FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Bitcoin exchange rate prediction using Twitter and Google Trends

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Vanja Ivljanin
Matrikelnummer 1428439

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Wien, 11. November 2019

Vanja Ivljanin

Wolfdieter Merkl

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Bitcoin exchange rate and trading volume prediction using Twitter and Google Trends

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Business Informatics

by

Vanja Ivljanin
Registration Number 1428439

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Vienna, 11th November, 2019

Vanja Ivljanin

Wolfdieter Merkl

Erklärung zur Verfassung der Arbeit

Vanja Ivljanin
Pichlergasse 2, 1090 Vienna

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 11. November 2019

Vanja Ivljanin

Acknowledgements

First of all, I would like to thank professor Dieter Merkl for his dedication, inspiration and continuous feedback.

Special thanks goes to my husband Ivan for his unconditional support and understanding during long nights of research and writing.

Furthermore, I would like to thank my parents, Lidija and Jovan for their constant support, not only during writing this thesis, but throughout my entire study period.

And finally I would like to thank my friends Nemanja, for all of those tasty meals during our master studies and Jovana for helping out with German translation.

Kurzfassung

Die Allgegenwärtigkeit des Internets macht es zu einer wertvollen Datenquelle. Produkte wie Twitter und Google Trends implementieren APIs die es erlauben diese Daten zu nützen. Da diese zwei Produkte große Mengen von Nutzerdaten speichern, kann man Sie nützen um Vorhersagen über das Benutzerverhalten in bestimmten Bereichen zu machen. In dieser Arbeit, beschreiben und implementieren wir statistische Method um Vorhersagen über den Bitcoin Wechselkurs mittels Stimmungsanalyse, der Frequenz von Tweets über Bitcoin und dem Interessenniveau zum Suchbegriff Bitcoin. Weiters zeigen wir wie das stündliche Handelsvolumen von Bitcoin mittels der stündlichen Rate von Tweets über Bitcoin und Google Trends Daten vorhergesagt werden kann.

In dieser Arbeit analysieren wir Daten aus dem Zeitraum Juli 2018 bis August 2018, d.h. zwei Monate. Erstens benützen wir die stündlichen Informationen über Tweet Stimmung, Tweet Frequenz und Google Suchdaten über Bitcoin um eine kausale Verbindung zwischen diesen Daten und dem aktuellen Preis von Bitcoin. Zweitens entwickeln wir ein vektorautoregressives Modell (VAR) dass erfolgreich den Bitcoin Wechselkurs und das Bitcoin Handelsvolumen vorhersagt. Weiters, verwenden wir zwei unterschiedliche Algorithmen, SentiStrength and Stanford Core NLP, um Stimmungsinformation zu erhalten. Wir zeigen dass mit dem SentiStrength Algorithmus welcher für kurze Nachrichten wie Tweets geeignet ist unser Modell bessere Resultate erzielt. Spezifisch, zeigen wir dass unser Modell wenn kombiniert mit SentiStrength eine Genauigkeit von 63% für Vorhersagen bezüglich dem Anstieg und Abstieg des Bitcoin Wechselkurses aufweist. Auf der anderen Seite, wenn man unser Modell mit Stanford Core NLP kombiniert erhalten wir eine Genauigkeit von 59%. Drittens, benutzen wir die Tweet-Frequenz und Suchinformationen um das Bitcoin Handelsvolumen mit einem mittleren quadratischen Fehler von 172.76 vorherzusagen. Zusammenfassend, zeigen wir in dieser Arbeit dass Informationen von Twitter und Google Trends sehr hilfreich sind wenn es um die Vorhersage zwei wichtiger Bitcoin Eigenschaften, Wechselkurs und Handelsvolumen, geht.

Abstract

Ubiquity of the Internet has made it a valuable data source, and products such as Twitter and Google Trends implement APIs that make that data available to us. Because these two services capture an abundance of data about their users, they can be used to predict user behavior in some areas. In this work, we describe and implement statistical methods that can be used to predict the bitcoin exchange rate by using sentiment analysis, frequency of tweets related to bitcoin, and level of interest in search term bitcoin. Also, we show how bitcoin hourly trading volume can be forecast by using hourly number of tweets about bitcoin and Google Trends data about bitcoin.

We analyze data for period of two months during July 2018 and August 2018. By observing hourly information about tweet sentiment, tweet frequency, and Google searches on bitcoin, we first show causality between these time series and impact of historical values on the current values. Also, we build a vector autoregression model that is used to successfully forecast bitcoin exchange rate and bitcoin trading volume. Furthermore, two different algorithms, SentiStrength and Stanford Core NLP, are used in order to extract the sentiment information. We show that SentiStrength which is optimized for short messages such as tweets does perform better. The model we developed using SentiStrength sentiment algorithm achieves accuracy of 63% in predicting increase or decrease in bitcoin exchange rate. On the other hand, VAR model with Stanford CoreNLP algorithm achieves an accuracy of 59%. Also, bitcoin trading volume is predicted by using tweet frequency and search information with root mean square error of 172.76. Our work shows that Twitter and Google Trends is very useful when predicting two important bitcoin properties, the exchange rate and trading volume.

Contents

CHAPTER 1

Introduction

1.1 Problem statement

Since decades people wanted to know what is the best way to invest their money. Minimizing uncertainty and maximizing one's profit was always important for investors. Many studies and analyses were done about investing in stock markets [BC87, Dam12, FSS93, Gor62, MM58, PMS13]. Before the Internet, market news and rumors, either false or true, needed weeks or sometimes even months to get proven [RS12]. With the emergence of Web 2.0, the way people communicate transformed. Thoughts, ideas and information can now be shared in seconds to a very large audience. However, people cannot always rely on what they read on the Internet [AG17]. Regardless of whether something is the truth or a lie, the information that is shared can contain good or bad sentiment about individual products or persons, or even the whole company [RS12].

Predicting stock market behavior was never an easy job. Some of the techniques that are used to predict market fluctuations are trend analysis, historical data analysis and cycle analysis [BC87]. Analysts may combine multiple techniques in order to obtain more reliable results. In addition to that, it is shown that macroeconomic factors can be used to forecast the stock market returns. These include industrial production index, interest rates, inflation, bonds [CRR86]. Nonetheless, bitcoin is a system of payment without any connection to these macroeconomic variables nor the monetary system. Cryptocurrencies do not have any real coverage behind, and supply and demand are the ones affecting the price. Therefore, many of the input signals that were used to analyze markets cannot be used now, and new techniques should be considered [MLM15].

