as mistakes (Kumar, 2009). In addition to the effect of overconfidence and disposition, biases about herding and blue-chip stocks also have an impact on risk perception (Almansour et al., 2023).

Behavioral finance studies show that emotions and biases influence decision-making. It has also been shown in many studies that news flows have an impact on the stock

market. It is possible to assume that the emotional states of individuals affected by news feeds also have an impact on stock market trading. Understanding public sentiment used to require difficult and expensive studies. Moreover, the accuracy of the methods used was limited due to the low correlation of the selected indicators with the public mood. Being able to measure public sentiment can be considered an important behavioral finance variable. Sentiment analysis studies, which started on blogs in recent years, have now become common on social media data (Bollen et al., 2011).

2.4. Machine Learning

Arthur Samuel (1959), a pioneer in the field of machine learning (ML) defined it as the “field of study that gives computers the ability to learn without being explicitly programmed”. Machine learning is a method that has been studied and used for many years. For example, this capability has been used for many years to analyze spam mail in mailboxes. Spam filters are a machine learning program that, given examples of spam and spam-free mail, can predict whether new incoming mail contains spam or not (Géron A., 2022).

Today, countless applications based on machine learning such as chatbots, diagnosis based on medical imaging, classification of large texts, detection of credit card fraud, intelligent bots in games, text summarization, sentiment analysis are already in life.

2.4.1. Definitions and Basic Concepts

Artificial intelligence has been an important field of study all over the world since Alan Turing introduced the Turing Test concept in 1950. This test opened up a debate on the methods of deciding whether a machine that can process information has intelligence or not. There have been great advances in artificial intelligence until today, but what this test questioned is valid. This debate still continues on a philosophical ground.

There is a lot of conceptual confusion about artificial intelligence, which is of interest to both academia and society today. Concepts such as artificial intelligence, machine learning, deep learning are sometimes used interchangeably or presented as something very newly invented. It is true that there is a lot of hype around AI, but there is also an

important reality of artificial intelligence that affects finance, academia, manufacturing and the service sector. Artificial intelligence has been rule-based for many years, whereas today's AI refers more to data-driven learning algorithms.

Although the development and growth of data-driven AI systems is witnessed today, the history of data-driven algorithms dates back 60 years. Today, the main factor driving research and development is the creation of a lot of data and the ability to process and store it. Data-driven algorithms existed in the past, but they could not produce useful outputs because there was not enough data (WEB-4, 2023).

Data-driven artificial intelligence systems can be defined as machine learning. Although machine learning algorithms are at the center of artificial intelligence today, it would not be correct to say that rule-based systems are completely out of use. Machine learning is like a black box, especially when used in the form of complex structures such as deep learning models. It is not possible to follow the program and understand which decisions are made and why. For this reason, it is risky to use completely data-driven artificial intelligence in some areas. For example, important components of systems used in cyber security continue to be designed as rule-based decision-making systems (Mende et al., 2022). Similarly, the use of rule-based systems is common in autonomous vehicles, especially in the driving parts. In cases such as information security, safety of life and property, hybrid structures such as these, where data-driven and rule-based systems are used together, will continue to be used (Bhattacharyya et al., 2022).

All of the cleaned and processable data that is subject to a data science process is called a "data set". In the data science process, this data set is divided into two parts: the "training set" and the "validation set". In machine learning, an algorithm is decided according to the type of data and problem and a "model" is developed. Machine learning algorithms are based on mathematical models or approximations. These models have no parameters before they are trained. As a result of the training phase, the parameters and variables of the model are formed according to the training set data. A validation set is used to test how accurate the model works and to measure its performance. In cases such as the model is complex due to the large number of variables, the training set is insufficient, the data contains too much noise, the model can explain the learning set very well, but if it is insufficient in generalization, this is called "overfitting". Figure 2.2 shows an example of overfitting (Géron A., 2022).

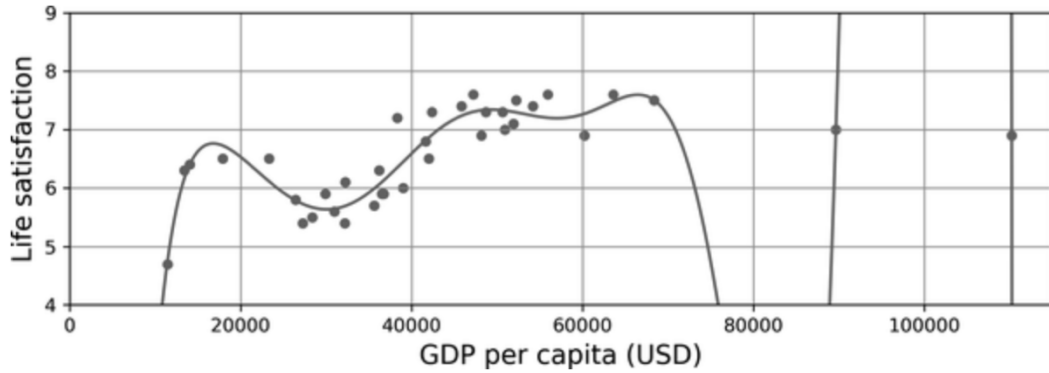


Figure 2.2: An Example of Overfitting (Géron A., 2022)

2.4.2. Machine Learning Types and Algorithms

Machine learning systems can be categorized into five main categories according to the type of training set; supervised, unsupervised, semi-supervised, reinforcement and self-supervised learning. If the types of data in the training set are specified, i.e., the data are labeled, this is called “supervised learning”. An example of such data is data stored in an excel file. The columns in the excel file determine the type of data, i.e., the label. If there is no distinctive labeling of variables in the data, the learning process using this data is called “unsupervised learning”. Learning on data such as text, audio, video is an example of unsupervised learning. The learning process on data with mixed labeled and unlabeled data is called “semi-supervised learning”. If the learning system selects an action based on the data in the environment and determines a new action based on the reward or punishment that results from the action, this is called “reinforcement learning”. The system learns based on past actions and their results. self-labeled learning is when the learning system working on unsupervised learning labels the data itself and turns it into supervised learning (WEB-5, 2023).

In the data science process, the type of learning is determined according to the structure of the data and the data problem. The type of learning is important for deciding which machine learning algorithm to use. For example, if the data set is a labeled data with inputs and outputs and the problem is to predict the output based on the input, regression algorithms are used. In unsupervised learning, the use of clustering algorithms is common. In unsupervised learning, the problem is to categorize the data and the use of clustering algorithms is common (WEB-5, 2023).

This study has two phases, obtaining sentiment scores from text data and predicting the stock price from the sentiment score. Semi-supervised deep learning algorithms are used for sentiment analysis and supervised learning regression algorithms are used for prediction phase.

2.4.3. Neural Networks and Deep Learning

Artificial intelligence is a broad field of study that encompasses both data-driven algorithms, i.e., machine learning, and rule-based algorithms. Machine learning is a field of artificial intelligence that takes data as input and models the data by using various mathematical approaches to reveal the parameters related to the data. Deep learning is a field of machine learning based on artificial neural networks that is used when the data is less labelable, complex and abundant. Today, deep learning algorithms are mostly used to model human activities such as vision, speech and writing. For this reason, it is common to understand deep learning algorithms when it comes to artificial intelligence today.

Deep neural networks, which form the mathematical models of deep learning algorithms, are structures based on artificial neural networks. In a real neural network, there are two parts responsible for communication; axons and dendrites. Axons transmit a signal from another neural network to the dendrite, which in turn transmits it to the axon of another cell. An artificial neural network (ANN) is a model based on a nerve cell, with inputs like axons and outputs like dendrites. Just as a nerve cell does not carry every signal, but only the signal that exceeds a threshold value, a neural network produces output by putting inputs into an activation function. Just as neural networks are connected to each other, it is also possible to create more complex neural networks by connecting the output of one neural network to the input of another (McCulloch and Pitts, 1943).

ANN has two main layers; “input layer” and “output layer”. Output layer also called “threshold logic unit” (TLU). ANNs with multiple TLUs are called “multi-layer perceptron” (MLP). TLUs between input layer and final TLU are called “hidden layers”. Neural networks with more than two hidden layers are considered “deep” and are called “deep neural networks” (DNN) (Géron A., 2022). Figure 2.3 shows basic architecture of a deep neural network.

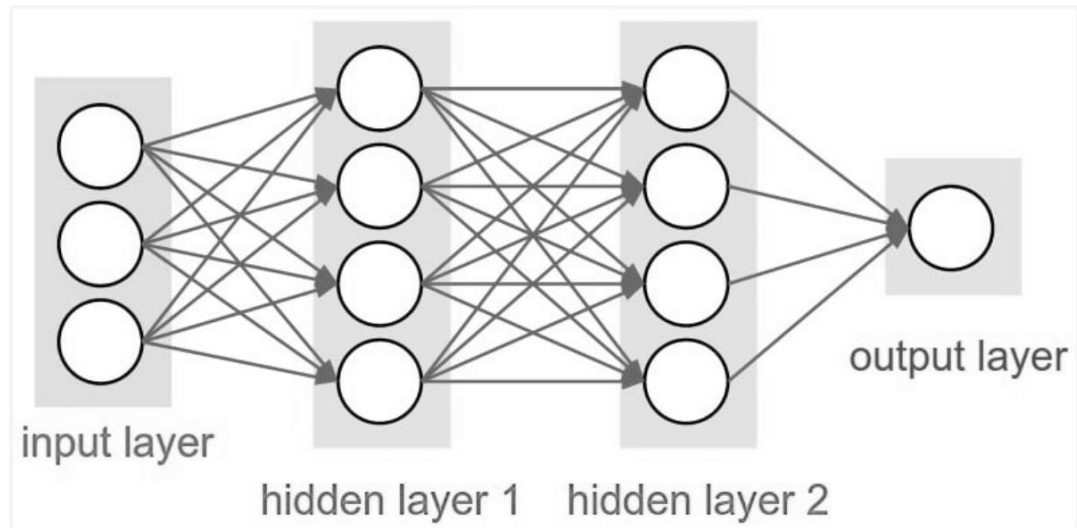


Figure 2.3: Basic Architecture of Deep Neural Network (WEB-6, 2020)

A single neural network is a single-variate linear regression model based on a function such as “ $w x + b = F(x)$ ”. x is the variable or feature, b is the bias, and a is the weight value. The goal of the neural network model is to calculate the weight and bias value that provides the optimal output for the input. Inputs, weights and outputs are all numerical values. Figure 2.4 shows a single artificial neural network structure.

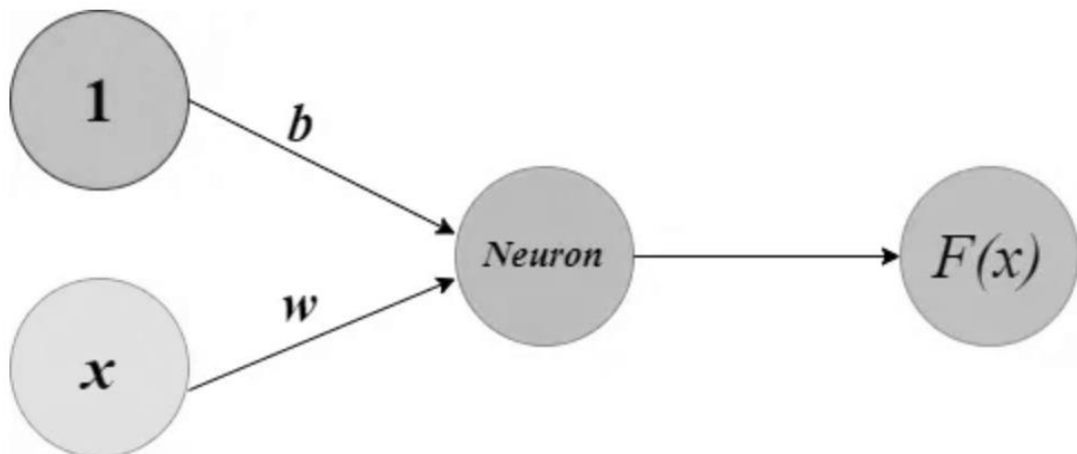


Figure 2.4: Single Artificial Neural Network (WEB-7, 2023)

The number obtained by multiplying the inputs by the weights and adding the bias value is put into an activation function. The activation function is a function that produces a number between 0 and 1 as a result of the number it receives as input. There

are different types of activation functions some are linear, step, sigmoid, ReLU, etc. Figure 2.5 shows a sigmoid activation function graph. It generates 0 for negative values and 1 for positives.

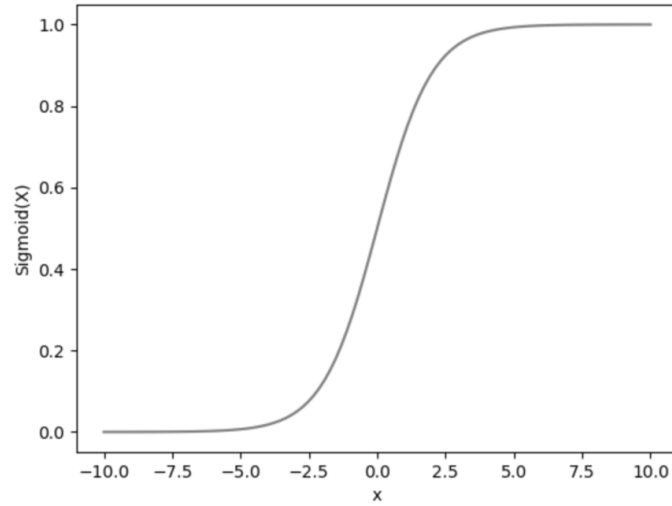


Figure 2.5: Sigmoid Function (WEB-8, 2023)

When many neural networks are connected together, the basic working principle does not actually change, only the number of variables and weights increases. When there are more variables, the inputs and weights are expressed in matrices and the calculations are performed on the matrices. Figure 2.6 shows mathematical model behind a neural network.

