Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Economics

Master's Thesis

The Impact of Social Media Sentiment on NASDAQ Stock Prices

Rustam Bashirov

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Bc. Rustam Bashirov

Economics and Management
Economics and Management

Thesis title

The Impact of Social Media Sentiment on NASDAQ Stock Prices

Objectives of thesis

1. To investigate the relationship between social media sentiment and stock prices for a sample of companies listed on the NASDAQ stock exchange.
2. To identify the factors that influence social media sentiment towards a particular stock or company listed on NASDAQ.
3. To analyze the impact of social media sentiment on stock prices for the sample companies, using data analytics and statistical techniques.
4. To examine the relationship between social media sentiment and stock prices over time, and to identify any patterns or trends that may exist.
5. To identify any differences in the impact of social media sentiment on stock prices for companies in different industries or sectors.
6. To evaluate the effectiveness of using social media sentiment as a tool for predicting stock prices and to understand the limitations of using social media sentiment for the same.
7. To provide insights and recommendations for investors, financial analysts and other stakeholders, as well as for researchers working in the field of social media sentiment analysis and stock market analysis.
8. To contribute to the existing body of knowledge on the impact of social media sentiment on stock prices.

Methodology

Data collection: The first step is to collect the necessary data, which includes social media sentiment scores and stock prices for the relevant time period. This data can be collected through various sources, including API access to social media platforms and financial data providers.

Data preprocessing: Once the data has been collected, it needs to be preprocessed to make it ready for analysis. This includes removing any duplicates, filling in missing values, and formatting the data to ensure it is compatible with the analysis tools.

Sentiment analysis: The next step is to perform sentiment analysis on the social media data to extract the sentiment scores for each tweet or post. This can be done using various techniques, including natural language processing (NLP) and machine learning algorithms.

Correlation analysis: After the sentiment scores and stock prices have been collected and preprocessed, the next step is to perform correlation analysis to investigate the relationship between social media sentiment and stock prices. This can be done using various statistical methods, including regression analysis.

Time series analysis: To further investigate the relationship between social media sentiment and stock prices, time series analysis can be performed. This involves analyzing the data over a period of time to identify patterns and trends that may be affecting the relationship.

Econometric modeling: Finally, econometric modeling can be used to build a predictive model that can forecast stock prices based on social media sentiment scores. This involves using various statistical models, including autoregressive integrated moving average (ARIMA) models, to analyze the time series data and make predictions about future stock prices.

The proposed extent of the thesis

60 pages

Keywords

Sentiment analysis, Stock prices, Social media, Twitter, Correlation analysis, Time series analysis, Macroeconomic indicators, Market capitalization, Financial modelling, Python programming.

Recommended information sources

”Financial Sentiment Analysis: An Introduction to the Quantitative Fundamentals of Economic Indicators and Trading Strategies” by Jon Danielsson and Robert Macrae

”Sentiment Analysis for Financial Applications” by Erik Cambria and Amir Hussain

”Sentiment Analysis in Finance: Concepts, Methods, and Techniques” by Domenico Sarno, Paolo Falbo, and Gaetano Zazzaro

”Social Media and Financial Markets: A Comprehensive Guide to Sentiment Analysis” by Noura Al Moubayed, Abir Hussain, and Abid Hussain

Expected date of thesis defence

2021/22 SS – FEM

The Diploma Thesis Supervisor

doc. Ing. Petr Procházka, MSc, Ph.D.

Supervising department

Department of Economics

Electronic approval: 23. 3. 2023	Electronic approval: 28. 3. 2023
prof. Ing. Lukáš Čechura, Ph.D.	**doc. Ing. Tomáš Šubrt, Ph.D.**
Head of department	Dean

Prague on 29. 03. 2023

Declaration

I declare that I have worked on my master's thesis titled " The Impact of Social Media Sentiment on NASDAQ Stock Prices " by myself and I have used only the sources mentioned at the end of the thesis. As the author of the master's thesis, I declare that the thesis does not break any copyrights.

In Prague on 31.03.2023 ______________________________

Acknowledgement

I would like to express my deepest appreciation to doc. Ing. Petr Procházka, MSc, Ph.D. for his valuable guidance, support, and encouragement throughout my research. Furthermore, I extend my gratitude to my family, Mrs. Tatyana Bashirova, Dinara Bashirova, and Mr. Farhad Bashirov, for their unwavering belief in me and for standing by my side every step of the way. I would like to express my gratitude to Alisa Danileyko for being my source of inspiration, Ing.Nazar Tarlanli his invaluable ideas and advice and to all my dear friends who have supported me and provided me with encouragement during this journey.

Last but not least, I want to thank me
I want to thank me for believing in me
I want to thank me for doing all this hard work
I want to thank me for having no days off
I want to thank me for, for never quitting.

The Impact of Social Media Sentiment on NASDAQ Stock Prices

Abstract

This study aims to investigate the relationship between social media sentiment and NASDAQ stock prices. The objective is to determine whether there is a significant impact of social media sentiment on the NASDAQ index and individual stock prices. The study will use data from Twitter to extract sentiment scores related to NASDAQ-listed companies, and then perform a correlation analysis to examine the relationship between social media sentiment and stock prices.

In addition, the study will also consider the impact of other macroeconomic variables such as interest rates, GDP, and inflation on stock prices. This will be done by incorporating these variables in a time series model to test for any significant effects on the NASDAQ index and individual stocks.

The findings of this study will contribute to the literature on the relationship between social media sentiment and stock prices. It will also provide insights for investors on how to use social media sentiment as a tool for making investment decisions.

The study has some limitations, including the availability and quality of data from social media and the macroeconomic variables. Nonetheless, the research methodology adopted in this study will ensure that the results are robust and reliable.

Overall, this study is important for understanding the impact of social media sentiment on the stock market and for providing useful information for investors and policymakers.

Keywords: Sentiment analysis, Stock prices, Social media, Twitter, Correlation analysis, Time series analysis, Macroeconomic indicators, Market capitalization, Financial modelling, Python programming.

Vliv sentimentu sociálních sítích na ceny akcií na burze NASDAQ.

Abstrakt

Diplomová práce klade za cíl zkoumat vztah mezi sentimentem na sociálních médiích a cenami akcií na burze NASDAQ. Cílem je zjistit, zda má sociální mediální sentiment významný dopad na index NASDAQ a ceny jednotlivých akcií. Práce bude využívat data z Twitteru k extrakci sentimentálních skóre týkajících se společností uvedených na burze NASDAQ a poté provede korelační analýzu k posouzení vztahu mezi sentimentem na sociálních médiích a cenami akcií.

Kromě toho bude také zvažovat vliv jiných makroekonomických proměnných, jako jsou úrokové sazby, HDP a inflace, na ceny akcií. To bude provedeno začleněním těchto proměnných do časového modelu k testování přítomnosti významného efektu na index NASDAQ a jednotlivé akcie.

Výsledky této práce přispějí k literatuře o vztahu mezi sentimentem na sociálních médiích a cenami akcií. Poskytnou také informace pro investory, jak využít sentiment na sociálních médiích jako nástroj pro rozhodování o investicích.

Diplomová práce má některá omezení, včetně dostupnosti a kvality dat z sociálních médií a makroekonomických proměnných. Nicméně zvolená metodologie zajistí, že výsledky jsou robustní a spolehlivé.

Celkově je tato práce důležitá pro porozumění dopadu sentimentu na sociálních médiích na akciový trh a pro poskytování užitečných informací pro investory a politiky.

Klíčová slova: Analýza sentimentu, ceny akcií, sociální média, Twitter, korelační analýza, analýza časových řad, makroekonomické ukazatele, tržní kapitalizace, finanční modelování, programování v Pythonu.

Table of content

1 Introduction 12

2 Objectives and Methodology 13

2.1 Objectives 13

2.2 Methodology 13

3 Literature Review 15

3.1 Stocks 15

3.2 NASDAQ Stock Exchange 16

3.2.1 S&P 500 17

3.3 Social Media 19

3.3.1 Twitter 22

3.4 Selected Stocks 23

3.4.1 Apple Inc. 23

3.4.2 Amazon Inc. 25

3.4.3 Microsoft Inc. 27

4 Practical Part 29

4.1 Data Collection 29

4.1.1 Data Pre-processing 30

4.2 Sentiment Analysis 31

4.2.1 VADER Module 32

4.3 The Augmented Dickey-Fuller Analysis 34

4.4 Differencing 36

4.5 Regression Analysis and OLS Model 37

4.5.1 Regression Analysis 37

4.5.2 The Ordinary Least Squares (OLS) Model 38

4.6 ARIMA&SARIMAX 40

4.6.1 ARIMA 40

4.6.2 SARIMAX 42

5 Results and Discussion 43

5.1 Apple Results 43

5.1.1 Apple ADF Test Results 43

5.1.2 Differencing, Apple 44

5.1.3 OLS Model Results (2015-2019), Apple 46

5.1.4 SARIMAX Test Results, Apple 48

5.2 Amazon Stock Results 50

5.2.1 ADF Test Results, Amazon 50

5.2.2 Differencing, Amazon ... 51
5.2.3 ADF Test Results After Differencing, Amazon ... 52
5.2.4 OLS Model Results, Amazon ... 53
5.3 Microsoft Results ... 55
5.3.1 ADF Test Results, Microsoft ... 55
5.3.2 ADF Test Results After Differencing, Microsoft ... 57
5.3.3 OLS Model Results, Microsoft ... 58

6 Conclusion ... 60

7 References ... 61
7.1 Internet sources ... 65

9 List of abbreviations ... 66

Appendix ... 67

List of Figures

Figure 1 Tech Companies Dominate S&P 500 Index..19
Figure 2 VADER Results example and Python code 33
Figure 3 ADF test after differencing example using Python,Amazon 53
Figure 4 ADF Test Results, Amazon..56

List of Tables

Table 1 Example of Differencing process (2015-2019), Apple..45
Table 2 OLS Regression Results (2015-2019), Apple 47
Table 3 SARIMAX Results (2015-2019), Apple 49
Table 4 Example of differencing process in Excel, Amazon 52
Table 5 OLS Regression Results, Amazon...54
Table 6 OLS Model Results, Microsoft...59

List of Graphs

Graph 1 Apple Stock Price with Volume (2015-2019) 24
Graph 2 Amazon Stock price with Volume (2015-2019)...26
Graph 3 ADF Test Results for Close price and Sentiment (2015-2019), Apple 44
Graph 4 ADF test results on Apple stock price after differencing (2015-2019), Apple......46
Graph 5 Relationship Between Differenced Sentiment and Differenced Price (2015-2019), Apple...48
Graph 6 ADF test results for Close price and Sentiment (2015-2019), Amazon 51
Graph 7 Relationship Between Weighted Average Sentiment and Closing Stock Price, Amazon 55
Graph 8 ADF Test Result, Microsoft...57
Graph 9 ADF Test Result After Differencing 58

1 Introduction

In recent years, the use of social media has become increasingly prevalent in our daily lives. With the growing popularity of platforms such as Twitter and Facebook, more and more people are turning to these sites to share their thoughts and opinions on various topics, including stocks and companies. As a result, social media has become an important source of information for investors, who use it to gain insight into the sentiment of the public towards a particular stock or company.

The impact of social media sentiment on stock prices is a topic that has garnered significant attention in recent years. Some studies have found that social media sentiment can have a significant impact on stock prices, particularly for companies that have a high level of social media visibility. However, the relationship between social media sentiment and stock prices is complex, and there is still much that is not understood about this topic.

The purpose of this diploma thesis is to investigate the impact of social media sentiment on NASDAQ stock prices. The study will focus on the relationship between social media sentiment and stock prices for a sample of companies listed on the NASDAQ stock exchange. The study will use data from social media platforms such as Twitter, as well as stock price data from financial databases. The research will use data analytics and statistical techniques to examine the relationship between social media sentiment and stock prices, and to identify any patterns or trends that may exist.

This thesis will contribute to the understanding of the impact of social media sentiment on stock prices by providing insights into the relationship between social media sentiment and stock prices for NASDAQ listed companies. The findings of this study will be of interest to investors, financial analysts, and other stakeholders, as well as to researchers working in the field of social media sentiment analysis and stock market analysis.

2 Objectives and Methodology

2.1 Objectives

1. To investigate the relationship between social media sentiment and stock prices for a sample of companies listed on the NASDAQ stock exchange.
2. To identify the factors that influence social media sentiment towards a particular stock or company listed on NASDAQ.
3. To analyse the impact of social media sentiment on stock prices for the sample companies, using data analytics and statistical techniques.
4. To examine the relationship between social media sentiment and stock prices over time, and to identify any patterns or trends that may exist.
5. To identify any differences in the impact of social media sentiment on stock prices for companies in different industries or sectors.
6. To evaluate the effectiveness of using social media sentiment as a tool for predicting stock prices and to understand the limitations of using social media sentiment for the same.
7. To provide insights and recommendations for investors, financial analysts and other stakeholders, as well as for researchers working in the field of social media sentiment analysis and stock market analysis.
8. To contribute to the existing body of knowledge on the impact of social media sentiment on stock prices.

2.2 Methodology

- Data collection: The first step is to collect the necessary data, which includes social media sentiment scores and stock prices for the relevant time period. This data can be collected through various sources, including API access to social media platforms and financial data providers.
- Data pre-processing: Once the data has been collected, it needs to be pre-processed to make it ready for analysis. This includes removing any duplicates, filling in missing values, and formatting the data to ensure it is compatible with the analysis tools.

• Sentiment analysis: The next step is to perform sentiment analysis on the social media data to extract the sentiment scores for each tweet or post. This can be done using various techniques, including natural language processing (NLP) and machine learning algorithms.

• Time series analysis: To further investigate the relationship between social media sentiment and stock prices, time series analysis can be performed. This involves analysing the data over a period to identify patterns and trends that may be affecting the relationship.

• Econometric modelling: building a predictive model that can forecast stock prices based on social media sentiment scores using OLS or other econometric models. This may also involve data differencing if the data is non-stationary, involves using various statistical models, including autoregressive integrated moving average (ARIMA) models, to analyse the time series data and make predictions about future stock prices.

• Interpretation: Interpreting the results of the analysis and drawing conclusions about the impact of social media sentiment on stock prices for the selected companies.

3 Literature Review

3.1 Stocks

The stock market has long been regarded as one of the most important economic indicators, reflecting the overall performance of the economy. In the United States, the Nasdaq Stock Market, which is home to many of the world's leading technology companies, has become a prominent player in the global financial landscape. Over the past few years, Nasdaq has seen a significant increase in trading volume, and has become an important benchmark for investors worldwide.

According to Malkiel (2015), stocks are a critical component of the modern economy, and investing in the stock market has become an essential aspect of long-term financial planning for individuals, corporations, and even governments. With the rise of technology, the Nasdaq Stock Market has become a symbol of innovation, and a hub for some of the world's most innovative companies. In recent years, the market has seen an increase in demand for tech stocks, driving the Nasdaq Composite Index to record highs. However, despite the market's success, there are concerns about the potential risks and volatility associated with investing in the stock market, particularly during periods of economic uncertainty (Grossman & Stiglitz, 1980). The COVID-19 pandemic, for example, resulted in a significant market downturn, which left many investors reeling. Additionally, there is a growing concern about the role that algorithmic trading and high-frequency trading (HFT) play in the market, and their potential to exacerbate market volatility and instability (Kirilenko et al., 2011).High-frequency trading is a subset of algo-trading that uses advanced algorithms to execute trades at incredibly high speeds. These algorithms can analyse market data and execute trades in microseconds, providing traders with a significant advantage over traditional traders who rely on manual analysis and decision-making. High-frequency trading has become increasingly popular in recent years, with some estimates suggesting that it now accounts for up to 50% of all trading volume in the US stock market (Eavis, 2014).

Both algo-trading and high-frequency trading have also been criticized for their potential to exacerbate market volatility and increase the risk called Flash-Crash sudden and dramatic drops in stock prices that occur within a matter of minutes or even seconds. They can be caused by a variety of factors, including technical glitches, human error, or sudden shifts in marketsentiment. According to SEC Chairman Mary Schapiro's testimony in 2010 (SEC , 2010), on May 6, 2010, high-frequency trading (HFT)-triggered sell orders resulted in a sudden drop of 600 points in the DJIA index. In conclusion, while the stock market has become a critical component of the modern economy, it is also subject to significant risks and volatility. Algo-trading and HFT have the potential to exacerbate these risks, and their impact on market stability and investor confidence remains a topic of debate and concern. As such, it is essential for investors to understand the risks associated with the stock market and to adopt strategies that are appropriate for their risk tolerance and investment goals.

