CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Examining the Interaction between the Cryptocurrency Market Development and Activity on Leading Social Networks

Bachelor's Thesis

Author: Jakub Doškář

Study program: Economics and Finance

Supervisor: Mgr. Nicolas Fanta

Year of defense: 2023

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

Prague, December 31, 2022 Jakub Doskar

Abstract

In this thesis, analyses are conducted to determine whether various measurements of social media activity, including the sentiment value of posts, can be drivers or even predictors of a change in a selected metrices of cryptocurrencies. The analyses are performed on data collected in one-hour and fifteen-minute time intervals from the February of 2021 until the November 2022. The results of the analysis show that variability of the closing price of Bitcoin can be to some extent explained by sentiment derivatives only. Furthermore, it was proven that sentiment derived from social media is significant when used as a predictor of a direction of a price change, under specific circumstances. These results oppose the previous studies, where sentiment was not recognized significant. Moreover, it was determined that considering one-hour intervals returns marginally better outcomes than in the case of shorter time intervals. The thesis outlines the challenges researchers can face when using this technique in their work.

Keywords

Cryptocurrencies, Bitcoin, Social sentiment, Sentiment analysis, Twitter, Social media, Cryptocurrency exchange

Title

Examining the Interaction between the Cryptocurrency Market Development and Activity on Leading Social Networks

Abstrakt

V této práci jsou prováděny analýzy s cílem zjistit, zda různá měření aktivity na sociálních médiích, včetně hodnoty sentimentu příspěvků, mohou být hnací silou nebo dokonce prediktorem změny vybraných metrik kryptoměn. Studie je prováděna na datech shromážděných v hodinových a patnáctiminutových časových intervalech od února roku 2021 do listopadu roku 2022. Výsledky výzkumu ukazují, že variabilitu ceny Bitcoinu lze do určité míry vysvětlit pouze deriváty sentimentu. Dále bylo prokázáno, že sentiment odvozený ze sociálních médií je za určitých okolností signifikantní, pokud se použije jako prediktor směru změny ceny. Tyto výsledky jsou v rozporu s předchozími studiemi, kde sentiment nebyl uznán za významný. Dále bylo zjištěno, že zohlednění hodinových intervalů přináší nepatrně lepší výsledky než v případě kratších časových intervalů. V práci jsou konečně uvedeny výzvy, kterým mohou výzkumníci čelit při používání této techniky ve své práci.

Klíčová slova

Kryptoměny, Bitcoin, Sociální sentiment, Analýza sentimentu, Twitter, Sociální média, Kryptoměnová burza

Název práce

Zkoumání Interakce mezi Vývojem Kryptoměnového Trhu a Aktivitou na Předních Sociálních Sítích

Acknowledgement

I would like to express my gratitude to the supervisor of this thesis Mgr. Nicolas Fanta for his dedicated help in conducting this work, valuable suggestions and guidance.

References

DOSKAR, Jakub: *Examining the Interaction between the Cryptocurrency Market Development and Activity on Leading Social Networks*. Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague, 2022. Advisor: Mgr. Nicolas Fanta.

Table of Contents

TABLE OF CONTENTS **1**
LIST OF FIGURES **2**
LIST OF TABLES **3**
INTRODUCTION **4**
1. CRYPTOCURRENCIES, A GENERAL VIEW **6**
1.1 Cryptocurrency Exchange 7
1.2 Bitcoin 8
1.3 Ethereum, Litecoin 9
2. SOCIAL MEDIA, A GENERAL VIEW **10**
2.1 Twitter 11
3. OVERVIEW OF EXISTING RESEARCH **13**
4. METHODOLOGY **18**
4.1 Social Media Sentiment Analysis 18
4.1.1 Sentiment Analysis Approaches Review 18
4.1.2 Twitter Sentiment Analysis 20
4.1.3 Social Media Sentiment Analysis Challenges 21
4.2 Models Used 21
5. DATA PROCESSING **23**
5.1 Twitter Data 23
5.2 Binance Data 24
5.3 Combined Data 25
5.4 Additional Supportive Datasets 29
6. RESULTS INTERPRETATION AND DISCUSSION **30**
6.1 Results for One-Hour Intervals 30
6.2 Results for Fifteen-Minute Intervals 35
6.3 Tested on More Cryptocurrencies 38
7. OPEN ISSUES, TOPICS FOR FURTHER RESEARCH **39**
7.1 The Role of Likes, Comments and Retweets 39
7.2 Tackling Bot Accounts 39
7.3 Topics For Further Research 40
CONCLUSION **41**
LIST OF REFERENCES **43**

List of Figures

Figure 1: Histogram of compound sentiment scores, excluding neutral 24
Figure 2: Development of Bitcoin price in examined time frame 25
Figure 3: Scheme of variables derived from Binance 26
Figure 4: One observation of variables derived from Binance 27
Figure 5: Scheme of variables derived from Twitter 27
Figure 6: One observation of variables derived from Twitter 27
Figure 7: Comparison of closing price and overall sentiment 29

List of Tables

Table 1: Overview of utilized variables 28
Table 2: Estimation using OLS 30
Table 3: Comparison of price change estimating models 32
Table 4: Results of logistic regression 34
Table 5: The results of OLS estimation 35
Table 6: Comparison of models estimating price change 37

Introduction

In this study, a method for describing and predicting changes in cryptocurrency market characteristics is proposed. Data from leading social media, Twitter, is used for the description of price shifts of the cryptocurrency. Twitter is selected thanks to the data being accessible, relevant, proven valuable by past research and dense. Past studies suggest only cryptocurrencies with high market capitalization are suitable for the research. That is the reasoning for Bitcoin is chosen for the analysis as the largest cryptocurrency in terms of market capitalization, as of November 2022. Additional hypotheses are tested on cryptocurrencies Litecoin and Ethereum to provide less narrowly focused viewpoint.

In November 2022, only the largest of the cryptocurrencies, Bitcoin, measured in terms of market capitalization, had a value of $324.6 billion. Yet at the time of Bitcoin's peak, its market capitalisation reached 1.23 trillion USD. That is, for comparison, a value of GDP in Mexico, for the same year. Due to the significant value of cryptocurrencies, the technological progress they represent, or as a diversification asset, they attract a substantial number of people and entities. Some of them see profit through their use as actual currencies, while others see them as investment opportunities. With any motives, it is desirable to understand the reasons of their market development and all the variables that affect it.

All cryptocurrencies, including those selected for the use in this paper, have experienced significant price swings on both short-term and long-term valuations. As an example, over the course of 2017, the value of one Bitcoin increased 20 times to a high of 17 500 USD in December. Two months later, in February 2018, the price of one Bitcoin more than halved to a value of 7 900 thousand USD.

As for the period analysed in this study, 10.24 thousand USD was a value one Bitcoin was traded for in the start of September 2020. Until April of 2021, the price of Bitcoin sixfolded in value, only to decrease by half of its value over the following months before climbing to its all-time high of over 65 thousand USD on 13th of November 2021. Since then, the price fell steadily to a value of 16 thousand USD as of November 2022.

In this paper, an explanation for such phenomenon is proposed with the use of data from Twitter. Twitter is selected as one of the leading social media in the world, with 450 million monthly active users as an average of 2022. The results of past studies,

where its undeniable qualities have been proven, place it highest among the media from which to derive useful information.

Sentiment analysis is utilised to derive opinions from textual attributes of posts collected from social media about selected cryptocurrency. The sentiment is converted into numerical values, extracted features are then analysed through descriptive statistics and predictive analysis using linear and logistic models. The predictive analysis is designed to investigate the relationship between sentiment derived from text data of social media posts and the change in the market price of cryptocurrency in time periods of various length. The evaluation of results is based on several indicators, including the accuracy of predicting the direction of the market price movements.

In past research, only data collected for time periods in the horizon of days, or at most months (Abraham, Higdon, Nelson, & Ibarra, 2018) was analysed. However, in this study, data collected in 20-month period is considered. Furthermore, only day to day closing prices were considered in previous studies regarding the topic (Inamdar, Bhagtani, Bhatt, & Shetty, 2019). However, due to high level of volatility of the cryptocurrencies as well as significant changes in the level of sentiment even on little time intervals, shorter time periods are considered in this thesis.

1. Cryptocurrencies, a General View

Cryptocurrencies were created for anonymous payments made completely independent of governments and banks. In recent years, cryptocurrencies have attracted a lot of attention on several fronts. Cryptocurrencies payments are based on an innovative technical solution and work differently from traditional payments. In certain payment situations, cryptocurrency can offer advantages over traditional payment methods in terms of lower costs, speed, or anonymity. However, their usage can also be riskier because cryptocurrencies are not directly subject to the laws that govern the payment intermediation. Weak consumer protection is also a reason why it may be difficult for cryptocurrencies to become a generally acceptable and viable means of payment.

