Corvinus University of Budapest
Applied Economics

Sentiment and volatility analysis of the stock prices of Intel and AMD

Bachelor's thesis

Written by: Marco Alan Hafid

Applied Economics (BSc)

2022, Budapest

Thesis advisor: Zoltán Madari

Table of Content

Declaration of authorship

“I hereby declare

- that I have written this writing sample (thesis, seminar paper or other) without any help from others and without the use of documents and aids other than those stated in the references,

- that I have mentioned all the sources used and that I have cited them correctly according to established academic citation rules,

- that the topic or parts of it are not already the object of any work or examination of another course unless this is explicitly stated,

- that I am aware that my work can be electronically checked for plagiarism and that I hereby grant the University of St.Gallen copyright as far as this is required for this administrative action.”

Date and Signature: 2022.04.23

Acknowledgement

I would like to thank Zoltán Madari, my thesis advisor and statistics/econometrics teacher for all his valuable advice during my thesis and for the amazing four semesters of classes.

I would also like to thank my friends and family members for proofreading my thesis several times. This could not have been done without them.

Abstract

The aim of this paper is to find the answer to whether the sentiment towards Intel and AMD has any effect on their respective stock prices. The sentiment values were obtained by sentiment analysing the comments on Intel's and AMD's subreddits, which were then were placed in a VECM model along with the volatility of Intel's and AMD's stocks. The volatility was gained from the T-GARCH models after the ARIMA models seemed insufficient to handle the volatility clustering. The final results show us that there is no connection between the sentiment towards Intel and AMD and their stock prices.

Keywords: Sentiment analysis, ARIMA, GARCH, VECM, Web-scraping, NLP, stock market, volatility

1. Introduction

The purpose of this paper is to compare the performance and volatility of Intel Corporation (Intel) and Advanced Micro Devices, Inc.'s (AMD) stocks, and whether the sentiment towards the companies has any effect on the stock prices. The two companies share a long history of rivalry, ever since their foundation. Intel was established in 1968 and AMD in 1969. Since then, both played a key role in the semiconductor industry.

This paper will discuss the semiconductor industry, its relation to other industries, and how supply chain issues, and supply shortages affect the industry.

After the introductions of the semiconductor industry, I will go through the rivalry between Intel and AMD and how this rivalry shaped the microprocessor market. Afterwards I will then give a short summary of the two companies.

Subsequently I will go through some of the previous literature regarding stock market predictions by the usage of sentiment analysis on different social media platforms. The different literature also uses a wide range of statistical and financial modelling for stock price forecasts.

Following the literature review, I will establish the theoretical framework of my research; I will introduce several econometric models that are used to predict and to forecast time series data. Such models include Autoregressive Integrated Moving Average Process (ARIMA), the Autoregressive Conditional Heteroskedasticity (ARCH-GARCH) model family, and the Vector Autoregression (VAR) and its version used for cointegrated time series, the Vector Error Correction Model (VECM). I will also examine some stylized facts regarding financial markets and mention the effective markets hypothesis.

After establishing the theoretical framework, I will present my datasets, which are the stock prices of Intel and AMD and the sentiment values of the respective companies. The sentiment values are gained by web-scraping the subreddits of Intel and AMD and sentiment analysing the comments I obtained.

After going through the datasets, I will present how my models are built and analyse the results, after which I will talk about some of the improvements and outlook that could improve the model and the methodology in the future.

My initial hypothesis is that the main users of consumer processors are also likely present on different social platforms, especially a subreddit created for the company. Some of these consumers and Reddit users can presumably be investors themselves, so while the consumer processors do not make up majority of the revenue for either company, their performance does affect the sentiment towards the company and possibly towards the stock price as well.

My aim is to answer the question whether the said sentiment really affects the stock performance and volatility of Intel and AMD, and if it does, to which extent does this effect occur.

2. Semiconductor Industry

The semiconductor industry is crucial for technological advancements; therefore, it plays a major role in the growth of the economy. The semiconductor market size is 410 billion dollars, and it is estimated to reach close to 600 billion by 2024. Numerous industries depend on semiconductors, such as consumer electronics -smartphones, computers, laptops-, automotive, healthcare, infrastructure, government and servers, data centers etc. (*Semiconductor Industry Worldwide by Application*, n.d.)

The creation of semiconductors is a very complex and expensive process that can involve several hundred steps. Each step can have a small bit of product loss, therefore, several hundred millions of dollars are spent on research in order to increase efficiency and minimise this loss. This brings us to the first larger segment in the manufacturing process, which is the creation of semiconductor manufacturing equipment. This segment is often overlooked; however, it is crucial in automating the production lines and is used by all other segments of the process. The five largest corporations in this segment are**:** ASML Holdings, Applied Materials, Lam Research, Tokyo Electrons and KLA Corp. (*Semiconductor Market Share by Company 2020*, n.d.)

The second major segment in the creation of semiconductors is design. This part of the process involves designing the semiconductors in software to be able to perform different functions, depending on where it will be used -processors, graphic cards-. There are some major companies that design semiconductors**,** such as AMD, NVIDEA, Broadcom Inc., MediaTek and Qualcomm. They produce the architecture and give an order to semiconductor foundry

companies, such as TSMC, where the chips are manufactured. These companies that do not manufacture their own semiconductors are also called fabless companies. (*Revenue of Top 10 IC Design (Fabless) Companies for 2020 Undergoes 26.4% Increase YoY Due to High Demand for Notebooks and Networking Products, Says TrendForce*, n.d.)

The last two segments are very closely connected to each other. Fabrication and assembly/packaging/testing. The fabrication process usually manufactures the semiconductors or 'chips' based on the design, using the previously mentioned semiconductor manufacturing equipment. The manufacturing results in wafers that contain several hundred individual integrated circuits, which must be cut apart and then packaged. A semiconductor package is usually made of a metal, plastic, glass, or ceramic casing containing one or more discrete semiconductor devices or integrated circuits. Individual components are fabricated on semiconductor wafers (commonly called silicon) before being diced into dye, tested, and packaged. The package provides means for connecting to the external environment, such as printed circuit board, via leads such as lands, balls, or pins**.** Also**,** protection against threats such as mechanical impact, chemical contamination, and light exposure. In addition, it helps dissipate heat produced by the device, with or without the aid of a heat spreader. (*Types and Basic Functions of Semiconductor Packaging*, n.d.) The largest semiconductor foundries are TSMC, Samsung, GlobalFoundries, UMC and SMIC. (*Top Semiconductor Foundries Quarterly Revenue 2021*, n.d.) There is a company that stands out, however Intel, while being a major player in the foundry business, they do not take outsourcing requests from companies like AMD or NVIDEA, they use their foundries for creating chips designed in-house.

The recent global supply shortage of semiconductors that has affected several other industries that rely on integrated circuits has splurged the major companies to expand their foundries. In a recent roadmap, Intel's new CEO Pat Gelsinger announced a yearly investment of 25-28 billion dollars of capital expenditure, most of which will be developing new foundries and a 15 billion dollars a year for research and development. (*Latest News*, n.d.) TSMC also announced a new foundry in Arizona that will manufacture 5nm chips, and a 3nm foundry is also being considered. The 5nm foundry will cost 10-12 billion dollars, while the 3 nm would be approximately 25 billion dollars. Samsung also announced a 17 billion dollars foundry in Texas, and as previously mentioned Intel will also build two new factories in Arizona near its existing Chandler facility. The three companies are also competing for a total of 50 billion dollars subsidy announced by Joe Biden due to the recent shortages highlighting a vulnerability

of domestic US manufacturing capabilities. The major American chip firms, such as Nvidia and Qualcomm rely heavily on Asian manufacturing. (*Newsroom Home*, n.d.)

3. History of The Microprocessor Industry

In this chapter I will present the history of the microprocessor market with the help of André Semler's (2010) paper and go through the key events that shaped not only the market structure but also the two competing companies as well. Afterwards I will talk about the history of Intel and AMD.

A duopoly is described as a market, in which two companies hold the majority of market share, however in our case there are companies with unequal market shares. Since there are no other major microprocessor producers, the microprocessor market can be defined as an oligopoly with an incumbent and a fringe firm where about 98.6% of the total market share is held by the two companies.

According to the author, the companies use many means to gain market share; the main one is technological innovation. The main form of technological innovation in the microprocessor sector comes from quality (speed), for which we need to understand Moore's law.

In the early stages of the processor industry there were a large number of competitors. In 1976 the competition was eliminated by Intel and AMD signing a cross-licensing agreement. The termination of the agreement in 1987 marked the beginning of heavy competition, as a result computer prices were driven under 1000$ in 1997. Also, Cyrix, another competitor exists in the market. In 1998 Intel and AMD began competing in every segment of the market, creating close substitutes. In 1999 AMD stops producing Intel-compatible microprocessors, which means an Intel based computer can not use AMD components and an AMD based computer can not use Intel components. Although this makes their products more differentiable, it also forms them to be less substitutable. As a result, consumer preferences are developed and a sort of "brand naming effect" is produced. Intel starts to use its vertical integration to achieve dominance in the market. (André Semmler, 2010)