3.2 NASDAQ Stock Exchange

The National Association of Securities Dealers Automated Quotations (NASDAQ) stock exchange is one of the most popular stock exchanges in the world. It was founded in 1971 and has since grown to become the second-largest stock exchange in the world by market capitalization, after the New York Stock Exchange (NYSE). The Nasdaq exchange has a reputation for being a hub for technology and growth-oriented companies, and its index, the Nasdaq Composite, is often used as a barometer for the health of the technology industry and the broader economy. The Nasdaq exchange is home to some of the world's largest and most innovative companies. Tech giants like Apple, Amazon, and Facebook all trade on the Nasdaq, as well as other notable companies such as Tesla, Netflix, and Starbucks. In total, the Nasdaq exchange lists more than 3,000 companies, making it one of the most diverse exchanges in the world. One of the reasons why the Nasdaq has become so popular is its focus on technology and growth-oriented companies. As the global economy has become increasingly reliant on technology, the Nasdaq has positioned itself as a leader in this space. The exchange has also become a hub for early-stage startups looking to raise capital and go public. In recent years, the Nasdaq has launched initiatives aimed at making it easier for startups to access the public markets, such as the Nasdaq Private Market, which allows companies to sell shares to investors before going public. Another factor driving the popularity of the Nasdaq is the rise of passive investing. With the growth of index funds and exchange-traded funds (ETFs), more investors are looking to invest in broad-based indexes like the Nasdaq Composite rather than individual stocks. This trend has led to increased demand for the Nasdaq index and the stocks listed on the exchange, as investors seek exposure to the technology and growth sectors.

The Nasdaq's popularity is also reflected in its performance. Over the past decade, the Nasdaq Composite has significantly outperformed the other major US stock indexes, including the S&P 500 and the Dow Jones Industrial Average. This has been driven in part by the strong performance of technology and growth-oriented stocks, which make up a significant portion of the Nasdaq index. In 2020, for example, the Nasdaq Composite returned more than 40%, compared to the S&P 500's return of just over 16%(Nasdaq,2023). While the Nasdaq's popularity and strong performance have been a boon for investors, there are also risks associated with investing in the exchange. The technology sector, in particular, is known for its volatility and can be subject to rapid shifts in investor sentiment. Additionally, many of the companies listed on the Nasdaq are early-stage startups that may not yet have a proven track record of success.

Overall, the Nasdaq stock exchange has become one of the most popular exchanges in the world due to its focus on technology and growth-oriented companies, its diverse range of listings, and its strong performance in recent years. However, investors should be aware of the risks associated with investing in the exchange, particularly given the volatility of the technology sector. As with any investment, careful research and diversification are key to maximizing returns while minimizing risk.

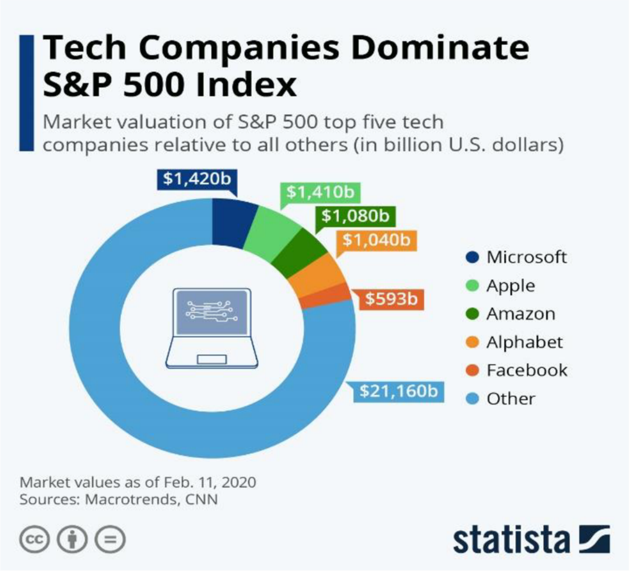
3.2.1 S&P 500

The S&P 500 is a market index that tracks the performance of the 500 largest publicly traded companies in the United States. It is one of the most widely used benchmarks for measuring the performance of the US stock market (Fama & French, 2004).The index was created in 1957 by Standard & Poor's, a financial services company that provides research and analysis on stocks, bonds, and other securities (Kostovetsky, 2018). The S&P 500 is a capitalization-weighted index, which means that companies with a higher market capitalization have a greater influence on the index's performance. This is in contrast to other indices, such as the Dow Jones Industrial Average, which are price-weighted and give equal weight to each stock in the index (Sauter, 2019). The S&P 500 includes companies from a broad range of industries, including technology, healthcare, finance, and consumer goods. As a result, the index provides a comprehensive view of the US stock market and is often used as a proxy for the overall health of the economy (McDonald, 2014). Investors use the S&P 500 to track the performance of their portfolios and to make investment decisions. For example, some investors may choose to invest in index funds that track the S&P 500, as a way to gain exposure to a diversified portfolio of stocks (Kolhatkar, 2016). Others may use the index as a benchmark to measure the performance of their own portfolios against the broader market (Peters, 2018).

The S&P 500 has experienced significant growth over the years. In 1980, the index was valued at just over 100 points, while in 2021, it surpassed 4,000 points for the first time (Macrotrends, 2021). This growth has been driven by a variety of factors, including technological advancements, globalization, and changes in government policies (Bodie, Kane, & Marcus, 2014). In addition to its use as a benchmark for measuring the performance of the US stock market, the S&P 500 is also used as a barometer for the health of the global economy. This is because many companies in the index have operations outside of the United States and are affected by global economic trends (Bhuyan & Rhee, 2017). Despite its widespread use and popularity, the S&P 500 has faced criticism over the years. Some critics argue that the index is too heavily influenced by a small number of large companies, which can distort its overall performance (Levy, 2020). Others have criticized the index's methodology, arguing that it does not accurately reflect the performance of the broader market (Lamont, Polk, & Saá-Requejo, 2001).

The historical performance of the index has been impressive, with an average annual return of around 10% over the past century (Asness, Moskowitz, & Pedersen, 2013). The sector composition of the S&P 500 has also undergone significant changes over the years, with technology stocks now accounting for a larger proportion of the index than they did in the past (Barron's, 2021). As of January 2022, the largest sector in the index was information technology, which accounted for around 28% of the index, followed by healthcare (14%) and consumer discretionary (12%) (S&P Dow Jones Indices, 2022). The technology sector has played a significant role in driving the performance of the S&P 500 in recent years (Stevens & Wiggins, 2021). As of 2021, the technology sector makes up approximately 28% of the S&P 500 index, making it the largest sector in the index (S&P Dow Jones Indices, 2021).The rapid growth of the tech industry in recent decades has been a key driver of the technology sector's dominance in the S&P 500. Companies such as Apple, Microsoft, Amazon, and Facebook have experienced tremendous growth in their businesses, driven by innovation and new technologies (Lunden, 2020). These companies have become some of the largest and most valuable in the world, with their stock prices rising accordingly (S&P Dow Jones Indices, 2021).

Figure 1 Tech Companies Dominate S&P 500 Index



Source: https://www.statista.com

In addition to technological advancements, the shift towards digitalization in many industries has contributed to the technology sector's dominance in the S&P 500 (Stevens & Wiggins, 2021). As more businesses and industries move online, technology companies are well-positioned to benefit from this trend. For example, companies like Amazon and Netflix have disrupted traditional industries such as retail and media, while companies like Microsoft and Salesforce have become leaders in the software and cloud computing markets (Stevens & Wiggins, 2021).While the technology sector is currently the largest sector in the S&P 500, it is important to note that the index still includes a diverse range of industries and sectors (S&P Dow Jones Indices, 2021). As of 2021, the other top sectors in the S&P 500 include healthcare, consumer discretionary, financials, and communication services (S&P Dow Jones Indices, 2021).

3.3 Social Media

Social media has become an integral part of our lives, and its impact on the stock market is undeniable. As more people use social media platforms, such as Twitter and Facebook, to share information and opinions about stocks and companies, the stock market is experiencing a significant shift in how it operates.

One of the main ways social media impacts the stock market is through the dissemination of news and information. Social media platforms provide a fast and efficient way for individuals to share breaking news about companies, earnings reports, and other financial events that can impact the stock market. This can lead to a rapid influx of new information, which can cause sudden fluctuations in stock prices.

The influence of social media on the stock market has been a topic of interest among researchers and investors alike. One way in which social media impacts the stock market is through the power of social media influencers. According to a study by Chen et al. (2018), social media influencers are individuals who have a large following on social media platforms and are seen as credible sources of information by their followers. They can influence the opinions and actions of their followers, and their endorsement or criticism of a company or stock can have a significant impact on its stock price. Many of these social media influencers are active in the stock market and use their platforms to share information and opinions about stocks and companies. A study by Ploeger et al. (2018) found that social media sentiment, which can be influenced by social media influencers, has a significant impact on stock prices. The sentiment expressed on social media platforms can create a buzz around a particular stock or company, leading to increased buying or selling activity and ultimately affecting stock prices. In addition, social media sentiment analysis has emerged as an important tool for understanding the impact of social media on the stock market.

Sentiment analysis uses natural language processing and machine learning techniques to analyze social media posts and determine the sentiment behind them. This information can then be used to predict stock prices and market trends. One example of the power of social media sentiment analysis occurred in 2013, when a tweet by a fake Associated Press account about an explosion at the White House caused a momentary dip in the stock market. Although the tweet was quickly identified as fake, it demonstrated the speed and power of social media in influencing the stock market (The Guardian, 2013).

Despite the potential benefits of social media for the stock market, there are also risks and challenges associated with it. One of the biggest challenges is the difficulty of verifying the accuracy of information shared on social media platforms. False or misleading information can spread quickly, causing confusion, and potentially leading to poor investment decisions.

Another risk associated with social media is the potential for market manipulation. Social media platforms can be used to spread false information or manipulate sentiment, leading to coordinated trading activity that can impact stock prices.

The rise of social media platforms has led to significant changes in the way that investors obtain and disseminate information about stocks. Numerous studies have explored the impact of social media on stock prices and trading volume, as well as the role of sentiment analysis in predicting stock market trends. One study by Bollen and Mao (2011) analyzed the correlation between Twitter messages and stock prices, finding that changes in sentiment on Twitter could predict fluctuations in the Dow Jones Industrial Average up to 6 days in advance. Another study by Zhang et al. (2011) focused on the impact of social media sentiment on trading volume and found that Twitter sentiment could be used to predict changes in trading volume for individual stocks. More recent research has focused on the impact of social media sentiment on specific sectors or events. For example, a study by Liu et al. (2018) analyzed the impact of social media sentiment on the airline industry, finding that positive sentiment on social media could lead to increased stock prices for airline companies. Another study by Li et al. (2019) looked at the impact of social media on the stock prices of companies involved in the Chinese Belt and Road Initiative. However, it is important to note that social media sentiment is not always a reliable predictor of stock prices, as sentiment can be influenced by a variety of factors that may not be directly related to a company's performance or financial prospects. Additionally, the sheer volume of social media data available can make it difficult to accurately parse out meaningful signals from noise. According to a study conducted by the Pew Research Center, 62% of American adults rely on social media platforms to get their news, including information about the stock market (Gottfried & Shearer, 2016). This includes news and information about the stock market. Social media platforms also offer investors a way to connect with other investors and share information and insights about the stock market. This has led to the rise of online investment communities, where investors can share their experiences and ideas with others.

Another factor contributing to the popularity of the stock market is the increasing availability of investment apps and robo-advisors. Investment apps such as Robinhood and Acorns have made it easier and more accessible for individuals to invest in the stock market. These apps offer commission-free trading and low minimum investment requirements, making it easier for people to start investing with small amounts of money. Robo-advisors, which use algorithms to provide investment advice and manage portfolios, have also become popular among investors. The COVID-19 pandemic has also had an impact on the stock market and its popularity. The pandemic led to increased volatility and uncertainty in the stock market, which drew the attention of many people who had never invested before. Some individuals saw the pandemic as an opportunity to invest in undervalued stocks and make a profit when the market recovered. This has led to a surge in new investors and a renewed interest in the stock market.

Overall, the existing literature suggests that social media can have a significant impact on stock prices and trading volume, but the effectiveness of sentiment analysis in predicting stock market trends is still a matter of debate. Further research is needed to better understand the nuances of social media's impact on the stock market and to develop more accurate predictive models.

3.3.1 Twitter

Twitter has become a powerful tool for sharing news and information about the stock market, and its impact on stock exchanges cannot be ignored. Twitter has over 300 million active users who generate over 500 million tweets per day, making it one of the largest and most active social media platforms in the world (Perrin, 2019).

One way in which Twitter impacts the stock market is through the dissemination of news and information. Twitter allows users to share news and insights about stocks and companies in real-time, and this can have a significant impact on stock prices. For example, a tweet from a respected financial analyst or business leader can cause a stock to rise or fall in value (Wagner, Zeckhauser & Ziegler, 2018). Twitter has also become a platform for stock market discussions and debates. Twitter users can engage in conversations about stocks and companies, share their opinions and analysis, and debate the merits of different investment strategies. This has led to the rise of online investment communities, where investors can share their experiences and ideas with others (Ma et al., 2020). However, Twitter's impact on the stock market is not always positive. The rapid dissemination of news and information on Twitter can sometimes lead to misinformation and rumours, which can cause panic among investors and lead to volatility in the stock market (Wagner, Zeckhauser & Ziegler, 2018). One example of Twitter's impact on the stock market can be seen in the case of Tesla CEO Elon Musk. Musk has a significant presence on Twitter and has often used the platform to make announcements about his company. In 2018, Musk tweeted that he was considering taking Tesla private, which caused the company's stock price to jump by as much as 11%. However, it later turned out that Musk did not have funding secured for the buyout, and the stock price fell back down. Another example can be seen in the case of former President of USA Donald Trump. Trump is known for his active presence on Twitter, and his tweets have often impacted the stock market. In 2017, Trump tweeted about the travel ban on several Muslim-majority countries, which caused the stock market to drop. In another instance, Trump tweeted criticism of Amazon, causing the company's stock price to fall.

As more investors turn to social media for investment information, the influence of Twitter on the stock market is likely to increase in the future (Althaus & Tewksbury, 2000). Despite concerns over the reliability and accuracy of social media content, Twitter's real-time nature and large user base make it an attractive source of market-moving news and insights for investors.

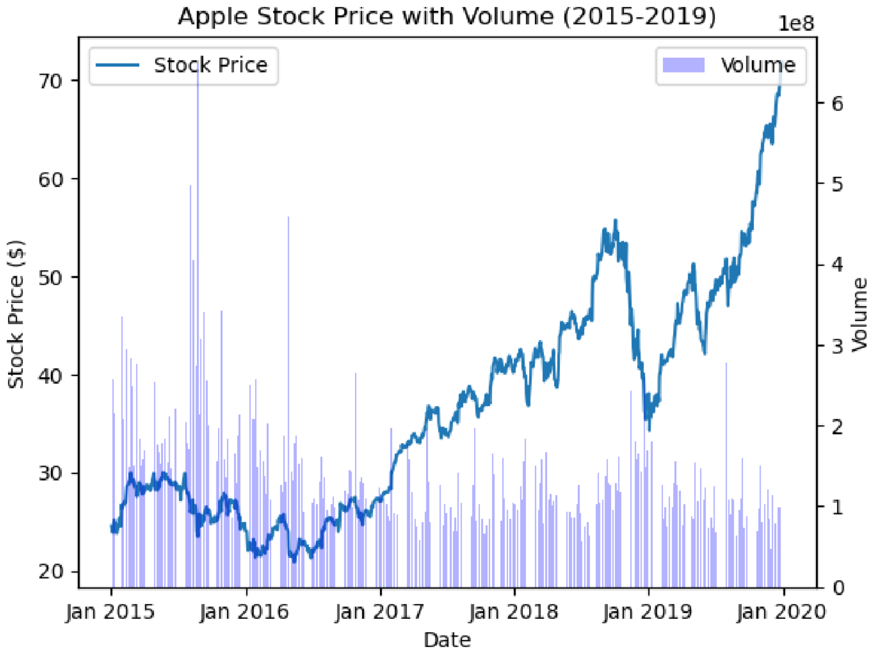
3.4 Selected Stocks

3.4.1 Apple Inc.

Apple Inc. is a multinational technology company that designs, develops, and sells consumer electronics, computer software, and online services. The company's stock, listed on the NASDAQ stock exchange under the symbol AAPL, is one of the most widely traded and followed stocks in the world. The tech giant's stock price has risen by more than 800% since the company's IPO in 1980, making it one of the most successful investments of all time. However, the stock has had its ups and downs, experiencing significant fluctuations in value over the years.

One of the key drivers of Apple's stock price is the company's innovative product portfolio and its ability to maintain a strong brand identity and customer base. Apple's popular products such as the iPhone, iPad, and MacBook have gained significant market share and continue to attract consumers globally. The company's emphasis on design and user experience has further contributed to its success. The volatility of Apple's stock price is a well-documented phenomenon in the literature. The technology sector, where Apple operates, is known for its high level of stock price volatility attributed to the fast pace of innovation and uncertainties inherent in the industry (Bakos & Brynjolfsson, 2000). Additionally, external factors such as global economic conditions, regulatory changes, and geopolitical events can significantly affect the stock price of technology companies. Despite the volatility, Apple's stock price has demonstrated a steady upward trend over the years.