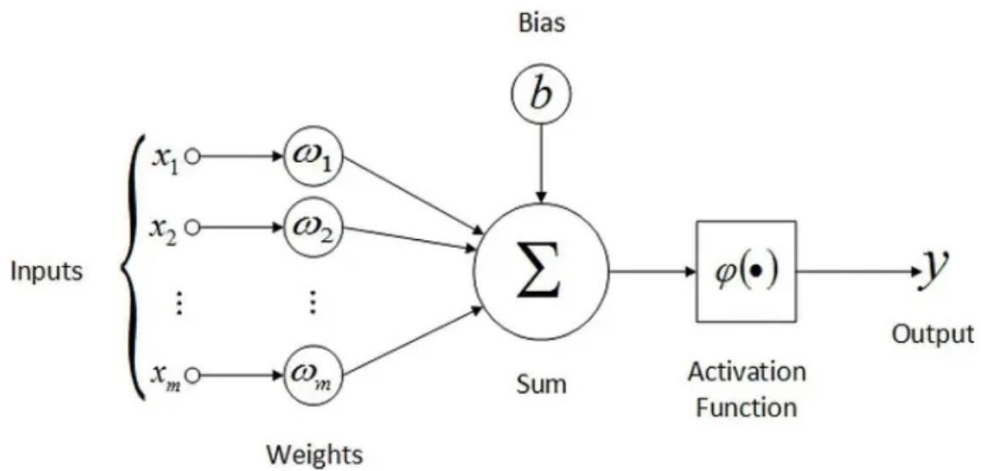


Figure 2.6: Multi-Variate Neural Network (WEB-9, 2023)

Deep neural networks contain both many variables and many layers. The neural nodes in each layer carry the values from the previous layer multiplied by the values and weights from the previous layer. This is a necessity for problems where the inputs consist of the most primitive data but increasingly complex features are needed to reach the result. Figure 2.7 shows a simple demonstration of image classification application. The deep learning model should be able to identify which animal the image belongs to, no matter what color or orientation it is.

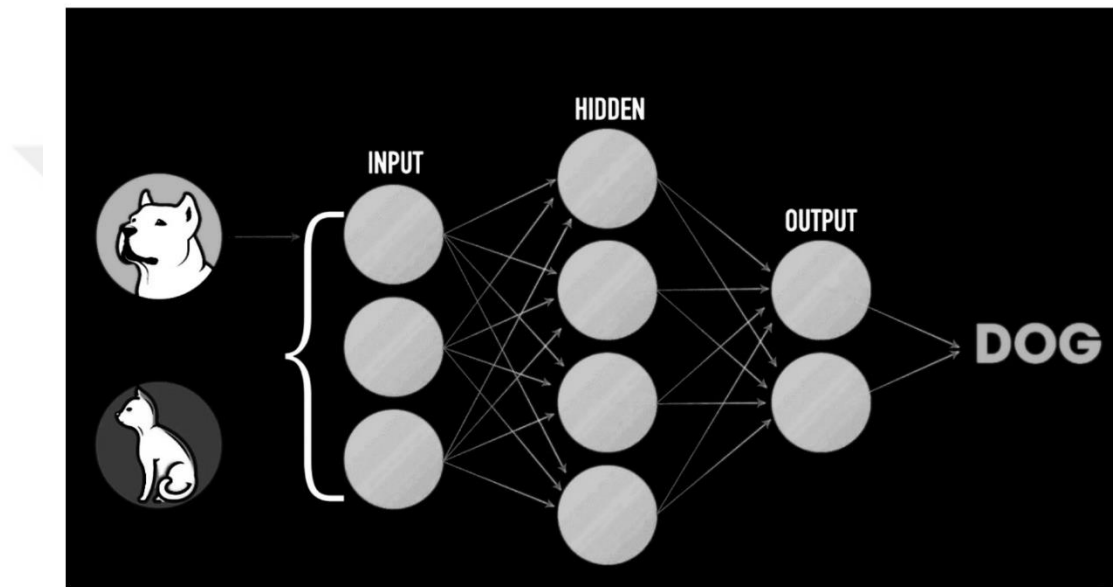


Figure 2.7: Image Classification is a Common Deep Neural Network Application
(WEB-9, 2023)

Initially, the only data that can be presented to the deep neural network are the pixel values of the image. Edges are detected in the first hidden layer. In the second layer, the patterns formed by the edges. The next layer may contain combinations of these patterns and subsequent layers may contain more complex combinations. The most important feature of deep learning is that it can extract more complex features from primitive features such as pixels without human intervention (Lecun et al., 2015). Figure 2.8 shows convolutional neural network phases, a specific deep neural network.

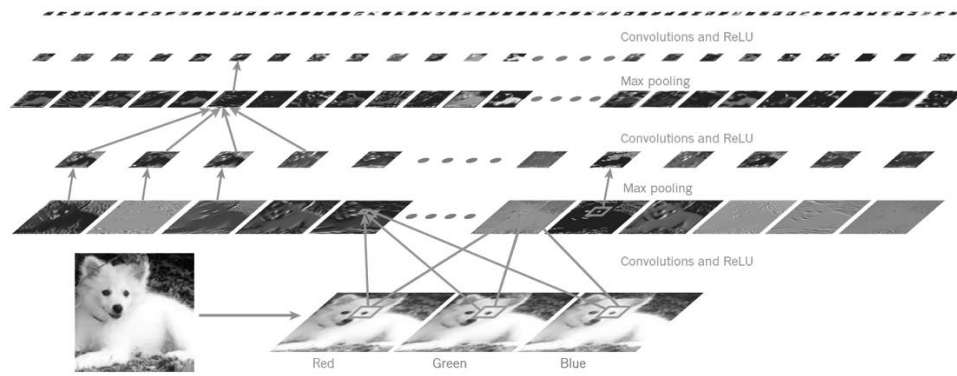


Figure 2.8: DNN Extracts Complex from Simple Features (Lecun et al., 2015)

In all machine learning applications, there is a difference between predicted results and actual results. For a single record this is called the “loss function”, for the whole dataset it is called the “cost function”. Minimizing these is an important step in the machine learning process. The methods for finding the minimum cost learning are called gradient descent. In the stages of minimizing the cost function in deep neural networks, the weights and bias values in the neural networks are adjusted again. This is called “back propagation” (Géron A., 2022).

2.5. Natural Language Processing

The study of natural language processing began in the 1950s as an intersection of artificial intelligence and linguistics. Chomsky (1956) categorized language grammars and his work on language formation for regular grammars was used to develop the “grep” tool for the UNIX operating system. Subsequently (1970s), lexical analyzers were developed. A lexical analyzer is a model that tokenizes text and checks the tokens for compliance with programmable grammar rules. Programming approaches based on grammar rules have some limitations. Language is vastly large and by its nature it is very difficult to put it into specific forms. On the other hand, it is not enough to analyze texts with a programmable grammar approach, the language processing process must also be able to extract meaning. It is not possible to say that daily language is completely grammatical. For these reasons, statistical natural language processing methods have been studied since the 1980s instead of rule-based methods. These methods are models obtained by training machine learning algorithms with data sets consisting of large amounts of text called "corpus" (Nadkarni et al., 2011).

Machine learning-based natural language processing methods have evolved considerably as a result of theoretical and technological developments, especially in the field of deep learning. Language can be considered sequential data. In the context of natural language processing (NLP), text data is typically treated as a sequence of tokens (words, characters, or sub words), where the order of these tokens matters. This sequential nature arises from the fact that the meaning of a sentence or a piece of text is often dependent on the arrangement of its constituent words or characters. Understanding sequential data, where information unfolds over time, is crucial for various tasks like language processing, speech recognition, and time series forecasting. This realm is dominated by powerful deep learning architectures like recurrent neural networks (RNNs), long short-term memory (LSTM) networks, transformers, and large language models (LLMs).

2.5.1. Recurrent Neural Networks and Long Short-Term Memory

Recurrent neural networks are widely used in sequence-to-sequence applications such as machine translation, speech recognition and video analysis. Basically, the difference of recurrent neural networks from other deep learning models is that they can remember the data from the previous step (Wang et al., 2019). RNN has proved its success in previous steps in learning sequential data, but it has been found to be quite unsuccessful in long-term prediction on loaded data. This is a phenomenon known as vanishing gradient (Cheelamanthula, 2022).

RNNs consist of consecutive cells. Each cell provides input to the next cell. The cells produce the output using the data from the previous cell and the hidden state information. The hidden state information is actually a limited memory and holds information about the previous steps. This information is critical to enable the network to understand the context and make predictions for the current input. Figure 2.9 shows architecture of RNN.

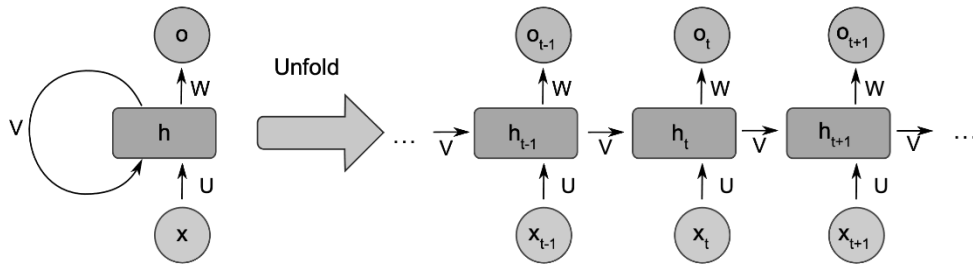


Figure 2.9: Architecture of RNN (WEB-10, 2024)

The limited memory of RNNs holding the hidden state data prevents them from long-term recall. To overcome vanishing gradient challenge, long short-term memory (LSTM) has been developed (Wang et al., 2019).

LSTM is a special type of RNN developed to solve the vanishing gradient problem, but unlike RNN, it uses two special structures called cell state and gate. A cell state is a memory where the information about the previous steps is stored, updated when necessary and used. In LSTM, unlike RNN, the data in this memory can be updated, that is, forgotten. Three type gates are used to control this cell state; input gate to determine which data will update the memory cell, output gate to determine which data will be used from the memory cell and forget gate to determine which data should be kept in the memory cell and which data should not be kept (Cheelamanthula K., 2022).

LSTM cells are more complex than RNN cells. The data from the previous cell passes through the control of the gates. Each gate puts the incoming data into one or more activation functions with the data in the memory cell. Figure 2.10 shows architecture of LSTM.

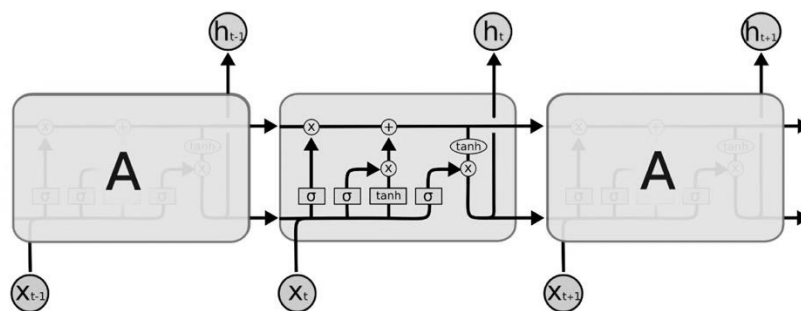


Figure 2.10: Architecture of LSTM (WEB-11, 2015)

Compared to RNNs, LSTMs can manage the data of previous steps more efficiently and store them longer. This makes them more effective for translating and generating long texts or for sentiment analysis. Table 2.6 shows comparison between RNNs and LSTMs.

Table 2.6: RNN and LSTM Comparison

Feature	RNN	LSTM
Architecture	Simple	Complex
Memory	Hidden state	Hidden state and cell state
Applications	Short texts and simple tasks	Long texts and complex tasks

Although LSTM is superior to RNN, it also has some limitations. Like RNN, it works sequentially and starts to struggle with very large content. The fact that natural language processing is both faster and able to process large content is thanks to transformers (Vaswani et al., 2017).

2.5.2. Transformers and Large Language Models

Transformers and large language models (LLM) are important AI developments that have led to the development of models such as GPT and BERT, which are revolutionizing natural language processing today. The RNN and its improved version LSTM models have made significant improvements in natural language processing, but they are inadequate or slow when the context is prolonged. In 2017, Google engineers presented the Transformer architecture. With this, a model that can manage both speed and long-distance contexts has been introduced.

Transformers have two main components called encoder and decoder. The encoder vectorizes the input sentence and processes it in layers. Each layer carries out the attention mechanism and layer normalization. This process continues until the layers are complete. The decoder receives the information from the encoder to generate the output sentence. Like the encoder, the decoder also operates in layers, each layer again using the attention mechanism and taking into account the previous word's information. In the last layer, the data is passed through an activation function and the

output sentence is generated (Vaswani et al., 2017). Figure 2.11 shows detailed and abstracted architecture of transformers.