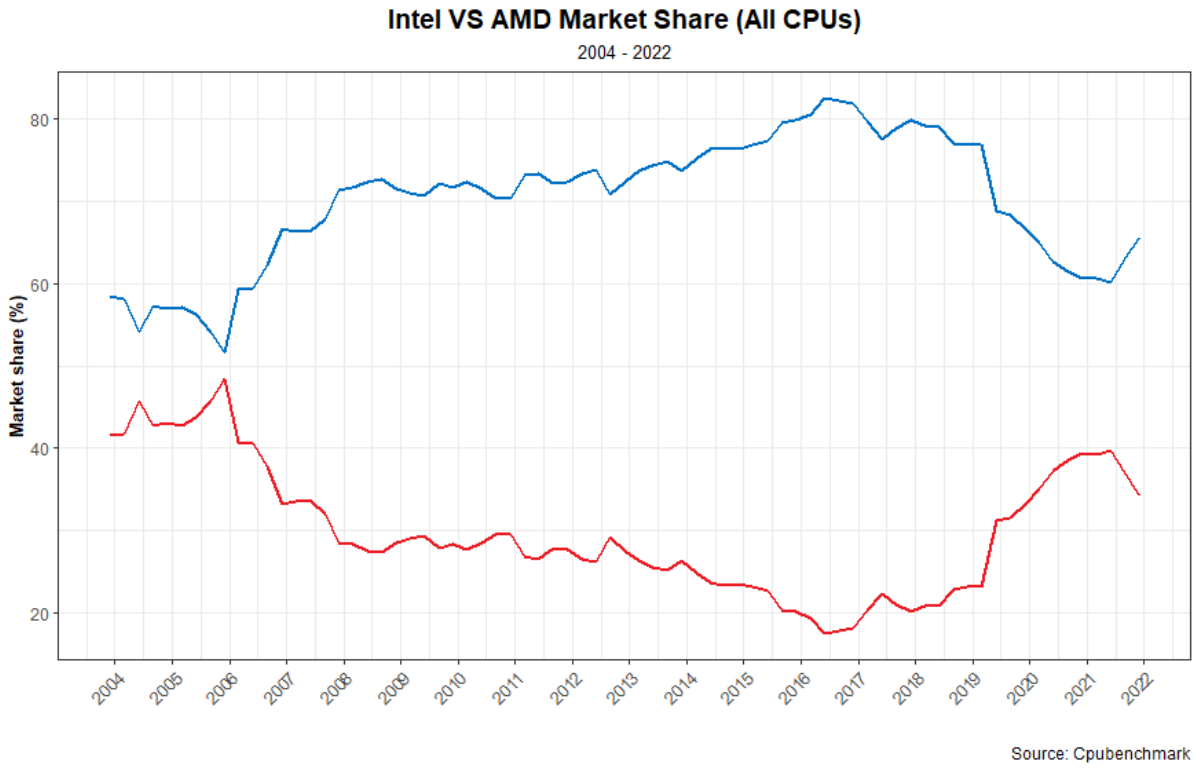

Figure 1: The market share of Intel and AMD between 2004 and 2022. (All CPUs)

This study was conducted in 2011, after which Intel continued to dominate the market until 2017-2018 when AMD started eating into its market shares, especially in the desktop market. While overall Intel still has the upper hand, AMD is slowly gaining market share in the laptop and server processor markets as well. (*PassMark CPU Benchmarks - AMD vs Intel Market Share*, n.d.)

3.1 Intel

Intel is an American multinational corporation founded in 1968 by Robert Noyce and Gordon Moore, both of whom had a name in the semiconductor industry. Prior to founding Intel, they also created Fairchild semiconductors in 1957. On top of this, Noyce co-invented integrated circuits and Moore articulated Moore's Law, which became an important principle in technology. (*Intel's Founding*, n.d.)

Intel had its Initial Public Offering (IPO) in 1971, three years after its foundation. Intel showed a steady growth and strong market positions throughout the years, with several stock splits and

stock buybacks behind their backs. They also started paying out dividends in 1992, which they have not interrupted since then. As of 2022.04.03 Intel has a market valuation of 195 billion dollars and a yearly revenue of 79 billion dollars in 2021. (*Annual Reports*, n.d.)

3.2 AMD

AMD is an American multinational corporation, formed in 1969 as a Silicon Valley start-up. From a dozen of employees AMD grew to be one of the biggest semiconductor and chip companies in the world with a wide range of technological products, such as CPUs, GPUs, FGPAs, Adaptive SoCs and deep software expertise. Like Intel, AMD also had its IPO few years after its foundation, in 1972. Unlike Intel, AMD never paid out dividends to its shareholders, even though they also had several stock splits and stock buybacks. As of 2022.03.04 AMD has a market valuation of 176 billion US dollars and had a yearly revenue of 16.4 billion dollars in 2021. (*About AMD,* n.d.)

4. Literature Review

Sentiment analysis is becoming more and more popular when it comes to stock market forecasting. There are many different papers with vastly differing methodologies. In this chapter I will outlie the basics of sentiment analysis, that is Natural Language Processing (NLP) and present a few research papers that sued sentiment analysis to predict stock movements. I tried to present papers with different methodologies.

Sentiment analysis uses NLP, text analysis and machine learning techniques to extract and classify sentiment from reviews and social media comments. (Doaa Mohey El-Din Mohamed Hussein, 2018)

Sentiments can take up the value of good and bad, or it can be on different scales (e.g: very bad, bad, good, very good). (Rudy Prabowo and Mike Thelwall, 2009)

In our age of growing online activity, it is very useful to be able to extract and analyse the large volume of opinions (sentiment analysis is also called opinion mining) from comments and

reviews. A study by Michael Lubitz discovered that using sentiment analysis on the economics subreddit, combined with analysing financial news could predict the future directions of the stock market with 56.49% accuracy, which was higher than analysing financial news alone. This proves that this method could add value to researching stock price movements and predictions. (Michael Lubitz, 2017).

In the paper of Thien Hai Nguyen, Kiyoaki Shirai and Julien Velcin (2015) they mention how on top of historical price, the overall social mood influences stock market movements, and that the overall social mood towards a company could play an important factor in its valuation. In our modern age a large amount of mood data is available, which could be used to predict stock movement with the help of historical data. The aim of the study was to develop a model that can forecast movements of the stock price using data from social media, with a model that predicts the value of the stock at t using information at t-1 and t-2 where t stands for transaction date. With this technique the model is trained by supervised machine learning. The authors acknowledged that many other factors, such as microeconomic and macroeconomic indicators play a key role on stock price movements, however they will only focus on the usage of mood data to forecast stock movements. The mood data will be gained by using sentiment analysis on social media and the sentiments will be integrated into the model that predicts stock price movements. There are shortcomings and hardships of this method however, including short texts, spelling and grammatical errors in addition to conflicting results of different research papers, for instance (Antweiler & Frank, 2004; Tumarkin & Whitelaw, 2001) reporting that social media sentiments have no predicting capabilities, while others, like (Bollen, Mao, & Zeng, 2011) reporting either weak or strong predicting capabilities. The authors of the paper investigated topic sentiment, unlike previous research, this represents sentiment towards a specific topic regarding a company, such as products, dividends etc. The data is gained by using an already existing topic model, called the joint sentiment/topic (JST) model and by the authors proposed sentiment method.

The authors of the paper used two datasets for their research, a historical price, and a mood information dataset. Both datasets use data of 18 stocks, with the help of Yahoo Finance historical price chart and message board for the mood dataset. The message board contains two types of messages, an original comment, and messages that are replies to those comments. The two types of messages are treated the same way. Each message contains a sentiment tag at the end with possible values such as: Strong Buy, Buy, Hold, Sell and Strong Sell. While previous work used sentiment extracted from twitter, the authors decided against it for several reasons;

information on Twitter is messier than on Message Board and collecting relevant data could be difficult with the different usage of hashtags. The other reason is a technical one; collecting Twitter data can be done in two ways: either gathering live data, which would mean collecting a year worthy of data would take a year, or through Twitter API which only goes back a week.

Although the research only achieved an average accuracy of 54.41%, the used method could attain more than 60% on a few stocks and provides a much higher accuracy for difficult stocks, that could only be predicted with past prices. Some limits of the paper included that the number of topics was pre-determined, and the model can only predict, if a stock moves either up or down. Some may want to forecast drastic movements for which the model could be extended by increasing the classes to "great up", "little up", "little down", "great down". Furthermore, the authors considered expanding upon their model by using additional indicators, such as covariance of stocks, macro/microeconomics indicators etc. (Thien Hai Nguyen, Kiyoaki Shirai and Julien Velcin, 2015)

Shangkun Deng et. al. (2011) applied a combined method of technical analysis and sentiment analysis for stock price prediction in their research. Technical analysis is based on technical indicators to identify trends and patterns to help traders foresee the direction of price changes and timing of transactions. A technical indicator for stock price returns a value for a given stock during a given period. These values could provide traders with an idea of whether the trend will continue, for example MACD (moving average convergence divergence) or BIAS which shows if a certain stock is oversold or overbought.

In the last decade we experienced the rise of machine learning methods applied in predicting movements in stock and FX markets. However**,** the human factor is of significant importance in the movement of stock prices. In the age of Internet**,** when vast amount of information, news and opinions are being posted online. Analysing the sentiment from said comments and news and then using the result to mine information on the correlation of stock price movements and sentiment from the Internet could prove to be beneficial to understand the relation between stock prices and the human sentiment factor.

The authors used a Multiple Kernel Learning (MLK) method to learn and predict the stock prices of Sharp, Panasonic and Sony. On top of using features from time series data, such as stock price and volume, the authors used numerical dynamics and sentiment analysis results, extracted from social networks. During the experiments a Shift-Periods method was applied

for training and forecasting to adapt for the changing dynamics of the stock market. The authors assumed, that the dynamics of the stock market change slowly, therefore, could be considered the same during the given period; however**,** some other research and evidence suggests otherwise, in fact money markets can move rather abruptly.

The technical indicators used are the previously mentioned MACD and BIAS and rate of change (ROC), which shows the difference between today's closing price/volume and the closing price/volume N days ago, which was set to N = 1 in the research. The time series data is downloaded from Google Finance Website and the news and comment data is gathered from Engadget, a special community focused on for Internet and information technology, which explains the choice of analysed stocks all of which are Japanese technological companies.