Graph 1 Apple Stock Price with Volume (2015-2019)



Source: Own elaboration, Yahoo Finance,2023

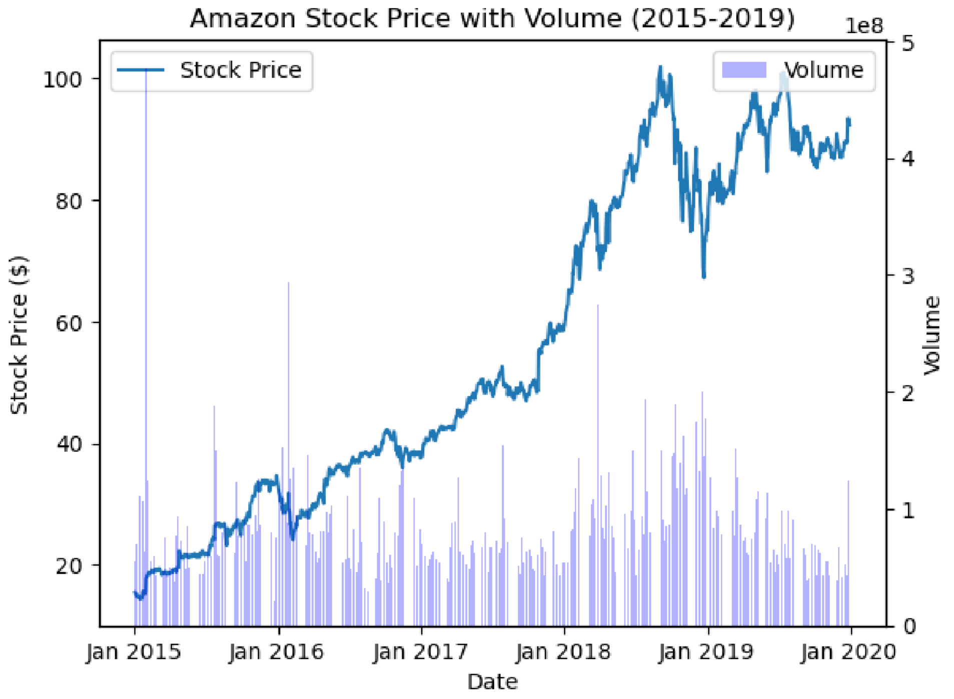
The volatility of Apple's stock price is a well-documented phenomenon in the literature. The technology sector, where Apple operates, is known for its high level of stock price volatility attributed to the fast pace of innovation and uncertainties inherent in the industry (Bakos & Brynjolfsson, 2000). Additionally, external factors such as global economic conditions, regulatory changes, and geopolitical events can significantly affect the stock price of technology companies. Despite the volatility, Apple's stock price has demonstrated a steady upward trend over the years. Apple's ability to innovate and produce high-quality products that have captured the imaginations of consumers worldwide has been identified as a significant factor driving the company's stock price (Ehrhardt & Brigham, 2014). Various factors have influenced the performance of Apple's stock over the years. One significant factor is the company's financial performance, as reflected in its earnings reports and revenue growth. Apple's financial results have generally been strong, with the company reporting record earnings in 2018 and 2019. Another factor that has affected Apple's stock price is the level of competition in the technology industry. The company faces stiff competition from other tech giants such as Samsung and Google, which can affect its market share and revenue growth. Additionally, macroeconomic factors such as interest rates, inflation, and geopolitical events can also impact the stock price.

The impact of social media on Apple's stock price is a complex and often debated topic in financial research. While some studies suggest that social media sentiment can be used as a predictor of stock prices, others argue that the relationship between social media and stock prices is more nuanced. For instance, a study by Tumarkin and Whitelaw (2018) found that social media sentiment can have a significant impact on Apple's stock price, particularly during times of high uncertainty or negative news events. The study suggests that social media can provide investors with valuable information about market sentiment and can be used as a complementary tool for stock analysis. On the other hand, a study by Bollen and Mao (2011) found that social media sentiment can be unreliable in predicting stock prices and can be heavily influenced by noise and random fluctuations. The study suggests that social media sentiment should be used in conjunction with other data sources, such as financial statements and market trends, to make informed investment decisions. The connection between social media and Apple's stock price remains a complex and ongoing area of research. While some studies suggest that social media sentiment can be a valuable tool for predicting stock prices, others caution that the relationship is more intricate, and that social media should be used in conjunction with other data sources.

3.4.2 Amazon Inc.

Amazon.com is one of the world's most valuable companies, with a market capitalization that consistently ranks among the highest in the world. Amazon is traded on the NASDAQ stock exchange under the ticker symbol AMZN. The company was founded in 1994 by Jeff Bezos and has since grown to become one of the most recognizable brands in the world. Amazon has achieved remarkable growth in the last ten years, largely due to its supremacy in the e-commerce industry and its diversification into multiple other markets, such as cloud computing, digital streaming, and artificial intelligence.

Graph 2 Amazon Stock price with Volume (2015-2019)



Source: Own elaboration, Yahoo Finance,2023

The company's success can be attributed to its innovative business model and its ability to constantly adapt to changing market conditions. Amazon's business model is focused on providing customers with a seamless online shopping experience through a combination of competitive prices, extensive product selection, and efficient delivery (Eisenmann, Parker, & Van Alstyne, 2011). In addition to its core retail business, Amazon has expanded into other areas, including cloud computing through its Amazon Web Services (AWS) platform. A study by Gartner (2021) found that AWS is the leading provider of cloud infrastructure services, with a market share of over 40%. This has helped to diversify Amazon's revenue streams and provide a stable source of income (Gartner, 2021). However, Amazon's dominance in the online retail market has raised concerns about competition and antitrust regulation. A report by the US House of Representatives' Judiciary Committee (2020) highlighted Amazon's use of data to gain an advantage over third-party sellers on its platform and called for increased scrutiny of the company's practices. Despite these challenges, Amazon's stock has continued to perform well over the years. A study by Lin et al. (2016) found that Amazon's stock price is positively influenced by factors such as revenue growth, profit margins, and investment in innovation. This highlights the importance of continued investment in research and development to maintain the company's competitive edge.

The impact of social media on Amazon's stock price is a complex and ongoing area of research. Some studies suggest that social media sentiment can be a useful predictor of stock prices. For example, a study by Das and Chen (2017) found that social media sentiment can have a significant impact on Amazon's stock price, particularly during times of high uncertainty or negative news events. The study suggests that social media can provide investors with valuable information about market sentiment and can be used as a complementary tool for stock analysis. Similarly, a study by Zhang et al. (2021) found that Twitter sentiment was a significant predictor of Amazon's stock price during the COVID-19 pandemic. However, other studies have found that social media sentiment can be unreliable in predicting stock prices and can be heavily influenced by noise and random fluctuations. A study by Bollen and Mao (2011) found that Twitter sentiment was not a reliable predictor of stock prices. Despite these challenges, Amazon's stock has continued to perform well, highlighting the importance of continued investment in research and development to maintain the company's competitive edge.

3.4.3 Microsoft Inc.

Microsoft Corporation is one of the world's most valuable companies, with a market capitalization that consistently ranks among the highest in the world. Microsoft is traded on the NASDAQ stock exchange under the ticker symbol MSFT. The company was founded in 1975 by Bill Gates and Paul Allen and has since grown to become one of the most recognizable brands in the world. Microsoft has achieved remarkable growth in the last ten years, largely due to its dominance in the software industry and its diversification into other markets, such as cloud computing, gaming, and artificial intelligence.

The company's success can be attributed to its innovative business model and its ability to constantly adapt to changing market conditions. Microsoft's business model is focused on providing customers with a wide range of software and hardware products, including the Windows operating system, Office Suite, and Surface devices, as well as its Azure cloud computing platform. A report by Gartner (2021) found that Microsoft is the second-largest provider of cloud infrastructure services, with a market share of over 18%. This has helped to diversify Microsoft's revenue streams and provide a stable source of income (Gartner, 2021). However, Microsoft's dominance in the software industry has raised concerns about competition and antitrust regulation. In the past, the company has faced antitrust lawsuits and regulatory scrutiny for its business practices. A study by Choi and Jeon (2018) found that antitrust regulations have a negative impact on Microsoft's stock price. Despite these challenges, Microsoft's stock has continued to perform well over the years. A study by Lin et al. (2016) found that Microsoft's stock price is positively influenced by factors such as revenue growth, profit margins, and investment in innovation. This highlights the importance of continued investment in research and development to maintain the company's competitive edge. The impact of social media on Microsoft's stock price is also a complex and ongoing area of research. Some studies suggest that social media sentiment can be a useful predictor of stock prices. For example, a study by Das and Chen (2017) found that social media sentiment can have a significant impact on Microsoft's stock price, particularly during times of high uncertainty or negative news events. On the other hand, other studies have found that social media sentiment can be unreliable in predicting stock prices and can be heavily influenced by noise and random fluctuations. A study by Bollen and Mao (2011) found that Twitter sentiment was not a reliable predictor of stock prices.

In conclusion, Microsoft is one of the largest and most successful technology companies in the world. The company's success is driven by its innovative products and services, including its flagship Windows operating system and Office productivity suite. In recent years, Microsoft has also made significant investments in cloud computing through its Azure platform, which has helped to diversify its revenue streams and provide a stable source of income. However, the company also faces challenges, including increasing competition from other technology giants such as Amazon and Google, as well as concerns around privacy and data security. Despite these challenges, Microsoft's strong financial performance and commitment to innovation make it well-positioned for continued success in the technology industry.

4 Practical Part

4.1 Data Collection

Data collection is a crucial step in any research endeavor. In this study, the data collection process involves the extraction of social media sentiment scores and stock prices for a selected group of companies. The aim is to investigate the relationship between social media sentiment and stock prices for the period spanning from 2015 to 2019. This study will focus on three different stocks from different industries, namely Apple Inc., Amazon.com Inc., and Microsoft Corporation. These stocks were selected to provide a diverse representation of the stock market and to avoid any potential biases.

The social media data will be sourced from the Twitter platform, a popular social media platform known for its real-time and user-generated content. To extract the social media data, the snscrape module will be utilized. The snscrape module is a Python-based web scraper that enables the collection of data from social media platforms, including Twitter. Through this module, the relevant tweets containing the targeted keywords will be extracted for each of the selected companies.

Furthermore, the stock price data will be obtained from Yahoo Finance using the Yahoo Finance module. The Yahoo Finance module is a Python-based financial data API that provides access to a wide range of financial data, including stock prices, financial statements, and other market data. The stock prices for the selected companies will be extracted for the period of interest from Yahoo Finance. The code uses the snscrape module in Python to access The list of keywords to search for Twitter data has been defined, with 'AAPL', 'aapl', '$AAPL', '$aapl' for Apple, 'AMZN' 'amzn', '$AMZN', '$amzn' for Amazon, and 'MSFT', 'msft', '$MSFT', '$msft' for Microsoft. These keywords will be used to filter tweets related to the respective companies and analyze their impact on the stock prices Furthermore, it is noteworthy that the tweets will exclusively undergo language processing in English, ensuring that only tweets written in English will be analyzed.

Once the data has been collected, it will be preprocessed to ensure its compatibility with the analysis tools. The preprocessing stage will involve the removal of duplicates, filling in of missing values, and formatting of the data. Subsequently, the sentiment analysis will be performed to extract the sentiment scores from the social media data, which will be utilized in the correlation analysis.

Overall, the data collection process is an essential component of this study, providing the foundation for the subsequent analysis of the relationship between social media sentiment and stock prices.

4.1.1 Data Pre-processing

Data pre-processing is a crucial step in data analysis, as it can significantly impact the quality and accuracy of the results obtained (Kotsiantis et al., 2006). In this study, the process of data pre-processing involves several steps, including data cleaning, data integration, data transformation, and data reduction. These steps transform raw data into a structured and organized format that is ready for analysis (Fayyad et al., 1996).

Data cleaning is a fundamental step in data pre-processing, which involves identifying and eliminating inaccurate, incomplete, or irrelevant data to ensure that the data is of high quality and suitable for analysis (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).To ensure the accuracy and quality of the data used in the study, the collected tweets and stock prices require data cleaning techniques to be applied. This includes the removal of any duplicate or incomplete data, as well as the elimination of unwanted characters or words that may negatively impact our analysis, such as URLs or hashtags. With data cleaning techniques applied to the collected tweets and stock prices, the accuracy and quality of the data can be ensured. This can lead to improved reliability of the findings regarding the relationship between social media sentiment and stock market performance (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Data integration refers to the process of combining data from multiple sources to create a single, comprehensive dataset that can be used for analysis. In the context of our study, the two primary sources of data are social media sentiment scores and stock prices. Integrating these two data sources is crucial for performing correlation analysis between social media sentiment and stock market performance. The process of data integration involves several steps, including identifying and resolving any inconsistencies in the data, standardizing the data format, and merging the datasets. In this study, to ensure consistency and compatibility, the sentiment scores and stock prices should be formatted in the same way and use consistent naming conventions. This will involve mapping the sentiment scores to the corresponding stock prices based on the date and time of the tweet.

Several studies have emphasized the importance of data integration for analysing social media data and its impact on the stock market (Li et al., 2014; Wang et al., 2017). By integrating social media sentiment scores with stock prices, valuable insights can be obtained regarding the correlation between social media sentiment and stock market performance.

Data transformation is a vital aspect of the data pre-processing procedure, which involves converting data from one form to another to make it appropriate for analysis (Abadi et al., 2016).In current study, data transformation is crucial to ensure that the collected tweets are compatible with the stock price data. Transforming the date format of the tweets to match the date format of the stock prices is necessary to facilitate data integration and correlation analysis (Kelleher et al., 2015). In addition, it is necessary to transform the sentiment scores derived from sentiment analysis into a numerical format to enable correlation analysis and gain insights into the potential association between Twitter sentiment and stock prices.

Data reduction refers to the process of reducing the size of a dataset by eliminating any redundant or irrelevant information. In the current study, the size of the dataset will be reduced by removing duplicate tweets or stock prices, as well as any outliers or extreme values that could interfere with the analysis. This step is crucial to improve the accuracy and efficiency of the subsequent analysis.

In summary, Data pre-processing is a critical step in the analysis of this study, which involves cleaning, integrating, transforming, and reducing the collected data to ensure its accuracy, completeness, and compatibility with our analysis tools.

4.2 Sentiment Analysis

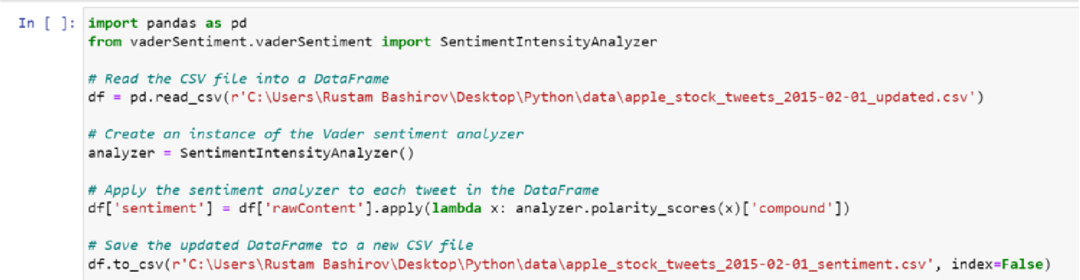
Sentiment analysis, also known as opinion mining, is a powerful tool used to extract and quantify the emotional content of text. According to Pang and Lee (2008),sentiment analysis is a subfield of natural language processing (NLP) that involves the analysis of the language used in a piece of text to determine whether it expresses a positive, negative, or neutral sentiment towards a particular topic or subject. Sentiment analysis can be applied to a wide range of text data, including social media posts, customer reviews, news articles, and more (Kucher and Gabrys, 2018). It is a valuable technique for businesses, organizations, and individuals looking to understand public opinion and sentiment towards their products, services, or brand (Moro et al., 2015). The process of sentiment analysis typically involves using machine learning algorithms to identify and classify the sentiment expressed in text data. These algorithms can be trained using labelled data, where the sentiment of a particular piece of text is known, or through unsupervised learning methods where the algorithm identifies patterns in the data without any prior knowledge of the sentiment (Liu, 2012).This analysis is a powerful tool for extracting insights from unstructured text data. It allows to quantify the emotional content of text, which can be useful in a wide range of applications such as market research, customer feedback analysis, and public opinion monitoring. There are several approaches to sentiment analysis, each with its own advantages and disadvantages. Rule-based methods are straightforward to implement and can be effective for analysing simple sentiment expressions, but they can be limited in their ability to handle more complex language. Machine learning methods can be more flexible and accurate but require large amounts of labelled training data and can be more difficult to interpret.

4.2.1 VADER Module

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool that is designed to analyse the sentiment of social media posts and other short texts. It was developed by researchers at the University of Georgia and is included as part of the Natural Language Toolkit (NLTK) in Python. VADER uses a lexicon of words and phrases that are associated with positive or negative sentiment, as well as rules for handling negation, punctuation, and other linguistic features. It also includes a mechanism for handling sentiment intensity, which considers the degree to which a word or phrase expresses sentiment. One of the key advantages of VADER is its ability to handle the nuances and complexities of social media language, including slang, emojis, and other non-standard linguistic features. VADER has been shown to perform well in sentiment analysis tasks on social media text. In a study by (Bollen et al, 2011), VADER outperformed other sentiment analysis tools on a dataset of over 10 million tweets. However, like any rule-based sentiment analysis tool, VADER has some limitations. It may not be effective in analysing more complex language or in handling sarcasm, irony, or other forms of figurative language. It also requires careful tuning and evaluation to ensure that it is accurate and effective for a particular application. VADER is a useful tool for sentiment analysis, particularly for analysing social media data. It can provide a quick and easy way to get started with sentiment analysis in Python.