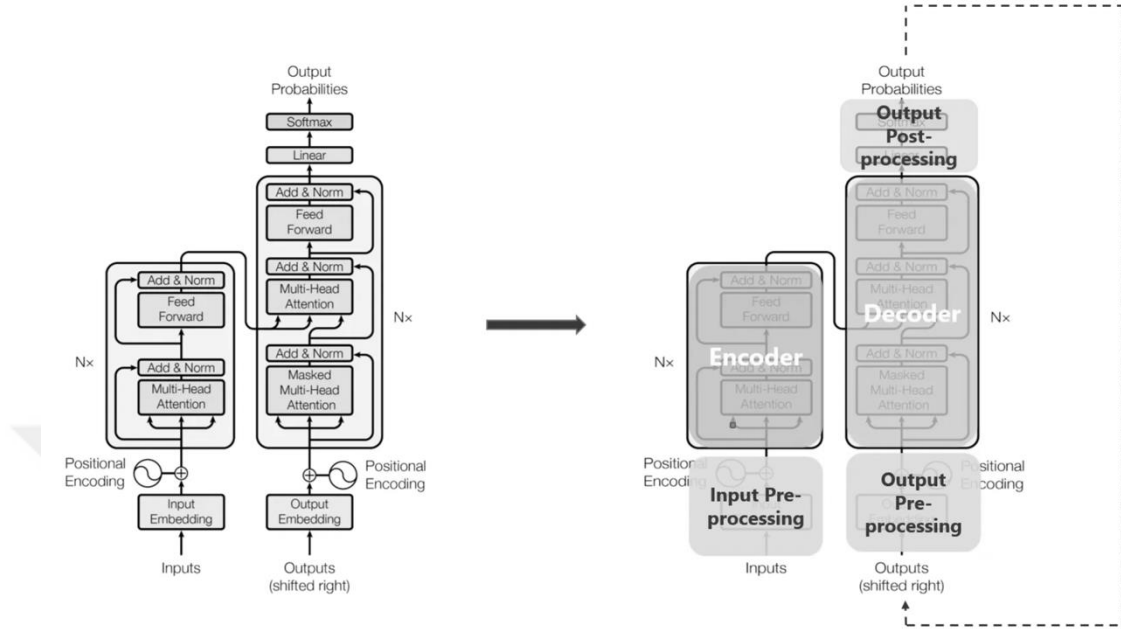


Figure 2.11: Detailed and Abstracted Architecture of Transformer (WEB-12, 2020)

The most important feature that increases the speed and context generation capacity of the converters is the "attention mechanism". It is based on vectorizing a subset of a long text rather than vectorizing the whole text. As the length of sentences increases, the performance of the encoder and decoder deteriorates. One of the most important features of transformers is that they are trained by parallelizing the parts of the corpus (Bahdanau et al., 2015).

Large language models are models that use transformers and can be trained using very large amounts of text. LLMs are not an alternative to transformers, but larger models using many transformers. LLMs are very complex and multi-layered structures involving billions of parameters. To compare LLM and transformer, a transformer is an engine, while a LLM is an automobile. An LLM does not have to have both encoders and decoders, which are components of transformers. Some of the most widely used LLMs may have only an encoder or only a decoder. Table 2.7 shows the most common used LLMs and their comparisons (Devlin et al., 2018; Lewis et al., 2020).

Table 2.7: Comparison of the Most Commonly Used LLMs

	BERT	GPT	BART
Model Type	Encoder only	Decoder only	Encoder-Decoder
Direction	Bidirectional	Unidirectional	Bidirectional
Pre-Training Objective	Masked language modeling (MLM)	Autoregressive language modeling	Span corruption
Fine Tuning	Task specific layer added on the top of pre-trained BERT model	Providing task specific prompts using few-shot or one-shot adaptation and adapting model's parameters	Versatile and can be used for various NLP tasks
Use Case	Sentiment analysis named entity recognition word classification	Text generation, text completion, creative writing	Translation, text summarization, Q&A
Original Organizations	Google AI	Open AI	Meta AI

LLMs are constantly being improved. Major technology companies frequently introduce faster and more successful models trained with a larger corpus.

2.5.3. Sentiment Analysis

Sentiment analysis or opinion mining are techniques for extracting emotions, attitudes, evaluations or opinions from written, audio or visual data produced by people about a product, a service or a topic. Although these techniques have been called by various names such as opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining over time, today all activities related to this subject are examined under the title of sentiment analysis. Today, the fact that people voluntarily produce data has led to an increase in natural language processing and its application area, sentiment analysis (Liu, 2012).

Sentiment analysis can be performed at document, sentence or entity and aspect level. With sentiment analysis at document level, the main idea of the whole document is extracted. At the sentence level, sentiment analyses of individual sentences that make up the text are performed. Entity and aspect level focuses on sentences that directly

express opinions. Sentiment analysis is a very challenging field of study. Human natural language is complex and vast. In daily life, people use words expressing sentiment in a very opposite or sarcastic way. Some sentences do not express a sentiment emotion even if they contain the sentiment words such as good, bad, terrible or amazing. In some sentences, there is no word expressing sentiment, but the whole sentence expresses a sentiment (Liu, 2012).

There are also sentiment analysis tools and libraries that do not use deep learning, or more specifically RNN, LSTM and Transformers. VADER, TextBlob, CoreNLP, Python NLTK, LIWC are some of them. Some of these can use classical machine learning methods. These systems extract tokens from the text and create a sentiment score by scanning the text based on the sentiment lexicon of the relevant language.

The manual generation of the sentiment lexicon is a very difficult task, so machine learning techniques such as Naive Bayes and SVM classifiers are used. There are also other drawbacks associated with machine learning sentiment lexicon. In order to train a quality machine learning model, a quality learning set is needed. In order to train the model with a large learning set containing all possibilities, high processing power is needed. Due to the black box feature of machine learning, it is also impossible to analyze the sentiment lexicon in detail and to understand how it decides the sentiment scores (Hutto and Gilbert, 2014).

2.6. Use of AI and Machine Learning in Finance

With the development of computer technologies, computers have started to be widely used in finance as in all fields. Computer science has proven itself many times in computationally intensive subjects such as stock analysis, portfolio management and risk management in finance. Both expert systems using classical machine learning methods and artificial neural networks have led to many successful software solutions and many researches in the field of finance. Genetic algorithms and fuzzy logic are other soft computing techniques used in financial applications other than machine learning (Dunis et al., 2016).

Portfolio management is the creation of an optimum portfolio that will provide the best gain according to risk perception and updating the portfolio according to prices over time. Artificial neural networks can be used in portfolio management. In the portfolio

forecasting applied on American stocks using neural networks and regression analysis, artificial neural networks provided higher returns (Motiwalla and Wahab, 2000).

Risk management is a decision-making process by taking into account social, economic and political factors. Bankruptcies, fraud, problems in credit returns can be evaluated within the scope of risk management (Dunis et al., 2016).

After the 2008 economic crisis, many studies were conducted and some of them utilized artificial intelligence methodologies. In a study conducted using financial statements and statistical data obtained from municipalities in the United States, classification models were created according to the credit ratings of municipalities. The credit rating classes obtained were found to be consistent with the values given to bond issuers (Hajek, 2011). As a result of the bank bankruptcy prediction study using AI methodologies, it is observed that the model produces accurate results with reasonable explanations (Ahn and Kim, 2009).

In recent years, soft computing methods have been widely used in both analysis and modelling activities for both pricing and valuation purposes in the stock market. Prices in the stock market are highly volatile and the price is affected by many factors. Therefore, forecasting in the stock market is quite challenging. Forecasting daily stock prices is not just a numbers problem, it is a real-world problem. Many non-numerical factors such as political and international relations and news also affect forecasting. There are many studies that take these factors into account (Dunis et al., 2016).

In an early study using artificial neural networks in stock forecasting, the price of IBM stock was tried to be predicted, and the potential of artificial neural networks to increase the ability in forecasting was mentioned (White, 1988). Another study, which presents a forecasting model using an artificial neural network in a low-risk, high-return portfolio which was managed with a buy-and-hold strategy, shows that artificial neural networks give better results than other algorithms. However, like other studies, this study also pointed out that stock forecasting will continue to remain a difficult issue due to the presence of non-numerical effects (Thawornwong and Enke, 2004).

Many studies emphasize that non-quantitative factors such as politics, international relations and news flow pose challenges to stock forecasting. This has highlighted the importance of including public opinion as a variable in research. In recent years, the number of sentiment analysis studies using social media data in finance has been

increasing. A study using news articles mentioned that sentiment analysis improves the accuracy of prediction, but the prediction time horizon should be shrunk due to the dynamism of the stock market (Li et al., 2014). Instantaneous price changes in the stock market may be more strongly correlated with social media, which is naturally more dynamic than news articles. A study analyzing the relationship between sentiment analysis scores obtained from news and twitter data and their relationship with the prices of some selected company stocks, reveals that twitter data has a more significant relationship with volatility than news data (Souza et al., 2015).

Since Twitter is an environment where instant news flow and social sentiment are instantly reflected, it has started to be used as a common data source in sentiment analysis-based stock forecasting research. A study by Rao and Srivastava (2012) shows that negative and positive dimensions of public mood have a strong causal relationship with price movements of individual stocks/indices. In financial sentiment analysis studies on news articles or social media data, deep learning algorithms are used as well as classical machine learning techniques. Deep learning techniques are more successful in evaluating the context of the whole text than classical approaches based on the sentiment analysis score of individual words (Sohangir et al., 2018).

3. METHODOLOGY

In the methodology section, information about obtaining the data used in the study, the types of data, the processes that the data went through and the data made ready for the study are given. The actual data used in this study is obtained by combining the sentiment scores obtained from tweet data with the price data of stocks and indices on a date basis. Figure 3.1 shows a brief flowchart of obtaining data and creating final datasets to be used in the study. Collection of three different types of data (tweets, stocks and index data) are not dependent stages. In the last stage before the analysis process starts, the three data sets composed of stock and index data are combined with the refined tweet data to obtain the final data sets.

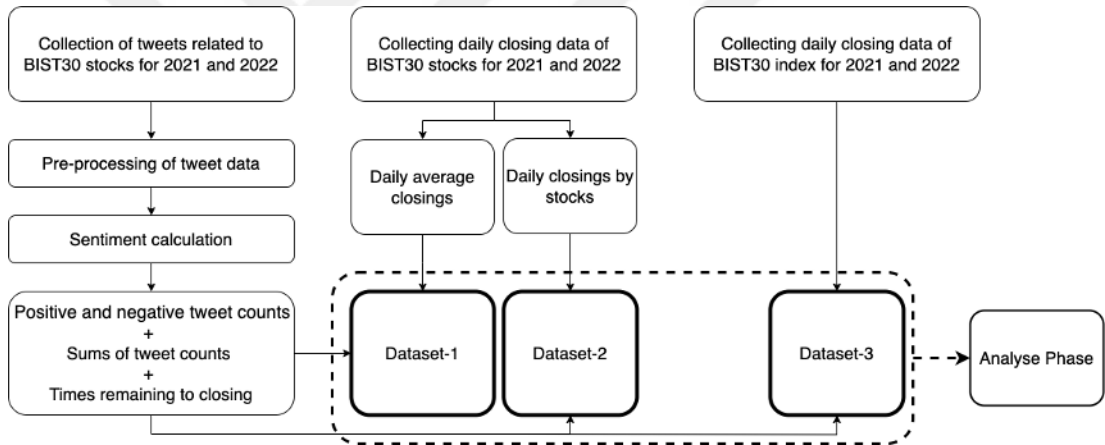


Figure 3.1: The Process of Obtaining Data Before Analysis

The following subsections describe the process of preparing the data for the analysis process. Obtaining and processing the data required the use of various tools. For this reason, programming languages, tools and libraries are mentioned before talking about how the data were obtained.

3.1. Data Types, Characteristics and Sources

As shown in Figure 3.1, there are three types of data initially obtained. Tweet data, closing data of BIST 30 constituents and closing data of BIST 30 index. Time

information plays an important role in merging all these data, so time data is included in all data.

Tweet data contains three types of data: the user who wrote the message, the text, i.e., the tweet, and the timestamp, i.e., the time when the tweet was sent. An acquisition of Twitter initiated on April 14, 2022 by Elon Musk and acquisition finalized on October 28, 2022. Musk confirmed the rebrand of Twitter on July 23, 2023 and it was rebranded as “X”. Following the rebranding, the company underwent significant changes both in terms of organization and services. Twitter was an important source of APIs for software developers and was one of the few social media platforms that provided a wide range of API endpoints. Between 2006 and 2010, Twitter offered developers a wide range of API opportunities. In the following years, some limitations were gradually introduced. Flexibility in authentication was reduced and limits were imposed on the amount of data that could be retrieved with the API. However, developers were still able to provide free data for a limited duration. In January 2023, access to APIs by third-party applications was closed, and in February 2023, the free use of APIs was terminated (WEB-13, 2024).

The data set consisting of daily closing data of BIST 30 stock components consists of the code of the stock, the date of the day and the closing price. Share closing data on a daily basis can be obtained from a wide variety of open sources. Many financial institutions in Türkiye are able to provide this data on a historical basis. The daily data for the stocks, which are the components of the BIST 30 index used in this study, for the years 2021-2022 were obtained from the public website of the investment subsidiary of a private bank. These sources can also provide more detailed data such as the number of trades and volume per day (WEB-14, 2024).

BIST 30 index data consists of the closing and closing date of the index. The index closing data is not a price as in the case of stocks, it is a composite of the performance of BIST 30 stocks as explained in section 2.1.2. The calculation method is provided by formula 2.1 given in the same section. Although the index is a mathematical expression, it is a composite of values such as price, volume and number of transactions, and gives an idea about the general trend. Using some public web sites, it is possible to obtain daily data of many financial assets such as stocks, commodities and foreign exchange. These sites offer a platform where both technical and fundamental analysis parameters, which are widely used in the investment world, can

be examined. In addition, many of them, such as the site from which the index data is taken in this study, provide economic news feeds that may affect investment assets. The platform used to obtain the BIST 30 index in this study allows report retrieval by giving start and end dates. Within the scope of this study, daily BIST 30 index data between the beginning of 2021 and the end of 2022 were extracted (WEB-15, 2024).

3.2. Technologies and Tools

Many programming languages can be used in scientific studies, but Python programming language maintains its widespread use due to its library richness and being a high-level language (Pedregosa et al., 2011).

The data sets used in data science studies are in a tabular structure, and Python's "pandas" library helps to process and query the data in the whole table or specific rows and columns. Using the "dataframe" object in the Pandas library, manipulation and querying of data is possible. In addition, pandas make it possible to read dataframes from the file and to write dataframes to the file. Thus, pandas is an important Python library for reading, processing and saving data.