It is very well known that social media can highly influence markets on economic level [BMZ11]. Microblogging is a new popular way of communication on social media.

People can share their thoughts, feelings and news in short messages. Status messages can be shared with followers instantly via mobile application or web [JSFT07]. One of the most popular microblogging networks nowadays is Twitter, and it has grown a lot since its launch in October 2006. It has over 320 million monthly active users. [1] Over the years Twitter has become an important tool in the business world. Individuals are sharing information as well as ideas about investment decisions.

One of the areas that have been recently gaining investors' attention is cryptocurrencies. Cryptocurrency is a kind of digital currency. It can be used in exchange for physical products or services. Cryptocurrency uses cryptography to control and secure transactions and creation of additional units [Chu15]. One of the most popular cryptocurrencies nowadays is bitcoin. Bitcoin is an electronic currency system, which is not managed by any bank or government. Everyone can use it no matter where in the world they live [RS13]. Large numbers of users are interested in its specific blockchain technology that contains a history of all public transactions [MLM15, Nak08].

Fluctuations in the bitcoin exchange rate were interesting ever since bitcoin appeared [Kam14]. It does not have a fundamental coverage and its exchange rate is very volatile. Because of that, it cannot be described with standard financial and economic theories. Its price can switch from surge to plunge very fast, and it is very hard to discover what is influencing its value [GTMP14]. In general, currencies are standard economic goods, which value is determined by supply and demand on the market. Macroeconomic variables like inflation, interest rates, GDP and others can influence market movements and thus a value of the currency, which is issued by an institution or country. These fundamentals of the market are missing when it comes to digital currencies because demand for digital currency cannot be determined by expected macroeconomic regulations. Because there are no interest rates on digital currencies, profits can be achieved only by recognizing when it is a good time to buy it, holding it and selling it later. Speculators and short- term investors are dominating the digital currency market. Consequently, investors' sentiment is now very important, as it is driving the investment decisions, and directly impacting the bitcoin exchange rate and trading volume.

This work focuses on analyzing data from Twitter and Google Trends in order to successfully predict two important bitcoin financial parameters, exchange rate and hourly trading volume. In terms of exchange rate, we will employ sentiment analysis to try to forecast the bitcoin exchange rate. By using automated sentiment analysis methods, we will try to discover if positive sentiment causes an increase in the bitcoin exchange rate and if negative sentiment leads to a decrease in the exchange rate. With regards to the second financial parameter that we would like to analyse, hourly trading volume, we will look into the number of web searches on bitcoin and the number of tweets that refer to

[1] https://investor.twitterinc.com/static-files/ (Accessed December, 2018)

bitcoin. We will explore how changes in these two metrics correlate to the bitcoin hourly trading volume.

Research questions:

- Can we successfully predict increase in the bitcoin exchange rate if the sentiment of tweets obtained from Twitter is positive?
- Correspondingly, can we predict decrease in the bitcoin exchange rate if the sentiment is negative?
- In addition to that, can we leverage the information from Google Trends and number of tweets to successfully predict bitcoin hourly trading volume?

1.2 Aim of the work

In this research, we will analyse whether information that is collected from the web search and social media activity can help investors in predicting bitcoin market parameters. Data from Twitter and Google Trends will be analysed. We will employ sentiment analysis in order to automatically process tweets and predict bitcoin exchange rate. Additionally, we will quantify number of web searches for bitcoin and number of tweets, and try to associate that information with the bitcoin hourly trading volume.

Online opinion mining has become an integral part of daily business decisions. Because of the number of its users, Twitter is an invaluable source for such analysis. However, messages are limited to 280 characters, and their length and structure posed challenges for the traditional sentiment analysis algorithms [SB17]. In recent years, advances in sentiment analysis have provided us with tools capable of successfully analysing content of tweets [PP10]. Sentiment analysis has been successfully used to predict stock market movements. In particular, application of sentiment analysis methods proved to be a reliable indicator of Dow Jones Industrial Average index, and analysing Twitter feeds has shown that overall population mood does correlate with the stock market changes [BMZ11].

In this work, by analysing tweets that refer to bitcoin, we are expecting to show that sentiment analysis can be used to successfully predict the bitcoin exchange rate. Overall positive sentiment will indicate an increase in the bitcoin exchange rate, while the negative sentiment will indicate a decrease in the exchange rate.

The web search data can be used to predict stock market daily trading volume. By using search frequency data from a search engine, researchers were able to forecast the daily trading volumes for NASDAQ-100 securities. They show that search trends originate from unorganized actions of many users, and it is confirmation of a hypothesis that general population actions are correlated with the stock market behavior [BBC+12]. Kristoufek in his research shows that the frequency of searched terms that are connected with digital currencies can represent the global interest in that currency [Kri13]. We are able to access information how often a term is searched for in correlation to the total number of searches on Google Search with Google Trends [2], and in this work we will correlate that information with the bitcoin trading volume.

In this research, we expect that examining popularity of web searches that refer to bitcoin and the number of tweets that contain the reference to bitcoin can predict the bitcoin hourly trading volume. An increased quantity of searches and tweets indicates rise in the bitcoin daily trading volume. Similarly, a decrease in quantity predicts drop in the trading volume.

1.3 Methodological approach

In order to answer our research question whether we can use sentiment analysis to successfully predict bitcoin exchange rate and also whether information from Google Trends and number of tweets can be used to predict bitcoin hourly trading volume, different approaches will be used.

- Data collection

  Three types of data will be necessary for the research. First, tweets will be gathered from the Twitter Application Programming Interface (API) [3]. All tweets that contain '#bitcoin' will be collected in the period between 1st July 2018 and 31st August 2018. Furthermore, frequency of searches for the term 'bitcoin' on Google Search will be gathered directly from Google Trends. Bitcoin market data, exchange rate and hourly trading volume, will be collected from `https://www.bitstamp.net`.

- Data analysis

  After the data collection phase, we will use quantitative analysis to better understand behaviour on social networks and the Internet, and how it can influence bitcoin exchange rate and trading volume. Data will be analysed as follows:

  – Automated sentiment analysis will be done on tweets containing the hashtag '#bitcoin'. There are different techniques on how to measure public mood directly from social media and to evaluate whether the message sentiment is positive or negative [PP10, KWM11]. In this research we will use two different sentiment analysis methods to process tweets. The first one is the Stanford

[2] `https://trends.google.com` (Accessed December, 2018)

[3] `https://developer.twitter.com/en/docs/api-reference-index` (Accessed December, 2018)

CoreNLP toolkit, set of tools specialized in the natural language processing with support for sentiment analysis [MSB+14]. The second algorithm that we will use was developed specifically to cope with short messages and informal spelling of words. It is called SentiStrenght, and it showed good results when analysing short messages from a social network site [TBP11]. Following that, results from both methods will be compared with the bitcoin exchange rate. We will apply statistical methods to find correlation between changes in the sentiment and changes in the exchange rate.