The conclusion of the research revealed that there was a correlation between stock prices and sentiment analysing news and comments. The coefficients of sentiment analysis in the MLK model were higher than the numerical dynamics, indicating that sentiment analysing text and news are better. The authors acknowledge that analysing different social networks, for instance Yahoo Finance could yield very different results and MLK coefficients. In the research an overall sentiment was calculated, and the order and context of the words was not evaluated further. (Shangkun Deng et. al.,2011)

In the paper of Saloni Mohan et. al. (2019) they mention that an automated system for analysing financial news and using the result to predict stock prices could be beneficial for investors. The automated system would gather financial news, related to the company, and combined with historical prices would predict stock price, with the help of a machine learning model. Compared to previous works that used Twitter, financial blog, or news data the authors decided to use financial articles from well-known sources to avoid false information. The analysis uses past stock prices and current days' financial news to predict current days' closing prices. Financial news has a significant effect on stock prices, so a predictive model applying this combined method could achieve better results than only considering past prices.

The research collected the closing stock prices and news articles of Standard and Poor's 500 (S&P500) companies between 2013 February and 2017 March. The articles were scraped from international daily newspaper websites and amounted for 265,463 articles. Due to the size of the data, the raw data was fed into a data pipeline which sent it to a machine learning model.

The data was evaluated with Mean Absolute Percentage Error (MAPE), which is a popular measure of forecasting accuracy. MAPE can be beneficial when only the difference between the predicted and actual values matter and the direction of the difference can be ignored. MAPE displays robustness with long tail datasets and overcomes the large deviation bias present in the Root Mean Square Error (RMSE).

The authors experimented with a few different models: ARIMA, Facebook Prophet and RNN LSTM. All of which are time series models used for forecasting. The RNN model with prices and text polarity input performed best across all the experiments, while the RNN with prices and text as input performed the best for stable stocks. The conclusion of the research suggested a strong relationship between stock prices and financial articles, although the models did not perform well for stocks with low prices and high volatility. The authors suggest using different models in the future, such as building domain-specific models, where companies are grouped according to their sectors. (Saloni Mohan et. al., 2019)

Several other researchers used similar approaches in predicting stock prices with Machine Learning and statistical methods, such as Robert P. Schumaker and Hsinchun Chen's research *Textual analysis of stock market prediction using breaking financial news: The AZFin text system* (Schumaker & Chen, 2009), Xue Zhang et.al.'s *Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"* (Zhang et al., 2011) and Robert Thumarkin and Robert F. Whitelaw's *News or Noise? Internet Postings and Stock Prices* (Tumarkin & Whitelaw, 2001)

5. Methodology

In this chapter I focus on some of the most important terminology regarding time series analysis and go through the methodology of the models I used during my research; autoregressive-moving-average (ARIMA), autoregressive conditional heteroskedasticity (ARCH) and vector autoregression (VAR). It is also important to talk about some of the properties of financial time series and theories of financial markets that will help explain the usage of the selected models.

5.1 ARIMA Model

In this sub-chapter I will explain the basics of time series data, talk about stationarity and introduce the AR (p), MA (q) and ARIMA (p,d,q) processes. The terms and notations presented in this part of the chapter are based on the book called *Introduction to Modern Time Series Analysis* by Gebhard Kirchgässner and Jürgen Wolters.

The difference between cross-sectional and time series data is that time series have an order. Time series is a collection of random variables that are indexed according to the order in which they are obtained in time. The variables are a collection of random samples at every observed time period, where the elements of the sample are probability variables realised at the given time. This means that unlike with cross-sectional data, we can not repeat the realisation from repeated sampling, therefore, we will not be able to draw a conclusion from taking one sample, since much information will be omitted from the data. Since it is not possible to statistically reproduce these observations, we are forced to examine the one sample we have. It is necessary to present several strong theoretical assumptions. Next we will go through these assumptions.

A full description of a time series is given by the joint distribution of all its $t \in N$ components - which we do not know- therefore, the usage stationarity is necessary. There is strict and weak stationarity. The requirements of strict stationarity are the following:

$$F(x_1, x_2, \dots, x_n) = F(x_{1+k}, x_{2+k}, \dots, x_{n+k}) \quad \forall\, n, k$$

This is more of a theoretical approach because its requirements are not fulfilled in economical time series.

With weak stationarity, also called covariance stationarity**,** we dissolve the assumption that the distributions must be identical. The following three criteria must be met for weak stationarity to be assumed; the expected value and variance of the time series must not change over time and the autocovariance between the distributions must only depend on the elapsed time between the two observations.

$$E(x_t) = \mu \qquad \forall t$$

$$Var(x_t) = \sigma^2 \qquad \forall t$$

$$cov(x_t, x_s) = \gamma(|t - s|) \qquad \forall t, s$$

From this point onward, when we talk about stationarity, we will refer to the fulfilment of these three assumptions. We will need to verify the autocovariance criterion.

We define autocovariance in the following way:

$$\mathrm{cov}(x_t\,,\,x_s) = \mathrm{E}\big((x_t - \mu)(x_s - \mu)\big),$$

and we estimate it consistently by:

$$\gamma(k) = \frac{1}{T}\sum_{t=k+1}^{T}(x_t - \mu)(x_t - k - \mu)$$

Autocorrelation is also an important term that we need to discuss. Autocorrelation means that the dependant variable has an effect on its own delayed (lagged) values. Autocorrelation is defined the following way:

$$\rho(x_t\,,x_s) = \frac{\mathrm{cov}(x_t\,,x_s)}{\sqrt{\mathrm{var}(x_t)\mathrm{var}(\,x_s)}} = \rho(k),$$

and can be estimated in the following way:

$$\rho(k) = \frac{\gamma(k)}{\sigma_x^2}$$

A stochastic u_t process is white noise (WN) if the following conditions are met:

$$\mathrm{E}(u_t) = 0 \qquad \forall t$$

$$\mathrm{Var}(u_t) = \sigma^2 \qquad \forall t$$

$$\mathrm{cov}(u_t\,,\,u_s) = \mathrm{E}(u_t\,,\,u_s) = 0 \qquad \forall t \neq s$$

We can only forecast with time series where the residuals are white noise. If the time series is stationary, then it is sufficient to examine the autocorrelation of the residuals. Our null hypothesis in this case is that the autocorrelation equals to zero from the first lag, which will mean that the autocovariance is zero as well, therefore, the process is white noise. If the residuals are not white noise then we will have to add ARIMA (p,d,q) terms to the time series.

The goal of Autoregressive Processes (AR) is to make the dependent variable depend on its lagged values. A p-th order AR process can be written as:

$$x_t = \delta + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \cdots + \gamma_p x_{t-p} + u_t$$

where u_t is the residual of the process in the t-th period.

Because we assume stationarity, we use the following conditions:

$$t \to \infty \quad \text{and}$$

$$|\gamma_i| < 1 \quad \forall i.$$

If these conditions are met, then the autocovariance function (ACF) will go down to zero in the long run.

The partial autocovariance function (PACF) shows us the value of autocovariance by filtering out the values between two time periods. In case of an AR(p) process:

$$PACF(k) = 0 \quad \forall k > p,$$

therefore, we can identify the order of the AR (p) process with the help of partial autocorrelation.

The goal of the Moving Average Process (MA) is to make the dependent variable depend on the residuals of previous periods. A q-th order MA process can be written as:

$$x_t = \delta + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \cdots + \beta_q u_{t-q} + u_t .$$

The autocovariance function for an MA(q) process is:

$$ACF(k) = 0 \quad \forall k > q,$$

So, the order of the MA process is determined by the ACF, while the partial autocovariance function will go down to zero.

If a model contains a dependent variable that depends on both its own lagged values and its residuals from the previous periods, then we get an Autoregressive-Moving Average Process (ARIMA). An ARIMA (p,d,q) process can be written as:

$$x_t = \delta + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \cdots + \gamma_p x_{t-p} + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \cdots + \beta_q u_{t-q} + u_t .$$

Due to the AR term, we can assume stationarity for an ARIMA process, in which case both the ACF and PACF will go down to zero.

We must also mention the lag operator, which, in the case of:

$$X_t = \delta + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + u_t$$

can be written as:

$$Lx_t = x_{t-1}(1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p)x_t = \delta + u_t$$

The condition of stationarity is that the roots of the lag polynomials are outside the unit circle. (Kirchgässner & Wolters, 2007)

5.2 Empirical Properties of Financial Markets

Before we move on towards the ARCH model, we have to explore some of the stylized facts about financial markets. These properties are important to understand, and they explain why ARIMA models are not always sufficient and why ARCH models are required. For the explanation of these properties, I will use the research paper *Empirical properties of asset returns: stylized facts and statistical issues* written by Rama Cont in 2001.

More than half a century of data in financial time series show us that a very wide range of asset classes -such as corn futures, IBM shares and Dollar/Yen exchange rate- demonstrate very similar properties if examined from a statistical point of view. These qualities, which are common across a wide range of instruments, markets, and historical periods, are referred to as stylized empirical facts. However, it is important to note that these properties are qualitative. Nevertheless, these stylized facts are very constraining and quite hard to reproduce, so while they do not apply to every single financial instrument, it is safe to conclude that the majority of them can be described by these properties.

The author mentions 11 stylized facts that are common to a wide range of financial assets:

(1) Absence of autocorrelation: Autocorrelation of asset returns is linearly insignificant, except for extremely brief intraday time periods, (~20 minutes) where microstructural effects can cause significant negative autocorrelation.
(2) Heavy tails (distributions): The distribution of the returns of financial instruments is wider on the tails and sharper compared to a normal distribution. This means that extreme events are more likely to happen and due to the sharper form less observations are around the expected value.
(3) Gain/loss asymmetry: Larger and more frequent downward movements can be observed in stock prices and index values, but equally large upward movements are

not as common. This property does not apply to currency exchange rates where a higher symmetry can be observed)

(4) Aggregational Gaussianity: The shape of the distribution changes if we change the time scales. The bigger the time scale we look at the closer the distribution of the return is to that of a normal distribution.

(5) Intermittency: The returns display a high degree of variability regardless of the time scale. This is explained by the presence of a wide range of volatility estimators appearing in irregular bursts.

(6) Volatility clustering: There is a positive autocorrelation of volatility over the period of several days. High volatility periods are followed by high volatility and low volatility periods are followed by lower volatility. This causes the cluster in volatility.