Another library that is widely used in data analysis and data science in Python is "numpy". Machine learning algorithms are run after the data has been digitized and put into the form of arrays of various sizes. An array with more than one dimension is called a "matrix". The library that facilitates working on matrices and arrays is "numpy".

Python offers many possibilities for visualization as well as digitization and processing of data, "matplotlib" is a very widely used Python library for graphing. More advanced or more visual graphics can be obtained by using the "seaborn" library developed on matplotlib. Libraries such as "scipy" for scientific calculations and "re" for running regular expressions on text data were also used in this study.

One of the reasons why the Python programming language is widespread in the world of data science and data analysis is the variety of libraries. These libraries allow complex operations to be done in fewer lines that can be done in more lines in other programming languages. Thus, data scientists and analysts can abstract complex mathematical operations and focus on the data problem. Python provides similar abstraction for machine learning algorithms. Whether supervised or unsupervised,

scikit-learn is one of the most well-known and proven libraries in the field of machine learning. It contains many machine learning algorithms and enables easy training and testing of models (Pedregosa et al., 2011).

Scikit-learn is a more specialized library for classical machine learning tasks, although it also includes modules for artificial neural networks. Other libraries such as Tensorflow, Keras, Pytorch are recommended for deep learning. In this study, a language model trained with BERT using transformer architecture is used to extract sentiment data from tweets. BERT is basically a deep learning model, but in this study a deep learning model is not trained, a trained transformer, i.e., a deep learning model is used. In order to use this model, no coding is done using Tensorflow, but the Tensorflow library needs to be installed.

3.3. Obtaining the Data

The three types of data subject to this study were obtained by different methods. Shell scripts and python scripts were used to obtain, filter and transform the data.

3.3.1. Collecting Tweets about Stocks in the BIST 30 Index

The tweet data for the years 2021-2022 used in this study were obtained when third-party libraries were able to access APIs. The "snsrape" Python library, which is widely used for this purpose, was used in this study to extract tweet data. This library made it possible to obtain tweet data by giving a hashtag value and date range. In this study, a script that uses this library and can work by taking the hashtag and date range as parameters was developed, and the tweets containing the codes of the stocks that make up the BIST 30 in the hashtag were obtained on the basis of months in 2021 and 2022. This query produces user, tweet message and tweet date data (WEB-16, 2023). However, especially tweet data is raw data. Before obtaining sentiment data from tweet data, some pre-processing is required. These processes are mentioned in the following sections.

3.3.2. Obtaining BIST 30 Index Closing Values

Day-to-day closing data of the BIST 30 index in 2021 and 2022 were obtained using investing.com web platform. This platform offers charts and news feeds on many assets such as stock markets, commodities, foreign exchange. In addition, the platform provides technical analysis by offering charting tools. By giving the index code, start and end dates as parameters to the relevant page of the platform, index closing data was obtained. Although platform provides other data such as average, maximum, etc. other than closing, only the required closing data was obtained by filtering (WEB-15, 2024). Table 3.1 shows an example for index data.

Table 3.1: BIST 30 Index Closing Data for the First Five Days of 2021

Date	Closing
2021-01-04	1.653
2021-01-05	1.640
2021-01-06	1.658
2021-01-07	1.678
2021-01-08	1.694

The obtained data was kept as data frames in the code using the facilities of the “pandas” library and the final data was stored in a file. Since stock exchanges do not operate on holidays, the dataset does not include data on holidays in Türkiye.

3.3.3. Obtaining Closing Prices of Stocks in BIST 30

The daily closing data of the BIST 30 component stocks are obtained from the open website of the investment subsidiary of a private bank. Closing data can be obtained by filtering the date and stock code on this website. Apart from the closing data, many other data are also available, but only the share code, date and closing data needed in this study are obtained (WEB-14, 2024). Index components may vary over the years. For this study, the stocks that make up the BIST 30 index in 2021 and 2022 are used. Table 3.2 shows an example for the stock data.

Table 3.2: Stock Closings Sample

Stock	Date	Closing
AKBNK	2021-01-04	6.69
AKSEN	2021-08-09	6.40
EKGYO	2022-03-17	2.17
GARAN	2021-02-16	9.34
KCHOL	2022-05-24	38.5

3.4. Data Pre-Processing

In this study, tweet data is the most unstructured data type that needs cleaning and pre-processing in tweet, index closing and stock closing data. Index closing data and stock closing data are completely numerical and noise-free.

Internet correspondence has reached an extraordinary traffic and user base with the increase of social media platforms. The natural language used in social media today has different characteristics that are constantly changing and it is very noisy. This characteristic is that it is informal, produces terminology and concepts of its own, and is widely used as slang. This creates a great challenge for sentiment analysis studies on microblog texts because it increases the probability of the emotion revealed by the text. Apart from this difficulty of social media texts, another difficulty is noise (Singh and Kumari, 2016). Advertisements, irrelevant messages and messages posted to trend a keyword are the main causes of noise.

Commonly, in order to reduce the difficulties of sentiment analysis in microblog texts, some techniques such as cleaning and, if necessary, modifications to the text before sentiment calculation are applied. These are not generalized techniques for all studies, and factors such as the language of the text, the importance of characters or texts to be changed or deleted for the study, may impact the decision on which techniques to be applied. There are many methods such as deleting numbers, removing repetitive phrases, managing all-capitalized words, converting all characters to lower case, removing stop words, stemming, managing negative affixes, correcting spelling and cleaning punctuation. There are also methods for deleting usernames in messages

written by people to talk about each other or to reply to each other, and for cleaning links in advertising messages (Effrosynidis et al., 2017).

In this study, phrases that does not express emotion were removed in the pre-processing process. In this category; hashtags, mentioned usernames, web links, expressions for retweeting tweets, and new lines were evaluated. In Turkish, the same character can appear next to each other, therefore, more than two identical characters were reduced to two. In the texts, uppercase and lowercase Latin and Turkish characters, numbers and punctuation characters that can only express emotion “.,!?” were taken into consideration. If more than one punctuation mark is taken into account, only one is taken into account. Due to the financial relationship of the studied data, the numbers were not cleaned, considering that the numbers may also reflect emotions. As a result of the text cleaning and replacement process, all characters were converted to lower case. In addition, some expressions which were written separately or left blank after being written separately, and which did not make sense on their own, were merged with the preceding words. Turkish dative, locative and ablative suffixes and punctuation marks are within this scope.

Emojis are small pictures or symbols that are widely used in social media texts to express emotions, objects, situations and concepts. In this study, Turkish letters “from a to z” are taken into account in the texts, all other characters are filtered. As a result, emojis were also not taken into consideration. This decision was based on the observation that emojis produced a dominant emotion score in the emotion analysis tests conducted with texts with emojis. Adding smileys to texts that do not express emotion is very effective in the emotional polarity of the text.

Preprocessing natural language texts to make them ready for sentiment analysis is a challenging task. Social media is not a formal correspondence platform and people do not pay attention to spelling and grammar. In sentiment analysis studies, noise is tried to be removed as much as possible, but it is impossible to remove it completely.

Figure 3.2 summarizes the processes applied to raw tweets until sentiment calculation.

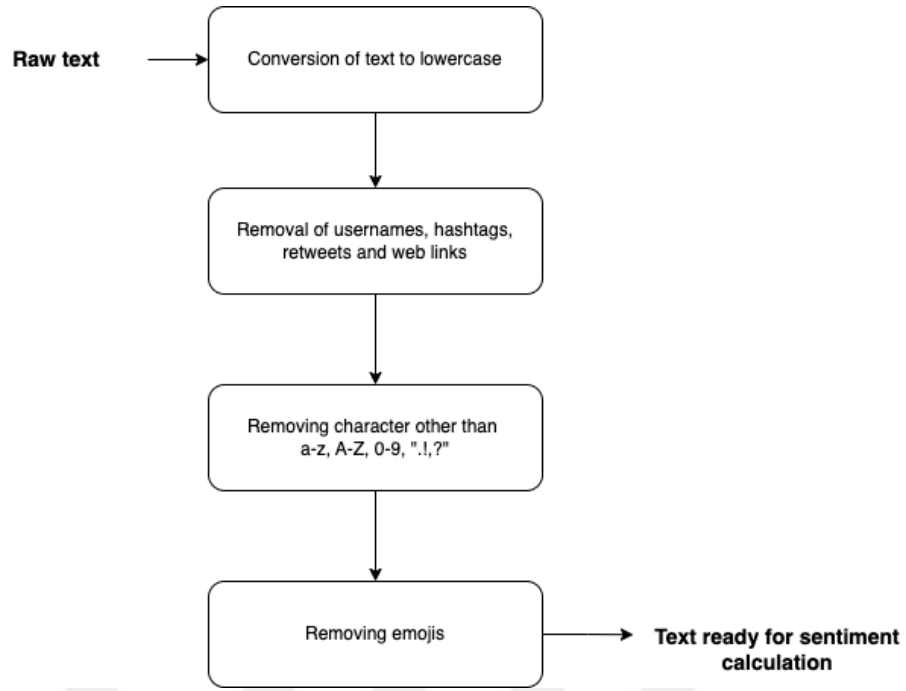


Figure 3.2: Data Cleaning and Pre-Processing Process Before Sentiment Calculation

3.5. Sentiment Calculation of Tweet Data

In this study, the BERT transformer model is used for sentiment analysis calculation. However, for this study, the model was not trained from scratch, but a trained model with Turkish language was used to calculate sentiment scores (Yıldırım, 2024). This model was trained using Turkish datasets from two different studies. One of these studies used film and shopping website reviews (Demirtas and Mykola, 2013) and the other one used tweet data (Hayran and Sert, 2017). This model generates a sentiment score between -1 and 1 from the sentences given as input. Scores less than zero indicate negative and scores greater than zero indicate positive sentiment.

Many AI developers publish and make their natural language processing models available on the Hugging Face platform established in 2016. This platform enables the models to be retrained, tested through the interface, and used on platforms like Amazon, Azure or personal computers through APIs. The developer of the model used in this study also made the model available through Hugging Face, and the model was used in this study through Hugging Face libraries (WEB-17, 2023). In this study, sentiment calculation of pre-processed raw tweets was performed using Hugging Face libraries that enable the use of this model in Python code.

Hugging Face also allows the models to be tested via a web interface. Figure 3.3 shows examples of testing a positive and a negative sentiment tweet.

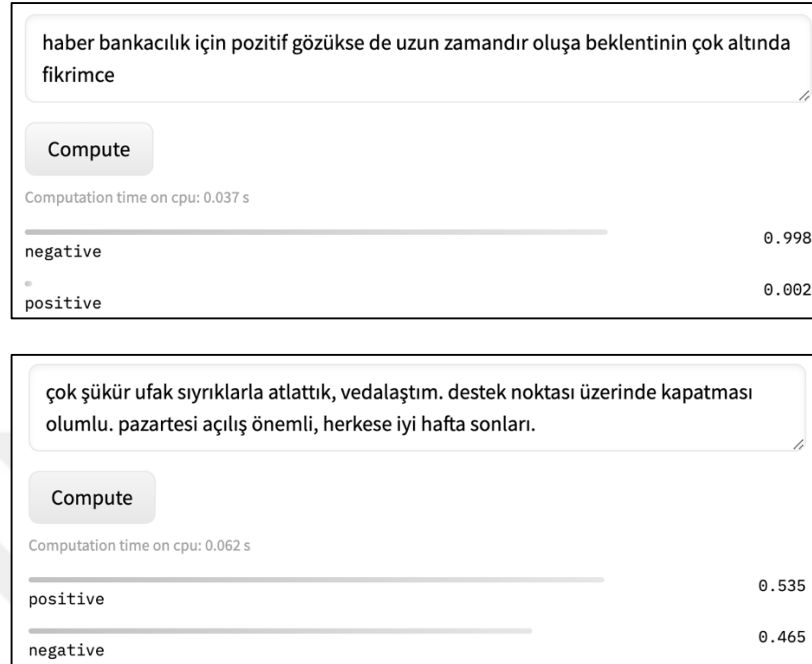


Figure 3.3: Testing Sentiment Analysis Model

3.6. Creation of Final Datasets

In parallel with the study questions, three main data sets were used in this study. These data sets were obtained by merging sentiment data with stocks and BIST 30 index data. In the merged datasets, the target (or dependent variable) is closure, the features (or independent variable) are the number of tweets with positive sentiment, the number of tweets with negative sentiment and the total number of tweets.

The main datasets before merging dataset containing the sentiment score of the tweets are; the closings of the stocks, and the closings of the BIST 30 index. From the data set related to stocks, two data sets were created on a daily basis, one for individual stocks and one for average closings. BIST 30 index data is the third basic data set. These three data sets are merged with the data set containing the sentiment score of the tweets to obtain the final data sets.

To create the sentiment data set, the raw sentiment data is obtained by creating the sentiment of the tweets downloaded between the beginning of 2021 and the end of

2022. Sentiment calculation of individual tweets is done programmatically. Technically, with the model used, a single sentiment score can be obtained for texts containing many words, but each tweet has its own context and it is deemed more appropriate to calculate sentiment scores individually. In the created data set, only the data between the opening and closing hours of the stock exchange were filtered. Figure 3.4 shows detailed process of creation final datasets and their fields.