- Going further we will concentrate on finding correlation between bitcoin hourly trading volume and general interest in bitcoin. We consider general interest to be comprised of the number of web searches for term 'bitcoin' and number of tweets that contain '#bitcoin' hashtag. The former one is obtained from Google Trends, while the latter is obtained from Twitter API. By applying data analysis methods we will correlate changes in general interest in bitcoin with changes in the bitcoin hourly trading volume.

1.4 Structure of the work

In **the first section** of the thesis we will investigate fundamentals of cryptocurrencies, what is causing considerable interest of public and what are the reasons for its high volatility. Furthermore, insights about stock markets and their movements will be explored. This will be followed by examining cryptocurrency markets. We will explain how social media works, how it influences the economy, and how those influences can be used. Social media platform that we will focus on is Twitter, one of the largest microblogging platforms nowadays with 320 million monthly active users. [4] [5] Furthermore we will explain what is Google Trends and how it will help us collect more insights in regards to web search media. Sentiment analysis methods will be analysed, and we will try to examine multiple methods in order to find the one that shows the best results with short messages such as tweets.

In **the second section** of the thesis previous research works related to the topic will be analysed. This will represent the basis for future research. It will enable us to get better insights about how social networks and web search engines work. Firstly, we will focus on how Twitter has been used in order to analyse public opinion. This will be followed by review of studies that tackle sentiment analysis of microblogging messages on Twitter. We will also research sentiment analysis methods, especially the ones that work well with the short texts. Furthermore, we will investigate existing approaches about stock market prediction using Twitter and Google Search queries. Detailed research about cryptocurrency markets will be done, with emphasis on bitcoin. Finally, previous work connected to this topic will be discussed.

[4] https://www.alexa.com/topsites (Accessed December, 2018)
[5] https://investor.twitterinc.com/static-files (Accessed December, 2018)

In **the third section** of the thesis firstly we will focus on gathering all the necessary data for the period of two months. This means that data will be collected from various platforms - Twitter [6], Google Trends [7], and Bitstamp [8]. From Twitter we will collect tweets that contain hashtag '#bitcoin'. These will be used for the sentiment analysis, and bitcoin exchange rate prediction. From Google Trends we will assemble data that is connected with search term bitcoin. Together with the collected tweets, this will be used to predict the hourly trading volume. Following that, we will use Bitstamp to collect actual bitcoin exchange rate and hourly trading volume. After all of the data has been gathered we will apply sentiment analysis and data analysis techniques to process the inputs, and to compare them with the actual values. Relevant data will be described and presented visually.

In **the fourth section** of the thesis, conclusion of the study will be presented. Limitations of the study will be considered, both the ones from the data collection and data analysis point of view. Following that, additional areas of improvement and suggestions for future research will be discussed.

[6] https://developer.twitter.com/en/docs/api-reference-index (Accessed December, 2018)

[7] https://trends.google.com (Accessed December, 2018)

[8] https://www.bitstamp.net/ (Accessed December, 2018)

CHAPTER 2

Theoretical Background

2.1 Digital currencies

A lot of things we use in our daily lives have gone digital. You can pay bills online, get a variety of exotic meals delivered to your home, watch the latest episode of your favorite show whenever you want or even get your haircut done without ever leaving the comfort of your home. If someone would pay attention to you while riding in a cab, they would not be able to spot when you actually paid for the ride. We became completely dependant on technology because it makes life easier. Very frequently we use our smartphone for the simplest tasks like playing our favorite playlist, finding the fastest route to bike to work, or exploring cheapest prices for summer destinations. As you can see, the world is rapidly changing, and so is one of the things we are using daily: money. Some countries have already been switching to cashless systems. For example, in Sweden, 59% of all transactions are completed without cash. In the United Kingdom 52% of all transactions are done contactless. [1] The world is moving away from the physical money, towards money that is stored electronically. This follows a general trend of digitalisation. In wake of these changes, digital currencies captured the interest of the public. In contrast to physical currencies, these currencies are stored electronically i.e. they are in digital form. Some examples of digital currencies are virtual currencies and cryptocurrencies. As banknotes and coins, digital money can also be used for buying goods. However, its use can be restricted e.g. virtual money used for payments in video games. But with cryptocurrencies you can also pay for products and services. Fiat currencies are the ones the majority of the population is using the most, but there are alternative currencies which represent different mediums of exchange. There are different reasons why alternative currencies are created. According to Hileman there are two main categories of alternative currencies: tangible and digital [Hil14]. They can be further divided into:

[1] `https://www2.deloitte.com` (Accessed May, 2019)

- Currencies with intrinsic utility - their value is not abstract and also they do not depend on any government or monetary instruments. Example of this would be how soldiers used metals or cigarettes for trading in Berlin after the Second World War.

- Token - this medium of exchange is usually driven by some social agreements or geographic location. Some of the examples can be local currencies like Brixton Pound in the United Kingdom. It is designed to encourage local production and trade.

- Centralized digital currency - good examples are loyalty points from various companies, air miles or currencies that are used in closed systems like games e.g. World of Warcraft Gold.

- Distributed and/or decentralized digital currency - all cryptocurrencies fall into this type. There is no legal entity or the government who can control their distribution. All transactions can be done without the third party due to blockchain technology. Some of the examples of cryptocurrencies are Bitcoin, Litecoin, Ripple, Ethereum [Hil14].

According to Nian et al., demand for alternative currencies is driven by various socioeconomic forces [NC15]. Some of them are:

- Technology - Our lives are becoming more digital and so is paying. It has become much more comfortable for us to pay via cards and online than in cash.

- Political economy - Economic uncertainty is in constant growth. Traditional banking is constantly facing possibilities to collapse and to cause financial crises.

- Environmentalism - There is a constant open question when we are going to reach the maximum point after which we will not be able to extract natural resources like oil.

- Inefficiencies - Most of financial services are overpriced.

- Financial freedom - Digital currencies will enable users to bypass big financial services and to avoid large fees they impose. Since they are not controlled by any government, they can provide security to users when there are fiat currencies crisis. [NC15].