(7) Conditional heavy tails: The heavy tail property of residual time series does not disappear even after correcting the volatility clustering with the usage of GARCH-type models. The tails are, however, less heavy compared to the unconditional distribution of returns.

(8) Slow decay of autocorrelation in absolute returns: The autocorrelation function of absolute returns slowly decays, typically as a power law with an exponent of $\beta \in [0.2, 0.4]$. This can be viewed as an indication of long-term reliance.

(9) Leverage effect: Most volatility measures of an asset are negatively correlated with the returns of the asset.

(10) Volume/volatility correlation: The trading volume of an asset is positively correlated with its volatility. (Rama Cont, 2000, p. 224)

5.3 Efficient Market Hypothesis

The efficient market hypothesis can be connected to Eugene Fama, who published one of his most famous articles *Efficient Capital Markets: A Review of Theory and Empirical Work* in the 'Journal of Finance' scientific journal.

Rama defines efficient markets, such as ones where market prices fully reflect available information at any given time, and they react to new information instantly and rationally. The

participants of the efficient markets must also act rationally by using all the available information to them when pricing assets (bonds, stocks, currencies). This means that it is not possible to gain higher returns than the risk-weighted average returns if the efficient market hypothesis prevails or as Fama stated it: “you can not beat the market”. Due to these reasons only the concept of random walk can provide a satisfactory explanation for the nature of price movements. Fama outlined three different variations that differ from each other in terms of the amount of information available on the markets and is reflected in the prices.

The weak form of the hypothesis states that the prices of all securities already have every publicly available information priced in. This implies that the use of technical analysis can not provide excess returns since past prices can not predict future performance. The possibility of fundamental analysis beating the markets is left open by this form.

The semi-strong form incorporates every assumption the weak form has and states that prices adjust fast enough so that fundamental analysis becomes futile as well.

The strong form of the hypothesis expands further on the semi-strong form, stating that every piece of information, even the ones not available to the public- such as insider information- are incorporated in the prices, which means that not even investors with insider information will be able to outperform the overall market returns. (Fama, 1970)

The efficient market hypothesis explains how the market valuation of Intel and AMD can be so similar, while Intel has roughly five times the yearly revenue of AMD. Growth forecasts, investor sentiment and outlook are all in AMD’s favour at the moment, with Intel currently experiencing lower growth forecasts for revenue and a drop in creativity when it comes to R&D.

5.4 ARCH – GARCH Models

The ARIMA model is capable to describe and forecast with a wide variety of time series. However, the ARIMA model can not deal with the previously mentioned stylized facts and properties of financial markets. Even though the ARIMA model can filter the autocorrelation of the residuals it can not deal with the clustering of volatility and several other stylized facts. The ARCH-GARCH model family is intended to fill the theoretical gaps of the ARIMA model.

The terms and notations presented in this part of the chapter are based on the book called *Introduction to Modern Time Series Analysis* by Gebhard Kirchgässner and Jürgen Wolters.

As previously mentioned, a stationary AR (1) process has an expected value and variance which does not change over time. In terms of an AR (1) process with an u_t residual, where the residual follows a normal distribution with 0 as expected value and σ^2 variance:

$$x_t = \delta + \gamma x_{t-1} + u_t$$

$$E(x_t) = \frac{\delta}{1 - \gamma}$$

$$Var(x_t) = \frac{\sigma^2}{1 - \gamma^2}$$

By introducing a set of information, it is possible to make the expected value time dependent, however this is not the case for the variance. If we look at the following set:

$$I_{t-1} = \{x_{t-1}, x_{t-2} \dots, u_{t-1}, u_{t-2}\}$$

which contains available information at every (t-1) time period.

$$E(x_t | I_{t-1}) = \delta + \gamma x_{t-1}$$

$$Var(x_t | I_{t-1}) = \sigma^2$$

Engle introduced the Autoregressive Conditional Heteroskedasticity (ARCH) models and its generalised form, the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model due to the time varying volatility in 1982. First the modelling of the return is required, which can be done with any of the previously introduced processes with the condition that the ϵ_t residual is white noise. The need to use the ARCH model family arises if the square of error terms is autocorrelated. If ϵ_t^2 can be estimated with an AR (p) model then an ARCH (p) model can help with analysing the volatility but if an ARIMA (p,d,q) model is required, then we need to use a GARCH (p,q) model. The variance of the error terms is analysed by the following models:

$$Var(\epsilon_t | I_{t-1}) = \sigma_t^2$$

ARCH (p)

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_p \epsilon_{t-p}^2$$

GARCH (p,q)

$$\sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \cdots + \alpha_p\epsilon_{t-p}^2 + \beta_1\sigma_{t-1}^2 + \cdots + \beta_q\sigma_{t-q}^2$$

The α_0 parameter reveals the value of the variance if the error term of the previous period is zero and in the case of GARCH model the volatility of the previous period is zero as well. The α_i parameters show the effects of the news from the previous periods, namely, it shows us the effect that the deviation from the estimated value has on the volatility. The β_i parameters from the GARCH models show the effects of volatility from the previous periods.

As one of the stylized facts states, the market reacts better to negative news than to positive ones. In the case of the same negative shock the volatility increases to a greater extent compared to the decrease in volatility caused by a positive news. We can use Threshold Generalized Autoregressive Conditional Heteroskedasticity (T-GARCH) if this occurs:

$$\sigma_t^2 = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \gamma\epsilon_{t-1}^2 d_{t-1} + \beta_1\sigma_{t-1}^2$$

$$d_{t-1} = \begin{cases} 1 \; if \; \epsilon_{t-1} < 0 \\ 0 \; if \; \epsilon_{t-1} \geq 0 \end{cases}$$

Investors expect higher returns if the volatility is higher, so it is not enough to simply model the volatility, but it must also be incorporated in the return forecasts, for which the GARCH in mean model is appropriate to use. Forecasting the volatility can be done by any of the previously mentioned models:

$$y_t = x_t\beta + \delta\sigma_t^2 + \epsilon_t ,$$

where the return can be explained by the x_t external variables and the volatility of the current period. (Kirchgässner & Wolters, 2007)

5.5 VAR and VECM Models

In this part of the chapter, I will talk about the properties of the VAR and VECM models and introduce the concept of impulse response functions and variance decomposition. The terms and notations presented in this part of the chapter are based on the book called *Introduction to Modern Time Series Analysis* by Gebhard Kirchgässner and Jürgen Wolters.

Vector Autoregressive Model (VAR) is used to explore the relation between more time series. It is based on the already mentioned AR(p) process. VAR models allow us to include not only

the variables' own lagged values and its lagged error term components, but also the previous values of other time series. The p lag value of the VAR(p) process specifies how many lags of the independent variables carry information about the time series.

Let $X_t = [x_t\ y_t\ z_t]$ denote the observed time series, in which case the VAR(p) models' equation and its lag operator can be written as:

$$X_t = \delta + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + u_t$$

$$I - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p$$

The model is stable if the roots of the lag operators are located outside the unit circle, in which case there exists an MA representation of the model, which allows us to analyse the impulse response function and variance decompositions.

It is worth noting that instead of the lag operator a characteristic polynomial is used in VAR models:

$$\lambda^p - \alpha_1 \lambda^{p-1} - \cdots - \alpha_p = 0\,.$$

In this case the condition of stability is that the roots are positioned within the unit circle.

The impulse response function allows us to see how the residuals of the analysed time series are affected by a shock of one standard deviation on the error terms. If the VAR model is stable the impulse response functions approach 0, meaning the shocks are not lasting.

Variance decomposition or forecast error variance decomposition indicates how much percentage change of the forecast error can be explained by an exogenous shock in the variables.

It is important to mention that if there is cointegration between the time series then we must use a Vector Error Correction Model (VECM) instead of VAR model. Cointegration refers to the event where two first-order integral processes' -which are not stationary, but their first-order differencing is stationary- linear combination is stationary.

To test cointegration between two time series we can use the Engle-Granger test, while to test more than two time series we need the Johansen test. The tests are based on eigenvalue calculations and the number of cointegrated vectors are given by iteration. (Johansen, 1991)

The VECM model is suitable to describe the short- and long-term comovements of cointegrating time series. Its basic equation is as follows:

$$y_t = C + y_0 x_t + y_1 x_{t-1} + \mu y_{t-1} + u_t$$

One of the requirements for cointegration is that the time series must be first order integrated, namely that their first order of differentials are stationary. This means that we get the connection between the time series by the differentiated form of the above equation:

$$y_t - y_{t-1} = C + y_0(x_t - x_{t-1}) - \alpha(y_{t-1} - \beta x_{t-1}) + u_t$$

Where $C + y_0(x_t - x_{t-1})$ represents the short-term while $\alpha(y_{t-1} - \beta x_{t-1})$ represents the long-term co-movements. The two equations are equivalent when $\alpha = 1 - u$ and $\beta = \frac{y_1 + y_0}{1-u}$. (Kirchgässner & Wolters, 2007)

6. Data

In the following chapter I will present the four time-series used in my research. The daily closing stock prices of Intel and AMD were downloaded from Yahoo Finance. I will present the two datasets and talk about some of the major events that had a significant impact on the price. For the sentiment values I web-scraped the respective subreddits and sentiment analysed the comments on 1000 threads, I will expand on the methodology of this at the end of the chapter.

6.1 Stock Prices of Intel and AMD

The 2. figure shows the daily closing price of Intel Corp.'s stock price between 2017.03.02 and 2022.03.04. It is noticeable that after the initial rise in price around 2017 November the stock price stayed in the 45-70 range. This is mostly due to the stagnating growth and lack of innovative period the company went through.