Cryptocurrencies are digital coins that can be transferred throughout the internet. Compared to alternatives, cryptocurrencies have several advantages. They are transmitted directly from one subject's wallet to another subject's wallet over a network without going through a bank or other third-party transaction provider. This also means that the transaction fees are significantly reduced. Aside from peer-to-peer transactions, there are several exchanges where one can sell or buy cryptocurrencies for dollars, euros and many other fiat currencies or even other cryptocurrencies. The cryptocurrencies are stored in a digital or external physical wallet that can be accessed via computer or mobile device. For most cryptocurrencies, the network is usually secured by the net of individuals called nodes and miners. Miners are rewarded with newly generated units of the related cryptocurrency for verifying transactions. Once verified, transactions are recorded in a transparent public ledger. This way it is protected from the users with malicious intentions. Almost everything is transparent in the world of cryptocurrencies, including their source code, just except information about users. Thanks to transparency, users and developers can try to improve their functionality, test their weaknesses, or even build up distinct software from parts of their source code. Even some institutions are aware of the benefits, and many central banks have launched their own national cryptocurrency initiatives (Bech & Garrat, 2017).

As with many new technologies, risks are present. In the nascent cryptocurrency market, one concern relates to the anonymous nature of transactions in some cryptocurrencies, which could allow for rogue actors to conduct illicit transactions, or even pose a broader threat to our society and institutions. However, this phenomenon was proven to be diminishing with the rising interest of mainstream public (Foley,

Karlsen, & Putnins, 2019). In recent years and with the growing interest of the general public, the transparency and ability of controlling the transactions along with growth in the community of users making substantial upgrades of the software have lowered the risk potential. Furthermore, few more disadvantages must be mentioned when discussing cryptocurrencies. One of the most significant shortcomings is the transaction time. As an example, for Bitcoin, the transaction can take up to higher tens of minutes. Moreover, cryptocurrencies have been heavily criticised for their environmental impact (Corbet & Yarovaya, 2020). It was estimated that the energy consumption of Bitcoin mining has augmented from 4.8 Terawatt-hours to 73.1 Terawatt-hours over the period of two years from 2017 to 2019, and the entire network consumes more energy than Austria. Estimated energy consumption per Bitcoin transaction exceeds 600 kilowatts, which is approximately equivalent to more than 300 thousand contactless payment transactions (Corbet & Yarovaya, 2020). However, there are more than 22 000 cryptocurrencies in existence as of the November 2022. Thus, even if they're not at the forefront of public interest like Bitcoin and are not so widely spread and utilized in the world, the environmental impact of cryptocurrencies is indisputably enormous.

Acknowledging the benefits and shortcomings, and the absence of a central regulatory institution, various countries took a different stance towards the use of cryptocurrency. While some attempt to regulate the use and track the transactions, others, as is the case of the People's Republic of China, have selected a much more restrictive approach and prohibit the use of cryptocurrencies. On the other hand, some countries try to popularize and regularize their use, as is the case of El Salvador. El Salvador became the first country to make Bitcoin legal tender in September 2021, requiring all businesses to accept the cryptocurrency.

Despite the undeniable shortcomings, it is acknowledged that cryptocurrencies are changing finance, and the innovations they bring became an important topic when discussing the future of the financial system (Szalay & Venkataraman, 2021).

1.1 Cryptocurrency Exchange

A cryptocurrency exchange or digital currency exchange is a business that enables customers to trade cryptocurrencies or digital currencies for other assets such as conventional fiat money or other digital currencies. Exchanges may accept credit card payments, wire transfers, or other forms of payment in exchange for digital currencies or cryptocurrencies. A cryptocurrency exchange usually works as a market maker that

typically takes the spread between supply and demand as a transaction commission for its service, or simply charges a fee as a matching platform.

Some exchanges that focus mainly on different assets such as stocks, for instance Robinhood and eToro, allow users to buy cryptocurrencies but not withdraw them into cryptocurrency wallets. Specialized cryptocurrency exchanges such as Binance, FTX, Kraken and Coinbase, however, allow cryptocurrency withdrawals. This fact, along with the values of fees per transaction may cause a slight deviations in the prices of cryptocurrencies on different exchanges.

For the research purpose of this thesis was selected a cryptocurrency exchange Binance as a source for market data about examined cryptocurrencies. According to actual data derived from CoinMarketCap (CoinMarketCap, 2022), a service that provides general information about cryptocurrencies, Binance is the largest cryptocurrency exchange in the world considering volume of cryptocurrencies traded measured in USD along with weekly visits by users. For that reason, the metrics of the cryptocurrencies derived from Binance are considered most accurate and are expected to reflect the behaviour of the whole market.

1.2 Bitcoin

Bitcoin, or BTC, is a decentralized digital currency that can be transacted on the peer-to-peer Bitcoin network. Bitcoin transactions are verified by network nodes using cryptography and are recorded in a public distributed ledger called the blockchain. This cryptocurrency was invented in 2008 by an unknown person or group of persons using the name Satoshi Nakamoto (Nakamoto, 2008). The currency came into use in 2009 when its implementation was released as open-source software.

Throughout the years, the value of Bitcoin has risen by tens of thousands of percent, along with the general public awareness about the cryptocurrency. From the basics of Bitcoin and blockchain technology, many new cryptocurrencies, called altcoins, were created. Along with the creation of the new market of cryptocurrencies, many investors and traders emerged, seeking new opportunities or diversification.

However, Bitcoins original purpose was to propose new system of payments, independent of institutions and managed by its users. There has not been much success in the use of Bitcoin as a currency, nevertheless, Bitcoin has been widely used as an investment asset.

1.3 Ethereum, Litecoin

Both cryptocurrencies belong among so-called altcoins, cryptocurrencies created after Bitcoin, based on the Bitcoin blockchain technology.

Ethereum is a decentralized open source blockchain with the capabilities of smart contracting. Ether is the platform's native cryptocurrency. Among cryptocurrencies, Ether is second only to Bitcoin in terms of market capitalization, as of 2022. Although the actual name of the cryptocurrency is Ether, it is listed on cryptocurrency exchanges and consensually referred to as Ethereum, with abbreviation ETH. For this reason, Ether will be referred to as Ethereum in this paper, although it is in fact not the name of the cryptocurrency.

Litecoin, with the abbreviated form LTC, is a decentralized peer-to-peer cryptocurrency and open-source software project released under the MIT/X11 license. Litecoin was inspired by Bitcoin and was one of the first altcoins to go on sale in October 2011. In terms of technical details, Litecoin's main chain shares a slightly modified Bitcoin codebase. The practical implications of these differences in the codebase are lower transaction fees, faster transaction confirmation, and faster changes in mining difficulty.

Additional tests are conducted on the data for these cryptocurrencies to provide less narrowly focused outcome.

2. Social Media, a General View

Social media are interactive digital platforms that provide a way to generate and share information, news, ideas, opinions, and other forms of expressions through virtual networks and communities. Although there are challenges in defining social media due to the variety of separate and integrated social media services that are currently available, there are several common features. Social media are Internet-based applications where content, such as text posts or comments, digital photos or videos is generated by users. To be able to create own content or interact with other users, one must create a profile specific to the service that is designed and managed by the social media provider. Social media help the development of online social networks by linking a user's profile to the profiles of other individuals or groups, creating communities with similar field of interest. Such communities are referred to as social networks, which, more specifically the one that connects people generating content regarding cryptocurrencies, are a field of study of this thesis.

Users typically access social media services via web applications on desktop computers or download services that offer social media features to mobile. By joining these electronic services, users are creating highly interactive platforms on which individuals, communities and organisations can share, co-create, discuss, participate in and edit user-generated content or content that they themselves create and publish online.

Social media, being one of the biggest content generators on the internet, are a great source of data for text mining. They have a large capacity, high data density and fast dissemination of information. In this paper, analyses of the relationship between selected posts related to cryptocurrencies in social media, namely Twitter and Reddit, and the attributes of the respective cryptocurrencies, are carried out.

According to user data activity, Twitter is among the most used social media platforms. It is second to Facebook as a platform based on sharing text messages and microblogging (DataReportal, 2022). And while Facebook is used mainly to communicate with friends and family or within a given group of interest, or to connect with users with similar interests on a specifically focused page, Twitter is used mainly to keep up to date with news and current events. Moreover, the posts are shared directly on the platform, not on the specific subparts or subpages of the social media, which makes it easier to collect the data and measure their impact. That is one of the reasons

why Twitter data is used in this thesis. Among the other reasons is the number of papers and research already having successfully conducted a sentiment analysis of Twitter on various topics, making it a legitimate source of information about public sentiment.

Reddit data is also tested in this thesis. Reddit is a social media forum with some special characteristics such as topic organization into smaller forums, so called subreddits. This feature makes it easier for the users to orientate and cumulate in groups of same interest, as they can easily find or create the community focused on certain subject. That also makes Reddit a great textual source for researchers and investors, as the relevant data can be found in one place. However, cryptocurrency relevant posts can be found on multiple cryptocurrency related subreddits, and single post can be shared on more subreddits as well. There is a problem, because on each subreddit, there is a different number of active users, thus on each subreddit, a post has a different reach. Additionally, users can partake on more subreddits and thus the community is blended, and the impact of the posts is difficult to interpret. Moreover, subreddits are moderated by its originators or selected users, meaning that not every post can be shared in a given subreddit. From that arises the main problem of the reddit data, because even when the data is collected from all the subreddits concerning cryptocurrencies, the density of the combined data is insufficient for considering even one-hour intervals.

2.1 Twitter

Twitter is a name of a social media platform where people communicate via short messages called tweets. It is mainly used as a platform for microblogging and has a huge amount of information available in the form of the tweets and the metadata around them.