Nowadays it is very easy to create a cryptocurrency and to make it available to everyone. Advancements in technology made it possible to do this almost for free. However, since many alternative currencies are in competition, not all of them will be adopted globally. There is a possibility that many of them will fall into oblivion very shortly after appearing on the market.

2.1.1 Cryptocurrencies

A cryptocurrency is a type of digital medium of exchange. In order to keep the creation of new units under control, verify and secure financial transactions, cryptocurrencies rely on cryptography. In comparison to centralised banking system which is under the control of state and governments, cryptocurrencies are completely decentralised. This means that cryptocurrencies are not connected to any monetary system, they do not have real coverage behind like gold and the only thing that affects the price is pure supply and demand [MLM15, CNC15, GTMP14, Kri13].

Beginning of cryptocurrencies goes back to 1983 when anonymous cryptographic electronic money - “eCash” was presented by Chaum [Cha83, CFN88]. The idea behind it was to have software which will store money locally in a digital format, but it required possession of the cryptographic signature from the bank. Cryptographic protocols were used to prevent double spending and also to keep users anonymous. This meant that the person who was using eCash could spend their digital money at the merchants who accepted “eCash” or on the Internet. Chaum’s company was called Digicash and because only this company was managing it, it was considered a centralised system. During the 90s, banks in the United States and Europe (Austria, Germany, Switzerland) were using this system for electronic payments. However, because of rising credit card popularity, Digicash went bankrupt in 1998 [NC15, Sch97]. Over the years that followed cryptocurrencies slowly developed into the form that are familiar to us today. Some of them are:

- E-Gold which was the first digital gold currency. Users could register on their website where their money was appointed in grams of gold. They could do instant transfers to other users of the platform. It was established in 1996 and lasted up to 2009 when it was closed due to legal issues. During its existance, it had 2 billion US dollars of spend per year. It was founded by Douglas Jackson and Barry Downey [NC15].

- Hashcash was first used to reduce spam emails and denial-of-service attacks. It was introduced in 1997 by Adam Back. It uses a proof-to-work algorithm that requires some amount of work to be computed which can be easily verified afterwards. Its algorithm is hash-based and cryptographic. Nowadays it is used as a proof-of-work for bitcoin mining [NC15, Bac02].

- B-money is predecessor of the bitcoin. It was introduced by Wei Dai in 1998. He proposed two protocols for "anonymous, distributed electronic cash system”. Even Satoshi Nakamoto was referencing Dai’s work when creating Bitcoin. Dai wanted to present paying system which will be used only between two concerned parties without need for trusted party to overlook. Dai’s concept has actually never been realised and it remained only a proposal [NC15, Dai98].

As the 2008 financial crisis came, interest in cryptocurrencies started to rise again. At that point of time it seemed that cryptocurrencies can offer solutions to issues that have arisen with fiat currency system. It was believed that cryptographic digital payment system could eliminate trusted third party from the online payments. Central banks were allowed to print money without coverage during crises, which caused inflation. With regards to that, cryptocurrencies offer limited supply and new coins are created only by predetermined rules. So no one could produce more than it was planned at the beginning. Huge government debts, people that could not repay their mortgages and stock market drops brought cryptocurrencies to attention even more. Trust in banks and governments was shaken. Hence, people wanted to consider alternative ways to save and invest their assets [NC15].

Cryptocurrencies provide a way to verify all transactions. This is accomplished by using distributed ledger. A distributed ledger can be represented as a database network that can be shared across multiple sites, institutions or countries. There is no centralised data storage nor central database administrator. Since everyone in the network will have their own identical copy of the ledger, each change to the database will be shown almost immediately in all copies of the ledger. Different types of data can be stored in the ledger like financial, physical and electronic assets. The security of stored data in a shared ledger is maintained by cryptography. Keys and signatures are used to control actions in the ledger. Depending on the rules of the network, some or all participants can make changes to the data in the ledger. [2]

One of the most widely used distributed ledgers is the blockchain, but mentions of cryptographically secured chain of blocks go back to 1991, when Stuart Haber and W. Scott Stornetta devised a system with capabilities to create a secure and trustworthy timestamp that cannot be changed [HS90]. After a couple tries to modify it and make it more efficient, first blockchain was created in 2008 by Satoshi Nakamoto [Nak08]. It is still not known whether Satoshi Nakamoto is one person or group of people under the pseudonym. They improved the design by adding Hashcash method for adding blocks to the chain and with this there was no need for a trusted party to sign for every block that was added to the chain anymore. No trusted authority nor central server was needed. The main purpose of creating a blockchain was to serve as a public transaction ledger of the bitcoin. This was the first time that double spending problem was solved only by using cryptography [NBF+16]. Blockchain is contained of linked records that can be added one after another. Records are called blocks and they are linked using the cryptographic function. Blockchain is designed in the way to resist any data modifications. Each of the blocks in the blockchain contains transaction data accompanied by a timestamp and cryptographic hash of the previous block [NBF+16]. Once block has been inserted in the chain, data cannot be altered without changing all previous blocks. This cannot be done without consensus of the network majority as, blockchain is using peer-to-peer network for validation of the new blocks and communication between the nodes. That is why blockchain is considered to be decentralised [NBF+16]. It is basically a shared database of records where all transactions that have been performed are distributed among all participating parties. Consensus of the majority of participants in the network

[2] `https://assets.publishing.service.gov.uk` (Accessed April, 2019)

is verifying each transaction that has been made. Once accepted by the network, any record that is part of the chain cannot be altered or erased. This means that all records of transactions are securely stored in the blockchain [CPV+16].