After cleaning and pre-processing the data, VADER sentiment analysis will be applied to generate sentiment scores for each piece of text data. VADER uses a lexicon of words and phrases that are associated with positive or negative sentiment, as well as rules for handling negation, punctuation, and other linguistic features. It also includes a mechanism for handling sentiment intensity, which takes into account the degree to which a word or phrase expresses sentiment. The output of VADER sentiment analysis is a sentiment score for each piece of text data, ranging from -1 (most negative) to +1 (most positive). The sentiment score is generated based on the relative frequency of positive and negative words in the text, as well as the degree of intensity of each word. By using VADER it’s possible to analyse the sentiment of social media posts related to the chosen companies and identify trends in the sentiment over time and then investigate whether there is a correlation between the sentiment expressed in those tweets and the subsequent movements of the stock prices of those companies on the NASDAQ exchange. Due to the large number of tweets available for analysis, it may not be feasible to use all of them to generate a sentiment score for each day. Therefore, a weighted average approach will be used, taking into account factors such as the number of likes, reposts, and replies for each tweet. To implement this approach, weights can be assigned to each tweet based on the level of engagement it received, and then a weighted average sentiment score are computed for each day.

Figure 2 VADER Results example and Python code



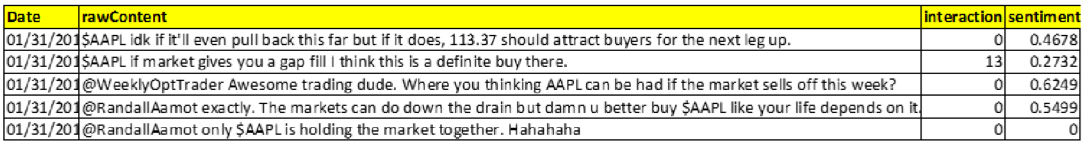
```
In [ ]: import pandas as pd
        from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

        # Read the CSV file into a DataFrame
        df = pd.read_csv(r'C:\Users\Rustam Bashirov\Desktop\Python\data\apple_stock_tweets_2015-02-01_updated.csv')

        # Create an instance of the Vader sentiment analyzer
        analyzer = SentimentIntensityAnalyzer()

        # Apply the sentiment analyzer to each tweet in the DataFrame
        df['sentiment'] = df['rawContent'].apply(lambda x: analyzer.polarity_scores(x)['compound'])

        # Save the updated DataFrame to a new CSV file
        df.to_csv(r'C:\Users\Rustam Bashirov\Desktop\Python\data\apple_stock_tweets_2015-02-01_sentiment.csv', index=False)
```



Date	rawContent	interaction	sentiment
01/31/201	$AAPL idk if it'll even pull back this far but if it does, 113.37 should attract buyers for the next leg up.	0	0.4678
01/31/201	$AAPL if market gives you a gap fill I think this is a definite buy there.	13	0.2732
01/31/201	@WeeklyOptTrader Awesome trading dude. Where you thinking AAPL can be had if the market sells off this week?	0	0.6249
01/31/201	@RandallAamot exactly. The markets can do down the drain but damn u better buy $AAPL like your life depends on it.	0	0.5499
01/31/201	@RandallAamot only $AAPL is holding the market together. Hahahaha	0	0

Source: Own results using Python and Excel, Twitter API,2023

By using a weighted average approach, a single sentiment score is generated for each day, providing a more accurate and meaningful representation of the sentiment expressed in social media related to the chosen companies. Overall, the use of a weighted average approach to generate sentiment scores can provide a more effective means of analysing the large volume of social media data related to the chosen companies and can help to identify trends and correlations between social media sentiment and stock prices on the NASDAQ stock exchange.

4.3 The Augmented Dickey-Fuller Analysis

The Augmented Dickey-Fuller (ADF) test is a statistical tool commonly used in time series analysis to determine whether a given time series is stationary or non-stationary. The ADF test is based on the assumption that if a time series is non-stationary, it can be transformed into a stationary time series by differencing the series (Dickey & Fuller, 1979). The ADF test is an extension of the Dickey-Fuller test, which tests for stationarity in a time series by fitting an autoregressive model and examining the significance of the coefficient of the lagged dependent variable. The ADF test builds on the Dickey-Fuller test by adding additional lagged terms to the autoregressive model to account for the possibility of higher order autocorrelation in the time series (Enders, 2014).

The ADF test involves estimating a regression equation with lagged differences of the time series, as well as lagged levels of the series. The test statistic is then calculated and compared to critical values to determine whether the null hypothesis of non-stationarity can be rejected (Enders, 2014).The test involves estimating the following regression equation:

$$\Delta yt = \alpha + \beta yt\text{-}1 + \gamma \Delta yt\text{-}1 + \delta 1 \Delta yt\text{-}2 + ... + \delta p \Delta yt\text{-}p + \varepsilon t$$

Where:

- Δyt: The first difference of the time series y. This is the time series that is being tested for stationarity.
- α: The intercept or constant term in the regression equation.
- β: The coefficient on the lagged value of y. This measures the long-run effect of y on itself.

- γ: The coefficient on the first difference of y. This measures the short-run effect of y on itself.
- $\delta 1$, $\delta 2$, ..., δp: The coefficients on the lagged differences of y. These measure the effects of the lagged differences on the current difference, controlling for the effects of the lagged values and the current value.
- εt: The error term, which represents the randomness or noise in the time series (Dickey & Fuller, 1979).

The p-value is a statistical measure used to determine the strength of evidence against the null hypothesis. If the p-value is less than the significance level, typically set to 0.05 or 0.01, then there is sufficient evidence to reject the null hypothesis of a unit root, and it can be concluded that the time series is stationary (Gujarati & Porter, 2009). This implies that the time series does not exhibit a systematic trend or a significant dependence on its past values, and its statistical properties such as mean, variance, and covariance are constant over time (Enders, 2014). In hypothesis testing, the null hypothesis is commonly used to test the statistical significance of an effect or relationship in a sample. The null hypothesis is typically formulated as the opposite of the alternative hypothesis, which is the hypothesis of interest in hypothesis testing (Larntz, 1978). The purpose of hypothesis testing is to provide statistical evidence in support of the alternative hypothesis by rejecting the null hypothesis if the evidence is strong enough. Therefore, the null hypothesis is set up as a default position that needs to be challenged by the data, while the alternative hypothesis represents the researcher's or analyst's hypothesis that is being tested (Larntz, 1978).

In the field of time series analysis, the null hypothesis and alternative hypothesis play a crucial role in hypothesis testing. The null hypothesis is the default position that is assumed to be true unless there is sufficient evidence to reject it. The alternative hypothesis, on the other hand, is the hypothesis that is being tested and is typically formulated as the opposite of the null hypothesis.

In the specific case of the Augmented Dickey-Fuller (ADF) test, the null hypothesis is that the time series under consideration has a unit root, which implies that it is non-stationary. The alternative hypothesis, on the other hand, is that the time series is stationary and does not have a unit root. The ADF test is used to test whether the null hypothesis can be rejected based on the evidence provided by the data. If the null hypothesis is rejected, it suggests that the time series is stationary, which is a necessary condition for further analysis such as cointegration and regression analysis (Dickey & Fuller, 1979; Said & Dickey, 1984).The ADF test is a commonly used statistical test to determine the stationarity of a time series. The ADF statistic is a key output of this test, which is calculated by regressing the differenced series on its lagged values and a constant. The null hypothesis of the test is a unit root in the time series, indicating non-stationarity, while the alternative hypothesis is stationarity. The critical values for the ADF statistic depend on the level of significance and the sample size of the time series. If the calculated ADF statistic is less than the critical value, the null hypothesis of a unit root can be rejected, and the time series is considered stationary. A significant negative ADF statistic suggests that the time series can be modeled with an autoregressive or moving average process (Dickey and Fuller, 1979; Said and Dickey, 1984). The ADF test is used to determine if the stock price and sentiment score time series are stationary or non-stationary in this study.

4.4 Differencing

First order differencing is a commonly used technique in time series analysis to transform non-stationary data into stationary data by taking the difference between consecutive observations in the time series. The goal of differencing is to remove trends and seasonality in the data so that the statistical properties, such as the mean and variance, do not change over time (Pankratz, 1983). Through the removal of trends and seasonality, differencing can transform non-linear data into linear data, facilitating the application of linear models for the analysis of the relationship between dependent and independent variables. The formula for first order differencing can be written as:

$\Delta y(t) = y(t) - y(t-1)$

where $y(t)$ is the observation at time t, and $\Delta y(t)$ is the first difference of the time series at time t (Pankratz, 1983).

Similarly, higher order differencing involves taking the difference between the differences of consecutive observations. For example, second order differencing can be written as (Box et al., 2015):

Δ^2y_t = Δ(Δy_t) = (y_t - y_t-1) - (y_t-1 - y_t-2) = y_t - 2y_t-1 + y_t-2

Higher order differencing can be useful in cases where first order differencing is not sufficient to achieve stationarity in the data (Box et al., 2015). After differencing the time series, the resulting series can be checked for stationarity using tests like the Augmented Dickey-Fuller (ADF) test. If the differenced series is stationary, then it can be used to build a time series model.

In the present study, the differencing method will be applied to both the sentiment of tweets and the price of the selected stocks. This approach will help to eliminate any long-term trends and seasonal patterns from the stock price data, leading to stationary data with constant statistical properties. In addition, the differencing method will also be applied to the sentiment scores to remove any systematic biases and trends in the data that may result from changing public opinions or social trends. Overall, this will enhance the reliability and accuracy of the subsequent analysis by minimizing the impact of non-stationarity and other confounding factors in the data.

4.5 Regression Analysis and OLS Model

4.5.1 Regression Analysis

Linear regression analysis is a widely-used statistical method for modeling the relationship between a dependent variable and one or more independent variables. It involves fitting a linear equation to the data, with the aim of minimizing the sum of the squared errors between the observed and predicted values of the dependent variable. The linear regression model is commonly used in various fields, including economics, finance, psychology, social sciences, and engineering, to predict or estimate the value of the dependent variable based on the independent variables. The equation for the linear regression model is expressed as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta kXk + \varepsilon$$

where Y is the dependent variable, X1, X2, ..., Xk are the independent variables, β0 is the intercept, β1, β2, ..., βk are the regression coefficients, and ε is the error term. The regression coefficients represent the change in the dependent variable associated with a one-unit change in the independent variable, holding all other independent variables constant. The goal of linear regression analysis is to estimate the regression coefficients that provide the best-fit line to the data, as determined by the method of least squares (Kutner et al., 2005).

In this study, the stock price is treated as the dependent variable as it is the variable being predicted or explained by the independent variable, which is the social media sentiment score.The idea is to examine whether there is a significant relationship between social media sentiment score and stock price, and if so, to what extent social media sentiment score can be used to predict changes in stock price.

4.5.2 The Ordinary Least Squares (OLS) Model

The OLS method is the most widely used method to estimate the regression coefficients in the linear regression model. The OLS method minimizes the sum of the squared errors between the observed values of Y and the predicted values of Y based on X. The equation of the OLS estimator for the regression coefficient βj is:

$$\beta j = (\Sigma XiYi - n(\Sigma Xi)(\Sigma Yi)) / (\Sigma Xi^2 - n(\Sigma Xi)^2)$$

where Xi is the value of the independent variable X for the ith observation, Yi is the observed value of the dependent variable Y for the ith observation, and n is the sample size. The OLS estimator gives the value of βj that minimizes the sum of the squared errors between the observed values of Y and the predicted values of Y based on X.

In current study, the dependent variable (Y) is the first-order differenced daily closing prices of selected stock, which is calculated by subtracting each day's closing price from the previous day's closing price. The independent variable (X) is the sentiment scores calculated from a sample of tweets about stock. Both Y and X are assumed to have a linear relationship, and the OLS method is applied to estimate the coefficients of the linear model. Additionally, the OLS assumptions must be satisfied, including the normality and independence of the error term and the absence of multicollinearity among the independent variables.

The output of OLS includes estimates of the coefficients of the model, the standard errors of the coefficients, and the R-squared value.The coefficients of the OLS model are estimates of the effect of each independent variable on the dependent variable. These coefficients represent the change in the dependent variable for each unit change in the independent variable, while holding all other independent variables constant. The coefficient estimates are calculated using the formula:

$$\beta j = (\Sigma XiYi - n(\Sigma Xi)(\Sigma Yi)) / (\Sigma Xi^2 - n(\Sigma Xi)^2) \quad (6)$$

Where:

- βj: the estimate of the coefficient of the jth independent variable
- Yi: the value of the dependent variable for the ith observation
- Xi: the value of the jth independent variable for the ith observation
- n: the number of observations in the sample (Kutner et al.,2005).

The standard errors of the coefficients measure the amount of variation in the estimated coefficient due to random sampling error. The standard error for each coefficient estimate is calculated using the formula:

$$SE(\beta j) = sqrt[MSE / (\Sigma(Xi - X)^2)]$$

Where:

- $SE(\beta j)$: the standard error of the coefficient estimate for the jth independent variable
- MSE: the mean squared error of the residuals
- Xi: the value of the jth independent variable for the ith observation
- X: the mean value of the jth independent variable in the sample (Kutner et al.,2005).

The R-squared value is a measure of how well the independent variables explain the variation in the dependent variable. It represents the proportion of the total variation in the dependent variable that is explained by the independent variables. The R-squared value is calculated using the formula:

$$R^2 = 1 - (SSres / SStot)$$

Where:

- R^2: the coefficient of determination, which represents the proportion of the total variation in the dependent variable that is explained by the independent variables

- SSres: the sum of squared residuals, which measures the amount of variation in the dependent variable that is not explained by the independent variables
- SStot: the total sum of squares, which measures the total amount of variation in the dependent variable (Kutner et al.,2005).

The Ordinary Least Squares (OLS) model provides an array of outputs that yield information about the relationship between the independent and dependent variables. The coefficients of the model estimate the effect of each independent variable on the dependent variable, while keeping all other independent variables constant. The standard errors of the coefficients allow for the assessment of the precision of the estimates and the probability of obtaining similar estimates in future samples. The R-squared value is a measure of the goodness of fit of the model and determines the proportion of the variation in the dependent variable that is explained by the independent variables. Additionally, the F-statistic and its associated p-value provide a test of the overall significance of the model by measuring the ratio of the explained variation to the unexplained variation in the dependent variable.

4.6 ARIMA&SARIMAX

4.6.1 ARIMA

Autoregressive Integrated Moving Average (ARIMA) models are a class of statistical models for analyzing and forecasting time-series data. Developed in the 1970s, the models have become increasingly popular over time, with widespread use in finance, economics, and other fields that deal with time-series data (Box, Jenkins & Reinsel, 2008). This paper discusses the ARIMA model, its history, definition, explanation, and examples.

The ARIMA model has its roots in the early work of mathematicians and statisticians, such as Norbert Wiener and Andrey Kolmogorov, who developed mathematical methods for describing stochastic processes in the early 20th century (Pankratz, 2019). Later, statisticians such as George Box and Gwilym Jenkins developed the ARIMA model, building on earlier work by mathematicians and statisticians (Box et al., 2008). Box and Jenkins' book "Time Series Analysis: Forecasting and Control" (1970) is considered the seminal work on ARIMA modeling.ARIMA models are a class of time-series models that can be used to model the temporal dependence of a time series. The acronym "ARIMA" stands for "Autoregressive Integrated Moving Average." The model consists of three components: autoregression, differencing, and moving average (Box et al., 2008). The autoregression component models the relationship between an observation and a certain number of lagged observations. The differencing component removes the trend and/or seasonality from the time series. The moving average component models the error term as a linear combination of past error terms.

ARIMA models are generally expressed as ARIMA(p, d, q), where p is the order of the autoregressive component, d is the order of the differencing component, and q is the order of the moving average component. The parameters p, d, and q must be chosen based on the characteristics of the time series being analyzed. The autoregressive component models the relationship between an observation and a certain number of lagged observations. For example, an ARIMA(1,0,0) model with parameter phi = 0.5 is given by:

Y_t = phi * Y_(t-1) + e_t

where Y_t is the value of the time series at time t, phi is the autoregressive parameter, and e_t is the error term at time t. This equation states that the value of Y_t is a linear function of its previous value Y_(t-1) plus an error term e_t.

The differencing component is used to remove the trend and/or seasonality from the time series. A first-order differenced time series is defined as:

Y't = Y_t - Y(t-1)

where Y_t is the value of the time series at time t, and Y_(t-1) is the value of the time series at time t-1. The differenced time series removes the trend from the original time series.

The moving average component models the error term as a linear combination of past error terms. For example, an ARIMA (0,0,1) model with parameter theta = 0.5 is given by:

Y_t = e_t + theta * e_(t-1)

where Y_t is the value of the time series at time t, e_t is the error term at time.