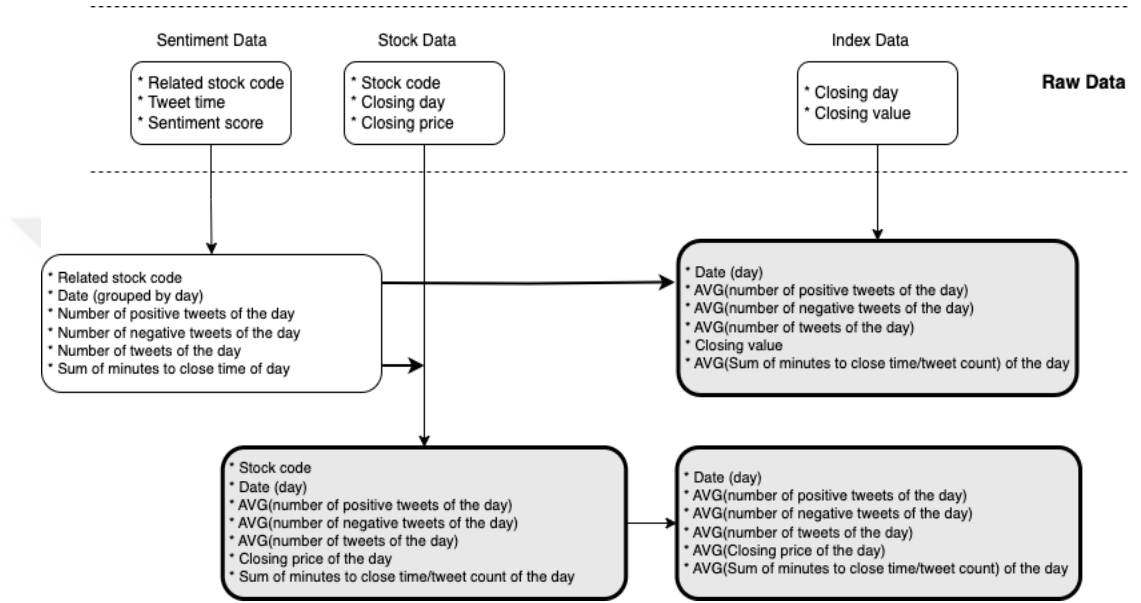


Figure 3.4: Detailed Process of Creation of Final Datasets, Features and Target Fields

Date and stock fields are not targets or features, they are just used for merging datasets. Table 3.3 shows usages of fields for each dataset.

Table 3.3: Usage of Fields for Each Dataset

	Stocks	Consolidated Stocks	Index
Average number of positive tweets of the day	Feature	Feature	Feature
Average number of negative tweets of the day	Feature	Feature	Feature
Average number of tweets of the day	Feature	Feature	Feature

Table 3.3: Usage of Fields for Each Dataset (Continued)

	Stocks	Consolidated Stocks	Index
Closing price of stock	Target	-	-
Average closing prices of the day	-	Target	-
(Sum of minutes to stock exchange close hour/Tweet counts) of the day	Feature	-	-
Average (Sum of minutes to stock exchange close hour/Tweet Counts) of the day	-	Feature	Feature
Closing value of index	-	-	Target

The stocks data set contains data on a stock-by-stock basis. The daily closing is already the closing value of the stock and averaging is unnecessary. The consolidated stock data set is the daily average of the closing prices of all stocks. One of the research questions is related to the effect of the timing of tweets on the price, and for this reason, the sum of the minutes remaining to the closing time of the tweets sent during the day is taken as one of the independent variables. However, this sum alone will be directly proportional to the number of tweets and will be dependent on the number of tweets. Therefore, a new variable was obtained by dividing this sum by the number of tweets.

In summary, features (independent variables) are number of tweets, positive tweets, negative tweets, time to close; target (dependent variable) is closing of stocks or index. All data sets were normalized before being given as input to the regression models and 25% of the data were randomly assigned as test data.

4. FINDINGS

The findings include some statistical data about the data, visual exploration of the data and results about the accuracy of the machine learning models. In this study, the largest data is the tweet data and until the analysis process, average metrics are obtained from the tweet data containing sentiment scores and combined with other data. Although the final datasets do not contain many rows, the raw and synthesized version of the tweets contains a large amount of data.

4.1. Data Exploration

The number of tweets between the beginning of 2021 and the end of 2022, which is considered raw in the study, is 1,906,966. Tweets containing share codes as hashtags were extracted separately and tweets were obtained on the basis of share codes.

After the tweets were pre-processed and filtered between the opening and closing hours of the stock market, the number of tweets subject to the study is 724,207. The first working day of 2021 is 01.01.2021 and the last working day of 2022 is 30.12.2022. There are 501 working days between these two dates (after subtracting weekend holidays, public and religious holidays). The average number of filtered tweets per day is approximately 1,446 and the number of tweets per stock is approximately 24,140. Table 4.1 shows the stocks with a number of tweets above the average number of tweets per stock.

Table 4.1: Stocks With a Number of Tweets Above the Average Number of Tweets

Stock	Number of Positive Tweets	Number of Negative Tweets	Number of Tweets	Ratio of Minutes to Closing to Number of Tweets
SASA	43,675	47,485	91,160	237.26
EREGL	26,663	29,143	55,806	232.79
THYAO	21,082	20,733	41,815	240.67
GARAN	19,296	19,630	38,926	239.87
ASELS	19,378	18,958	38,336	239.48
HEKTS	17,448	17,259	34,707	231.16
SISE	17,486	16,726	34,212	229.60
KOZAL	12,924	14,299	27,224	235.084
PETKM	12,886	11,719	24,605	235.24

Figure 4.1 shows number of tweets and stocks. The average line and the shares above it are shown. It is observed that a few stocks, in particular, have increased the average, while most of them have remained below the average.

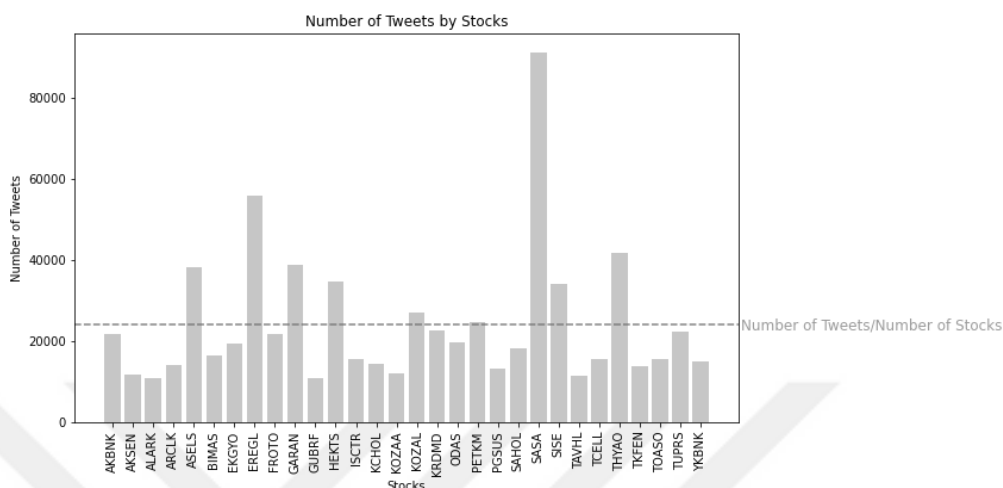


Figure 4.1: Stock Codes and Number of Tweets for Each Stock

Table 4.1 shows that most of tweets between opening and closing hours are sent 4-5 hours before closing time. Figure 4.2 shows same data as distribution of the ratio of the number of tweets to the number of minutes before the closing time of the stock exchange. This graph clearly shows that tweet activity is low near the opening and near the closing hours, with most of the activity concentrated in the middle of the trading hours. This shows that investors' convictions are largely formed near the middle of the timeframe that can be traded.

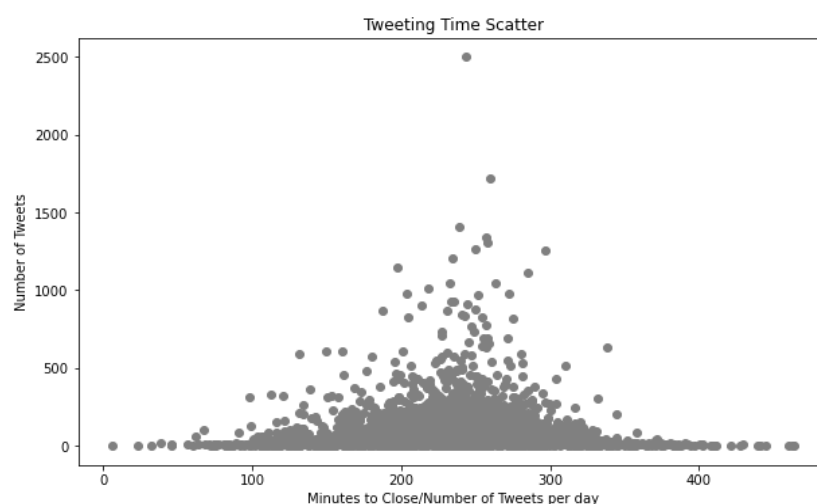


Figure 4.2: Average Remaining Rime to Close Distribution of Tweets

The COVID-19 pandemic has had some macro-economic effects all over the world. During the pandemic, countries' monetary liquification to expand social policies and aids caused high inflation rates that could not be prevented despite the high interest rate policy in many countries. Countries are still trying to eliminate the economic effects of the pandemic with different approaches. Macro-economy has been one of the most determining factors in the stock markets of countries after the COVID-19 pandemic (Jawad and Naz., 2023). Figure 4.3 shows 2021-2022 behavior of BIST 30 components. It is seen that all stocks have a similar pattern, entering a major upward trend as of the beginning of 2022. When analyzed in sectoral detail, the similarities of the trends are more significant. This is an indication of macro-economic dynamics strong enough to affect all companies in a sector.

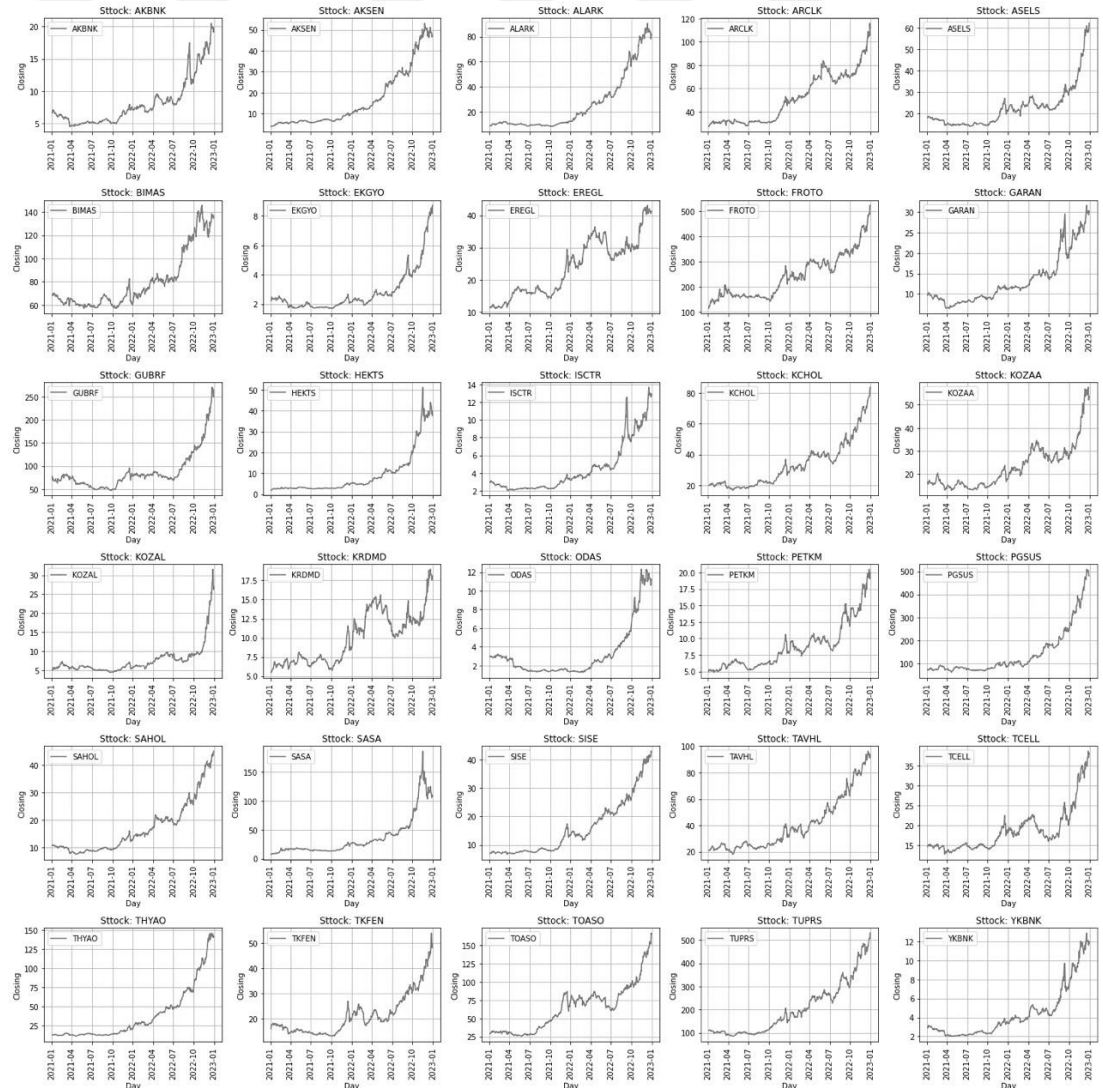


Figure 4.3: Trends of BIST 30 Components Between 2021 and 2022

Example shares for sectoral similarities; EREGL and KRDMR, iron and steel industry stocks; AKBNK, ISCTR and YKBNK, banking stocks; FROTO and TOASO, automotive sector stocks; PGSUS and THYAO, aviation stocks.

A similar trend can be seen on the basis of the index and the average closing of stocks. Figure 4.4 and Figure 4.5 show that macroeconomic factors affecting all stocks have an impact on the overall index.