These features make Bitcoin, and other cryptocurrencies, more safe for its users. Digital payments are often connected to the trusted third parties. Because there is a high probability of frauds, payment transaction fees are high. Cryptocurrencies, on the other hand, use cryptography instead of the trusted third parties. This means that fees for such transactions will be much less compared to traditional ones where third parties are involved. Each transaction that is done over the blockchain will contain digital signature [CPV+16]. Digital signature consists of a public key and a private key. Every transaction needs to be sent using the public key to the receiver. In addition, every transaction has been previously digitally signed with the private key of the sender. In order for the sender to spend money, he or she needs to prove that they are the owners of the private key. Receiver verifies if the public key on the transaction corresponds to the private key of the sender. All transactions are broadcasted to every node in the network. After a transaction has been verified it will be recorded to the public ledger. However, before the immutable transaction data is persisted, it needs to be verified. Nodes that are verifying transactions have to make sure that spender actually owns that cryptocurrency and that spender has enough cryptocurrency in their account for that particular transaction [CPV+16]. Problem of double spend is arising, because not all of the transactions come in the same order they are generated. So now the system needs to make sure that all transactions are all linked chronologically. All transactions considered to happen at the same time are gathered in the same block. Every block has a reference to the block added to blockchain before it, which implies that the transactions in that block happened after the transactions in the referred block. There is still a problem of how the network decides which of the blocks will be next in the chain because any of the nodes can create a block from still unconfirmed transactions. Order of the nodes is not reliable in this case because different blocks can come from different nodes at the same time. This problem is solved in the blockchain technology by using mathematical puzzle. Each of the blocks will be added to the blockchain only after solving specific mathematical problem - “proof of work”. Node in the network that has created a block needs to prove that it has put enough computing resources to solve mathematical problem. Usually it would take about ten minutes for a node to solve the puzzle. First of the nodes who solve the mathematical problem will share block to the rest of the network. However, it is possible that more than one block is broadcasted at the same time. In that case, both blocks will be accepted and more branches will be formed. Because mathematical problems are rather complicated, it is very unlikely that both branches will continue to progress with the same pace, and one of them will eventually become longer. Once that happens, only the longest chain is valid and accepted by the network. Because of that, it is impossible for an attacker to create a fraud. An attacker would have to generate a fraudulent block by solving the math problem and to generate subsequent blocks faster than good nodes in order to make the branch containing the bad block the longest. This becomes even more difficult because all of the blocks are cryptographically linked [CPV+16].

The double-spending problem is very well known problem in digital payment systems. It refers to a problem where the same money can be spent more than once. For example A would like to send 100 euros to B. A can make the transaction, and B will receive 100 euros. The transaction is executed by A by creating a copy of 100 euros, and sending that copy to B. However, there is no guarantee that A reduced the amount of money in their account. Hence, A is able to spend them again. On the other side, B already received the transaction. At the end of this transaction, the entire system will have 100 euros more than at the beginning, and this difference does not have any real coverage behind it. This issue is solved by trusted third party who will manage the transaction properly, e.g. a bank. By providing this service, the third party will take a percentage of the transaction to cover their costs. And this simple transaction becomes more expensive [NC15]. The blockchain technology and entire network verification have made digital transactions of cryptocurrencies secure and cheaper by avoiding the need for a trusted third party that controls the transactions. Hence, after every successful verification of transaction, only one record will be added to the blockchain, and that record cannot be changed or removed. All new transactions afterwards will be checked against the already existing records in the blockchain and therefore it is not possible to spend the same money twice [NC15].

In order for a new block to be created, miners have to solve a mathematical problem which represents the proof of work. Participants are racing against each other to solve the puzzle. First participant who successfully decodes it, gets a reward in the form of bitcoin. With solving the proof of work, record of transactions is created. This record cannot be changed without redoing the proof of work [NC15].

According to Crosby et al., digital economy will force all industries to welcome blockchain technology [CPV+16]. In the world where we already use big data to predict behaviour and to improve e-commerce, blockchain can become new “growth engine” of digital economy and the industry itself. Even though there is a lot of controversy about the bitcoin, blockchain technology that stands behind the bitcoin cryptocurrency had worked without any problems. That is why it has found a wide range of usage in both financial and non-financial sectors [CPV+16]. Financial sectors used to see cryptocurrencies and blockchain as competition and threats to traditional business models. Nowadays financial institutions begin to see the benefits and potential usage of blockchain technology in financial sectors. The world’s biggest banks are becoming open to opportunities that blockchain might bring [CPV+16]. As Crosby et al. analyse in their paper, usages can be very different. For example blockchain can be used to securely store stocks, bank account balances, bonds, derivatives and mortgages. Secondly, both physical or digital assets can be stored in the blockchain. For example houses, cars, laptops and other valuables that cannot be easily destroyed or replicated can be saved in the blockchain. In this way ownership can easily be proven and transaction history can also quickly be verified. Likewise, one of the bitcoin exchanges that is based in the US is using blockchain to reduce time that is necessary to settle and clear financial transactions. The result is reflected in the reduction of time needed for transaction from a few days to just a few minutes [CPV+16]. Smart contracts were first introduced in 1994 by Nick Szabo [Fra14]. The idea behind was that contracts automatically execute when all conditions are met. However, this did not find usage until the invention of blockchain and cryptocurrencies. Now contracts are able to execute fully or partially without human interaction. When conditions in smart contracts are met, software behind it can automatically make payments to all involved parties. In this way payments become more transparent. However, legal status of smart contracts is still not defined and therefore their usage is still not widespread. [Fra14] Furthermore, there are many opportunities in the non-financial areas as well. Legal services, health institutions, private securities and even music industry can benefit from using the blockchain. For example document authenticity can be verified by using blockchain technology. In this case there will be no more need for centralised authority to control all the documents. Blockchain would provide trustworthy Proof of Ownership, Proof of Existence and Proof of Integrity of any document [Swa15]. This could not be forged and no one could be bribed, hence, this could provide us with a more trustworthy system than we have now. In the music industry, information of all music rights could be stored in a distributed database. In that way all artists, labels, songwriters and producers will be able to receive fair payments. This will enable more transparency who should be reimbursed, because this became really complex with the emergence of the Internet and streaming services [CPV+16]. Even some of the computer games are based on blockchain technology. One of the first such games was “Huntercoin” [3]. A couple of years after “CryptoKitties” [4] were introduced. Going further, it can even be used for decentralized voting [TT16].

Currently there are three types of blockchain: private, public and consortium blockchain. In private blockchain systems only nodes that come from one specific organisation are allowed to take part in the consensus process. Because all nodes come from the same organisation this is considered a centralised network. Unlike private blockchain, in public blockchain all public nodes can participate in consensus process. That is why this is defined as decentralised network. In between those two types, there is consortium blockchain which implies only pre-selected nodes to take part in consensus process. Nodes from several organisations can participate in this network and that is why this is specified as partially decentralized. According to Zheng et al. Table 2.1 shows how three types of blockchain are compared [ZXD+17].