As of July 2022, Twitter has 83.4 million users in the United States of America only (DataReportal, 2022). That is very relevant to the topic of cryptocurrencies, as it is estimated via the history of transactions, that majority of the available Bitcoin is owned by the inhabitants of the USA.

Users generate content using short messages called tweets, each containing at maximum 240 characters. Those tweets are displayed to users' followers, as well as to people that search for keyword included in a given tweet or have an interest in the topic that given tweet covers. Users can interact with the tweets by marking that they like it, or as will be from this point referred to, like it, write or reply to a comment to the tweet, reply to it with their own tweet, or use a function called retweet, that creates a new

tweet that displays the reference to retweeted. Every comment to a certain tweet can be liked, disliked, retweeted, or commented on by the users. Each tweet has attributes displaying the number of users that like the tweet, the number of people that commented on it or replied to that comment, and the number of times the post was retweeted.

Hereafter the tweet will be referred to as a tweet or as a post. Retweets are in the second part of this study analysed as tweets themselves because they have same attributes as if they were newly created tweets.

Despite the enormous amount of the data available on Twitter, its policies limit their free of charge retrieval. Thus, per researcher, only a limited amount of information can be derived. However, its quality of provided services and large developers database of guidance is undisputable. Furthermore, most recent, or even live data can be obtained, possibly making Twitter ideal tool to derive insights for potential market development from.

3. Overview of existing research

Recently, an increasing number of academic articles and studies have begun to focus on the influence of social sentiment and its importance in explaining economic reality. Even during the elaboration of the thesis, new studies were released. However, this study does not take them into account nor is influenced by them, and the analyses are conducted on a unique dataset.

Among the topics of related studies can be included the prediction of prices of individual assets in the stock market (Mittal & Arpit, 2011). It was proven that changes in the company's stock price correlate with the public's views of the company as expressed in tweets. However, the study considered rather the mood of the users' posts, including calmness, happiness, alert and kindness, than the positive and negative sentiment, as computed in this study. Yet the domain where the social sentiment studies have started being utilized even further is while studying the cryptocurrency market. The influence of the sentiment has been tested using Google trends tools (Inamdar, Bhagtani, Bhatt, & Shetty, 2019) with little to no impact revealed, but especially social media, in particular Twitter (Abraham, Higdon, Nelson, & Ibarra, 2018) and reddit (Gui, 2019).

However, in their final model, Abraham, Higdon, Nelson and Ibarra in their study conducted in 2018 didn't use sentiment metrics at all, arguing that the sentiment is overally positive, or possitively biased. This problem, further adressed in methodology, can be eliminated by using proper derivatives of the metrices. This problem is possibly caused by the fact that the words in the lexicons of the sentiment analysis methods are positively biased. At the end, in their study, solely number of tweets in an interval and google trends are utilized to estimate the closing price. Even then, the residual standard error remains significantly high, and the model is thus unable to be used to precisely predict the direction of a price change, which is considered as one of the main objectives in this thesis. Another shortcomming of the study might be considering overly long time intervals. Despite the fact that most cryptocurrencies can be volatile even on short time intervals, the predictions are made on a day to day basis. As an example, Bitcoins value have rissen overnight by 42 % on 26 of the October, 2019. Yet not only one direction price swings appear in long intervals. Taking into account longer time periods can result in covering a period where both significant decrease and increase in price occur together. Thus, in this study, shorter time frames are taken into

consideration, to cover more detailed price development. Besides that, likewise was used in aforementioned study, linear model is utilized to estimate the closing price.

Although the sentiment analyses have started being used to describe phenomena on the financial markets not long ago, with the beginnings dating after the introduction and large-scale adoption of social media along with development of computer science, they have since then been established as reliable source of information.

The information is then used mainly for making business or marketing decisions by various institutions, companies, individuals and especially marketing companies in order to obtain the most accurate information on public needs or opinions. Among the sentiment analysis, as an alternative research technique, advantages, it needs to be included that data collection and data analysis from a large sample and without time delays is allowed, as along the fact that the data show real time opinions and attitudes (Meena & Gama, 2013).

The results of the sentiment analysis can then be used either to determine the general opinion of a subject, to describe the target group in subject marketing, or to target personalised advertisements to individual users or groups. Be it any of those, the public sentiment is accurate and powerful information. However, one has to take into account the means of obtaining such information, as some legal policies might be violated when obtaining or manipulating the data. The information about users and their activities are usually by the user agreement a belonging of a given social media company. Anyone who seeks to collect the data and information about the users or their activity, either has to create a contract with a selected social media or use free of charge services that not many social media services decide to provide. Furthermore, possibly to encourage potential customers to obtain paid version of such a service, only basic and amount-limited information are available to be acquired for free.

Thus, one can obtain search the information about individual posts directly on the specific social media, yet with no overall information possibly available to be conducted about overall sentiment. Or information about specific posts can be obtained via services such as APIs from individual service providers. In the case of Twitter and Reddit, the information about the posts can be retrieved very specifically, related to a very narrow topic, for example by using keywords that is the post required to contain. In most studies, a sentiment analysis is performed on the posts obtained in this way. From the posts retrieved, the sentiment values are derived, based on which an attempt is then made to predict the market evolution of the chosen commodity. However, in this paper

the focus will aim also on the process in opposite direction, thus how the market development influences the general opinion about the selected assets on social media. For both parts of the research, sentiment analysis is necessary to be performed.

Various studies have been conducted on the topic of sentiment analysis for the domain of blogs and social media and numerous methods were introduced.

Yet even though the methods of obtaining a sentiment from the text, even specifically from the posts from microblogging platforms, are already in existence, more work is needed to properly describe the influence of the content of the posts, as was proven by researchers who have done work on examining the effects of individual posts' features on sentiment (Kouloumpis, Wilson, & Moore, 2021). It was proven that various Twitter posts text features do influence the sentiment of the tweet differently. Consequently, specific features are often connected to the topic the posts are related to. For instance, the tweet reporting Bitcoin prices contains several numeric characters, while the tweet created to express a mood of its author might contain only words or emoticons. This does play a role when selecting a right lexicon for the sentiment analysis, since different dictionaries may be useful only in a specific field of study. Even when a dictionary is created that takes into account topic of a given tweet, it is debatable whether sentiment from posts on different topic are comparable. Furthermore, it remains unclear whether the sentiment of the post, meaning the sentiment expressed by the author of the post, can be considered same as sentiment perceived by the reader of the post.

Despite the popularity of social media based on microblogging being quite recent, papers that analyse machine learning techniques in the specific domain of Twitter already exist in quite an amount. Multiple methods of obtaining the sentiment were compared by various factors, including irony detection, emotion detection or opinion retrieval ability and then categorized according to the techniques they use or field of study they are related to (Giachanou & Crestani, 2016). Work has been done on the topic of how pre-processing, or removing non-sentiment colored characters and their recognition affect the overall sentiment analysis quality (Jianqiang & Xiaolin, 2017). Overall, text classification using machine learning techniques is a well-studied field, the effects of several machine learning methods have been computed, namely Naive Bayes Maximum Entropy, and Support Vector Machines on numerous domains (Inamdar, Bhagtani, Bhatt, & Shetty, 2019). Relevant still is the work that has been done in using emoticons as labels for positive and negative sentiment (Read, 2005). This is very

relevant to Twitter and other social media, since many users use emoticons to express their attitude or even a certain reality, as even emoticons representing certain objects are in existence.

Related to cryptocurrencies and cryptocurrency market description and analyses, researchers have examined statistical properties of the largest cryptocurrencies (determined by market capitalization). The exchange rates versus the USD were characterized by fitting parametric distributions to them (Chan, Chu, Nadarajah, & Osterrieder, 2017). It was shown that returns are undoubtedly non-normal, and no single distribution fits jointly to every cryptocurrency examined, however, the characteristics of log returns of various cryptocurrencies were introduced. In other papers, the interdependencies among a panel of the most traded digital currencies are explored and evaluated from statistical and economic points of view (Candila, 2021). The relationship between the most popular cryptocurrencies and a range of selected fiat currencies, in order to identify any pattern or causality between the series was analysed (Corelli, 2018). The findings show very interesting results in regard to how the causal effect is concentrated on some specific cryptocurrencies and fiat currencies. The outcome is a relationship between some smaller cryptocurrencies and Asian markets. It is thus recommended to only research the properties of largest cryptocurrencies, as of market capitalization, or take into account possible correlations or causalities concerning fiat currencies. Along that, various aspects of a cryptocurrency market have been inspected, among which the volatility and returns relationship was examined with the use of GARCH-in-mean models. Additionally, spillover effects from the cryptocurrency market to other financial markets of leading economies was observed, yet little to none in the case of opposite direction, concerning large cryptocurrencies only (Liu & Serletis, 2019). Cryptocurrency market has been attempted to be predicted in various studies, most of which focused on using deep learning and machine learning techniques such as ARIMA, FBProphet and XG Boosting (Inamdar, Bhagtani, Bhatt, & Shetty, 2019). These machine learning techniques take into account the past development of the explanatory variables to predict the future values. However, these cannot be used in this study, since the collection of the Twitter data is not continuous.