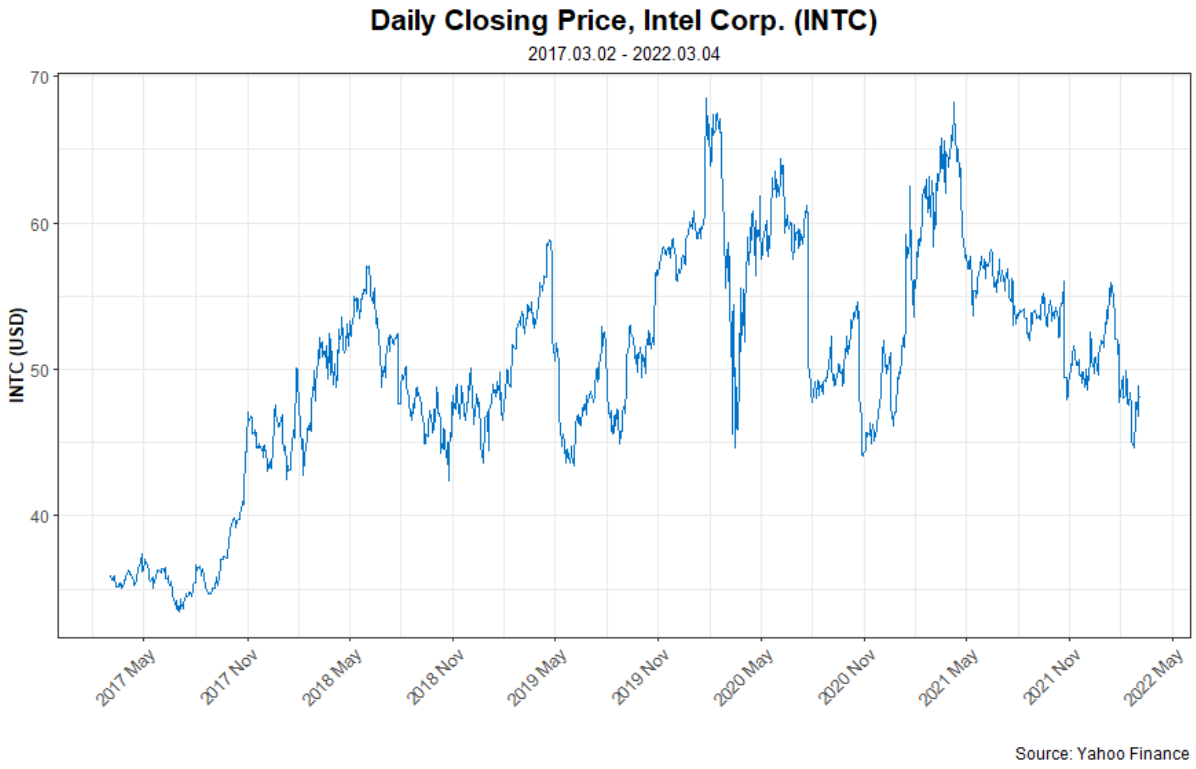

Figure 2: Daily closing price of Intel Corp. between 2017.03.02 and 2022.03.04 (own edit)

Intel experienced an excellent year in 2018, beating other semiconductor companies and having a strong growth forecast for the next year, however the macroeconomic conditions and the US-Chinese trade war had a negative impact on the outlook for the sector, hence the decline of the price in 2018 May until the end of the year. (*Annual Reports*, n.d.) Intel also experienced scandals and the appointment of a new CEO, Bob Swan during these times. The stock price climbed back up in 2019 May to the 2018 May levels, after the company reported weak revenue forecasts for 2019. This drop was the steepest the company had experienced since January 2016. (Bursztynsky, 2019) The stock price continued to repeat this cycle of rise and fall, where the reasons for decline were the COVID crash of 2020 and weak earnings forecasts in 2021. The arrival of the current CEO, Pat Gelsinger in February 2021 was received with positivity, which is reflected in the climb in price, up until 2021 May where Intel announced weaker earnings, and delayed growth from all the investments made in new factories and research.

The other stock price used in my research is Intel's only competitor in the duopolistic microprocessor market, AMD. As we can see on the 3. figure the stock prices of AMD appear

very different compared to Intel's. After stagnating for about a year it started to rapidly climb in the second half of 2018, having an exponential growth until the end of 2021 after which we can see a slight pullback.

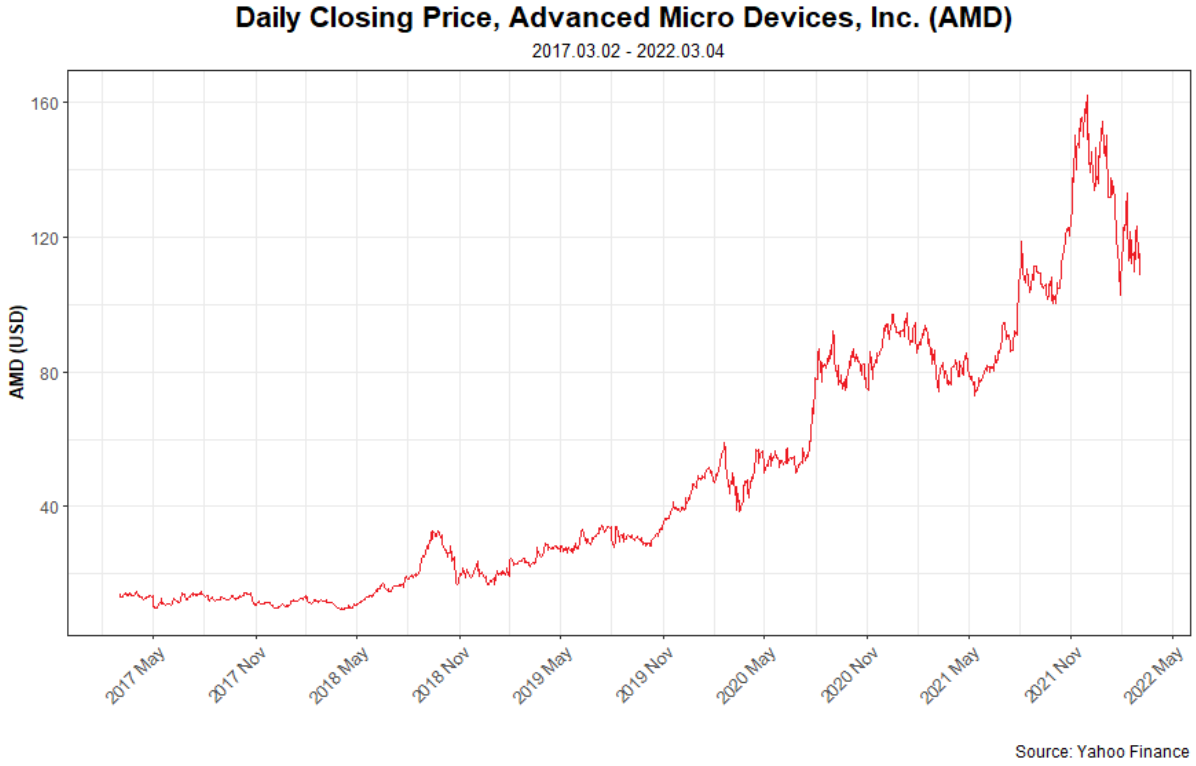

Figure 3: Daily closing price of Advanced Micro Devices, Inc. between 2017.03.02 and 2022.03.04 (own edit)

The main reason for AMD's stock not experiencing the same sideways trend compared to Intel can be accounted for the higher growth and revenue forecasts and as we saw on the 1. figure AMD started chipping away from Intel's market share both on the consumer and server microprocessor markets. It will be interesting to see how this is reflected in the sentiment towards the two companies and their products. AMD's stock was not without issues on its road to the top, it did experience a correction in 2018 October after weaker forecasts for the 4. financial quarter. The stock experienced a period of sideways movements between 2020 August and 2021 October, where COVID put strains on supply-chains that hurt the semiconductor industry. After this period the price kept climbing until the end of 2021 where Barclays downgraded the stock, which, combined with pressure from competitors and supply-chain problems kept the stock price going downward most of 2022. (*Annual Filings*, n.d.)

6.2 Sentiment Value

To get the sentiment values for the 2017.02.03-2022.03.04 period I web scraped the subreddits of Intel and AMD. Unfortunately, due to the limitations of the reddit Application Programming Interface (API) only the posts from the last 1000 pages could be web scraped. Afterwards I extracted the comments from the topics, which resulted in 89,874 comments from the Intel and 301,434 comments from the AMD subreddits. The difference is not surprising, since as of 2022.03.15 the AMD subreddit has double the subscribed users compared to Intel's subreddit (1.4 million compared to 700 thousand); however, both samples are large enough for us to work with.

The web scraped comments had to go through some pre-processing steps, which transforms the text into a form that can be sentiment analysed. The first step is tokenization and converting to lower case. Tokenization refers to breaking down the text into smaller chunks, 'tokens'. The second step involves the removal of stop words such as 'how', 'to', 'a', words that carry no positive or negative sentiment and are not important to sentence structure. The last step of the process was lemmatization and normalization. Normalization involves the removal of emoticons, links, while lemmatization refers to grouping together the inflected forms of words so they can be analysed as one single item.