One of the downsides of cryptocurrencies is that its number is rising daily. Currently there are approximately 2022 different cryptocurrencies. [5] New cryptocurrency can be created anytime. Because there are so many of them, the majority will not be able to stay long on the market [NC15]. If the coins are mined too early and quickly there is higher possibility that they will be shorter on the market and that eventually they will disappear. Miners will not be interested in that cryptocurrency anymore, because the coin itself is not a reward. Different kinds of rewards can be introduced, like transaction fees, but

[3] https://www.huntercoin.org (Accessed May, 2019)

[4] https://www.cryptokitties.co (Accessed May, 2019)

[5] https://www.coinlore.com (Accessed March, 2019)

Property	Public blockchain	Consortium blockchain	Private blockchain
Consensus determination	All miners	Selected set of nodes	One organization
Read permission	Public	Could be public or restricted	Could be public or restricted
Immutability	Nearly impossible to tamper	Could be tampered	Could be tampered
Efficiency	Low	High	High
Centralized	No	Partial	Yes
Consensus process	Permissionless	Permissioned	Permissioned

Table 2.1: Comparison of private, public and consortium blockchain

again it is disputable how much would miners be interested in that kind of reward. On the other hand, that approach would definitely increase the price of transaction. Going further, blockchains without users are more susceptible to attacks. This can happen when interest in certain cryptocurrency goes down and subsequently the number of miners also drops. This allows bad nodes to cooperate and to attack the network. In that case they can create different longest chain, which will cause doubts in the trustworthiness of the cryptocurrency itself [NC15].

2.1.2 Bitcoin

Bitcoin was first introduced in 2008 in the paper "Bitcoin: A peer-to-peer electronic cash system" by Satoshi Nakamoto [Nak08]. Up until now, identity of Satoshi Nakamoto has not been revealed. It is also not known whether only one person or group of people are standing behind the pseudonym. There are suggestions that name might be composed of names of four big technology companies: SAmsung, TOSHIba, NAKAmichi, and MOTOrola. However this has never been proven [Wal11]. Bitcoin represents one of the most popular examples of cryptocurrencies today. System on which this cryptocurrency is running is open source, decentralised peer-to-peer network. Besides that, all nodes in the network are connected to each other which means that the network is fully distributed. However, in spite of this transparency, there is still a lot of mistrust in bitcoin and other cryptocurrency systems. This happens because typically they are not connected to any legal entity, it is not always clear who stands behind them, and quite frequently it is not obvious how they operate without a trusted third party [NC15]. The doubt grew even more after the shut down of one of the world's biggest bitcoin exchanges, Mt. Gox. Mt. Gox operated from 2010 when it was launched until 2014 when it declared bankruptcy because 850.000 bitcoins were missing or stolen [DW14]. However, exchanges that trade bitcoin are not the part of the bitcoin system itself, so that is why it should not be correlated. Besides, cryptocurrency system is a complex system, with no financial regulations controlling it. Its complexity is hard to be understood even by professionals. This is why this area is of special interest for researchers from different fields like financial or technology fields and also for investors who see the opportunity to quickly raise their gains. Hence, the bitcoin attracts much attention around the world. In their research Nian et al. argues that there are 11 “general arguments for a successful distributed cryptocurrency ” - open source software, decentralisation, peer-to-peer network, global, fast, reliability, security, sophisticated and flexible, automated, scalable and platform for integration [NC15].

Bitcoin is a form of a digital currency. It does not have any trusted institution like a bank or a government backing it. Also it does not rely on a third party that controls transactions. Instead it uses decentralised and peer-to-peer network system to verify and process all transactions. Bitcoin technology relies on cryptography for processing transactions, and it is implemented as an open-source system published in 2009 [Nak08]. For the first time in the history of online payments, it is possible to make a transaction without trusted third party involved and without paying fees to centralised authority [NC15]. All transactions are stored in a public distributed ledger - blockchain. Also all nodes in the network are in charge for verifying transactions using cryptography. Miners are rewarded with bitcoins for solving proof of work which represents finding a solution to a given mathematical problem. Currently a reward for miners for creating one block of transactions is 12.5 bitcoins. [6] Bitcoin can be used to buy products and services, and it can also be exchanged for other cryptocurrencies or fiat currencies [NC15]. Money can be defined as a medium of exchange, a unit of account and store of value [KW89]. Because bitcoin satisfies definition of the money and it is in digital form it can be defined as digital currency [KW89, NC15]. Bitcoin currency code which is widely used is “BTC”, however, there are some exchanges that use “XBT” instead [NC15]. Bitcoin avoids having trusted third party in online transactions and it solves the double spending problem by using peer-to-peer network. To provide this, the network hashes all of transactions and puts them into the chain. Chain is created from hash based proof-of-work. The longest chain in the network presents proof of the sequence. As long as good nodes are controlling the most of the CPU power, they will create the longest chain, thus attacker nodes will not have a chance to make fraudulent blocks in the chain. Everything is shared with all the nodes in the network, however nodes can leave at any time and rejoin again. In that case in order to see what happened while they were away they will have to accept the longest chain [Nak08]. Bitcoin system is built as an open-source software which means that anyone on the Internet can access it and can also inspect it, modify it and enhance it. Going further, Satoshi Nakamoto left the Bitcoin project in 2010 thus leaving the project in the hands of community, and removing any doubt that there is someone controlling it from the background. Besides, even if some of the software developers would wish to change the source code, this change cannot be accepted unless full consensus of all nodes that are currently in the network agrees on it. If there is some change made to the system, all users, including developers would have to agree on it [NC15].

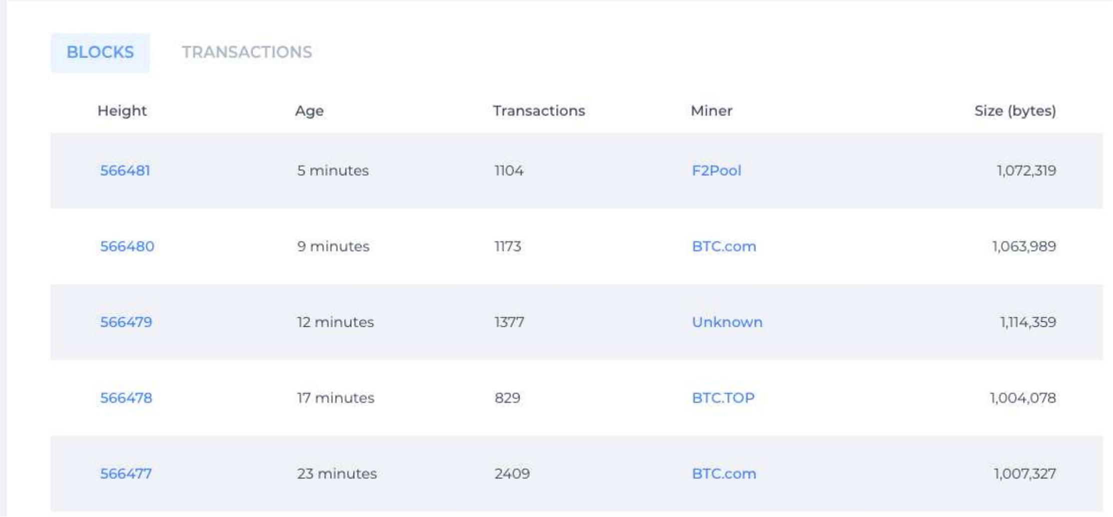
For most people, who are not familiar with cryptocurrencies, bitcoin is just another digital currency that is stored electronically. One can purchase it either through cryptocurrency