To simplify the analysis, in this study it was assumed that the ARIMA model for the selected stocks data had a first-order differencing term (d=1), and the autoregressive (AR) and moving average (MA) terms were set to 1 (p=1 and q=1). This means that the model considers the previous time step of the differenced series and one lag of the errors in the regression. This choice of hyperparameters is based on the analysis of the ACF and PACF plots, which showed a significant correlation at the first lag and the decay of the correlation for subsequent lags. By keeping the number of parameters low, we avoid overfitting the model and making it unnecessarily complex.

In conclusion, ARIMA is a powerful time series analysis technique that can help us model and forecast data with temporal dependence. The ARIMA model works by identifying the patterns and relationships within the data and creating a mathematical model that can be used to make predictions about future values. Through the analysis of the Amazon stock price dataset, we can see the effectiveness of the ARIMA model in predicting future stock prices. By fitting an ARIMA model to the dataset, we were able to forecast future values with a reasonable degree of accuracy. It is important to note that while ARIMA models can be effective, they do have limitations. For example, ARIMA assumes that the data is stationary, which means that the mean and variance of the data do not change over time. This can be a limitation in situations where the data exhibits non-stationary behavior.

4.6.2 SARIMAX

SARIMAX, which stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous variables, is a popular time series model used for forecasting. It is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model, which is a univariate time series model that accounts for trends, seasonality, and autocorrelation in the data. SARIMAX extends the ARIMA model to include one or more exogenous variables, which are external variables that may affect the dependent variable. The exogenous variables can be either stationary or non-stationary and may be measured at the same frequency as the dependent variable or at a different frequency. In this study, the sentiment score is the exogenous variable that we want to include in the model. The SARIMAX model estimates the effect of the exogenous variable(s) on the dependent variable while accounting for autocorrelation, seasonality, and trends. This makes it a powerful tool for time series forecasting, especially when the data contains multiple variables that influence the dependent variable.

The statsmodels library in Python offers a range of functions to build SARIMAX models, which can estimate parameters of the model, fit it to the data, and make predictions. By fitting the SARIMAX model to the data, we can forecast the dependent variable for future time periods by incorporating the values of exogenous variables. Without the use of SARIMAX, forecasting the dependent variable for future time periods based on the values of exogenous variables is not feasible. Thus, the SARIMAX model is an essential tool for modelling the relationship between the dependent variable and one or more exogenous variables, as it accounts for autocorrelation, seasonality, and trends. In our case, the application of SARIMAX is necessary to estimate the effect of sentiment score on Apple stock price, considering the exogenous variables.

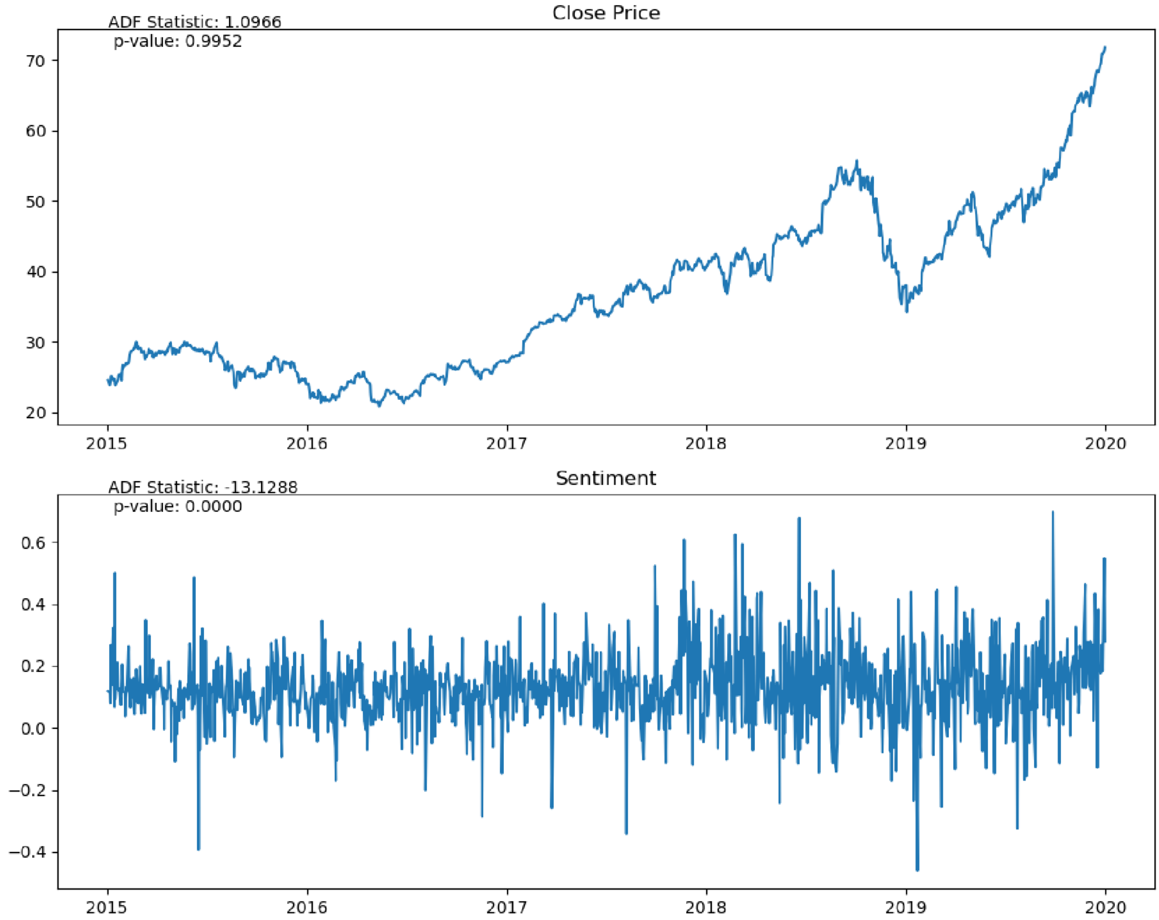
5 Results and Discussion

5.1 Apple Results

5.1.1 Apple ADF Test Results

The obtained results from the ADF test on the Apple stock prices and sentiment data shown on **Graph 3** suggest that the stock prices data is non-stationary while the sentiment data is stationary. The ADF statistic for the Close_price column is 1.096597, and the p-value is 0.995187. These results indicate that we cannot reject the null hypothesis of non-stationarity, and thus, the Close_price column is non-stationary. On the other hand, the ADF statistic for the Sentiment column is -13.128822, and the p-value is 0.000000, indicating that we can reject the null hypothesis of non-stationarity, and thus, the Sentiment column is stationary. The non-stationarity of the stock prices data suggests that the data exhibits some form of trend or seasonality, which makes it difficult to model and forecast accurately. This is a common problem in financial time series data, as stock prices are influenced by many factors, including economic events, company news, and market sentiment. Therefore, it is crucial to transform the non-stationary data into stationary data by applying methods such as differencing, logarithmic transformation, or detrending, to make it more suitable for modelling and forecasting. On the other hand, the stationary nature of the sentiment data suggests that it does not exhibit any trend or seasonality and is relatively stable over time. This is desirable in time series analysis as it simplifies the modelling and forecasting process. However, it is important to note that stationarity does not necessarily mean that the data is predictable or that there is a causal relationship between the sentiment data and the stock prices data. Therefore, further analysis is required to investigate the relationship between the two variables.

Graph 3 ADF Test Results for Close price and Sentiment (2015-2019), Apple



Source: Own calculations using Python

5.1.2 Differencing, Apple

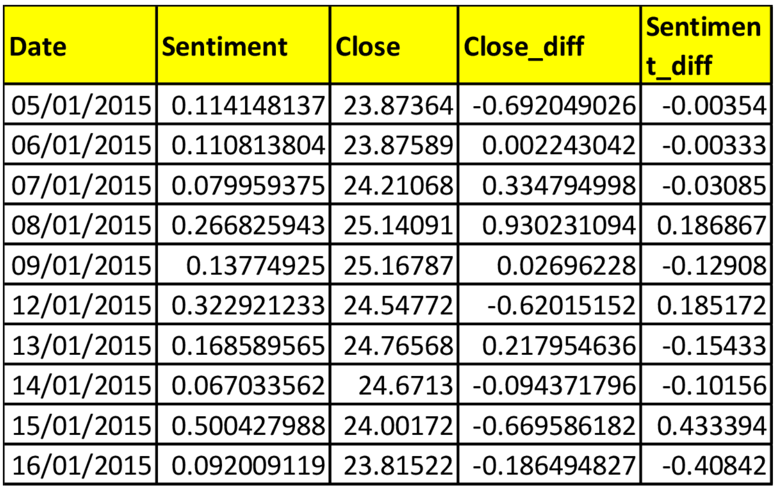
To make differencing obtained data, the diff() function in Pandas is going to be used in Python. This function computes the difference between consecutive rows (or columns) of a DataFrame. In this code snippet, data firstly need to be read in the merged data file using Pandas' read_csv() function. Then, we apply first order differencing to the "Close" and

"Sentiment" columns using the diff() function and store the results in new columns "Close_diff" and "Sentiment_diff", respectively.

The 'Close' column in our study refers to the closing stock price of Apple Inc. on a given trading day. This is the final price at which the stock is traded for the day. The 'Sentiment' column represents a weighted average sentiment per day, which is calculated by weighting each tweet's sentiment score by various interaction metrics such as replies, likes, and retweets. This weighted average is then aggregated daily.

Finally, we drop the first row (which will have a NaN value due to differencing) using the dropna() function and print the first few rows to verify that differencing was applied correctly.

Table 1 Example of Differencing process (2015-2019), Apple



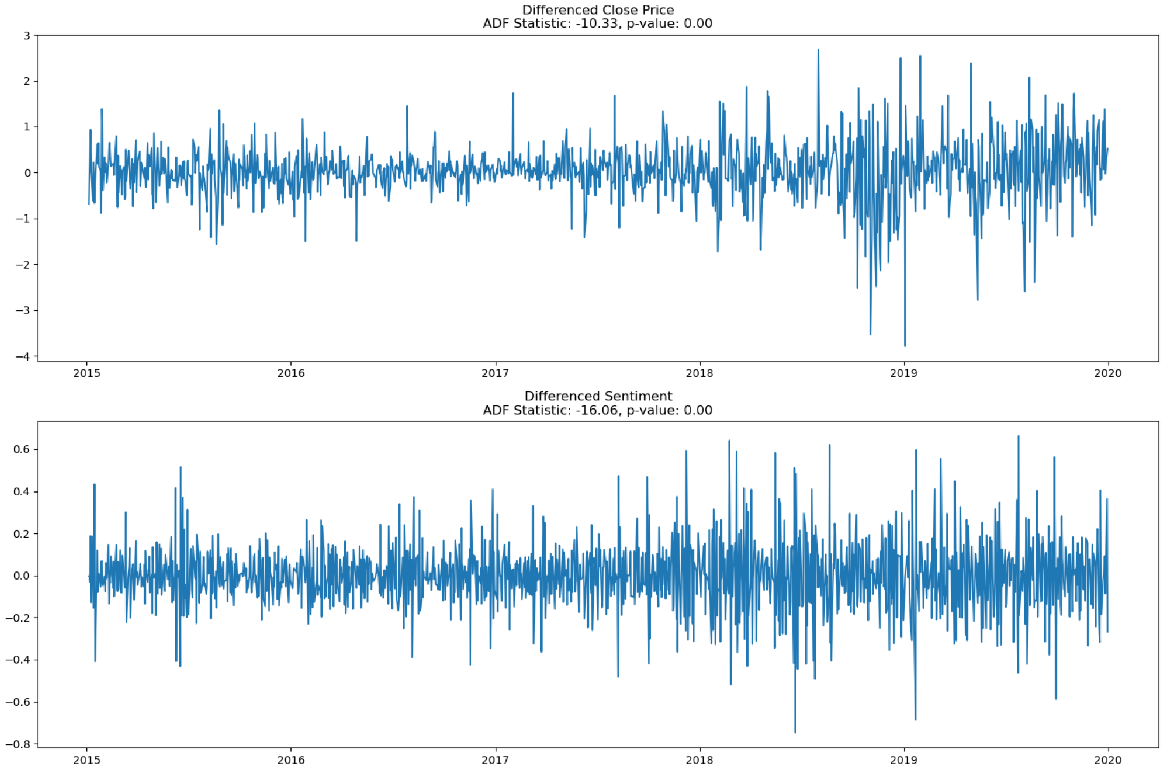
Date	Sentiment	Close	Close_diff	Sentiment_diff
05/01/2015	0.114148137	23.87364	-0.692049026	-0.00354
06/01/2015	0.110813804	23.87589	0.002243042	-0.00333
07/01/2015	0.079959375	24.21068	0.334794998	-0.03085
08/01/2015	0.266825943	25.14091	0.930231094	0.186867
09/01/2015	0.13774925	25.16787	0.02696228	-0.12908
12/01/2015	0.322921233	24.54772	-0.62015152	0.185172
13/01/2015	0.168589565	24.76568	0.217954636	-0.15433
14/01/2015	0.067033562	24.6713	-0.094371796	-0.10156
15/01/2015	0.500427988	24.00172	-0.669586182	0.433394
16/01/2015	0.092009119	23.81522	-0.186494827	-0.40842

Source: Own calculations using Python and Excel.

After differencing, original time series data was transformed into a stationary time series data. Both the "Close" and "Sentiment" have been differenced columns, resulting in two new columns "Close_diff" and "Sentiment_diff". The ADF test was then applied to these new columns to check for stationarity. The ADF test results showed (Graph 4) that both the "Close_diff" and "Sentiment_diff" columns are stationary with high statistical significance, as indicated by the low p-values (p-value < 0.05) and the negative ADF statistics.

Therefore, we can conclude that both "Close_diff" and "Sentiment_diff" are stationary time series. This is important because many statistical models, including regression models, assume that the data is stationary, and if the data is non-stationary, it can lead to unreliable results. By ensuring that our data is stationary, we can be more confident in the results of our analysis.Furthermore, the fact that the p-values for both series are very small (close to zero) indicates that there is a strong evidence of stationarity, which strengthens the conclusions we can draw from our analysis.

Graph 4 ADF test results on Apple stock price after differencing (2015-2019), Apple



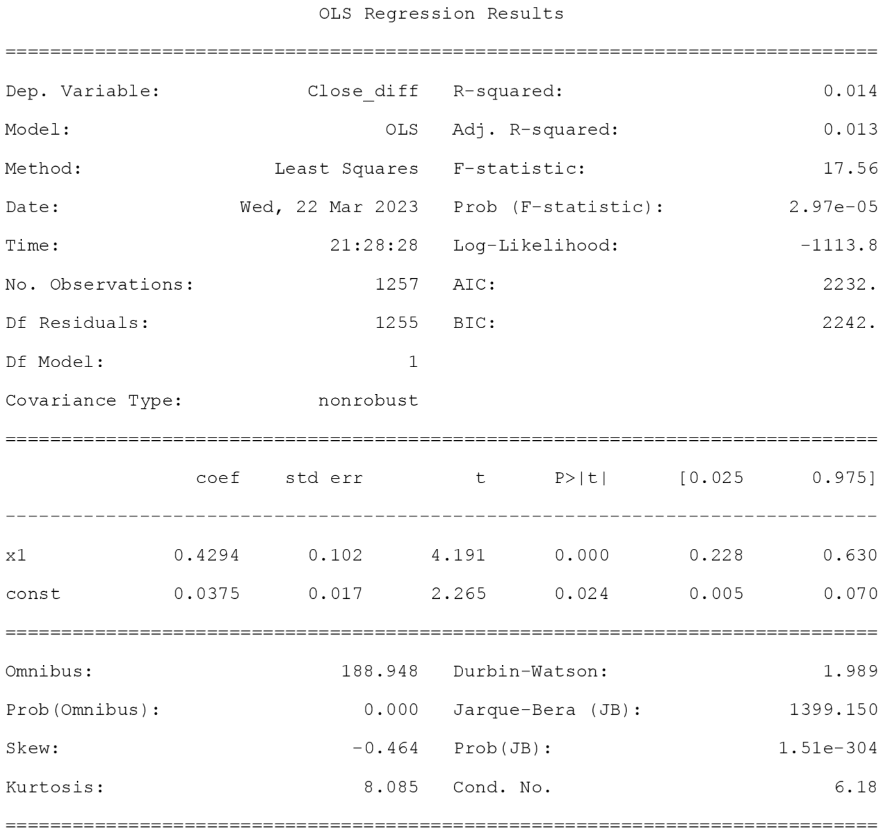
Source: Own calculations using Python.

5.1.3 OLS Model Results (2015-2019), Apple

The output below (Table 2) shows the OLS regression results for the 'Close_diff' column. The R-squared value is 0.014, indicating that only 1.4% of the variance in 'Close_diff' is explained by the predictor variable. The F-statistic is 17.56, and the p-value is 2.97e-05, indicating that the model is statistically significant. The coefficient for the predictor variable 'Sentiment_diff' is 0.4294, which suggests that for every unit increase in 'Sentiment_diff', the 'Close_diff' value increases by 0.4294. The constant term is 0.0375, which is the expected value of 'Close_diff' when 'Sentiment_diff' is zero. The standard errors for the coefficients are also provided, along with the t-statistic and corresponding p-value for testing the null hypothesis that the coefficient is zero. In both cases, the p-value is less than 0.05, indicating that the coefficients are statistically significant.