Overall, sentiment analyses methods are described and classified to same extent, however, multiple topics are pointed out to be insufficiently covered, such as the meaning of same expressions when used in different field of study or context.

Nevertheless, specific methods for sentiment analyses in the financial field do exist and were tested even concerning cryptocurrencies.

Numerous studies were conducted with an aim to describe various properties of a cryptocurrency market. Among the findings is included possible dependency between the development of specific cryptocurrencies, generally smaller in terms of market capitalization, on certain fiat currencies or other commodities. Furthermore, different cryptocurrencies have distinct properties, such as approximate distribution of log returns. Thus, when conducting studies on the topic of cryptocurrencies, every cryptocurrency should be treated individually and only larger in terms of market capitalization are recommended to be considered as otherwise additional, yet non-defined influences might affect the results.

4. Methodology

4.1 Social Media Sentiment Analysis

Social media analysis is the process of analysing the data collected from social media to derive insights for either most commonly making business decisions or, as in the case of this study, research field. Sentiment analysis on social media is an area that has recently attracted the interest of academics. It addresses the problem of analysing user generated content in terms of the opinion it expresses. In this study, algorithms are used to perform a sentiment analysis on Twitter. In this part, selected approaches are presented, and their functionality is explained. Areas related to social media sentiment analyses, including quality of opinion mining, sentiment tracking over time, irony detection, emotion detection, and quantifying the sentiment, are addressed, and compared.

4.1.1 Sentiment Analysis Approaches Review

Various methods of subtracting the sentiment from the text have been used in past research. Furthermore, work have been done on classifying the sentiment analyses approaches and describing their optimal use along with possible shortcomings. However, there is no consensus approach for sentiment analysis, not even for a specific field of study. There is rather large variety of approaches one can select from and the choice of the method can be depending on the research question. As an example one of the most commonly utilized tools when subtracting sentiment for a topic regarding financial field, there are analyses that measure sentiment using VADER (Valence Aware Dictionary for Sentiment Reasoning) (Hutto & Gilbert, 2014).

Yet to determine whether that is the best option in the case of this study, various other methods, classified as suitable for financial field of study were put into comparison.

Thus, a dictionary henceforth referred to as DictionaryGI is tested for computations of sentiment. This is a dictionary with a list of positive and negative words according to the psychological Harvard-IV dictionary as used in the General Inquirer software (Stone, Dunphy, & Smith, 1966). It is a general-purpose dictionary developed and enriched by the Harvard University. The DictionaryGI for sentiment is a consensus and worldwide acknowledged dictionary, not only used for obtaining the sentiment of a social media posts, but widely used for text sentiment classification in numerous fields of study, addressing texts of various lengths and contents.

Furthermore, a dictionary hereafter referred to as DictionaryLM is tested in calculations. It is a dictionary with a list of positive, negative and uncertainty words according to the Loughran and McDonald finance-specific dictionary (Loughran & Mcdonald, 2011). This dictionary was first presented in the Journal of Finance and has been widely used in the finance domain ever since.

Both aforementioned methods, DictionaryLM and DictionaryGI do have similar method of application, yet they differ in the number of words contained in the lexicons, as well as in their content and the intended purpose of the usage.

For the text of each post, analyses of sentiment are performed based on the DictionaryGI and DictionaryLM. Each word of the text of every post, is searched for in the dictionary. When it's found, it is assigned either a positive value, if it is noticed among positive word, or a negative value, if it is discovered among negative. Neutral value is assigned to words not found in the dictionary. Each word can be thus assigned either value of 1, -1, or 0, based on whether the word was discovered in the positive section (1), negative one (-1), or was not discovered at all or noticed in a neutral section (0). A sentiment of each post is then computed by two methods. The first one takes into account number of all words in the post, by computing the overall sentiment of the post as the difference between the number of positive words and the number of negative words divided by the total number of words. The second method computes overall sentiment of the text as the difference between the number of positive words and the number of negative words divided by the total number of positive and negative words combined. Furthermore, ratio of uncertainty is computed for each tweet as the number of uncertainty words divided by total number of words.

However, both aforementioned methods have shown to be unhelpful in case of deriving the sentiment from Twitter messages. They might still be valid indicators of overall sentiment of a given text or article, nevertheless in case of examining selected tweets, those methods were proven to not be useful. That is because they only are sensitive to polarity, that is, either positivity or negativity, but not the intensity of a positivity or negativity. Additionally, those methods do not count for abbreviations or emoticons, widely used in social media communication. Along that, the hashtags, the URL addresses and even numerical characters, are labelled as a word, which can negatively affect the overall results of a sentiment computations.

Thus, in this paper, VADER is used for text sentiment analysis. VADER is sensitive to both polarity (positivity/negativity) and their intensity. Additionally, it can

understand the context in which the words are used. This method was specifically designed and accustomed to sentiments expressed in social media. Using VADER, it is accounted for sentiment-laden slang words, emoticons, sentiment-laden initialisms, acronyms and even the contractions as negations. The use of punctuation to signal increased sentiment intensity and the use of word-shape to signal emphasis are also accounted for.

For each text classified using VADER sentiment analysis, multiple scores are calculated. The normalized, weighted composite score, or shortly the compound score is computed by summing the valence scores of each word in the lexicon, and then normalized to be between -1 (most negative) and +1 (most positive). This is the only metric used in this thesis, as it is the only unidimensional measure of sentiment for a given sentence. Furthermore, positive, neutral, and negative scores are computed as ratios for proportions of text that belong to each of the categories. The sum of these should thus add up to one.

Among the positives of using VADER sentiment analysis is as well the processing speed of the operation, being multiple levels higher when compared to other methods, however, that might be caused by the continuous adjustments by the developers. In any case, when processing real time data to predict future values of examined variables, the speed of processing plays an enormous role.

4.1.2 Twitter Sentiment Analysis

Twitter has several characteristics that make it easier for researcher to collect and interpret the data, along with many challenges that must be overcome in order to obtain the desired outcome. For instance, almost every tweet contains so-called hashtag or hashtags, which are listed in the head of each tweet, and help users to label and identify what is the tweet related to. For researchers or a data analytics, hashtags are useful way when identifying topic tweets are related to at a scale of large datasets.

The text of each examined post is explored using VADER sentiment analysis. Using this analysis, the text is compared with lexicon of positive and negative words and each word, emoticon or other character the text contains is assigned a value of sentiment from the lexicon. The values range from -4, denoting the most negative sentiment, and 4, denoting the most positive sentiment, including decimals. Parts of the examined text that are located possessing neutral value or are not contained in the dictionary are assigned a neutral sentiment value of 0. The compound, positive, neutral,

and negative scores are then derived. However, it is debatable whether, when pre-processing the data, the text of each post should be striped of any hashtags, URLs or other non sentiment-coloured characters. In this study, the text of each post is preserved in the same form as was obtained.

4.1.3 Social Media Sentiment Analysis Challenges

One of the major issues with using the techniques to determine sentiment is that some of the dictionaries only include words with a sentiment value of either positive or negative integer. This means that the sentiment value using DictionaryGI or DictionaryLM for the word "good" is the same as for the word "excellent", namely positive or 1. The same is also the case for a negative value. Another limitation is the number of positive and negative words in the dictionary. For instance, the DictionaryGI dictionary that is tested contains 2005 words with negative sentiment value and merely 1637 words with positive sentiment value. Moreover, it is also noticeable that not all sentiment-laden words can be included in dictionaries. Another shortcoming is that if a word is incorrectly spelled in the text or only expressed by an abbreviation, these variations may not appear in the lexicon.

Furthermore, using each of the methods, sentiment values in the research are found to be positively biased and average sentiment per each period is almost always positive. The probable cause of this phenomenon is the positive biasness of the lexicons used to compare the textual data with, rather than overall positivity in sentiment towards the subject, as even when the price of the examined commodity is only rapidly depreciating in the given period of time, the overall sentiment remains positive. However, further research is needed to investigate the certain reasoning for the bias, perhaps by testing sentiment analyses on various topics. In this study, the problem is dealt with by deriving new variables which negate the influence of the positive bias on the model.

4.2 Models Used

Primarily multiple linear regression model is used, and OLS to obtain the estimated values of coefficients as was used in previous research. Variable transformation is used to deal with high coefficients of skewness and kurtosis and obtain near normal distributions of variables. The variables for each model are chosen not to have close to perfect collinearity. The disturbance term is tested to be independent of the regressors,

and for each model approximately follows a normal distribution with zero mean. The assumption of homoskedasticity is also tested for and met in every model proposed. Correlation between disturbances and the regressed variable are measured and possible explanations provided in results interpretation. The best fit model is selected by comparing multiple model accuracy metrics. Those are, adjusted R-squared, Akaike's Information Criteria (AIC) and Bayesian information criteria (BIC). Akaike's Information Criteria is a metric a metric which is designed to penalize the inclusion of additional variables to a model. The lower the AIC, the better the model. Bayesian information criteria is a variant of AIC with a stronger penalty for including additional variables to the model. Accuracy metrics penalizing the inclusion of more variables are utilized because the total number of variables, including the lagged values, is over one hundred. Using combinations of all the explanatory variables, the most optimal model for each number of explanatory variables is chosen, and followingly the overall best model is selected by comparing those models with one another.