Once the text is in its processed form, we can perform sentiment analysis. I experimented with a few sentiment dictionaries, but their output was not significantly different, they all had ~95% correlation, so I selected the one that produced the lowest Bayesian Information Criterion (BIC) in the VECM model. The chosen sentiment dictionary was the psychological Harvard-IV dictionary (GI) which is a general-purpose dictionary developed by Harvard University. This dictionary can be found in the 'SentimentAnalysis' R library and it contains 1316 positive and 1746 negative words. The dictionary gives a sentiment score to each word, which can take up -1 for negative, 1 for positive and 0 for neutral words. (Naldi, 2019)

The sentiment value of the words are then aggregated by each row of the data frame, where each row represents a reddit comment. At the end we aggregate the sentiment value of the comments for each day -if there was no data for a given day, I assigned a neutral, 0 sentiment for a day, this was about 25% of the days- and get the following two times series, representing the sentiment values for Intel and AMD:

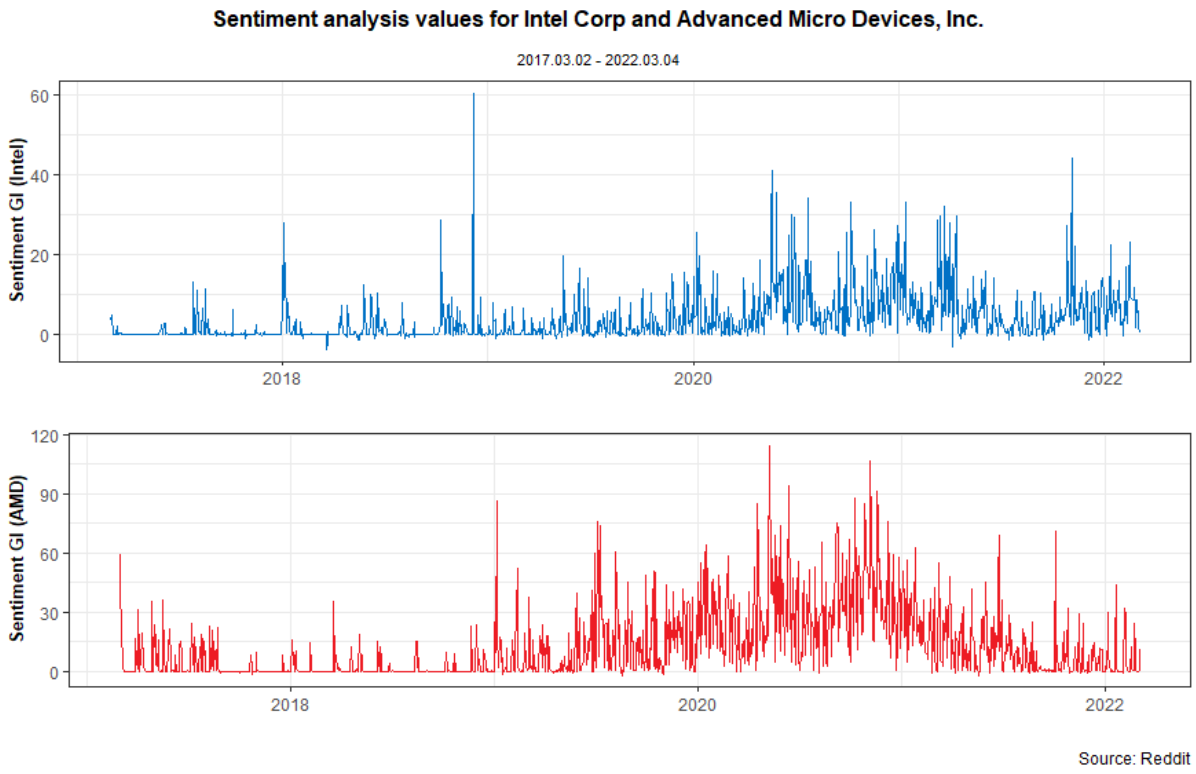

Figure 4: The Harvard GI Sentiment value for the web-scraped comments of Intel and AMD subreddits

These two time series will be used in the VECM model as exogenous variables.

7. Model building, analysis

In this chapter I will go through the steps I took to build the models and analyse some of the most important outputs. I will start with the ARIMA model, followed by the ARCH-GARCH volatility model family, from which I take the volatility of Intel's and AMD's stock prices and use it in a VECM model, adding the sentiment values as exogenous variables.

7.1 ARIMA Model

The first step was to transform Intel's and AMD's stock prices into their logarithmic returns the following way:

$$r_t = \ln(P_t) - \ln(P_{t-1})$$

where P_t represents the given days' stock prices and P_{t-1} represents the previous days' prices.

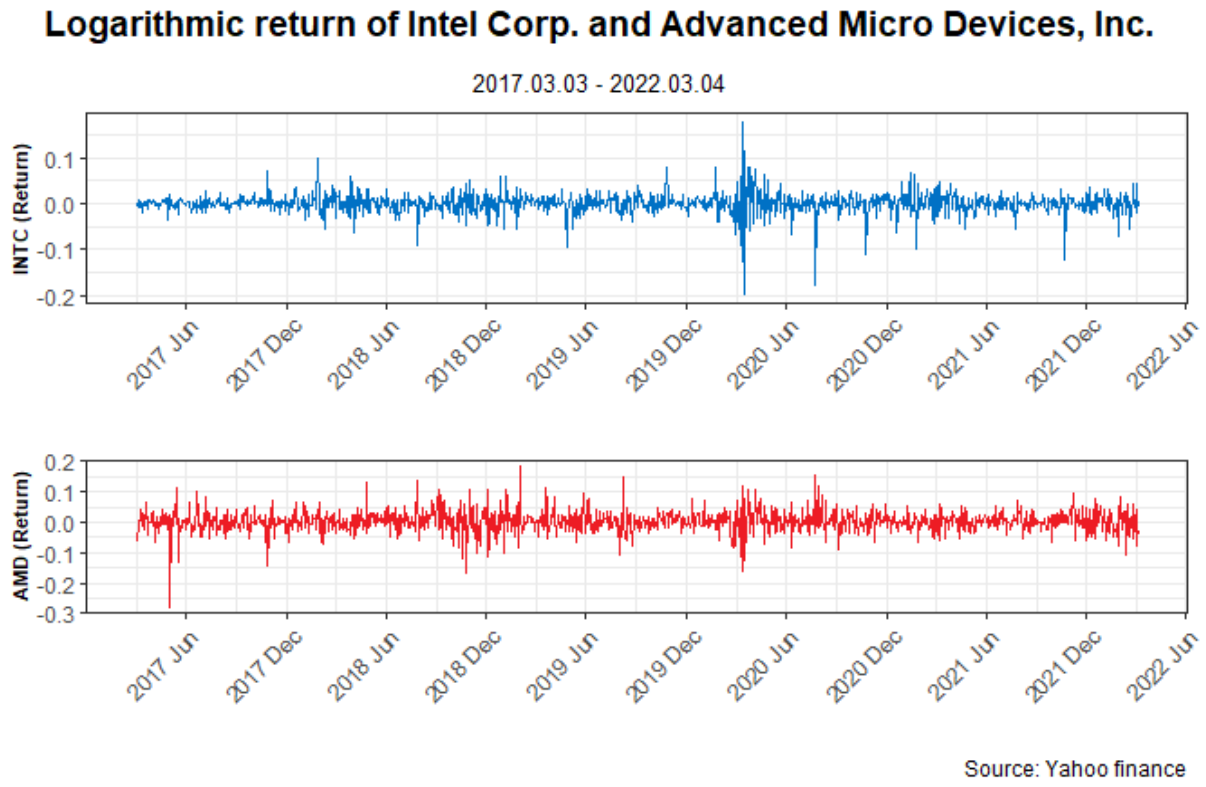

Figure 5: The logarithmic return of Intel and AMD between 2017.03.03-2022.03.04

In figure 5 we see that both returns have an expected value of 0, however the variance changes over time, so ARIMA modelling is required. To test the stationarity of the returns, an Augmented Dickey-Fuller (ADF) test had to be used, where the null hypothesis states that the time series is unit root, meaning it is not stationary. A significance level of 5% was used for both tests. The p-value was less than 0.01 in both cases; therefore, we reject the null hypothesis, which means that both time series are stationary.

The next step is to examine whether the residuals follow a white noise process. This was carried out by looking at the ACF and PACF to get an idea on which p and q parameters to use for the ARIMA (p,d,q) model and by a Ljung-Box test. The null hypothesis for the Ljung-Box test states that there is no autocorrelation between the residuals of the model, which means that it is a white noise process.

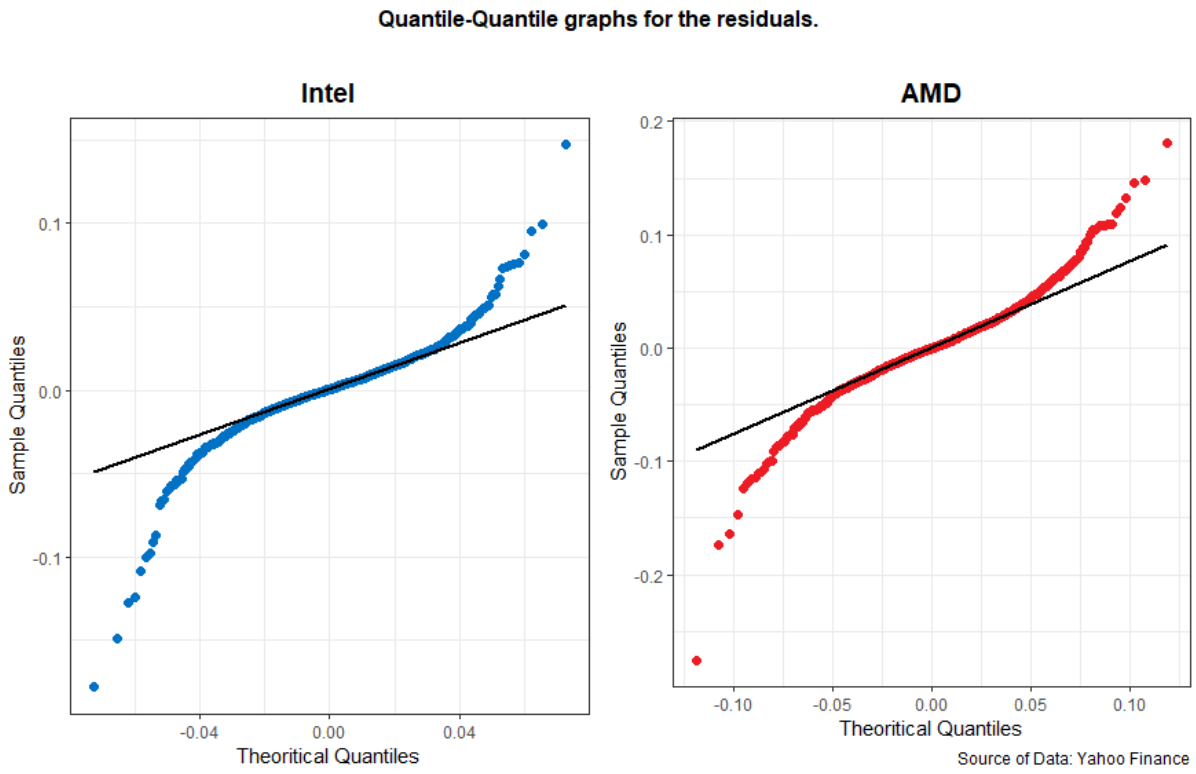

Figure 6: The Quantile-Quantile plots for the residuals of Intel and AMD.