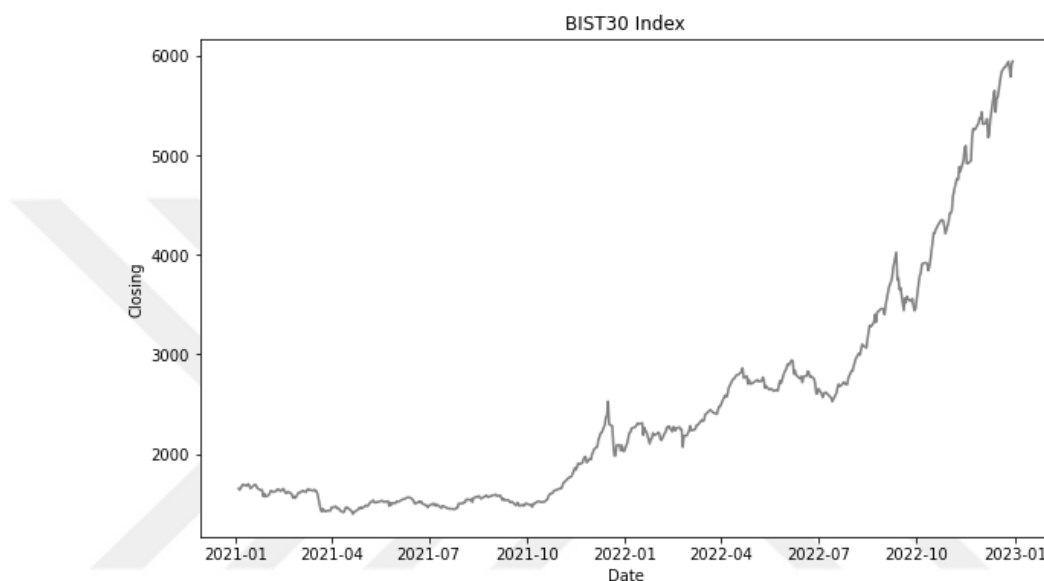


Figure 4.4: BIST 30 Index Closing Trend Between 2021 and 2022



Figure 4.5: Average Closing of BIST 30 Components Between 2021 and 2022

Although there is a main trend, Figure 4.3 shows that there are sectoral volatilities in some periods. Figure 4.6 shows tweeting activity with years and months.

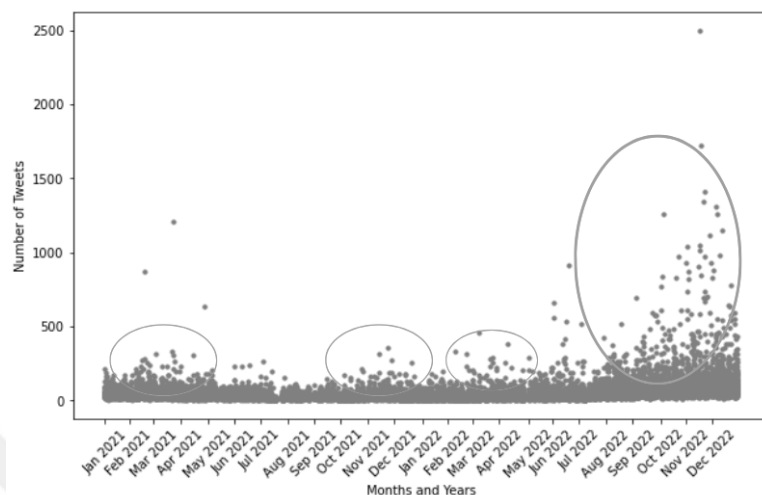


Figure 4.6: Tweeting Activity over Months and Years

The distribution of the number of tweets overlaps with the volatility of some stocks at the beginning of 2021, some at the end of 2021 and the first quarter of 2022. Likewise, it overlaps with the trend rise starting with the summer of 2022 in most of them.

Figure 4.3 shows that SASA, the most tweeted stock in Figure 4.1, rose very high in a short period of time towards the end of 2022 and fell sharply. Figure 4.7 shows that there is a social media activity parallel to the price volatility in SASA.

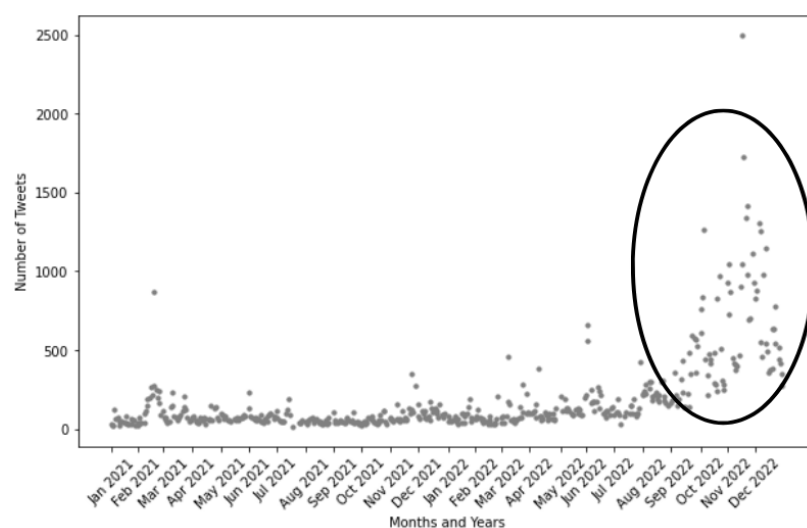


Figure 4.7: SASA Specific Tweeting Activity

Normal distribution analysis indicates that on average, the sentiment about the tweets is not overly negative or positive. Figure 4.8 shows the distribution of emotion scores.

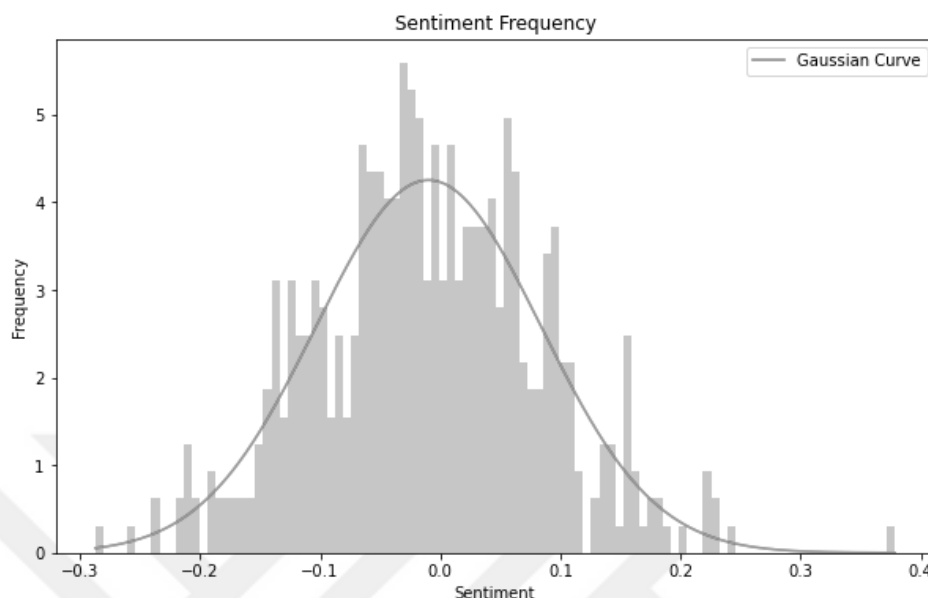


Figure 4.8: Distribution of Tweet Sentiments

A low standard deviation in the sentiment score indicates that there is a certain intensity of emotion in the tweets about that stock. Table 4.2 shows the stocks with a number of tweets above the average, standard deviation of sentiment scores, standard deviation rank in ascending order of standard deviations of all stocks.

Table 4.2: Most Tweeted Stocks and Standard Deviation of Sentiment Scores

Stock	Number of Tweets	Standard Deviation of Sentiment Score	Standard Deviation Rank in Ascending Order
SASA	91,160	0.13	1
EREGL	55,806	0.15	3
THYAO	41,815	0.14	2
GARAN	38,926	0.15	4
ASELS	38,336	0.17	6
HEKTS	34,707	0.19	8
SISE	34,212	0.16	5
KOZAL	27,224	0.19	9
PETKM	24,605	0.19	7

Stocks with above-average number of tweets were found to have the lowest standard deviation in the sentiment score distribution. This shows that tweeting activity is

directed towards a specific sentiment, and that the sentiment about the much talked about stock has become a generalized opinion.

Clustering of the average closing prices and sentiment data reveals that the closures associated with the widest spectrum sentiment score are lower, while the average closures between narrower sentiment scores are higher. Figure 4.9 shows clustering study of data.

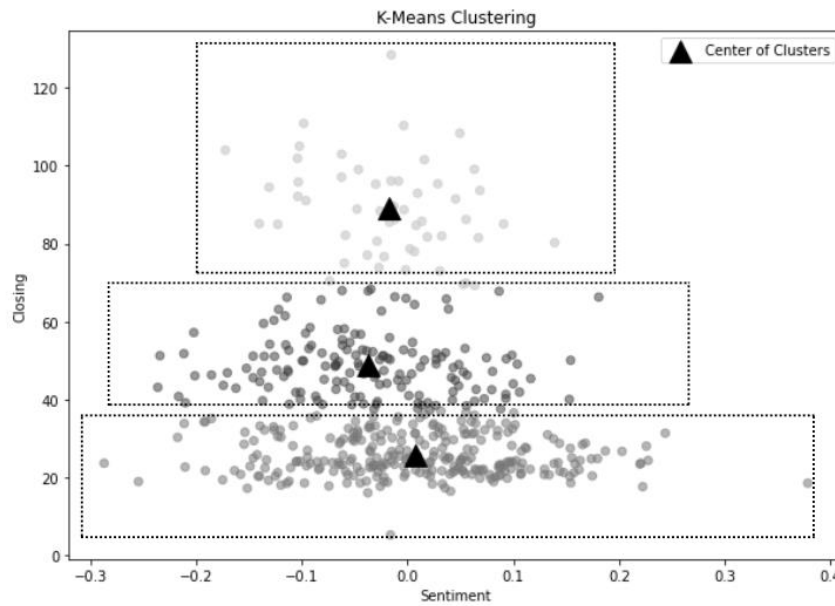


Figure 4.9: Clustering Model Between Sentiment and Average Closing

4.2. Regression Model

Regression is to obtain the best independent variable coefficient and bias value expressing the dependent variable. By creating a regression model, it is tried to fit the available data to a function such as $ax + b = y$. SVR (Support Vector Regression) algorithm was used in this study. SVR is an algorithm that can manage outlier data better than classical linear regression.

While linear regression models select the best dependent variable corresponding to the independent variable, they show lower tolerance because they try to produce a line function. SVR, on the other hand, works in a higher tolerance bandwidth, has a higher tolerance for errors, and has a lower risk of overfit. Figure 4.9 shows comparison linear regression and SVR.

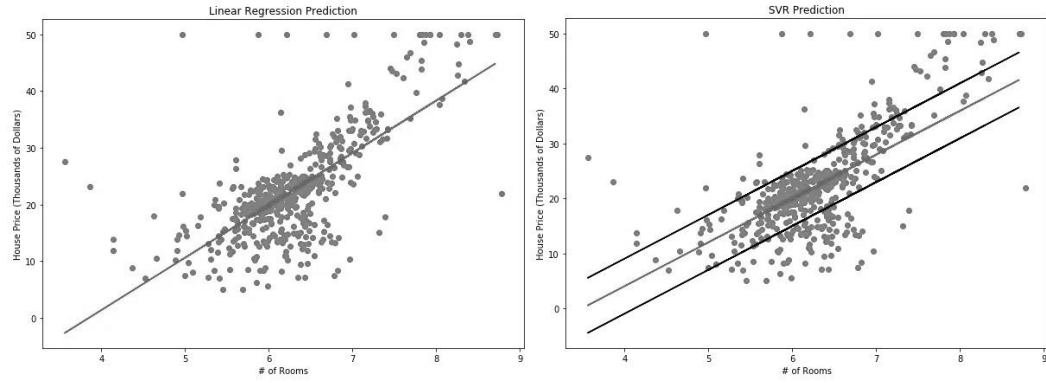


Figure 4.10: Linear Regression and SVR

In this study, regression models are developed and performance metrics of the models are obtained by using data sets created with tweet sentiment score data, stock closing data and BIST 30 index closing data. There are various metrics to measure the performance of the model. The most commonly used are R-Squared, MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error).

A significant R-Squared produces a value between 0 and 1. A value close to 1 indicates that the model explains the relationship between the dependent and independent variable very well and that the independent variable has a significant effect on the dependent variable. R-Squared can be a negative number but that means relationship between independent and dependent variable does not fit on model. A different model may explain this relationship, or there may be no significant relationship between the two variables. MAE (Mean Absolute Error) is a performance measure that quantifies how far the predictions of a regression model are from the true values. MAE is the mean of the absolute value of the error value of each prediction. MAE calculates absolute value of errors. MSE (Mean Squared Error) aims to measure the same metric as MAE, but instead of an absolute value, it is the mean square of the error value of each estimate. RMSE (Root Mean Squared Error), used for the same purpose, is the square root of the mean of the error squares. There was no ideal number for these three metrics and a low number indicates a low error rate and high performance of the model.

In the study, firstly, the data set formed by combining the average sentiment score with the BIST 30 index data was analyzed with the regression model. The number of tweets, the number of positive tweets, the number of negative tweets and the ratio of the average of the minutes remaining until the closing time of the tweets to the number of tweets (this will be called as “time ratio”) are analyzed as independent variables and

the closing value as dependent variable. The relationship between the independent variables and the dependent variable was analyzed one by one.

Table 4.3 shows the performance of the regression model constructed with the closing value for each feature using BIST 30 index data.