Additionally, the logistic regression model and multiple logistic regression models are tested for estimation. It is used to estimate the direction of the price change and sentiment change. The data is adapted to fit the model by removing strongly influential outliers.

5. Data Processing

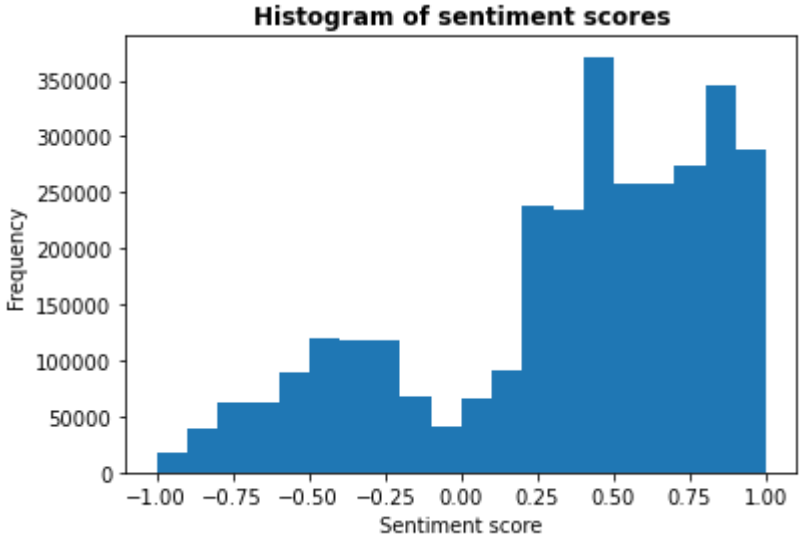
5.1 Twitter Data

For the cryptocurrency Bitcoin, data was collected about posts from the social media Twitter that contained the keyword Bitcoin, or the abbreviated term BTC. The official Twitter API is used to collect the data. The data was collected in a time period from the February of 2021 until the November of 2022.

Since there are limitations on the amount of data that can be retrieved this way, data from social media Twitter are not continuous, they are collected in multiple time frames. Those missing values make the analysis more difficult overall. Firstly, the data was cleaned from missing values. Posts with incorrectly attributed values or ones with missing essential information, were either removed or completed. The data from a total of 4 585 471 Bitcoin related posts was collected. For each tweet, information was obtained about tweet creator account attributes, such as number of other users that follow it, along with user identifications. The information about usernames or identification numbers is not used in this study. Moreover, the information about each tweet contains two key attributes, the text of the tweet and the time the tweet was created.

For each posts text, a value of sentiment was computed. It was calculated using the VADER sentiment analysis described in methodology. Each tweet has been thus assigned with a value of either positive, negative or zero compound sentiment score. The compound sentiment score ranges from -1 denoting most negative sentiment and 1, standing for most positive sentiment. If a post did not contain words that express sentiment according to the VADER, the post was assigned a neutral, or zero sentiment value. Tweets having the compound sentiment score equal to zero, account for 31 % of the dataset. They are not removed from the model because they express the neutral stand to the topic. Furthermore, almost 53 % of the sample consists of tweets of positive sentiment values.

Figure 1: Histogram of compound sentiment scores, excluding neutral



It can be observed that, similarly to the past research, that there is three times more positive sentiment post than negative ones. Additionally, variable to account for the influence of the tweet is introduced and denoted as sentiment power. The variable is created to take into account the possible impact of a given tweet and is computed by multiplying the sentiment value of the tweet by the number of followers the user had when the post was created.

5.2 Binance Data

From the cryptocurrency exchange Binance, data on the development of the market for the trading pair Bitcoin and Binance dollar (BUSD) is extracted. That is due to non-existent historical data for trading USD for Bitcoin in the database of Binance historical trading data. Binance dollar is a so called stablecoin, a cryptocurrency with its value pegged to the value of a reference asset. Binance dollar has its value fixed to the value of USD and that in the rate of 1 BUSD = 1 USD. BUSD is thus considered as an equivalent to USD. Despite the fact that the trading pair might be the best option to pick, since it has the most volume traded and the highest number of trades, with its value permanently equivalent to USD, not deviating, which is a recent problem of the another considered stablecoin, tether, BUSD still compared to other cryptocurrencies with their value fixed to USD, represents only a part of all trades for dollar equivalent currencies.

For the first part of the research, the historical market data was obtained for one-hour and fifteen minutes time intervals. The data is then combined into continuous data frames containing data for the period from the February 2021 until November 2022. The data frames, both for fifteen-minutes and one-hour periods, contain the data about opening time, opening price, highest value, lowest value, closing price, volume traded, closing time, quote asset volume and number of trades for each interval.

Figure 2: Development of Bitcoin price in examined time frame

5.3 Combined Data

For both datasets, one with hourly data and one with data in fifteen minutes time intervals, a value of cumulative sentiment was assigned to each observed interval, computed as the sum of the sentiments of the posts that were created during the given period. Followingly, to each period was assigned number of tweets posted in the selected time periods, including neutral ones. Along that, new variable, named sentiment per tweet is derived. It is computed as overall sentiment divided by the number of tweets in each time frame. By introducing this variable, the problem of overall positively biased values of sentiment is accounted for. The difficulty was, that with generally positive sentiment scores, number of tweets is correlated with total sentiment value in each period, thus with more tweets in a given period, comes higher sentiment. The sentiment per tweet variable or more precisely an average sentiment of a post in a given period, is close to being normally distributed. The problem of seasonality of the data for number of tweets, and consequently overall sentiment, is eliminated as

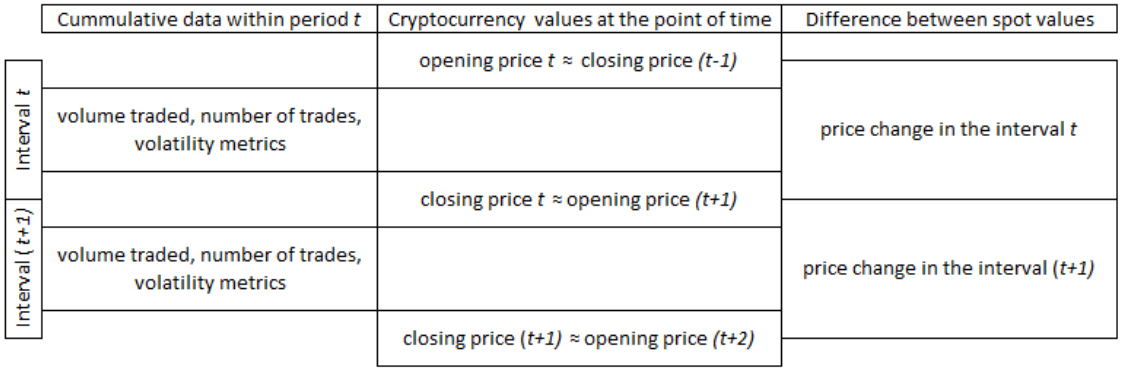
well. Additionally, the sentiment power variable for each period is computed as total sum of the sentiment powers.

The combined data frame contains data from both social media Twitter activity and cryptocurrency exchange Binance. Moreover, multiple metrices were added to the dataset. Firstly, the variables for price change for each interval are calculated as a difference between the closing price and the opening price along that, the percentage difference of those prices is computed, along with difference in logarithms of successive values, denoted as log returns. Additionally, variables for volatility measurement are added as a difference and percentage difference between the highest and the lowest price in the given time interval. Alongside that, for each observation, lagged values until lag 5 were added for each of the variables. These past values of each variable serve for testing of the possible predictive power of the data, nonetheless on the cost of losing degrees of freedom. Moreover, for future research, values of variables from the more distant past are recommended to be added to the dataset.

Furthermore, dummy variables were added to the data frame that indicate whether the price was in growth or in decline for the last 3 and 10 time periods taking a value of 1 if yes and 0 if otherwise. These variables are introduced to account for the change in the trend, as well as to provide deeper insights about the data and its long-term development.

In addition, the dataset has been enriched with variables describing the changes between successive time periods to describe the changes in sentiment per tweet.

Figure 3: Scheme of variables derived from Binance



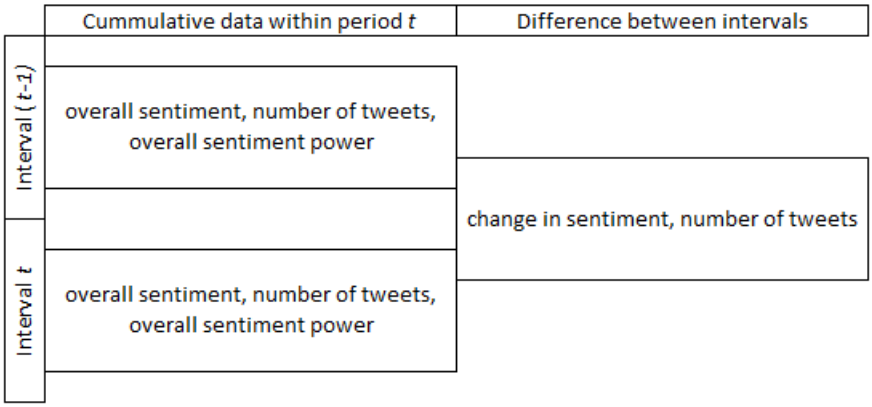
The previous scheme displays the continuity of variables derived from cryptocurrency exchange Binance.