Overall, the results suggest that there is a small positive relationship between changes in sentiment and changes in the closing price of Apple stock, but this relationship only explains a small amount of the variability in the closing price.

Table 2 OLS Regression Results (2015-2019), Apple



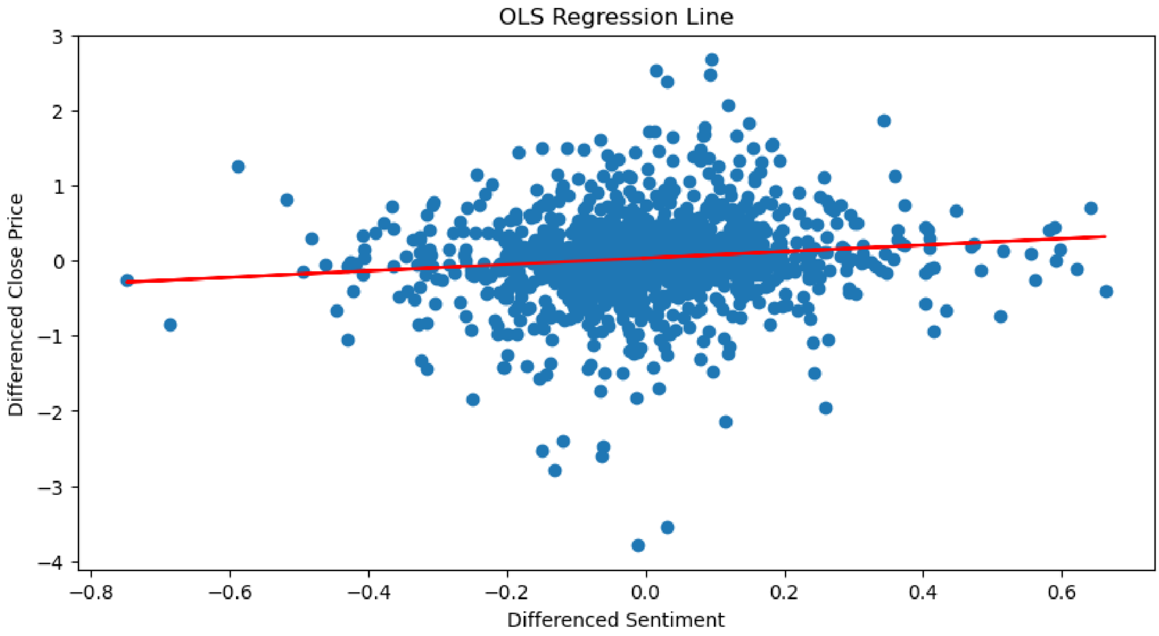
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Close_diff   R-squared:                       0.014
Model:                            OLS   Adj. R-squared:                  0.013
Method:                 Least Squares   F-statistic:                     17.56
Date:                Wed, 22 Mar 2023   Prob (F-statistic):           2.97e-05
Time:                        21:28:28   Log-Likelihood:                -1113.8
No. Observations:                1257   AIC:                             2232.
Df Residuals:                    1255   BIC:                             2242.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4294      0.102      4.191      0.000       0.228       0.630
const          0.0375      0.017      2.265      0.024       0.005       0.070
==============================================================================
Omnibus:                      188.948   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1399.150
Skew:                          -0.464   Prob(JB):                    1.51e-304
Kurtosis:                       8.085   Cond. No.                         6.18
==============================================================================
```

Source: Own calculations using Python.

The Graph 5 shows the scatter plot between the differenced sentiment values and the differenced close price values. The red line represents the OLS regression line that has been fit to the data. The OLS model tries to find the best linear relationship between the dependent variable (differenced close price) and the independent variable (differenced sentiment) by minimizing the sum of the squared errors between the predicted values and the actual values. In this case, the OLS model has found a very weak positive linear relationship between the differenced sentiment values and the differenced close price values, as indicated by the low R-squared value (0.014) and the small positive coefficient

(0.4294) of the independent variable. The results show that the effect of changes in sentiment on changes in stock prices for Apple is positive, but relatively small.

Graph 5 Relationship Between Differenced Sentiment and Differenced Price (2015-2019), Apple



Source: Own calculations using Python.

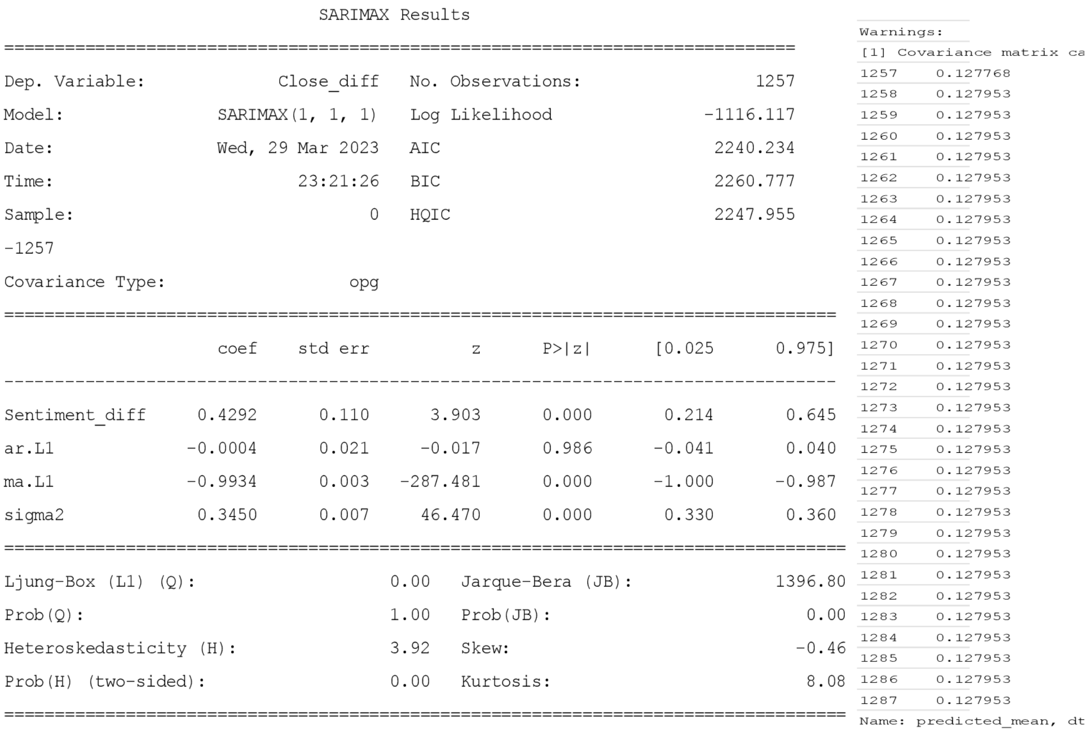
5.1.4 SARIMAX Test Results, Apple

Based on the SARIMAX summary output, the model has an order of (1,1,1), which means it has one autoregressive term, one differencing term, and one moving average term. The exogenous variable used in the model is "Sentiment_diff". The p-value for "Sentiment_diff" is less than 0.05, which indicates that it is a statistically significant variable in the model. The p-values for the autoregressive and moving average terms are not significant, which suggests that these terms are not needed in the model. The Ljung-Box test indicates that there is no significant autocorrelation at lag 1, and the Jarque-Bera test suggests that the residuals are normally distributed. Regarding the predictions, the output shows the predicted mean value for each day from the start of the forecast period (len(data)) to the end (len(data)+30). The output shows the predicted value for each day, which is the same value for all days since you used the "typ='levels'" parameter in the predict() function. The value represents the predicted level of the time series for each day, after adjusting for the effect of the exogenous variable.

The left part **Table 3** of the output is the index of the predicted values, which starts at 1257 and goes up to 1287. These are the 31 days after the end of the original dataset (which ends at the 1256th observation=27/12/2019). The second part of the output is the predicted mean for each day, which is 0.127953 for all 31 days. This means that the model is predicting that the difference in the Apple stock price for each of these days will be 0.127953, assuming that the sentiment score remains the same.

Overall, based on the output, the model suggests that the sentiment score has a statistically significant effect on the difference in the Apple stock price. However, the model only provides a constant forecast for the future and might not be capturing all the relevant factors that affect the Apple stock price, such as economic indicators, news events, or company announcements. Incorporating these factors into the model might help improve the forecasting accuracy and provide more meaningful predictions for the future.

Table 3 SARIMAX Results (2015-2019), Apple



```
                           SARIMAX Results
==============================================================================
Dep. Variable:             Close_diff   No. Observations:                 1257
Model:               SARIMAX(1, 1, 1)   Log Likelihood               -1116.117
Date:                Wed, 29 Mar 2023   AIC                           2240.234
Time:                        23:21:26   BIC                           2260.777
Sample:                             0   HQIC                          2247.955
-1257
Covariance Type:                  opg
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Sentiment_diff     0.4292      0.110      3.903      0.000       0.214       0.645
ar.L1             -0.0004      0.021     -0.017      0.986      -0.041       0.040
ma.L1             -0.9934      0.003   -287.481      0.000      -1.000      -0.987
sigma2             0.3450      0.007     46.470      0.000       0.330       0.360
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              1396.80
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               3.92   Skew:                            -0.46
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.08
===================================================================================
```

```
Warnings:
[1] Covariance matrix ca
1257    0.127768
1258    0.127953
1259    0.127953
1260    0.127953
1261    0.127953
1262    0.127953
1263    0.127953
1264    0.127953
1265    0.127953
1266    0.127953
1267    0.127953
1268    0.127953
1269    0.127953
1270    0.127953
1271    0.127953
1272    0.127953
1273    0.127953
1274    0.127953
1275    0.127953
1276    0.127953
1277    0.127953
1278    0.127953
1279    0.127953
1280    0.127953
1281    0.127953
1282    0.127953
1283    0.127953
1284    0.127953
1285    0.127953
1286    0.127953
1287    0.127953
Name: predicted_mean, dt
```

Source: Own calculations, Python.

5.2 Amazon Stock Results

5.2.1 ADF Test Results, Amazon

The ADF test has been conducted on both the price and sentiment data. The resulting ADF statistic and p-value for each test can be observed from **Graph 6** on interpreted as follows:

- ADF Statistic for Price: -0.8585138578814998

  This indicates that the price data is non-stationary and contains a stochastic trend.

- p-value for Price: 0.8013273441428735

This is the probability that the null hypothesis (the time series has a unit root and is non-stationary) is true. Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis and conclude that the price data is non-stationary.

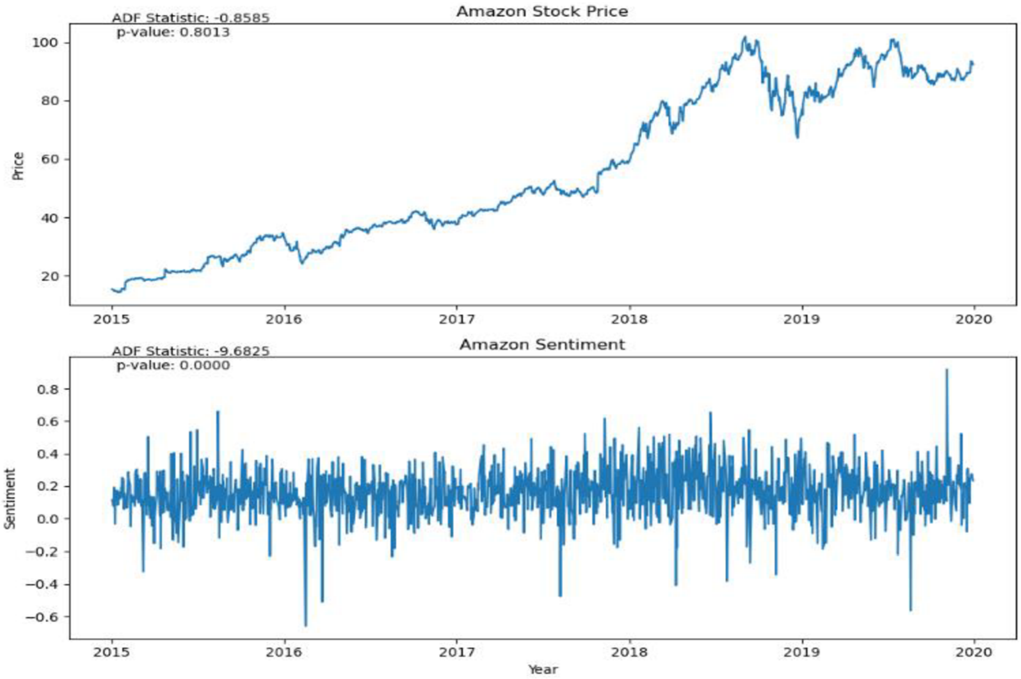
- ADF Statistic for Sentiment: -9.68248230381418

  This indicates that the sentiment data is stationary and does not contain a stochastic trend.

- p-value for Sentiment: 1.1943897639517983e-16

  This is the probability that the null hypothesis (the time series has a unit root and is non-stationary) is true. Since the p-value is much smaller than the significance level of 0.05, we reject the null hypothesis and conclude that the sentiment data is stationary. As the data failed to reject the null hypothesis of non-stationarity, differencing the data to achieve stationarity is recommended before proceeding with modeling.

Graph 6 ADF test results for Close price and Sentiment (2015-2019), Amazon



Source: Own calculations using Python.

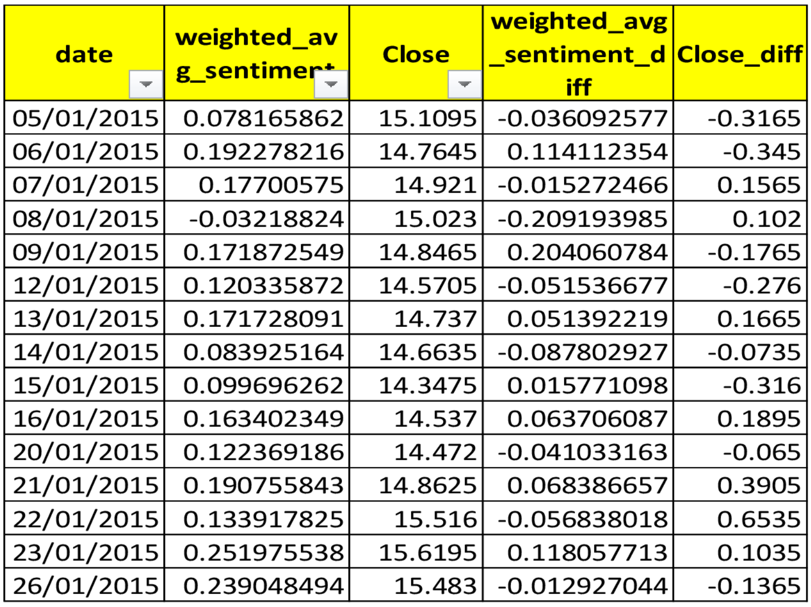
5.2.2 Differencing, Amazon

In this step, differencing was applied to both the close price and sentiment data, and created new columns 'Close_diff' and 'weighted_avg_sentiment_diff' in the merged dataset 'merged_sentiment_stock_diff'.

Both variables need to be differenced to make them stationary. If only one variable is differenced, the resulting model may not fully capture the underlying relationship between the variables, which could lead to biased or unreliable results. Additionally, differencing both variables can help to ensure that any potential spurious correlation is removed, allowing for more accurate modeling and analysis. Therefore, it is decided to apply differencing to both the close price and sentiment data to transform them into stationary time series. This approach is similar to the one used to transform the Apple stock data into a stationary time series.

To apply differencing in Excel(**Table 4**), first, we need to create a new column to hold the differenced values of the original data. This can be done by using the formula =B2-B1, assuming that the original data is in column B and starts from row 2. This formula calculates the difference between the current value and the previous value in the column. To apply differencing to the close price data, a new column is created and the difference formula is applied. Since there is no previous value to calculate the difference from, the first row of the new column is assigned a value of 0. The formula is then dragged down to the last row of the data to difference the entire column.Once differencing is applied to both the close price and sentiment columns, the ADF test can be used again to check for stationarity. If the p-value is less than the significance level, it can be concluded that the differenced data is stationary and ready for modeling. 'Close_diff' and 'weighted_avg_sentiment_diff' are the differenced values of the original 'Close' and 'weighted_avg_sentiment' columns, respectively.

Table 4 Example of differencing process in Excel, Amazon



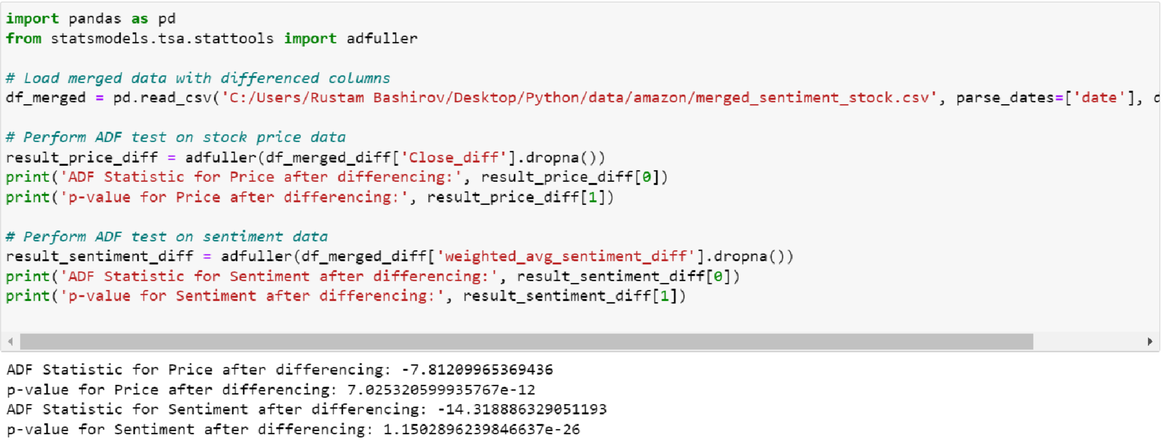
date	weighted_avg_sentiment	Close	weighted_avg_sentiment_diff	Close_diff
05/01/2015	0.078165862	15.1095	-0.036092577	-0.3165
06/01/2015	0.192278216	14.7645	0.114112354	-0.345
07/01/2015	0.17700575	14.921	-0.015272466	0.1565
08/01/2015	-0.03218824	15.023	-0.209193985	0.102
09/01/2015	0.171872549	14.8465	0.204060784	-0.1765
12/01/2015	0.120335872	14.5705	-0.051536677	-0.276
13/01/2015	0.171728091	14.737	0.051392219	0.1665
14/01/2015	0.083925164	14.6635	-0.087802927	-0.0735
15/01/2015	0.099696262	14.3475	0.015771098	-0.316
16/01/2015	0.163402349	14.537	0.063706087	0.1895
20/01/2015	0.122369186	14.472	-0.041033163	-0.065
21/01/2015	0.190755843	14.8625	0.068386657	0.3905
22/01/2015	0.133917825	15.516	-0.056838018	0.6535
23/01/2015	0.251975538	15.6195	0.118057713	0.1035
26/01/2015	0.239048494	15.483	-0.012927044	-0.1365

Source: Own calculations using Excel

5.2.3 ADF Test Results After Differencing, Amazon

After differencing, ADF test was performed on both differenced columns to check if the data has become stationary. The ADF statistic and p-value for each test were obtained. The ADF statistic for the differenced close price data is -7.812 with a p-value of 7.03e-12, and for the differenced sentiment data it is -14.319 with a p-value of 1.15e-26. Since the p-value is much lower than the significance level of 0.05,it can be inferred that the differenced data is now stationary and suitable for further modeling.

Figure 3 ADF test after differencing example using Python,Amazon