The ARIMA model with the best BIC value was selected out of those in which the Ljung-Box test and the ACF showed no autocorrelation for the residuals. An ARIMA (1,1,3) model was chosen for Intel and an ARIMA (1,1,1) for AMD. Figure 6 shows us an inverted S form for the residuals on the quantile-quantile plot. This, and the visible volatility clustering from figure 5 shows us, that past values for the two time series are not enough to explain today's values, we will need to use GARCH models.

7.2 GARCH Model

To model the volatility of Intel and AMD, I used the following four models: GARCH (1,1), T-GARCH (1,1), GARCH (1,1) in mean and T-GARCH (1,1) in mean. A Student-T distribution was used for all models due to the fat tail distribution of the ARIMA models, which resulted in two additional parameters for the models, which describes the shape and kurtosis of the distribution. The coefficients in the model were examined with a 5% significance level, and the combination of BIC, AIC and likelihood value were used to select the best model.

As we previously saw in figure 5, there is an observable clustering of the volatility, which means it is not constant over time, so we had to use the GARCH models. The best model turned out to be the T-GARCH (1,1) model for both Intel and AMD, where the return was described by an ARIMA (3,3) for Intel and an ARIMA (1,1) for AMD. The general formula for the two models are as follows:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \gamma \epsilon_{t-1}^2 \, d_{t-1} + \beta_1 \sigma_{t-1}^2$$

$$d_{t-1} = \begin{cases} 1 \; if \; \epsilon_{t-1} < 0 \\ 0 \; if \; \epsilon_{t-1} \geq 0 \end{cases}$$

ARIMA (3,1,3) version for Intel:

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \gamma_3 y_{t-3} + \delta_1 \epsilon_{t-1} + \delta_2 \epsilon_{t-2} + \delta_3 \epsilon_{t-3} + \epsilon_t.$$

ARIMA (1,1,1) version for AMD:

$$y_t = \mu + \gamma_1 y_{t-1} + \delta_1 \epsilon_{t-1} + \epsilon_t.$$

The two models contain 13 and 9 parameters respectively. The parameters of the volatility: constant (α_0), the coefficient of the square of the error term (α_1), the asymmetry coefficient (γ) and the previous periods volatility's coefficient (β_1) and the two parameters due to the Student-t distribution that describe the skew and kurtosis of the distribution. On top of these the model for Intel's return contains the constant (μ) and the 6 coefficients ($\gamma_1, \gamma_2, \gamma_3, \delta_1, \delta_2, \delta_3$) of the ARIMA (3,3) model and the ARIMA (1,1) for AMD's return has the 2 (γ_1, δ_1) parameters. The gamma parameter displays the fact, that investors react more strongly to negative news than to positive ones, which means the decrease in returns is higher for negative news than the increase for positive news. Every parameter, except the asymmetry parameter(gamma) was significant for both models.

The results of the models are the following:

$$y_t = 0.000712 - 1.299572 y_{t-1} - 1.133004 y_{t-2} - 0.716913 y_{t-3} + 1.249597 \epsilon_{t-1} + 1.091430 \epsilon_{t-2} + 0.709126 \epsilon_{t-3} + \epsilon_t$$

$$\sigma_t^2 = 0.000464 + 0.121004 \epsilon_{t-1}^2 + 0.036687 \epsilon_{t-1}^2 \, d_{t-1} + 0.894289 \sigma_{t-1}^2$$

$$d_{t-1} = \begin{cases} 1 \; if \; \epsilon_{t-1} < 0 \\ 0 \; if \; \epsilon_{t-1} \geq 0 \end{cases}$$

For Intel, and:

$$y_t = 0.001791 - 0.750895y_{t-1} + 0.715491\epsilon_{t-1} + \epsilon_t$$

$$\sigma_t^2 = 0.001319 + 0.094686\epsilon_{t-1}^2 + 0.065678\epsilon_{t-1}^2 d_{t-1} + 0.895284\sigma_{t-1}^2$$

$$d_{t-1} = \begin{cases} 1 \; if \; \epsilon_{t-1} < 0 \\ 0 \; if \; \epsilon_{t-1} \geq 0 \end{cases}$$

for AMD.

A 1 unit increase in the previous days' volatility has an affect of 0.89 unit increase for today's volatility for both Intel and AMD. This shows the previously mentioned volatility clustering, how a higher volatility period is followed by a higher volatility period. The 0.89 coefficient shows that the increased volatility needs time to go down. This is also visible in figure 7, which presents the volatility of both Intel and AMD.

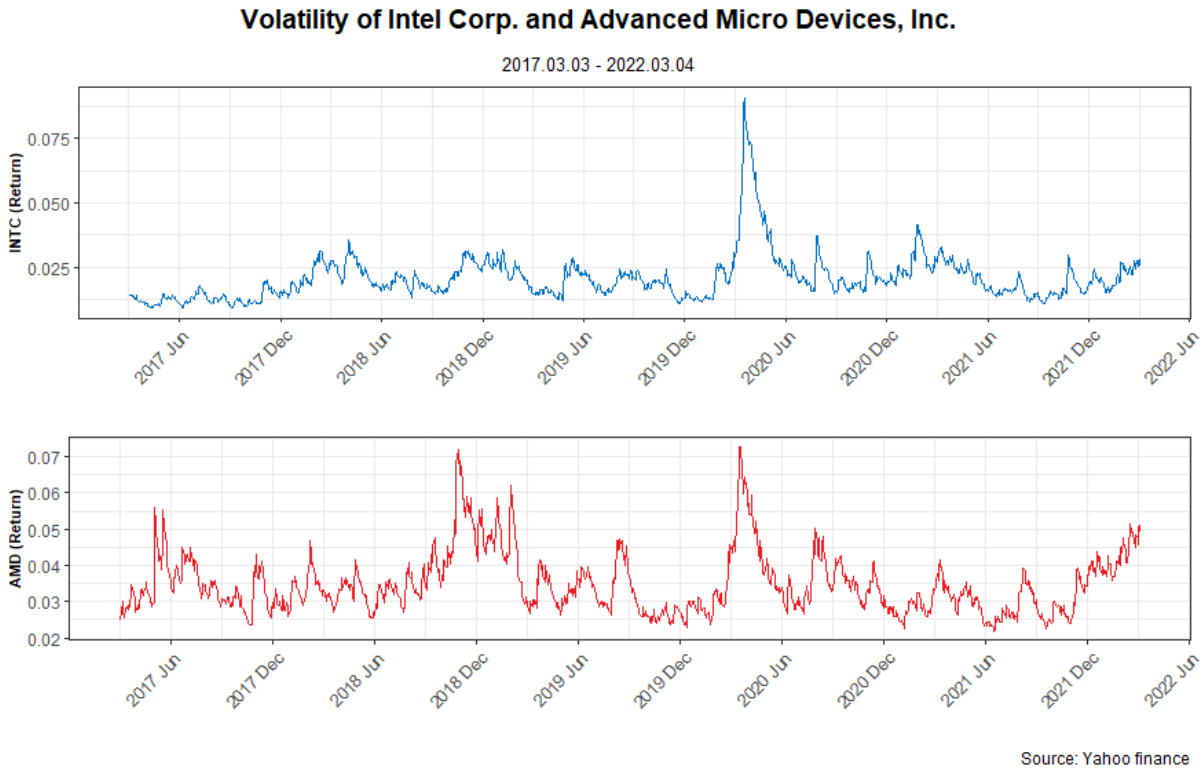

Figure 7: The volatility of Intel and AMD.

Although the clustering is not as visible as before, the GARCH model was unable to perfectly filter it out. The spike in volatility around May 2020 is explained by the COVID crash, and the higher volatility period for AMD in 2018 is when the stock price started to increase in an exponential trend. This is around the time when they started taking away Intel's market share and forecast higher earnings.

Positive news has an impact of 0.121004 increase in volatility for Intel and 0.094686 increase for AMD, which goes against our previous assumption that positive news would decrease volatility. This could be explained due to the duopolistic structure of the market, where R&D is key to gain market share, so any positive news would result in an increase in volatility. Negative news increases volatility by 0.157691 (0.121004+0.036687) for Intel and 0.160364 (0.094686+0.065678) for AMD. As we can see Intel's volatility increases more with positive news, and less with negative news, this can be explained with the previously mentioned lower growth forecasts, and generally lower sentiment towards the company. The lack of innovation Intel showed in recent years could mean that the stock price is already undervalued, thus negative news does not affect the volatility as much as in the case of AMD.

7.3 VECM Model

The last step in my research involved the usage of VECM models. I used the sentiment values as exogenous control variables and the volatility of the stock prices as dependent variables. I applied a Johansen test for cointegration for the two volatility time series. The null hypothesis of the test states that there is no cointegration between the two equations. The null hypothesis was rejected, the volatility of Intel and AMD cointegrate, which means that a VAR model would not be stable, so VECM had to be used.

The linear combination of volatility was tested with an ADF is stationary. The following equations describe the connection between the two volatilities:

$$x_t = 1 * i_t - -1.678835 a_t$$

$$y_t = 1 * a_t - 0.5956513 i_t ,$$

where i_t is the volatility of Intel and a_t is the volatility of AMD.