[6] https://www.blockchain.com/about (Accessed December, 2018)

exchanges or from vending machines. It can also be obtained as a payment for goods or services. One can keep it on the exchange, which is not very secure, knowing what had happened to one of the biggest bitcoin exchanges at the time, Mt. Gox [DW14]. The other way to store them is to save them in a wallet. More specifically, private keys are stored in wallets itself [NC15]. Since these private keys are used to access bitcoins, there is a need for them to be stored securely. There are different types of wallets like online, desktop, mobile, hardware (cold) wallets which can secure bitcoin private keys. Online wallets are accessible anywhere, from every device, as long as there is an internet connection. In this type of a wallet, a service provider stores bitcoin addresses. In addition, service providers try to provide some additional security through two factor authentication or extra encryption. Desktop wallets, on the other hand are wallets that are kept only on users' computers. Similar to desktop wallets, mobile wallets are installed on users' mobile phones. They are usually an application with functionality of a wallet. Compared to desktop wallets, they are more convenient, since they can be used on the go like in a store. Bitcoin Wallet and Mycelium are some of the applications that only exist on the mobile platform. Whereas applications like Blockchain.info have both mobile and desktop versions. Lastly, cold wallets or hardware wallets are complete opposite to online wallets. They represent a special type of device that can store bitcoin private keys electronically and most importantly offline. Cold wallets are immune to computer viruses and also they store private keys in a protected area, which makes it harder to extract them in plaintext. Bitcoins and other cryptocurrencies can be also sent and received through these devices. One example of the hardware wallet is Tresor. [NC15] Each of them have its positive and negative sides. Services that provide online wallets are vulnerable to theft or they can experience attacks during the bitcoin transfers. However, they are much more user friendly compared to hardware wallets and they can be accessible from anywhere. Hardware wallets, on the other hand, are more secure, since they are not connected to the Internet, and once currencies are stored there, no one except the owner has access to it. But significantly less people use them. One of the examples of the hardware wallet is Tresor [HR17, NC15]. As mentioned earlier, one of the biggest changes that bitcoin introduced is solving a double spending problem. This problem is known in computer science, as problem that digital money can be spent more than once. Bitcoin solves this problem by decentralising the responsibility to all nodes in the network, instead of having one single trusted party who would control it. Bitcoin system keeps a record of all transactions and balances in the public ledger called blockchain. All transactions that have ever happened are stored in the blockchain, and they are easily accessible for verification. Entire network gets information about all transactions. All new transactions that come, are checked against already existing transactions in the blockchain, so that money that is spent, cannot be spent again. In that way double spending problem is solved [NC15].

Here is one example of a block in the bitcoin blockchain:

In the Figure 2.1, we can see most recent blocks that have been added to the Bitcoin blockchain. Under the 'Height' we see position of the block in the blockchain itself.



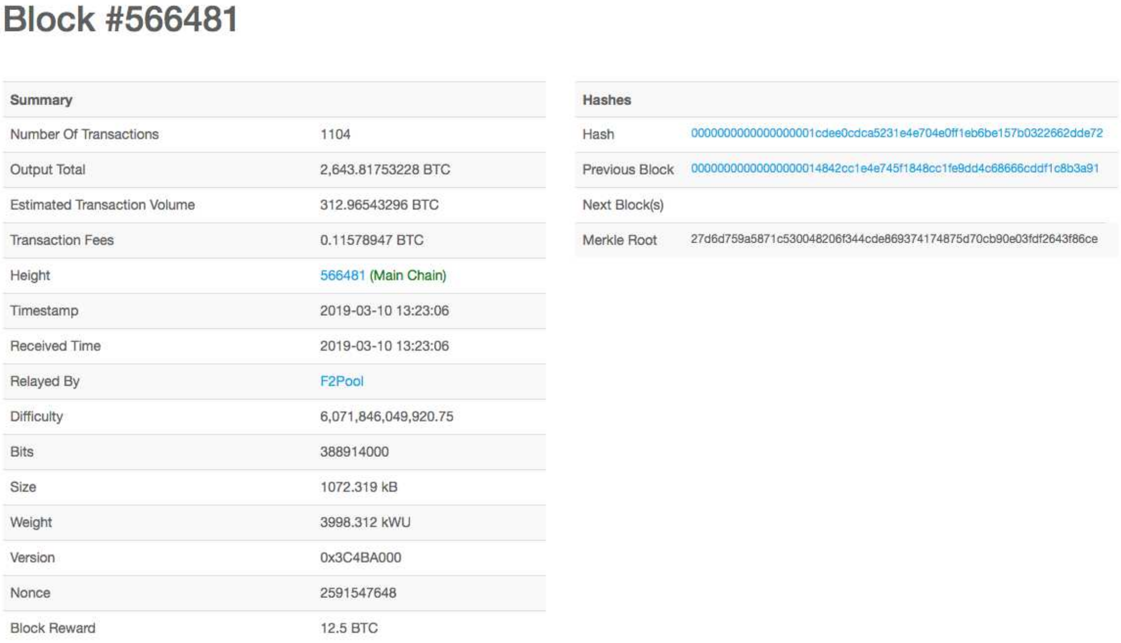
Height	Age	Transactions	Miner	Size (bytes)
566481	5 minutes	1104	F2Pool	1,072,319
566480	9 minutes	1173	BTC.com	1,063,989
566479	12 minutes	1377	Unknown	1,114,359
566478	17 minutes	829	BTC.TOP	1,004,078
566477	23 minutes	2409	BTC.com	1,007,327

Figure 2.1: Blocks in the Bitcoin blockchain (`https://www.blockchain.com`)

‘Age’ represents how long ago was the block added. ‘Transactions’ tell us the number of transactions that are contained in each block. Who “mined” the block is under the ‘Miner’ column and we can also see the size of the entire block.