Table 4.3: SVR Metrics of BIST 30 Index

Metric	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
R-Squared	0.7144	0.6964	0.7022	-0.0593
MAE	0.0878	0.0904	0.0885	0.1312
MSE	0.0115	0.0121	0.0123	0.0332
RMSE	0.1076	0.1103	0.1110	0.1824

Table 4.3 shows that the independent variables (number of positive and negative and all tweets) explain approximately 70% of the variance in the dependent variables (index close). Time ratio shows a very poor fit with the model. MAE, MSE and RMSE values are also higher than those of other independent variables. This means that the error rate is higher than the others.

When a regression model is constructed with the same variables using the average closings of BIST 30 components, the values in Table 4.4 are obtained. Table 4.4 shows that the metrics are quite close to the index result shown in Table 4.3. Model fits approximately 70% with number of positive and negative and all tweets. The time ratio still does not fit with model significantly. This result shows that the closing of the index represents the closing average of the stocks very well.

Table 4.4: SVR Metrics of Consolidated Data of BIST 30 Component Stocks

Metric	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
R-Squared	0.7007	0.6798	0.6898	-0.0664
MAE	0.0876	0.0887	0.0891	0.1391
MSE	0.0105	0.0104	0.0115	0.0369
RMSE	0.1026	0.1024	0.1075	0.1921

Table 4.5 shows the R-Squared values of individual stocks. It is observed that the model fails to explain significantly for ARCLK, GARAN, GUBRF, TAVHL stocks. For 11 stocks, the R-Squared value was below the average of all stocks. 15 stocks are

above average and AKSEN, ALARK, HEKTS, ISCTR, SAHOL, SASA, SISE, YKBNK is above 50%.

Table 4.5: R-Squared Metric of Individual Stocks

Stocks	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
AKBNK	0.3451	0.3699	0.2909	-0.0558
AKSEN	0.5797	0.5721	0.5222	-0.1221
ALARK	0.5025	0.5298	0.4149	-0.0922
ARCLK	0.0909	0.1145	0.0309	-0.1311
ASELS	0.3579	0.3482	0.3366	-0.0254
BIMAS	0.1669	0.2542	0.0866	-0.0228
EKGYO	0.1449	0.1690	0.1032	-0.0060
EREGL	0.2812	0.2883	0.2657	-0.0118
FROTO	0.3233	0.3085	0.3014	-0.0357
GARAN	-0.0101	0.0211	-0.0421	-0.0273
GUBRF	0.0486	0.0557	0.0337	0.0025
HEKTS	0.6445	0.6722	0.6051	-0.0200
ISCTR	0.6395	0.6192	0.5966	-0.0740
KCHOL	0.3808	0.3066	0.3907	-0.0478
KOZAA	0.3571	0.3438	0.3248	-0.0263
KOZAL	0.2534	0.2905	0.0221	-0.0130
KRDMD	0.2129	0.1698	0.2398	-0.0515
ODAS	0.4394	0.4691	0.0355	-0.0470
PETKM	0.2940	0.2657	0.3111	-0.0449
PGSUS	0.2613	0.2289	0.2062	-0.0244
SAHOL	0.6213	0.5867	0.5745	-0.0547
SASA	0.6422	0.6548	0.6069	-0.0118
SISE	0.5421	0.5426	0.4939	-0.1199
TAVHL	0.0735	0.0540	0.0605	-0.0820
TCELL	0.2259	0.2365	0.1648	0.0049
THYAO	0.4071	0.3893	0.4081	-0.0677
TKFEN	0.2644	0.1999	0.2787	-0.0253
TOASO	0.1129	0.1370	0.0780	-0.0232
TUPRS	0.2920	0.2569	0.2873	-0.0715
YKBNK	0.5341	0.5064	0.4941	-0.0652
Average	0.3461	0.3322	0.2841	-0.464

R-Squared values of the number of positive and negative tweets are close. When the stocks with insignificant R-Squared values are excluded, the R-Squared value of positive tweets is higher in 21 out of 26 stocks. This is approximately 80% of significant stocks. Model fails to explain relationship between time ratio and closings.

4.3. Correlation Coefficients and p-values

Correlation and p-value data can be obtained using the libraries of the Python programming language. The effects of independent variables on dependent variables on the final data sets were analyzed with correlation and p-value matrices.

Table 4.6 shows that there is a high correlation between the number of positive and negative and all tweets and the index close. Parallel to this, p-values of these variables are highly significant. The fact that the time ratio has a negative correlation can be interpreted as tweets are more effective as the closing time approaches, but this is a very weak correlation and the p-value is above alpha value 0.05.

Table 4.6: Correlation Coefficients and P-Values for BIST 30 Index

Metric	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
correlation	0.8466	0.8363	0.8406	-0.0379
p-value	3,91E-134	6,49E-128	1,74E-130	4,05E-01

Table 4.7 shows that there is a high correlation between the number of positive and negative and all tweets and the average closing of stocks. p-values of same variables are highly significant. The correlation coefficients and p-values are quite close to the index analysis, but only slightly lower. Similar to Table 4.6, the time ratio has a very low negative correlation with the close and the p-value is greater than 0.05.

Table 4.7: Correlation Coefficients and P-Values for BIST 30 Component Stocks

Metric	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
correlation	0.8385	0.8259	0.8329	-0.0786
p-value	3,22E-129	4,45E-122	5,84E-126	8,38E-02

Both analyses show that the effect of positive and negative sentiment scores on closure is quite close. Index and consolidated stock analysis are close to each other, similar to the findings in the regression analysis. As in the regression analysis, this indicates that the index is a fairly good representation of the stocks it includes.

Table 4.8 shows the correlation analysis of individual stocks. It seems, number of tweets, positive and negative tweets are highly correlated with the closing variable.

Table 4.8: Correlation Coefficients of Individual Stocks

Stock	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
AKBNK	0.5943	0.6145	0.5512	-0.0478
AKSEN	0.7634	0.7568	0.7235	-0.0104
ALARK	0.7118	0.7282	0.6500	0.0167
ARCLK	0.3436	0.3693	0.2668	0.0625
ASELS	0.6068	0.5936	0.5983	0.0058
BIMAS	0.4384	0.5082	0.3652	-0.0269
GUBRF	0.2621	0.2599	0.2335	-0.0812
HEKTS	0.8219	0.8327	0.7970	0.0165
SASA	0.8087	0.8133	0.7903	0.0469
EKGYO	0.4276	0.4571	0.3717	-0.0242
THYAO	0.6580	0.6372	0.6489	-0.1706
FROTO	0.5690	0.5564	0.5494	0.0834
ISCTR	0.8027	0.7908	0.7740	0.0114
GARAN	0.1744	0.2443	0.1163	-0.1421
KCHOL	0.6275	0.5742	0.6293	-0.0402
KOZAL	0.5694	0.5684	0.5407	-0.0028
KRDMD	0.4937	0.4378	0.5177	0.1378
ODAS	0.6659	0.6891	0.5990	-0.0319
PETKM	0.5701	0.5441	0.5763	-0.0497
PGSUS	0.5484	0.5260	0.5274	-0.0121
SAHOL	0.7893	0.7670	0.7583	-0.0156
SISE	0.7366	0.7380	0.7041	0.0343
TAVHL	0.3422	0.3199	0.3323	-0.0143
TCELL	0.5087	0.5220	0.4678	0.1457
TKFEN	0.5296	0.4543	0.5438	0.0126
TOASO	0.3754	0.3963	0.3185	0.0580
TUPRS	0.4494	0.5317	0.5538	0.0423
YKBNK	0.7406	0.7217	0.7124	-0.0815
EREGL	0.5364	0.5392	0.5194	0.0962
KOZAA	0.6144	0.6062	0.5750	0.0431
Average	0.5693	0.5699	0.5437	0.0021

Table 4.9 shows distribution of stock numbers, gives number of stocks that have greater correlation coefficient than average, 50% and 40%.

Table 4.9: Distribution of Correlation Coefficient Values of Stocks

Comparison	Number of Tweets	Number of Positives	Number of Negatives
> Average	16	14	18
> 0.5	21	22	22
> 0.4	25	25	23

All tweets, positive and negative tweets of all stocks have p-values below 0.05. Again, all of them have time ratio values above 0.05. There is a high significant relationship between all variables except time ratio and closing. For this reason, it is unnecessary to give the entire p-value table, Table 4.10 is shared for 15 sample stocks.

Table 4.10: P-Values for Sample Stocks

Stock	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
AKBNK	1.88E-47	1.71E-51	9.58E-40	2.94E-01
BIMAS	4.11E-24	4.40E-33	1.10E-16	5.54E-01
GUBRF	4.91E-09	6.69E-09	2.09E-07	7.45E-02
SASA	5.95E-113	3.04E-115	2.05E-104	3.03E-01
EKGYO	6.73E-23	2.53E-26	2.83E-17	5.95E-01
THYAO	2.99E-56	2.17E-56	4.51E-59	1.00E-04
ISCTR	1.22E-104	1.22E-104	1.51E-97	8.01E-01
KCHOL	3.07E-54	1.03E-43	1.23E-54	3.76E-01
KRDMD	4.71E-31	4.91E-24	1.79E-34	2.30E-03
TCELL	3.75E-33	4.11E-35	1.21E-27	1.30E-03
TKFEN	2.76E-36	5.88E-26	1.55E-38	7.82E-01
TOASO	1.30E-17	1.26E-19	7.50E-13	2.03E-01
TUPRS	3.98E-41	1.34E-36	3.51E-40	3.53E-01
YKBNK	4.25E-85	6.71E-79	5.53E-76	7.35E-02
EREGL	2.45E-37	8.81E-38	9.95E-35	3.40E-02

5. DISCUSSION AND CONCLUSION

This study attempts to understand the impact of Twitter (now called X), a widespread social media and more specifically a microblog, on the stock market. This section discusses whether the findings contribute to the theoretical background of the study and what practical implications and contributions the findings have. In addition, the difficulties and limitations that may necessitate revisiting the interpretations and conclusions are also mentioned.

5.1. Theoretical Contributions

In this study, a transformer model based on deep learning was used to obtain sentiment scores. Transformers are architectures that are successful in modeling sequential inputs that have relationships with each other, like natural language. Deep learning has matured later in natural language processing than in fields such as image processing. Before transformers, different approaches such as Long-Short Term Memory (LSTM) and Recurrent Neural Network (RNN) were also put forward. A study using moving average, linear regression, autoregressive integrated moving average (ARIMA) and LSTM in stock forecasting with sentiment analysis compared the performances of the methods. It showed that LSTM, a deep learning approach, performed better than other techniques (Shubham et al., 2024). Another study examining pre-transformers deep learning models on the example of sentiment analysis and stock market forecasting also revealed that LSTM performed better than other models (Pattanayak et al., 2024).

The shortcomings of LSTM in modeling longer natural language texts have been overcome with the transformer. Since the technical architecture of the model was only revealed in 2017 (Vaswani et al., 2017), it has been used as a method in fewer studies than other techniques. In this study, a transformer model called Bidirectional Encoder Representations from Transformers (BERT) is used. In a study using BERT in stock sentiment analysis, message data received from an investor platform was labeled as bearish and bullish and the BERT model was tuned. Unlike this study, the model, which aims to classify the message rather than stock prediction, performed quite well (Lee et al., 2020).

This study shows the impact of people's activities reflected on social media on the market by analyzing digital data with known machine learning techniques. The model, in which the number of tweets representing social media activity are used as feature, can obviously explain prices and closings. A study examining whether social media or conventional media is more effective on companies found a stronger relationship between sentiment data and companies' returns and risks (Yu et al., 2012).

The number of tweets with positive and negative sentiments did not show significantly different performance in explaining the average closing prices of indices and stocks. However, when individual stocks were examined, it was found that the number of tweets with positive sentiment was able to explain price increases better than those with negative sentiment in the majority of them. A study examining the relationship between the frequency of words expressing happiness, fear, sadness, and excitement in tweet data and the closing prices of the Dow Jones, S&P 500, and NASDAQ indices found that, similarly, the index rose on days following days with a higher prevalence of positive sentiment expressions compared to days with a higher prevalence of negative sentiment expressions, specifically for the Dow Jones index (Zhang et al., 2011).

It performs slightly better in explaining the average number of negative tweets for the general index and stocks. In a study where sentiment data of articles was studied, the better performance of negative polarity was evaluated as an indicator that negative emotions may have a more dominant effect (Schumaker et al., 2012). However, for this study, the difference between the performances of the two polarities is not significant. Another study examining the impact of news feeds on stock prices using sentiment analysis concluded that although sentiment analysis increases prediction accuracy, models that rely on polarity-based forecasting, scored as positive and negative, do not perform well (Li et al., 2014).

In this study, social media activity represented by the number of tweets is a more effective factor than sentiment scores. A study that predicts NASDAQ stocks by obtaining sentiment scores using deep learning on tweet data has revealed that although positive sentiment is an important factor in explaining the price, it cannot be generalized on all stocks. It is explained that having more news flow about some stocks results in more social media activity and activity moves in parallel with positive sentiment (Guo and Xie, 2024).

5.2. Practical Contributions

This study is an example of the use of artificial intelligence and machine learning techniques in finance. In the domain of finance, minimum information has been used and the study is mainly data-driven. It is an example that similar studies can be done in other domains with these techniques. This study is an indication that these techniques can be easily combined in more than one domain and holistic results can be obtained. In this study, the use of third-party software was kept to a minimum, and visuals and analyses were obtained with the help of programming as much as possible. The findings obtained are also an indication that further studies on data can be carried out using cheap and open sources.