Figure 4: One observation of variables derived from Binance

Open time	Close time	Open price	Close price	Volume traded	Number of trades	Price change	% price change	Log price change
12.05.2022 14:00:00	12.05.2022 14:59:59	28009.07	29419.26	2177.2300	60574	1410.19	5.034762	4.912118e-02

The above pictured observation contains the sample of information about one interval. Cumulative data about the interval is represented in number of trades and volume traded. Various measures of change in value of the price within the time period are listed. In the regression, it is tested which of these metrics can be explained the best.

Figure 5: Scheme of variables derived from Twitter



The displayed scheme demonstrates the continuity of variables derived from cryptocurrency exchange Binance.

Figure 6: One observation of variables derived from Twitter

Open time	Close time	Overall sentiment	Number of tweets	Sentiment per tweet	Number of trades	% sent per tweet change	Log sent per tweet change
12.05.2022 14:00:00	12.05.2022 14:59:59	436.2087	2274	0.19182441	60574	34.9442656	0.299691660

The above pictured observation contains the sample of information about one interval. Cumulative data about the interval is represented in number of tweet and overall sentiment. Variable ‘sent per tweet change’ descripes a change in sentiment per tweet from previous time interval either as percentage change, or as a difference in natural logarithms.

The data have been then cleaned and separated from the observations with missing or partially missing values. The dataset used in models observing the development in hour-to-hour periods has 3377 observations with the fully available and complete data. The dataset with fifteen-minute time intervals has 14941 observations.

If the variables are inspected in more detail, it can be observed that the distributions of some are skewed. That is tested for, and high coefficients of skewness and kurtosis are revealed. Thus, the transformed variables are derived using variable transformations such as taking squared roots or logs of the values and added to the dataset. The transformed values of overall sentiment, overall sentiment power and number of tweets and trades are derived. The disadvantage is, however, the loss of degrees of freedom as few negative values could not be transformed.

Furthermore, a lot of outliers were identified in various variables. However, they were included in the dataset, and the models were tested on both data with and without them.

Table 1: Overview of utilized variables

Statistic	N	Mean	St. Dev.
open_price	14,941	39,157.940	13,214.350
highest_value	14,941	39,267.790	13,244.960
lowest_value	14,941	39,044.240	13,179.820
close_price	14,941	39,157.200	13,214.090
abs_diff_betw_perc_price_ch	14,941	−0.0001	0.392
price_ch	14,941	−0.630	170.550
neg_trend_dummy_3	14,941	0.112	0.315
pos_trend_dummy_3	14,941	0.111	0.314
pos_trend_dummy_10	14,941	0.0005	0.022
neg_trend_dummy_10	14,941	0.0005	0.022
log_diff_betw_low_high	14,941	0.581	0.499
perc_num_of_trades_change	14,941	9.604	62.745
number_of_tweets	14,941	295.263	207.492
overall_sentiment	14,941	69.935	58.393
overall_sent_power	14,941	408,641.300	787,081.500
sent_per_tweet	14,941	0.222	0.087
perc_num_of_tweets_change	14,941	3.623	40.282
perc_sent_per_tweet_ch	14,941	4.551	661.454
sq_overall_sentiment	14,845	7.620	3.515
sq_number_of_tweets	14,941	16.002	6.261
sq_overall_sent_power	13,484	22.820	8.056
sq_number_of_trades	14,941	8.676	0.760
log_returns	14,941	−0.00002	0.004
log_sent_per_tw_ch	14,759	0.001	0.451
log_ovr_sent_ch	14,759	0.001	0.494

The Table 1 shows fundamental description of the variables used, for fifteen-minute time intervals. The log values are computed as the natural logarithm of a value

in current period minus the natural logarithm of value in previous period. Percentage values denote difference between values as a percentage of the former one.

Figure 7: Comparison of closing price and overall sentiment

The figure 5.7 shows the development of the overall sentiment values and the close price in fifteen-minute time intervals, from 17. 7. until 31. 7. 2022. The overall sentiment values are inflated mostly in the times when large number of tweets occur. Furthermore, signs of seasonality can be observed. It is evident, that despite the apparent steep increase of Bitcoin price, overall sentiment does not follow the same trend.

5.4 Additional Supportive Datasets

Additional data was collected for the cryptocurrencies Bitcoin, Litecoin and Ether from Twitter including retweets. Thus, not only original messages were obtained, but also the same messages shared by other users. The originating date of the post, tweet or retweet is when a given post is created. Thus, when a retweet is posted, the date of its origination is not the same as the date of the origination of given tweet. This should possibly cause greater continuity of the dataset. However, the data was collected for a period of units of days only. That is caused by the restrictions on number of tweets one can retrieve from Twitter via API. The data has multiple times higher density than the original dataset, which should result in more accurate measurements of the sentiment, thus more accurate results.

6. Results interpretation and discussion

6.1 Results for One-Hour Intervals

Firstly, using linear regression, close price in each time period was estimated only with the use of values obtained from social media for the same period. The model, after optimization, consists of five explanatory variables, overall sentiment, number of tweets, their square roots and sentiment per tweet. Each variable in this model including the intercept is considered significant on a 0.001 level of significance. However, in this model, explanatory variables account only for approximately 36.7% of variability of the close price. Furthermore, the residual standard error is extremely high.

Table 2: Estimation using OLS

	Dependent variable:
	close_price
sent_per_tweet	−91,616.760***
	(9,291.044)
number_of_tweets	40.949***
	(1.945)
overall_sentiment	−81.037***
	(7.313)
sq_number_of_tweets	−5,420.386***
	(227.083)
sq_overall_sentiment	7,970.358***
	(480.527)
Constant	86,107.320***
	(1,936.913)
Observations	3,371
R^2	0.368
Adjusted R^2	0.367
Residual Std. Error	10,442.930 (df = 3365)
F Statistic	391.532*** (df = 5; 3365)
Note:	*p<0.1; **p<0.05; ***p<0.01

This might be caused by the fact that the data was collected in different time periods, in which the price was significantly different, however, the sentiment values were similar.

Thus, more variables were added to the original model of estimating closing price. At first, lagged variables of sentiment metrices were added to the model, and the explanatory power did rise significantly, with the model explaining slightly over 46.4 percent of the variability of the closing price. However, the residuals standard error remains extremely high. Furthermore, it was tested whether the model has more desirable outcome when using the original variables of overall sentiment and number of tweets, or their transformed versions removed of excessive skewness and kurtosis. It was determined that variables transformed and freed from superfluous skewness and kurtosis are marginally better for the model. Ultimately, using the maximum likelihood estimation is desirable in future research, since the original variables for number of tweets and overall sentiment might follow other than normal distribution and their transformation might cause a loss of degrees of freedom.

Another assumption that is tested for, is that the sentiment variables can be a good choice when estimating the close price in periods of solely positive price change or negative price change. However, the restricted models show worse results.

Followingly, the close price is estimated only with the use of variables describing reality of the Bitcoin market in the past time frames. However, since the data is not continuous, the viability of time series models is limited. The best model was identified for estimating close price with the use of value of close price in last period and variables accounting for price changes in past periods. Residual standard error was computed to be 336.5. Additionally, variables accounting for sentiment were added to the model, and the residual standard error lowered, yet not even by one. Only few of the variables derived from Twitter were significant in the best fitting model. Those were, square root of overall sentiment in previous period, square root of number of tweets in last period, and percentage change in number of tweets between previous period and the one before.

Along estimating the closing price, the price change metrics were estimated. Nevertheless, values derived from Twitter have close to none explanatory power regarding these variables, using linear model. The significant F-statistics still say, that the models fits the data better than the intercept-only ones. The worst results were obtained when estimating the price change in base units (variable *price_ch*), USD, and

the best results are obtained when estimating the log returns (*log_pch*), computed as the logarithm of close price in a given period minus a logarithm of open price in a same period. The results for estimating the percentage price change (*perc_price_ch*) are simmilar to the log returns.

Table 3: Comparison of price change estimating models

	Dependent variable:		
	perc_price_ch	log_pch	price_ch
	(1)	(2)	(3)
number_of_tweets	0.0003*	0.00000*	0.111
	(0.0002)	(0.00000)	(0.071)
overall_sentiment	−0.002***	−0.00002***	−0.612**
	(0.001)	(0.00001)	(0.269)
sent_per_tweet	−3.811***	−0.039***	−898.632**
	(0.949)	(0.009)	(372.699)
log_sent_per_tw_ch	0.129	0.001	21.775
	(0.091)	(0.001)	(35.761)
sq_number_of_tweets	−0.082***	−0.001***	−22.763***
	(0.022)	(0.0002)	(8.596)
sq_overall_sentiment	0.192***	0.002***	51.401***
	(0.047)	(0.0005)	(18.409)
Constant	0.728***	0.007***	181.524**
	(0.193)	(0.002)	(75.828)
Observations	3,006	3,006	3,006
R^2	0.010	0.010	0.004
Adjusted R^2	0.008	0.008	0.002
Residual Std. Error (df = 2999)	0.843	0.008	331.142
F Statistic (df = 6; 2999)	4.978***	5.252***	2.032*
Note:			*p<0.1; **p<0.05; ***p<0.01

Though, the aim could be to simply predict, whether the price change will be negative or positive in future time period. To obtain that result, a logistic regression is used. A dummy variable is added to the dataset, taking a value of one, if the price change is positive in a given time interval, and zero if otherwise. This dummy variable in now estimated using metrices form Twitter and even variables from past intervals.