The optimal lag for the VECM model was 2, where both Intel and AMD contribute approximately 12.7% of the information to explain the other stocks volatility. The equation for the VECM model showed us that the exogenous sentiment values are not significant, which means they have no effect on the volatility of the stock prices of Intel and AMD. The granger causality was also investigated, where the null hypothesis states that if a variable 'a' does not Granger-cause variable 'b' that means that the past values of 'a' contain information that helps

in the prediction of ‘b’. All combination of variables were tested, none of the tests rejected the null hypothesis, which means that there is no Granger causation between any of the variables.

8. Limitations, Outlook

In this chapter I will talk about some of the limitations I encountered during my research and suggests some improvements that can be done in the future. Most of the limitations of the research came from collecting and analysing the sentiment of the text data.

My aim was to find out whether there is any connection between the sentiment towards Intel and AMD and their stock prices. For this I web scraped every comment from the social platform called Reddit and since Reddit is an open forum, available to anyone with an account it is very likely, that there were many comments in my dataset that were irrelevant to the research. Although a comment can be irrelevant, it can still carry a negative or positive emotion, so it can be reflected in the results. One of the improvements that could be done to treat this issue is to add certain keywords to comments that have to be included in the comment, otherwise the comment will not be downloaded.

Another issue is that the Intel and AMD subreddits are subforums, which are created for people interested in the companies, likely customers, or future customers. This means that the sentiment data is biased to be more positive, as figure 4 confirms this, there are very few days where the daily aggregated sentiment takes up a negative value. A way to improve this could be to search for relevant forums -for example technology related ones- where you can find comments about Intel and AMD without bias. Due to the API limitations of Reddit, this was out of scope for my paper, due to the limited number of comments I could have gathered that way.

A third improvement that could have been made is to include different social networks to gather comments from, since I had to fill up approximately 25% of the sentiment data with 0 values, using more social media platforms would have provided a more accurate representation. This would have allowed the research to be done in a wider time frame and the larger volume of data could have improved the results. Unfortunately, there are technical limitations in this regard as well, for example gathering data from Twitter requires a real-time web scraping method, so due to time limitations this was not an option either.

The last suggestion for improvements is towards the choice of sentiment dictionary. The sentiment dictionary I used can only take up three possible values (-1,0,1) which is a rather simple approach. It is not hard to admit that there can be a difference of sentiment, for example: between the words ‘great’ and ‘amazing’. The dictionary I used would assign the value 1 for both words, so a way to improve the results could be the usage of a dictionary that used a wider scale and has the sufficient number of words.

9. Conclusion

The goal of my study was to investigate whether the sentiment towards Intel and AMD has any relation to their stock prices. My initial hypothesis stated that there will be a connection, since both companies create products - consumer processors - that are available to the public, and while these products do not make up most of their revenues, they are the closest ones to the consumers.

Intel and AMD are the biggest microprocessor producers, which is why it was important to understand the basic market structure to help me better analyse the results I gained from the models. The two companies take up 98.6% of the industry, making this market a duopoly. The microprocessor industry is part of the semiconductor industry, so it was also important to look at the position the two companies take up relative to other members of the industry.

In previous papers I found several other cases where the relation of sentiment and stock prices was researched. They used a wide range of approaches in the way they collected the sentiment data, and the statistical and econometrical methodology differed a lot as well. However, none of the papers I read collected sentiment related to the companies themselves, they were all focusing on the sentiment of the stocks of the companies.

The first part of my methodology involved creating ARIMA models for the logarithmic returns of the stock prices. The results of the ARIMA models showed signs of volatility clustering, which meant that a GARCH model had to be used. Out of the four GARCH models, the best one proved to be the T-GARCH model for both companies according to the BIC and likelihood values. This showed us that a 1 unit increase in volatility of the previous day affects today’s volatility by 0.89 units. The results also showed us that while negative news has a higher effect on volatility, contrary to our anticipation positive news also increases it.

I gathered the sentiment data by web scraping the subreddits of Intel and AMD. After obtaining the comments, the text had to be transformed into a format that can be sentiment analysed; this was done by converting every letter into lower case and then separating the words into tokens, removing stop words, emotes, URLs. The processed format of the text was then sentiment analysed by the general-purpose Harvard-IV dictionary, which assigns a value of (-1,0,1) to each word depending on whether it is negative, positive, or neutral. Due to limitations of the Reddit API, around one quarter of the researched days had no data, where I also had to assign a neutral sentiment of 0.

The last part of the research involved using the volatility I obtained in the GARCH models in a VECM model, where the sentiment values were the exogenous control variables. The results showed that the sentiment values do not have any effect on the volatility of the stock prices, their coefficients were not significant. The two stock prices did however contribute approximately 12.7% to each other's volatilities.

The answer to my initial question is that there is no connection between the sentiment towards Intel and AMD and their stock prices, so my initial hypothesis has been wrong. There are, however, ways to improve the results in the future. Most of the improvements are aimed towards collecting the sentiment data. A wider search of different subreddits or different social platforms could be worth looking into, to increase the volume of available data. The way the text was sentiment analysed could also be changed, the dictionary I used can only assign 3 different sentiment values to a word. A dictionary with a wider scale could also improve the results. Last but not least, it can also be worth looking into ways to combine sentiment towards a company and their respective stocks.

Sources

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, *1*(2), 223.

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011, December). Combining technical analysis with sentiment analysis for stock price prediction. In *2011 IEEE ninth international conference on dependable, autonomic and secure computing* (pp. 800-807). IEEE.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383-417.

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, *30*(4), 330–338. https://doi.org/10.1016/j.jksues.2016.04.002

Johansen, S., 1991. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica, 59(6), pp. 1551-1580.

Kirchgässner Gebhard, & Wolters Jürgen. (2007). *Introduction to modern time series analysis.* Springer.

Lubitz, M. (n.d.). *Who drives the market? Sentiment analysis of financial news posted on Reddit and Financial Times*. 39.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 205-208). IEEE.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. https://doi.org/10.1016/j.joi.2009.01.003

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, *27*(2), 12:1-12:19. https://doi.org/10.1145/1462198.1462204

Semmler, A. (2010). Competition in the Microprocessor Market: Intel, AMD and Beyond. *Univeristy of Teier*. https://www.academia.edu/1860422/Competition_in_the_Microprocessor_Market_Intel_AMD_and_Beyond

Tumarkin, R., & Whitelaw, R. F. (2001). News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal*, *57*(3), 41–51. https://doi.org/10.2469/faj.v57.n3.2449

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear." *Procedia - Social and Behavioral Sciences*, *26*, 55–62. https://doi.org/10.1016/j.sbspro.2011.10.562

Online sources

About AMD. (n.d.). Retrieved April 18, 2022, from https://www.amd.com/en/corporate/about-amd

Advanced Micro Devices, Inc. (AMD) Stock Price, News, Quote & History—Yahoo Finance. (n.d.). Retrieved April 18, 2022, from https://finance.yahoo.com/quote/AMD/

Annual Filings. (n.d.). Advanced Micro Devices, Inc. Retrieved April 18, 2022, from https://ir.amd.com/sec-filings/filter/annual-filings

Annual Reports. (n.d.). Intel Corporation. Retrieved April 18, 2022, from https://www.intc.com/filings-reports/annual-reports

Bursztynsky, J. (2019, April 26). *Intel shares suffer steepest plunge in over three years after disappointing revenue forecast*. CNBC. https://www.cnbc.com/2019/04/26/intel-shares-down-more-than-10percent-on-revenue-forecast.html

Intel Corporation (INTC) Stock Price, News, Quote & History—Yahoo Finance. (n.d.). Retrieved April 18, 2022, from https://finance.yahoo.com/quote/INTC/

Intel's Founding. (n.d.). Intel. Retrieved April 18, 2022, from https://www.intel.com/content/www/us/en/history/virtual-vault/articles/intels-founding.html

Latest News. (n.d.). Taiwan Semiconductor Manufacturing Company Limited. Retrieved April 18, 2022, from https://pr.tsmc.com/english/latest-news

Naldi, M. (2019). A review of sentiment computation methods with R packages. ArXiv:1901.08319 [Cs]. http://arxiv.org/abs/1901.08319

Newsroom Home. (n.d.). Intel. Retrieved April 18, 2022, from https://www.intel.com/content/www/us/en/newsroom/home.html

Reddit. (n.d.). Reddit. Retrieved April 18, 2022, from https://www.reddit.com/r/Amd/

Reddit. (n.d.). Reddit. Retrieved April 18, 2022, from https://www.reddit.com/intel/

Revenue of Top 10 IC Design (Fabless) Companies for 2020 Undergoes 26.4% Increase YoY Due to High Demand for Notebooks and Networking Products, Says TrendForce. (n.d.). Design And Reuse. Retrieved April 18, 2022, from https://www.design-reuse.com/news/49698/revenue-ranking-of-top-10-ic-design-companies-2019-2020.html

Semiconductor industry worldwide by application. (n.d.). Statista. Retrieved April 18, 2022, from https://www.statista.com/statistics/498265/cagr-main-semiconductor-target-markets/

Semiconductor market share by company 2020. (n.d.). Statista. Retrieved April 18, 2022, from https://www.statista.com/statistics/266143/global-market-share-of-leading-semiconductor-vendors/

PassMark CPU Benchmarks—AMD vs Intel Market Share. (n.d.). Retrieved March 10, 2022, from https://www.cpubenchmark.net/market_share.html

Top semiconductor foundries quarterly revenue 2021. (n.d.). Statista. Retrieved April 18, 2022, from https://www.statista.com/statistics/867210/worldwide-semiconductor-foundries-by-revenue/

Types and Basic Functions of Semiconductor Packaging. (n.d.). Retrieved April 19, 2022, from https://www.researchdive.com/blog/semiconductor-packaging-basic-types-functions-and-covid-19-impact