Individual investors, financial institutions, and portfolio managers use tools for technical and fundamental analysis to evaluate stocks. Technical analysis focuses on price changes over time using statistical algorithms, while fundamental analysis relies on balance sheet data released quarterly by companies. Both methods often overlook organic data, such as news and psychology, influencing investor behavior. Technical analysis is typically used for short-term trading, and fundamental analysis for long-term strategies. Professional portfolio managers use tools that combine both analyses. This study reveals that these traditional techniques of modeling investor behavior and sentiment can improve and strengthen stock analysis

The findings of this study show the impact of public opinion on social media on stock prices and indices, and suggest that the research can be integrated into existing tools as a software solution. Decision support systems using short-term technical analysis can also be fed into decision support systems using short-term technical analysis by measuring investor sentiment in a continuous news feed and social networks. On the other hand, the metrics of these models can be added to tools that recommend long-term investment decisions based on parameter ratios in balance sheet data.

Sentiment analysis for financial forecasting requires significant computational resources and real-time data processing. While high-frequency trading demands sophisticated AI systems, simpler models can be used for longer-term investments. Sentiment analysis can complement traditional fundamental analysis by providing insights into market sentiment and company perception. This combined approach can help investors make more informed decisions.

5.3. Limitations and Future Research Suggestions

There are some technical difficulties and limitations in this study, and the findings show traces of these limitations. The difficulty of stock forecasting studies due to their nature, the fact that many parameters are involved, reveals the difficulty of the study in general.

The most challenging technical part of this study is the obtaining of sentiment scores. In order to obtain the sentiment scores, a large number of tweets were captured, cleaned and scores were calculated. The preprocessing of tweets is a very challenging and difficult to generalize process, which causes the data to always have noise. It is not an easy task to extract numerical data from a structure as complex as natural language, where even punctuation marks have a meaningful effect. The main data on which the study is based is sentiment score data, and it is not possible to analyze hundreds of thousands of tweets individually. At this point, although tested, the artificial intelligence model that produces the sentiment score was relied on. Artificial intelligence is inherently a black box and it is impossible to reverse engineer what it decides and why.

Since the study required processing a large number of tweets, it required a large amount of computing resources, and it took a long time to obtain the sentiment scores. The artificial intelligence model used here is an improved model learnt on another text. Tuning the model to learn with texts containing stocks would provide more accurate results for future studies. However, the creation of such data and the high GPU resource required would be the biggest challenges here. The reason why the number of positive and negative tweets had close effects in both regression models and correlation analyses may be due to the difficulties in obtaining sentiment scores.

In this study, daily closure data were used and emotion scores were also grouped daily. In future studies, working with more granular time periods would give more significant results. The relationship between tweet times and closing could not be significantly revealed in this study. This may be due to the fact that the study was conducted with a wide interval of one day.

The fact that the stocks and the index move with very similar trends and the volatility is low may have caused the study not to obtain more significant results between sentiment and closings. A significant relationship similar to the one between tweet

activity and closings did not reveal the effect of positive or negative sentiment in a continuously rising market. This shows that in future AI-supported finance studies, in addition to solving the technical constraints mentioned above, taking macro and even micro-economic factors into account may reveal more significant relationships.



REFERENCES

- Ahn H., Kim K. J., (2009). Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing*, 9(2), 599–607.
- Almansour B. Y., Elkrghli S., Almansour A. Y., (2023). Behavioral finance factors and investment decisions: A mediating role of risk perception. *Cogent Economics & Finance*, 11(2).
- Bahdanau D., Cho K., Bengio Y., (2015), “Neural Machine Translation by Jointly Learning to Align and Translate”, 3rd International Conference on Learning Representations (ICLR), San Diego, California, USA, 7-9 May.
- Bentes S. R., Navas R., (2013). The Fundamental Analysis: An Overview. *Int. J Latest Trends Fin. Eco. Sc.*, 3(1), 389-393.
- Bhattacharyya R., Jung S., Kruse L., Senanayake R., Kochenderfer M. J., (2021). A Hybrid Rule-Based and Data-Driven Approach to Driver Modeling through Particle Filtering. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 13055-13068.
- Bildik Z., (2020), “Predicting Stock Returns with Macroeconomic Indicators & Google SVI & News Sentiment”, Msc. Thesis, Istanbul Technical University.
- Bollen J., Mao H., Zeng X., (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(2011), 1-8.
- Cheelamanthula K., (2022), "Implementation of Long Short Term Memory Neural Network Model for Electrical Load", 2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russian Federation, 09-11 November.
- Constantinescu A., (2012). The Business Cycles and the Influence of Economic Confidence Indicators in the European Region. *Romanian Economic Journal*, 15(44), 201-224.
- Demirtas E. and Mykola P., 2013. “Cross-lingual polarity detection with machine translation”. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '13), Chicago, IL, USA, 11 August.
- Devlin J., Chang M. W., Lee K., Toutanova K., Manning, M. D., (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL North American Chapter of the Association for Computational Linguistics*, 2018(1), 417-428.
- Dunis C. L., Middleton P. W., Theofilatos K., Karathanasopoulos A., (2016), “Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics”, Palgrave Macmillan

Effrosynidis D., Symeonidis S., Arampatzis A., (2017). “A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis”. 21st International Conference on Theory and Practice of Digital Libraries, Thessaloniki, Greece, 18-21 September.

Géron A., (2022), “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, O'Reilly Media, Inc.

Gidofalvi G., Elkan C., (2001). Using news articles to predict stock price movements. *Department of computer science and engineering, University of California, San Diego* 17 (2001).

Gold S., (2015). The Viability of Six Popular Technical Analysis Trading Rules in Determining Effective Buy and Sell Signals: MACD, AROON, RSI, SO, OBV, and ADL. *Journal of Applied Financial Research*, 2(2015), 8-29.

Guo K., Xie H., (2024). Deep learning in finance assessing twitter sentiment impact and prediction on stocks. *PeerJ Computer Science*, 10(2018).

Hayran A., Sert M., (2017), "Sentiment Analysis on Microblog Data based on Word Embedding and Fusion Techniques", IEEE 25th Signal Processing and Communications Applications Conference (SIU 2017), Antalya, Turkey, 15-18 May.

Hájek P., (2011). Municipal credit rating modelling by neural networks. *Decision Support Systems*, 51(1), 108–118.

Hutto C. J., Gilbert E.E., (2014). “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”. Eighth International Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1-4 June.

Jawad M., Naz M., (2023). Impact of Covid-19 pandemic on macroeconomic aspects. *Journal of Open Innovation: Technology, Market, and Complexity*, 9(3), 100126.

Kumar A., (2009). Hard-To-Value Stocks, Behavioral Biases, and Informed Trading. *Journal of Financial and Quantitative Analysis*, 44, 1375-1401.

Lecun Y., Bengio Y., Hinton G., (2015). Deep Learning. *Nature*, 521(7553), 436-444.

Lee C., Gao Z., Tsa, C., (2020). “BERT-Based Stock Market Sentiment Analysis”. 2020 IEEE International Conference on Consumer Electronics, Taoyuan, Taiwan, 4-6 November.

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer, L., (2020). “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. 58th Annual Meeting of the Association for Computational Linguistics, Online Conference, 5-10 July.

Li X., Xie H., Chen L., Wanga J., Deng X., (2014), News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69(2014), 14-23.

Liu Z., Yang L., Takada T., (2023), “Predicting Stock Prices Using Tweet Frequency and AI: Leveraging Social Media Insights to Forecast Tomorrow's Market Trends”, 2th International Conference on Digital Image Processing and Vision, California, USA.

Liu B., (2012), “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers.

McCulloch M. S., Pitts W., (1943), A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5, 115-133.

Mende H., Peters A., Ibrahim F., Schmitt R. H., (2022), “Integrated decision-support system for diagnosis, maintenance planning, and scheduling of manufacturing systems”, 32nd CIRP Design Conference, Online Conference, 28-30 March.

Motiwalla L., Wahab M., (2000). Predictable variation and profitable trading of US equities: A trading simulation using neural networks. *Computers & Operations Research*, 27, 1111–1129.

Nadkarni P. M., Ohno-Machado L., Chapman W. W., (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), 544-551.

Naseer M., Tariq Y. B., (2015). The Efficient Market Hypothesis: A Critical Review of the Literature. *The IUP Journal of Financial Risk Management*, 12(4), 48-63.

Nasir M. A., Huynh T. L. D., Nguyen S. P., Duong D., (2019). Forecasting cryptocurrency returns and volume using search engines. *Financial Innovation*, 5 (1), 2. doi: 10.1186/s40854-018-0119-8.

Pattanayak A. M., Swetapadma A., Sahoo B., (2024). Exploring Different Dynamics of Recurrent Neural Network Methods for Stock Market Prediction - A Comparative Study. *Applied Artificial Intelligence*, 38(1).

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., and others., (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(2011), 2825-2830.

Ranco G., Aleksovski D., Caldarelli G., Grčar M. and Mozetič I., (2015). The Effects of Twitter Sentiment on Stock Price Returns. *PLoS ONE*, 10(9).

Rao T., Srivastava S., (2012). “Analyzing Stock Market Movements Using Twitter Sentiment Analysis”. International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Türkiye, 26-29 August.

Saltz J. S., Grady N. W., (2017), "The Ambiguity of Data Science Team Roles and the Need for a Data Science Workforce Framework", IEEE International Conference on Big Data (BIGDATA), 2355-2361, Boston, MA, USA, 11-14 December.

Schumaker R. P., Zhang Y., Huang C. H., (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3), 458-464.

Shah D., Isah H., Zulkernine F., (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studie*, 7(2), 26.

Singh T., Kumari M., (2016). Role of Text Pre-Processing in Twitter Sentiment Analysis. *Procedia Computer Science*, 89(2016), 549-554.

Sohangir S., Wang D., Pomeranets A., Khoshgoftaar T. M., (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, 5(1), 1-25.

Souza, T.T.P., Kolchyna, O., Treleaven, P.C., Aste, T., (2015). Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry. ArXiv:1507.00784.

Thawornwong S., Enke D., (2004). The adaptive selection of financial and economic variables for use with artificial neural networks. *Neurocomputing*, 56, 205–232.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., (2017), “Attention Is All You Need”, 31st International Conference on Neural Information Processing Systems, 5998-6008, Long Beach, California, USA, 4-9 December.

Wang M., Wang Z., Lu J., Lin J., (2019). E-LSTM: An Efficient Hardware Architecture for Long Short-Term Memory. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2), 280-291.

WEB-1, <https://www.borsaistanbul.com/en/index>, (Eriřim Tarihi: 25./02/2024).

WEB-2, (2020), <https://www.borsaistanbul.com/files/bist-stock-indices-ground-rules.pdf>, (Eriřim Tarihi: 25./02/2024).

WEB-3, (2024), <https://tr.tradingview.com>, (Eriřim Tarihi: 25./02/2024).

WEB-4, (2023), <https://www.forbes.com/sites/forbestechcouncil/2023/12/01/why-ai-is-popular-now-and-two-ways-to-use-it-better/>, (Eriřim: 29.03.2024).

WEB-5, (2023), <https://www.ibm.com/blog/machine-learning-types>, (Eriřim: 03.04.2024).

WEB-6, (2020), <https://www.bmc.com/blogs/deep-neural-network>, (Eriřim Tarihi: 02.04.2024).

WEB-7, (2023), <https://medium.com/@asifurrahmanaust/lesson-3-neural-network-is-nothing-but-a-linear-regression-e05a328a0f23>, (Eriřim Tarihi: 08.04.2024).

WEB-8, (2023), <https://www.geeksforgeeks.org/activation-functions-neural-networks>, (Eriřim Tarihi: 08.04.2024).

WEB-9, (2023), <https://medium.com/@Coursesteach/deep-learning-part-1-86757cf5a0c3>, (Eriřim Tarihi: 08.04.2024).

WEB-10, (2024), https://en.wikipedia.org/wiki/Recurrent_neural_network, (Eriřim Tarihi: 21.04.2024).

WEB-11, (2015), <https://colah.github.io/posts/2015-08-Understanding-LSTMs>, (Eriřim Tarihi: 21.04.2024).

WEB-12, (2020), <https://towardsdatascience.com/transformers-89034557de14>, (Eriřim Tarihi: 22.04.2024).

WEB-13, (2024), <https://en.wikipedia.org/wiki/Twitter>, (Eriřim Tarihi: 30.04.2024).

WEB-14, (2024), <https://www.isyatirim.com.tr/tr-tr/analiz/hisse/Sayfalar/Tarihsel-Fiyat-Bilgileri.aspx>, (Eriřim Tarihi: 30.04.2024).

WEB-15, (2024), <https://tr.investing.com/indices/ise-30>, (Eriřim Tarihi: 30.04.2024).

WEB-16, (2023), <https://github.com/JustAnotherArchivist/snsrape>, (Eriřim Tarihi: 10.03.2023)

WEB-17, (2023), <https://huggingface.co/savasy/bert-base-turkish-sentiment-cased>, (Eriřim Tarihi: 10.03.2023)

White H., (1988). "Economic prediction using neural networks: The case of IBM daily stock returns". Second IEEE annual conference on neural networks, San Diego, CA, USA, 24-27 July.

Yıldırım S. (2024). "Fine-tuning Transformer-based Encoder for Turkish Language Understanding Tasks.", arXiv:2401.17396.

Yu Y., Duan W., Cao Q., (2012). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4), 919-926.

Zhang X., Fuehres H., Gloor P. A., (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia- Social and Behavioral Sciences*, 26(2011), 55-62.

BIOGRAPHY

Mustafa Kemal MAYUK graduated from Karadeniz Technical University Computer Engineering in 2007. For 16 years, he has worked in various areas of the IT sector such as software and system management. He is still working as a manager in IT infrastructure related to information technologies in a corporate company. In 2022, he started the M.Sc programme at Gebze Technical University, Department of Management.



PUBLICATIONS AND PRESANTATIONS FROM THE THESIS

Mayuk M. K., (2023), “Examining the effect of social networks on the stock exchange with machine learning methods”, 6th International Antalya Scientific Research and Innovative Studies Congress, 815-821, Antalya, Türkiye, 02-04 December.