There the disadvantage of sentiment values being predominantly positive becomes evident. The price change is close to normally distributed and its mean value close to zero, along with the fact that 49.9 percent of the values are positive. However, the sentiment per tweet value, even when close to normally distributed as well, has its mean value of 0.22 and the value of standard deviation of 0.076. Thus, introducing a dummy variable that takes a value of 1 when the sentiment per tweet or another sentiment value is positive is not very useful. What could be done is, creating a dummy variable taking a value of 1 when the value of sentiment per tweet is higher than mean or, since the number of positive to negative price changes ratio is almost one to one, even median value. The downside of this process is that for different time periods of the data collection, mean and median value of sentiment per tweet varies.

While estimating the price change, or rather the direction of the price change using logistic regression, variable accounting for the change of sentiment per tweet between following time periods was found significant. Thus, logistic regression was run where the dependant variable accounts for price change and the independent variable is equal to 1 when the change in sentiment per tweet value from the previous period is positive and 0 otherwise. Yet the results don't show clear dependence. However, after restricting the data for the price change to be greater than 1 %, it clearly helps to increase the explaining potential of the model. The proposed formula is:

$$positive\_price\_change = -0.3165 + 0.395 * positive\_sentiment\_per\_tweet\_change$$

The *p*-values for both intercept and explanatory variable are below 0.01 level, indicating statistical significance. The same result is obtained when using the original change in sentiment per tweet between two following periods as percentage value and as a difference of logarithms.

From the model, it can be observed that the log-odds of positive price change occurring, if the sentiment per tweet has changed in positive direction since past period, are approximately 0.0785, which corresponds to 52 percent probability. Yet the log-odds of negative price change occurring, if the sentiment per tweet has changed in negative direction from past period, are estimated to be -0.3165, which corresponds to 58 percent probability. Thus, it can be reflected that when the negative change in sentiment from the past interval is happening, it is more likely that the price change is

negative as well then when the positive change in sentiment is occurring and the price change is also positive, yet this only was proven when the change in price is higher than 1 %.

Table 4: Results of logistic regression

	Dependent variable:		
	pos_pc_ch		
	(1)	(2)	(3)
pos_sent_ch	0.395**		
	(0.195)		
perc_sent_per_tweet_ch		0.012**	
		(0.005)	
log_sent_per_tw_ch			1.154**
			(0.506)
Constant	−0.317**	−0.120	−0.100
	(0.136)	(0.098)	(0.098)
Observations	427	427	427
Log Likelihood	−293.059	−292.676	−292.464
Akaike Inf. Crit.	590.118	589.353	588.927
Note:			*p<0.1; **p<0.05; ***p<0.01

Furthermore, models were tested on various restricted datasets, revealing little to none descriptive power of the explanatory variables on the price change. Followingly, the direction of the price change was estimated using not only variables derived from Twitter, but also the variables accounting for the past development of the Bitcoin market, yet with the use of the past values only. It was discovered that rather than on values describing the reality of sentiment, the variables explaining the past price trend are better for the model predicting the direction of the price change.

Aside the price indicators, the relationship between the aggregate data from Twitter with and volume traded, or number of trades is measured. The results show that aggregated data from Twitter account for 39 percent of variability of volume traded and for 36 percent of number of trades. Furthermore, better results are obtained when estimating the squared roots of number of trades and volume traded.

Besides estimating the variables extracted from the cryptocurrency exchange, the values of sentiment are estimated using the data from the exchange. Nevertheless, no significant relationship is found by the estimations.

6.2 Results for Fifteen-Minute Intervals

Firstly, variables describing close price and the magnitude of price movements were estimated. When estimating the close price with the use of the of variables derived from Twitter only, weaker relationship was obtained than in the case of one-hour intervals.

Table 5: The results of OLS estimation

	Dependent variable:
	close_price
number_of_tweets	84.901***
	(2.757)
overall_sentiment	−87.997***
	(11.328)
overall_sent_power	0.001***
	(0.0002)
sent_per_tweet	−58,298.580***
	(3,921.371)
sq_number_of_tweets	−6,506.911***
	(180.411)
sq_overall_sentiment	8,582.051***
	(387.916)
sq_overall_sent_power	−92.039***
	(28.059)
Constant	73,957.540***
	(904.251)
Observations	13,430
R^2	0.283
Adjusted R^2	0.282
Residual Std. Error	11,121.630 (df = 13422)
F Statistic	754.977*** (df = 7; 13422)
Note:	*p<0.1; **p<0.05; ***p<0.01

While estimating using only values of variables collected within a given interval, the optimized model accounts for 28 percent of variability of the closing price. As opposed to the one-hour intervals, variables accounting for sentiment power were found significant, yet with small coefficient. Followingly, more variables were added to the model describing reality in past intervals. Lagged values of explanatory variables by lag 5 were introduced to the model. After optimizing the model, described variability has increased to over 42 percent. That is slightly lower number than in the case of one-hour time intervals. Yet the residual standard error remains notably high. Next, the value of close price was predicted with the use of closing price in previous time interval and the variables accounting for the description of price movements and trend. The predicted price changes are too differing compared to the actual price changes, that the predictive power of the model is insufficient. After the variables covering the sentiment development and Twitter activity statistics were introduced to the model, and some proven significant even on 0.001 level of significance, the residual standard error declined, yet only on the scale of tenths of a percent. Thus, the overall ability of the model to precisely predict the close price is still negligible, yet higher than in the case of predicting the close price with the use of past price development values only. Moreover, after comparing the values of real price changes and the estimated price changes, computed as a difference between the fitted value and the close price of previous interval, it was calculated that the model did predict the direction of the price change in 51.2 % cases out of 14 607.

Secondly, multiple models estimating the price change and percentage price change were tested. However, alike in the case of one-hour intervals, the results don't show great explanatory power, even with the use of historical values of the explanatory variables. The interesting thing is, that the variables accounting for the changes in the values of sentiment from previous time period are not found significant, same as in the case of considering the one-hour periods.

Table 6: Comparison of models estimating price change

	Dependent variable:		
	perc_price_ch	log_price_ch	price_ch
	(1)	(2)	(3)
number_of_tweets	0.0003***	0.00000***	0.121***
	(0.0001)	(0.00000)	(0.040)
overall_sentiment	−0.001***	−0.00001***	−0.510***
	(0.0004)	(0.00000)	(0.161)
sent_per_tweet	−0.310**	−0.003**	−63.493
	(0.141)	(0.001)	(55.541)
sq_number_of_tweets	−0.024***	−0.0002***	−7.168***
	(0.006)	(0.0001)	(2.523)
sq_overall_sentiment	0.049***	0.001***	14.666***
	(0.014)	(0.0001)	(5.454)
Constant	0.079**	0.001**	17.062
	(0.032)	(0.0003)	(12.438)
Observations	14,845	14,845	14,845
R^2	0.001	0.001	0.001
Adjusted R^2	0.001	0.001	0.001
Residual Std. Error (df = 14839)	0.433	0.004	170.422
F Statistic (df = 5; 14839)	3.877***	4.122***	2.789**
Note:			*p<0.1; **p<0.05; ***p<0.01

Followingly the price change, or rather the direction of the price change was estimated using logistic regression. It was found, that when estimating the data as whole, the values of sentiment play no significant role in estimating the direction of the price change. However, after restricting the data for the absolute value of percentage price change to be more than 0.5 %, one variable was found to be significant. The variable accounts for whether the sentiment per tweet value was greater than the median value of the whole sample. It was estimated, that when the sentiment per tweet is higher than the median value, the log odds of the positive price change in a given interval rise by 0.21. Furthermore, it was studied whether the sentiment per tweet change from previous periods can estimate the direction of the price change, for data restricted on the size of the sentiment change. It was estimated, that for absolute changes in sentiment per tweet values greater than 15 %, which is, slightly more than half of the observations, the variable denoting the sentiment change is significant on 0.05 level of significance

when estimating the log odds of the direction of the price change and affects it in the same direction as is the sentiment change.

Subsequently, predictive power of direction of the price change on the change in sentiment per tweet was tested. The results show strong relationship. For when the price change in previous period was positive, the log odds of the positive sentiment per tweet change between the actual and the previous periods are higher, and when the price change in previous period was negative, the log odds of the negative sentiment per tweet are higher.

Furthermore, strong relationship was discovered for predicting the direction of price change only by variables accounting for trend of price. In a simple logistic model, where only two explanatory variables are used, first taking the value of 1 when the price change was positive in three consecutive previous periods, and the second taking the value of 1 when the price change was negative in three consecutive previous intervals. The log odds of the positive price change are estimated to be higher by 0.26 when the price change was negative in previous time periods and to be lower by 0.23 when the price change was positive in previous time periods.