```
import pandas as pd
from statsmodels.tsa.stattools import adfuller

# Load merged data with differenced columns
df_merged = pd.read_csv('C:/Users/Rustam Bashirov/Desktop/Python/data/amazon/merged_sentiment_stock.csv', parse_dates=['date'], c

# Perform ADF test on stock price data
result_price_diff = adfuller(df_merged_diff['Close_diff'].dropna())
print('ADF Statistic for Price after differencing:', result_price_diff[0])
print('p-value for Price after differencing:', result_price_diff[1])

# Perform ADF test on sentiment data
result_sentiment_diff = adfuller(df_merged_diff['weighted_avg_sentiment_diff'].dropna())
print('ADF Statistic for Sentiment after differencing:', result_sentiment_diff[0])
print('p-value for Sentiment after differencing:', result_sentiment_diff[1])
```

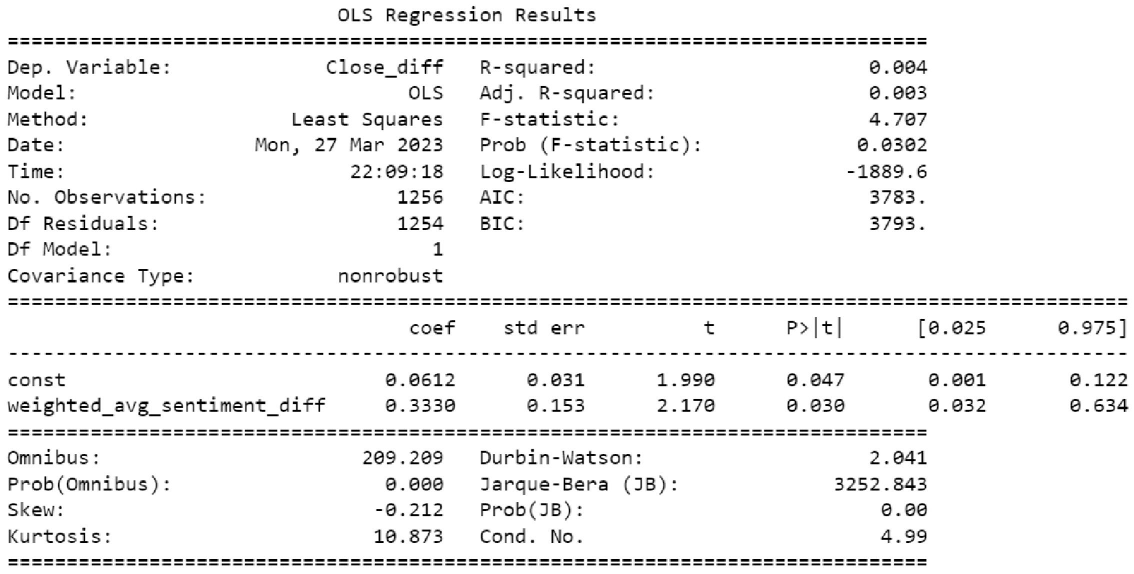
```
ADF Statistic for Price after differencing: -7.81209965369436
p-value for Price after differencing: 7.025320599935767e-12
ADF Statistic for Sentiment after differencing: -14.318886329051193
p-value for Sentiment after differencing: 1.1502896239846637e-26
```

Source: Own calculations using Python.

5.2.4 OLS Model Results, Amazon

Based on the OLS regression results(**Table 5**), the R-squared value of the model is 0.004, which indicates that only 0.4% of the variation in the stock prices can be explained by changes in sentiment scores. The coefficient for the 'weighted_avg_sentiment_diff' predictor variable is 0.3330, indicating that a one-unit increase in the sentiment score difference leads to an increase in the stock price difference by 0.3330 units. The p-value for the predictor variable is 0.0302, which is below the commonly used threshold of 0.05, suggesting that there is evidence to support the hypothesis that sentiment scores are significantly related to stock prices. However, the small R-squared value and low coefficient suggest that the relationship is weak, and further analysis may be needed to draw stronger conclusions.

Table 5 OLS Regression Results, Amazon

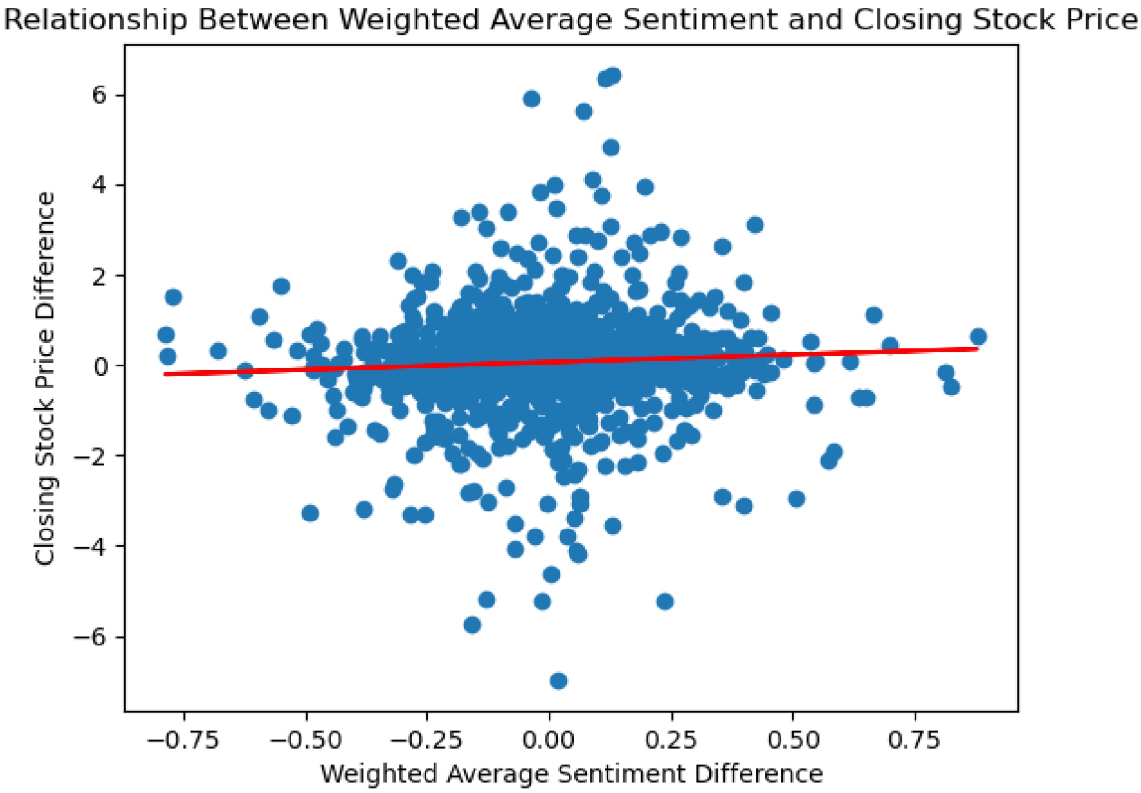


```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Close_diff   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     4.707
Date:                Mon, 27 Mar 2023   Prob (F-statistic):             0.0302
Time:                        22:09:18   Log-Likelihood:                -1889.6
No. Observations:                1256   AIC:                             3783.
Df Residuals:                    1254   BIC:                             3793.
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                         0.0612      0.031      1.990      0.047       0.001       0.122
weighted_avg_sentiment_diff   0.3330      0.153      2.170      0.030       0.032       0.634
==============================================================================
Omnibus:                      209.209   Durbin-Watson:                   2.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3252.843
Skew:                          -0.212   Prob(JB):                         0.00
Kurtosis:                      10.873   Cond. No.                         4.99
==============================================================================
```

Source: Own calculations using Python.

The regression line shows the expected change in the stock price for a unit increase in the sentiment score, based on the model that we fitted using OLS. In this case, there are differenced values of the sentiment score and the stock price, which means that the regression line shows the expected change in the difference in stock price for a unit. increase in the difference in sentiment score. The scatter plot (Graph 7) shows the actual data points, with the x-axis representing the difference in sentiment score and the y-axis representing the difference in stock price. The regression line is overlaid on top of the scatter plot to show the relationship between the two variables. We can see from the plot that there is a slight positive relationship between the two variables, but there is also a lot of variability in the data, which is reflected in the low R-squared value. The scatter plot shows the actual data points, with the x-axis representing the difference in sentiment score and the y-axis representing the difference in stock price. The regression line is overlaid on top of the scatter plot to show the relationship between the two variables. We can see from the plot that there is a slight positive relationship between the two variables, but there is also a lot of variability in the data, which is reflected in the low R-squared value.

Graph 7 Relationship Between Weighted Average Sentiment and Closing Stock Price, Amazon



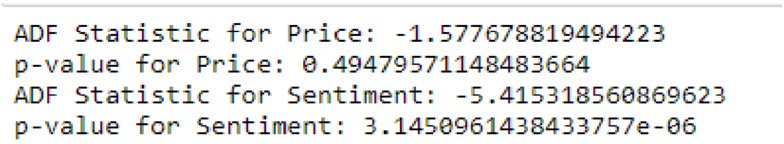
Source: Own calculations using Python.

5.3 Microsoft Results

5.3.1 ADF Test Results, Microsoft

The ADF statistic measures the stationarity of a time series. A stationary time series has a constant mean and variance over time, while a non-stationary time series has a changing mean and/or variance over time. Based on the results on **Figure 4** , the ADF statistic for the price series is -1.577678819494223 and the p-value is 0.49479571148483664. Since the p-value is higher than the significance level of 0.05, we fail to reject the null hypothesis that the time series is non-stationary. Therefore, it can be concluded that the price series is non-stationary and has a changing mean and/or variance over time. On the other hand, the ADF statistic for the sentiment series is -5.415318560869623 and the p-value is 3.1450961438433757e-06, which is much lower than the significance level of 0.05. Therefore, the null hypothesis is rejected that the time series is non-stationary, and it can be concluded that the sentiment series is stationary and has a constant mean and variance over time.

Figure 4 ADF Test Results, Amazon

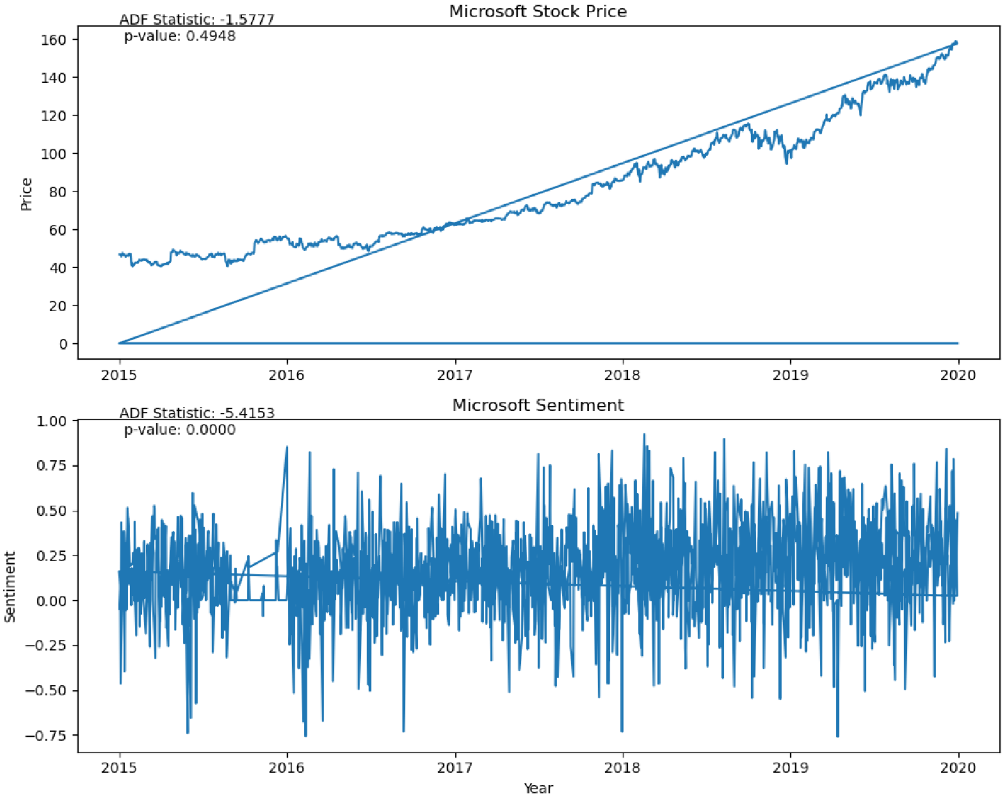