6.3 Tested on More Cryptocurrencies

The analyses of supportive datasets including retweets for Bitcoin, Ethereum and Litecoin were conducted. It was observed that the models considering only the data for the original tweets are better than when considering retweets in the case of Bitcoin. When estimating Ethereum, similar results were obtained as in the case of Bitcoin, yet even slightly worse. That might be caused by the fact that not every post addressing Ethereum is actually related to the cryptocurrency, but rather to the technology platform. In the case of Litecoin, a shortcoming was detected as not enough tweets or retweets were posted in a researched period and with low density of the tweets, the data is not found useful to measure the effects of sentiment in selected time frames. However, that might not be problem when considering day to day intervals. The results of the study indicate that cryptocurrencies with lower market capitalization might be proportionally less mentioned on social media platforms. For future research, it is thus recommended to consider longer time intervals for cryptocurrencies with low market capitalization, to preserve relevancy of the data.

Furthermore, signs of dense bot activity as described in related literature were observed.

7. Open Issues, Topics for Further Research

7.1 The Role of Likes, Comments and Retweets

In this study, number of retweets, comments, and likes of the tweet is not taken into account. This might not be relevant in measuring the expressed sentiment of users, however the reactions of other users to the original post may declare the overall attitude to the tweets content. Thus, it may be of importance to include such metrices in future research, when studying the perceived impact of tweets. Nevertheless, as these metrices may vary gradually as new users add their reaction to the tweet, it could arise difficulties when reflecting or determining the tweets impact in time.

7.2 Tackling Bot Accounts

A common problem arising when processing the data might be biased data by bot activity. So called bot is an automated program, that is written by a person to interact with the Twitter environment. Commands to a bot account can be given via the Twitter API. A bot account can be for instance programmed to like or retweet post that contain a certain keyword or hashtag. That post is thus falsely getting relevant, even when no real person actually reacted to it. This is a problem that was addressed in few studies, such as study on Identifying Correlated Bots in Twitter (Chavoshi, Hamooni, & Mueen, 2016). In this study, it was shown that completely unrelated accounts can produce perfectly correlated content, such as posting the same messages in an exact same time, repeatedly. However, this problem is constantly being tackled by the evolution of the bot account detecting algorithms. Further research was conducted on about whether bot accounts impact Twitter activity (Gilani, Farahbakhsh, & Crowcroft, 2017). By setting up bot account, it was shown, that despite the fact that bots account may be in minority, they are able to have more impact than account operated by real people. Techniques to identify abnormally correlated user accounts in Twitter, which are very unlikely to be human operated, were created. The conducted study shows daily reports on bots at a rate of several hundred bots per day (Chavoshi, Hamooni, & Mueen, 2016). Those techniques are recommended to be implemented in future research. However, with the evolution of bot detecting procedures, the bots, or the algorithms that operate them, do evolve as well.

7.3 Topics For Further Research

It was proven that sentiment and metrics derived from social media are useful when estimating, or even predicting the cryptocurrency characteristics even on short time intervals. Forecasting and modelling the prices of various assets have been conducted multiple times in the past. However, work needs to be done on the evolution of sentiment analysis methods. That is necessary because some widely used methods are either outdated, since they do not recognize modern means of expressions, or have other shortcomings, such as insufficient vocabulary in the dictionary. Even more concerning, not yet addressed problem, is the absence of multilingual sentiment analysis means. Almost exclusively, the language they operate in is English.

Furthermore, work needs to be done on the field of how many tweets in a time interval are enough to derive the overall sentiment from. Moreover, to solve the problem of high positive correlation between the number of tweets and overall sentiment in a given time period, only a random sample of given number of tweets might be selected.

Another topic that must be addressed is the influence of the bot accounts on the behaviour, or opinions expressed of real users. Until work on the topic is done, it is not recommendable to include retweets in future studies.

When inspecting the sentiment about a given cryptocurrency, it may be driven by the sentiment towards cryptocurrencies as a whole. Thus, in future research, it is recommended to test for such phenomenon by the inclusion of metrics for activity from social media regarding the cryptocurrencies as a whole.

Conclusion

In summary, the results of the thesis indicate that social sentiment as derived from leading social media does play a role when describing the development of the cryptocurrency Bitcoin. Not only moderate part of the variability of Bitcoin close price in a selected time period is explained by its derivatives, but it helps to predict the price movements as well. Furthermore, when considering one-hour time intervals, at certain conditions, the direction of the price change can be described by the sentiment variables only and with a robust result. In case of using social media activity on shorter time periods, its influence diminishes, along with the predictability power.

Furthermore, it was proven that direction of price change in one time interval, if the absolute value of the percentage price change is high enough, can be a good predictor for the sentiment change between that interval and the following one.

Additionally, the results show that the direction of price change in shorter time periods is more dependent on the trend of directions of price changes than in longer time intervals.

It is recommended that only cryptocurrencies with high market capitalization are considered in the follow-up research, since they are usually not affected by external effects or different market reality. Other reason is, that the density of the data produced on social media about them is not high enough to describe the reality in short time periods. However, for daily intervals, they might still be decent option. Furthermore, not every social media is suitable to derive the data from. As an example, social media Reddit was tested for computations, yet the density of the posts and even comments to those posts was not sufficient to derive the sentiment from on short time intervals, even for largest of cryptocurrencies.

It was proven, that considering shorter time intervals, is possible and in some cases desirable in future studies. That is, because of the high volatility and price changes of cryptocurrencies even in short periods of time. Nevertheless, the drawback is, that by considering shorter and shorter intervals, the distribution of the price changes is on average nearing zero value. As computed in the thesis, the smaller changes in price are almost indescribable by the explanatory variables, as opposed to the greater changes.

Subsequently, new method for deriving the sentiment from the tweets is appropriate to be used for future analyses. The sentiment values should be zero cantered, as the positively biased values cause a vital problem for many computations,

since it cannot be precisely determined, whether the sentiment was positive or negative. New metrices of sentiment variables were introduced in this thesis to tackle this problem.

Followingly, when considering the data including retweets, it is necessary to clean the dataset from the tweets posted by bot accounts. In past research, it was proven that bot accounts, or accounts that are not operated directly by a human user, but rather operate as an algorithm, can influence sentiment significantly in the period they operate in.

A large disadvantage needs to be pointed out when considering the use of the social media data for deriving insights for future market development. That is the limited amount of data one can obtain by free of charge methods. Furthermore, the larger the data samples, the more time it takes to derive sentiment from, making it difficult to make real time decisions based on the data.

List of References

Bech, L. M., & Garrat, R. (2017, 9). Central Bank Cryptocurrencies. *BIS Quarterly Review*.

Candila, V. (2021, 9). Multivariate Analysis of Cryptocurrencies. *Econometrics*, p. 28.

CoinMarketCap. (2022, 11). Retrieved from https://coinmarketcap.com/

Corbet, S., & Yarovaya, L. (2020). The environmental effects of cryptocurrencies. *Cryptocurrency and Blockchain Technology*.

Corelli, A. (2018, 6). Cryptocurrencies and Exchange Rates: A Relationship and Causality Analysis. *Risks*, p. 111.

DataReportal. (2022, 11). Retrieved from DataReportal: https://datareportal.com/social-media-users

Foley, S., Karlsen, J. R., & Putnins, T. (2019). Sex, drugs, and Bitcoin: How much illegal activity Is financed through cryptocurrencies? *Review of Financial Studies*, pp. 1798-1853.

Giachanou, A., & Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv. 49*.

Gilani, Z., Farahbakhsh, R., & Crowcroft, J. (2017). Do Bots impact Twitter activity? *Proceedings of the 26th International Conference on World Wide Web Companion*, (pp. 781–782).

Gui, H. (2019). Stock Prediction Based on Social Media Data via Sentiment Analysis, A Study on Reddit. *Faculty of Information Technology and Communication Sciences*.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *8th International AAAI conference on weblogs and social media.*

Chan, S., Chu, J., Nadarajah, S., & Osterrieder, J. (2017, 10). A Statistical Analysis of Cryptocurrencies. *Journal of Risk and Financial Management*, p. 12.

Chavoshi, N., Hamooni, H., & Mueen, A. (2016). Identifying Correlated Bots in Twitter. *Social Informatics*.

Inamdar, A., Bhagtani, A., Bhatt, S., & Shetty, P. M. (2019). Predicting Cryptocurrency Value using Sentiment Analysis. *International Conference on Intelligent Computing and Control Systems*, (pp. 932-934).

Jethin, A., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review*.

Jianqiang, Z., & Xiaolin, G. (2017). Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access*, pp. 2870-2879.

Kouloumpis, E., Wilson, T., & Moore, J. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the International AAAI Conference on Web and Social Media*, (pp. 538-541).

Liu, J., & Serletis, A. (2019). Volatility in the Cryptocurrency Market. *Open Econ Rev*.

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*.

Meena, R., & Gama, J. (2013). Marketing research: The role of sentiment analysis. *Universidade do Porto, Faculdade de Economia do Porto, No. 489*.

Mittal, A., & Arpit, G. (2011). Stock prediction using twitter sentiment analysis. *Standford University, CS229*, p. 15.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *Proceedings of the ACL Student Research* (pp. 43–48). Association for Computational Linguistics.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. M.I.T. Press.

Szalay, E., & Venkataraman, S. K. (2021). What are cryptocurrencies and stablecoins and how do they work? *Financial Times*.