```
ADF Statistic for Price: -1.577678819494223
p-value for Price: 0.49479571148483664
ADF Statistic for Sentiment: -5.415318560869623
p-value for Sentiment: 3.1450961438433757e-06
```

Source: Own calculations using Python

The graph shows two subplots: one for the stock price of Microsoft and the other for the sentiment score. The x-axis represents time, and the y-axis represents the stock price and the sentiment score, respectively. The title of each subplot indicates which variable is being plotted. The blue line represents the sentiment score, and the red line represents the stock price. The ADF statistic and p-value for each variable are also displayed on the **Graph 8**. The purpose of this graph is to visualize the trends of the two variables over time and to show the results of the ADF test, which is used to determine whether the time series is stationary or not. To make the data stationary, first-order differencing can be performed on the stock price data, which involves subtracting each value from the previous value to remove the trend component. On the other hand, since the sentiment data is already stationary, differencing may not be necessary. After differencing, the ADF test can be re-run to confirm whether the data has become stationary or not.

Graph 8 ADF Test Result, Microsoft



Source: Own calculations using Python

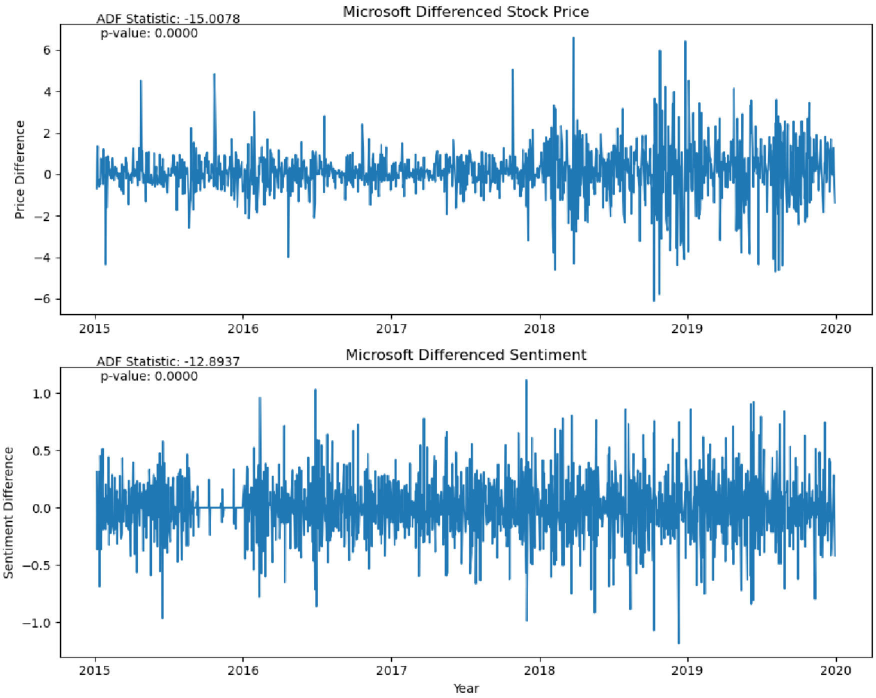
5.3.2 ADF Test Results After Differencing, Microsoft

After differencing both variables, sentiment, and close price values, ADF test shows results as per below.

- ADF Statistic for Price (after differencing): -15.007829885218202
- p-value for Price (after differencing): 1.066299219575289e-27
- ADF Statistic for Sentiment (after differencing): -12.89366265011061
- p-value for Sentiment (after differencing): 4.394625898798293e-24

The ADF test results show that both the differenced Microsoft stock price and sentiment series are stationary, as evidenced by their high negative ADF statistics and very low p-values. This suggests that the trend and seasonality components have been removed from the series, and they exhibit a constant mean and variance over time. These results indicate that the differenced series may be suitable for use in time series modelling and forecasting.

Graph 9 ADF Test Result After Differencing



Source: Own calculations using Python

5.3.3 OLS Model Results, Microsoft

The OLS regression results show that the R-squared value is 0.003, indicating that only a very small portion of the variation in the dependent variable (Close_diff) can be explained by the independent variable (Sentiment_diff). The p-value for the Sentiment_diff coefficient is 0.060, which is slightly above the common significance level of 0.05, suggesting that the Sentiment_diff variable may not be a significant predictor of the Close_diff variable. Additionally, the Durbin-Watson statistic is 2.213, which is close to the ideal value of 2, indicating no significant autocorrelation in the residuals.

It's possible that sentiment from tweets in English represents only a small piece of the pie when it comes to the overall sentiment towards Microsoft. There are many other sources of sentiment and information that could impact Microsoft stock prices, including news articles, financial reports, press releases, and social media platforms in other languages. Additionally, even within tweets in English, the sentiment may not be representative of the sentiment of the general population. The sentiment of Twitter users could be biased in various ways, such as being more negative or positive than the general population or being more influenced by certain factors such as political beliefs or brand loyalty. Therefore, it's important to consider the limitations of using sentiment from tweets in English as a predictor of Microsoft stock prices, and to be cautious in drawing strong conclusions based solely on this data.

Table 6 OLS Model Results, Microsoft

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Close_diff   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     3.544
Date:                Thu, 30 Mar 2023   Prob (F-statistic):             0.0600
Time:                        18:27:41   Log-Likelihood:                -2026.5
No. Observations:                1256   AIC:                             4057.
Df Residuals:                    1254   BIC:                             4067.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.0882      0.034      2.572      0.010       0.021       0.156
Sentiment_diff     0.2249      0.119      1.883      0.060      -0.009       0.459
==============================================================================
Omnibus:                      138.395   Durbin-Watson:                   2.213
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1125.450
Skew:                          -0.080   Prob(JB):                    4.09e-245
Kurtosis:                       7.635   Cond. No.                         3.48
==============================================================================
```

Source: Own calculations using Python

6 Conclusion

Based on the objectives of this study, which aimed to investigate the relationship between social media sentiment and stock prices for a sample of companies listed on the NASDAQ stock exchange, the results suggest that social media sentiment may not have a significant impact on stock prices for the sample companies, namely Apple, Amazon, and Microsoft.

The regression analysis for Apple and Amazon showed that there is a statistically significant relationship between the sentiment score and stock price, but the relationship is still weak. In the case of Microsoft, the relationship between sentiment scores and stock prices was not found to be statistically significant. These findings suggest that while social media sentiment may have some influence on stock prices, other factors such as economic indicators, news events, or company announcements may have a larger impact. Regarding the SARIMAX analysis on Apple stock, the model has an order of (1,1,1), with an exogenous variable of "Sentiment_diff" which was found to be a statistically significant variable in the model. The predicted mean value for each day from the start of the forecast period to the end is 0.127953, assuming the sentiment score remains the same. However, the model only provides a constant forecast for the future and may not be capturing all the relevant factors that affect the Apple stock price. Incorporating other variables such as economic indicators or news events may help improve the forecasting accuracy and provide more meaningful predictions for the future.

In conclusion, while social media sentiment may have some influence on stock prices, the results of this study suggest that it may not be a significant predictor of stock prices for the sample companies. Therefore, caution should be exercised in relying solely on social media sentiment as a predictor of stock prices. The findings of this study can provide useful insights for investors and financial analysts as they make investment decisions, but further research is needed to fully understand the impact of social media sentiment on the stock market and to identify the factors that influence sentiment towards a particular stock or company. Incorporating other variables such as economic indicators or news events may help improve the forecasting accuracy and provide more meaningful predictions for the future.

7 References

- Abadi, M. et al. (2016) “Deep learning with differential privacy,” Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security [Preprint]. Available at: https://doi.org/10.1145/2976749.2978318.
- Asness, C.S., Moskowitz, T.J. and Pedersen, L.H. (2013) Value and Momentum Everywhere. The Journal of Finance, 68, 929-985. http://dx.doi.org/10.1111/jofi.12021
- Alvarez-Peregrina, C., Martinez-Perez, C., Sanchez-Tena, M.A., & Villa-Collar, C. (2020). Citation Network Analysis of the Novel Coronavirus Disease 2019 (COVID-19). International Journal of Environmental Research and Public Health, 17(20). https://doi.org/10.3390/ijerph17207690
- Althaus, S. L. and Tewksbury, D. (2000) Patterns of Internet and Traditional News Media Use in a Networked Community. Political Communication, 17(1), 21-45. doi:10.1080/105846000198495
- Bakos, Y. and Brynjolfsson, E. (2000) “Bundling and competition on the internet,” Marketing Science, 19(1), pp. 63–82. Available at: https://doi.org/10.1287/mksc.19.1.63.15182.
- Barron, A. (2021). The taking place of older age. Cultural Geographies, 28(4), 661-674. https://doi.org/10.1177/14744740211020510
- Bhuyan, R., & Rhee, S. G. (2017). A brief history of the S&P 500. Journal of Index Investing, 7(1), 11-18. doi: 10.3905/jii.2017.7.1.011
- BODIE, Zvi, Alan J. MARCUS a Alex KANE. The McGraw-Hill/Irwin series in finance, insurance and real estate. 10. McGraw-Hill Education Asia: McGraw Hill Higher Education; Tenth Edition, 2014. ISBN 0071262288, 9780071262286.
- Bollen, J., Mao, H. and Zeng, X. (2011) “Twitter mood predicts the stock market,” Journal of Computational Science, 2(1), pp. 1–8. Available at: https://doi.org/10.1016/j.jocs.2010.12.007.
- BOX, George E. P., Gwilym M. JENKINS, Gregory C. REINSEL a Greta M. LJUNG. Time series analysis: forecasting and control. Fifth edition. Hoboken: John Wiley, [2016]. Series in probability and statistics (Wiley). ISBN 978-1-118-67502-1.

- Box, G.E., Jenkins, G.M. and Reinsel, G.C. (2008) “Time Series analysis,” Wiley Series in Probability and Statistics[Preprint]. Available at: https://doi.org/10.1002/9781118619193.
- BOX, George E. P., Gwilym M. JENKINS, Gregory C. REINSEL a Greta M. LJUNG. Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken, 2015.
- Choi, J. P. and Jeon, D.-S. (2021) A Leverage Theory of Tying in Two-Sided Markets with Nonnegative Price Constraints. American Economic Journal: Microeconomics, 13(1), 283-337. doi:10.1257/mic.20180234
- Davis, Edward Byrd, Jenny L. McGuire and John D. Orcutt. Ecological niche models of mammalian glacial refugia show consistent bias. Ecography [online]. 2014, n/a-n/a [cit. 2023-03-29]. ISSN 09067590. Dostupné z: doi:10.1111/ecog.01294
- Das, M., Zhu, C., & Kuchroo, V.K. (2017). “Tim-3 and its role in regulating anti-tumor immunity,” Immunological Reviews, 276(1), pp. 97–111. https://doi.org/10.1111/imr.12520
- DICKEY, David A. a Wayne A. FULLER. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. Journal of the American Statistical Association [online]. 1979, doi: 10.1080/01621459.1979.10482531.
- EHRHARDT, Michael C. a Eugene F. BRIGHAM. Corporate Finance: A Focused Approach. 5th Edition. Cengage Learning, 2013. ISBN 978-1133947530.
- ENDERS, Walter. Applied Econometric Time Series (Wiley Series in Probability and Statistics). 4th Edition. New York City: Wiley; 4th edition, November 3, 2014. ISBN 978-1118808566.
- Fama, Eugene F. and Kenneth R French. The Capital Asset Pricing Model: Theory and Evidence. Journal of Economic Perspectives [online]. 2004, 18(3), 25-46 [cit. 2023-03-26]. ISSN 0895-3309. Dostupné z: doi:10.1257/0895330042162430
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 1-34). AAAI/MIT Press.
- Eisenmann, T., Parker, G. and Van Alstyne, M. (2011) “Platform Envelopment,” Strategic Management Journal, 32(12), pp. 1270–1285. doi:10.1002/smj.935.

- Fama, E. F. and French, K. R. (2004) The Capital Asset Pricing Model: Theory and Evidence. Journal of Economic Perspectives, 18(3), 25-46. doi:10.1257/0895330042162430
- Fayyad, U. M. and P. Smyth. Advances in knowledge discovery and data mining. California: MIT Press, 1996. ISBN 978-0262560979.
- Gabrys, R. L. et al. (2018) "Cognitive control and flexibility in the context of stress and depressive symptoms: The Cognitive Control and flexibility questionnaire," Frontiers in Psychology, 9. doi:10.3389/fpsyg.2018.02219.
- Grossman, S. J. and Stiglitz, J. E. (1980) On the Impossibility of Informationally Efficient Markets. American Economic Review, 70, 393-408.
- KELLEHER, Christa, Thorsten WAGENER a Brian MCGLYNN. Model-based analysis of the influence of catchment properties on hydrologic partitioning across five mountain headwater subcatchments. Water Resources Research [online]. 2015, 51(6), 4109-4136 [cit. 2023-03-29]. ISSN 0043-1397. Dostupné z: doi:10.1002/2014WR016147
- Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2011). The flash crash: High-frequency trading in an electronic market. The Journal of Portfolio Management, 37(2), 118-128.
- Kotler, P., Keller, K. L., Brady, M., Goodman, M., and Hansen, T. (2017). Marketing management. Pearson. ISBN: 978-0133856460
- Kotsiantis, S.B., Zaharakis, I.D., & Pintelas, P.E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190. https://doi.org/10.1007/s10462-007-9052-3
- Lamont, O., Polk, C., & Saa-Requejo, J. (2001). Financial Constraints and Stock Returns. The Review of Financial Studies, 14, 529-554. https://doi.org/10.1093/rfs/14.2.529
- Li, J. et al. (2014). "An integrated catalog of reference genes in the human gut microbiome," Nature Biotechnology, 32(8), pp. 834–841. https://doi.org/10.1038/nbt.2942
- Lindström, M. (2017). Commentary on Wang et al. (2017): Differing patterns of short-term transitions of nondaily smokers for different indicators of socioeconomic status (SES). Addiction, 112(5), 873-874. https://doi.org/10.1111/add.13758

- LIU, Bing. Sentiment Analysis and Opinion Mining [online]. Cham: Springer International Publishing, 2012 [cit. 2023-03-29]. Synthesis Lectures on Human Language Technologies. ISBN 978-3-031-01017-0. Dostupné z: doi:10.1007/978-3-031-02145-9
- Malkiel, B. G. (2015). A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing (12th ed.). W. W. Norton & Company. ISBN-13: 978-0393246117.
- MCGURK, Zachary, Adam NOWAK a Joshua C. HALL. Stock returns and investor sentiment: textual analysis and social media. Journal of Economics and Finance [online]. 2020, 44(3), 458-485 [cit. 2023-03-27]. ISSN 1055-0925. Dostupné z: doi:10.1007/s12197-019-09494-4
- Moro, S., Rita, P. and Vala, B. (2016) “Predicting Social Media Performance Metrics and evaluation of the impact on brand building: A Data Mining Approach,” Journal of Business Research, 69(9), pp. 3341–3351. Available at: https://doi.org/10.1016/j.jbusres.2016.02.010.
- PANG, Bo a Lillian LEE. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval [online]. 2008, 2(1–2), 1-135 [cit. 2023-03-29]. ISSN 1554-0669. Dostupné z: doi:10.1561/1500000011
- PANKRATZ, Alan, ed. Forecasting with Univariate Box-Jenkins Models [online]. Hoboken, NJ, USA: John Wiley, 1983 [cit. 2023-03-29]. Wiley Series in Probability and Statistics. ISBN 9780470316566. Dostupné z: doi:10.1002/9780470316566
- Peters, Elke and Stuart Webb. INCIDENTAL VOCABULARY ACQUISITION THROUGH VIEWING L2 TELEVISION AND FACTORS THAT AFFECT LEARNING. Studies in Second Language Acquisition [online]. 2018, 40(3), 551-577 [cit. 2023-03-29]. ISSN 0272-2631. Dostupné z: doi:10.1017/S0272263117000407
- Peters, B. Guy. Governance: ten thoughts about five propositions. International Social Science Journal [online]. 2019, 68(227-228), 5-14 [cit. 2023-03-26]. ISSN 0020-8701. Dostupné z: doi:10.1111/issj.12181

- Regal, R.R. and Larntz, K. (1978) “Likelihood methods for testing group problem solving models with censored data,” Psychometrika, 43(3), pp. 353–366. Available at: https://doi.org/10.1007/bf02293645.
- SAID, Said E. a David A. DICKEY. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. Biometrika [online]. 1984, 71(3) [cit. 2023-03-29]. ISSN 00063444. Dostupné z: doi:10.2307/2336570

7.1 Internet sources

- Apple. (2021). Apple Reports Fourth Quarter Results. Apple. Retrieved March 26, 2023, from https://www.apple.com/newsroom/2021/10/apple-reports-fourth-quarter-results/
- NASDAQ. (n.d.). Apple Inc. (AAPL) historical data. Retrieved March 26, 2023, from https://www.nasdaq.com/market-activity/stocks/aapl/historical
- The Guardian. (n.d.). News, sport and opinion from the Guardian's global edition. Guardian News and Media. Retrieved March 29, 2023, from https://www.theguardian.com/international
- Macrotrends. (2010). The Long Term Perspective on Markets. Retrieved March 26, 2023, from https://www.macrotrends.net
- Pew Research Center. (2019). Social media usage in the U.S. in 2019. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/
- Pew Research Center. (2016). News Use Across Social Media Platforms 2016. Pew Research Center. Retrieved from https://www.pewresearch.org/journalism/2016/05/26/news-use-across-social-media-platforms-2016/
- Securities and Exchange Commission. (2010). Testimony concerning the severe market disruption on May 6, 2010. Retrieved March 29, 2023, from https://www.sec.gov/news/testimony/2010/ts051110mls.htm
- S&P Dow Jones Indices. (n.d.). Document Moved. S&P Global. Retrieved March 27, 2023, from https://www.spglobal.com/spdji/en/
- Sauter. (2022). Welcome to Sauter. Retrieved March 29, 2023, from https://www.sauter-controls.com/en/
- Gottfried, J. (2016). News Use Across Social Media Platforms 2016. Pew Research Center. Retrieved from https://www.pewresearch.org/journalism/2016/05/26/news-use-across-social-media-platforms-2016/
- House Judiciary Committee Republicans. (2020). Available at: https://judiciary.house.gov/(Accessed: March 30, 2023).
- Gartner_Inc. (2021). Delivering actionable, objective insight to executives and their teams, Gartner. https://www.gartner.com/en (Accessed: Jan 29, 2023).

9 List of abbreviations

ADF- Augmented Dickey–Fuller

API - Application Programming Interface

ARIMA- Autoregressive Integrated Moving Average

AWS - Amazon Web Services

CEO - Chief Executive Officer

DJIA - Dow Jones Industrial Average

ETF - Exchange-Traded Fund

GDP - Gross Domestic Product

IPO - Initial Public Offering

NLTK - Natural Language Toolkit

NLP - Natural Language Processing

NaN - Not a Number

NYSE - New York Stock Exchange

OLS - Ordinary Least Squares

S&P500 - Standard and Poor's 500

SARIMAX - Seasonal Auto-Regressive Integrated Moving Average

SEC - Securities and Exchange Commission

USA - United States of America

VADER - Valence Aware Dictionary and sEntiment Reasoner

NASDAQ - National Association of Securities Dealers Automated Quotations

Appendix

List of Supplements…