In the Figure 2.2, we can see a summary of one block. In this case we are looking into the first block shown in the Figure 2.1, block with height 566481. Here we can see if the block is in the Main Chain. Also we can see what is the estimated transaction volume that is in this block in particular. Among other things we can see the weight of this block which is represented by kWU, or the difficulty for the proof of work that miners had to find. The miner, in this case “F2Pool”, got rewarder for mining this block by 12.5 BTC. Hash of this block and the previous block are also shown in the Hash section. However, since this was the last block added to the blockchain, at the time of observation, there is no hash from the Next block stored. This will be populated as soon as next block is accepted to the chain. One can also see all transactions that are gathered in this block, and the amount of each transaction that has been executed.

Bitcoin is designed with a limited supply of bitcoins. That limit is 21 million bitcoins. It is expected that the last bitcoin will be mined around 2040 [NC15]. In order to generate new coins, miners collect unprocessed transactions in a block and they try to solve a proof of work in order for their block to be accepted in the blockchain. This process is called mining. As a reward if their block passes verification of the network, they receive bitcoins. Currently that reward consist of 12.5 bitcoins per block. Size of each bitcoin block is 1 MB and it can contain transactions up to its maximum size. Through this process not only new bitcoins will be generated and issued, but new transactions will be processed and added to the blockchain. This is how miners are contributing with their computer power to maintain the whole network, and this is why they get rewarded.



Block #566481

Summary	
Number Of Transactions	1104
Output Total	2,643.81753228 BTC
Estimated Transaction Volume	312.96543296 BTC
Transaction Fees	0.11578947 BTC
Height	566481 (Main Chain)
Timestamp	2019-03-10 13:23:06
Received Time	2019-03-10 13:23:06
Relayed By	F2Pool
Difficulty	6,071,846,049,920.75
Bits	388914000
Size	1072.319 kB
Weight	3998.312 kWU
Version	0x3C4BA000
Nonce	2591547648
Block Reward	12.5 BTC

Hashes	
Hash	0000000000000000001cdee0cdca5231e4e704e0ff1eb6be157b0322662dde72
Previous Block	00000000000000000014842cc1e4e745f1848cc1fe9dd4c68666cddf1c8b3a91
Next Block(s)	
Merkle Root	27d6d759a5871c530048206f344cde869374174875d70cb90e03fdf2643f86ce

Figure 2.2: Summary of a block (https://www.blockchain.com)

Computer power is important also for verification purposes of transactions since only legitimate transactions can be recorded to the blockchain [NC15]. On average new block can be created every 10 minutes and problem difficulty that miners need to solve in order for block to be created is usually adapted for that period of time. If new blocks are generated too fast, the difficulty will increase, and if blocks are generated too slow, the difficulty will decrease. Miners collect new and unverified transactions, and by using the computer program try to guess a “nonce” number which will enable their block with transactions to be accepted to the chain. The hash function that is used in bitcoin system is SHA-256. It is used for digital signatures in transactions, bitcoin addresses and verification of payments. Furthermore, this hash function represents the basis for the mathematical proof-of-work problem that needs to be solved. SHA-256 is successor of SHA-1 hash function which was first represented by the NSA in the United States and which was used in the Hashcash [NC15, Bac02]. Going further, public-key cryptography is used in the bitcoin system. With the help of public and private keys, the authenticity of transactions can be easily identified. In this type of cryptography, public key is created from the private key, but it is impossible to recreate the private key from the public one. Thus, public key can be accessed and shared publicly, while on the other hand, private key has to be secured. Hence, a private key is used to create digital signatures, while the public key is used only to validate them [NC15].

As a new technology, bitcoin brings a lot of fresh points of view in already established monetary system. However, because there is a lack of government regulations there are a lot of risks of using it as well [NC15].

Some of the positive aspects include [NC15]:

- You can easily and quickly transfer your money, free of charge or for a very low transaction fee. It can easily be transferred even across borders which with current digital payments is very expensive. That is made possible with bitcoin because there is no third party involved in the whole process.
- Bitcoin represents an alternative digital payment for merchants. It is cheaper for them to accept bitcoin than to pay fees to credit or debit card providers. These fees often increase cost of accepting credit cards in the shops. Furthermore, reduction in fees that merchants need to pay to credit card providers, will enable them to accept small payments without having a minimum transaction restriction.
- Transactions with bitcoin tend to be more secure, compared to transactions with standard credit or debit cards. This is because transactions do not have any personal information. The same cannot be said for credit or debit cards which contain personal information that can be used in fraudulent transactions or theft. Also users are in control how much money merchants are able to debit from their account, and there is no possibility for unwanted charges later.
- Since bitcoin does not have a real coverage behind, and it is not connected to any government, its volatility is not affected by global financial crises nor government fiscal policies.

There are also negative aspects which include [NC15]:

- Even though it represents quite secure method of digital payment, there are some challenges for bitcoin in that area as well. For example, if a private key is lost, it is not possible to retrieve access to the bitcoin.
- Current online digital wallets and exchanges are not very secure, so digital thefts can happen.
- At the moment, the price of bitcoin is very volatile so even users and merchants tend to transfer bitcoins to fiat currency as soon as possible. There are speculations that these price differences remind of financial bubbles [GTMP14]. However, the price of bitcoin may become more stable in the future, as it becomes more widely used and more mature.
- Because of its pseudo anonymity, bitcoin can be used for paying illegal products or services. One of the examples is that it was used on the dark web Silk Road website which is a black market for selling drugs and fraudulent passports. One of the major government concerns is that it can be used for money laundering and financing terrorist activities.

2.2 Stock markets

Shares, or in American English stocks, represent divided company ownership. One share, in proportion to the total number of shares, represents partial company ownership. Stocks can be bought or sold on stock exchanges. Any stock trading is highly regulated by different laws and governments in order to prevent malversations, frauds and money laundering [BHM12, PW97]. Stock exchanges represent a place where stocks, bonds and other financial securities can be bought or sold. Traders or stock brokers are the ones who are actually doing the trade. Another way of trading securities is over-the-counter. This means that stocks are traded for the companies that are not listed on the formal exchanges. Stock exchanges are actually a subset of stock markets. All the trading that is happening on different stock exchanges represent stock market [BHM12, PW97]. Some of the best known stock exchanges are the New York Stock Exchange, NASDAQ, Japan Exchange Group, London Stock Exchange Group, Deutsche Börse. [7]

Since a lot of companies' shares are traded every day on exchanges, people try to investigate what is the best way to invest their assets. All assets have value, regardless if they are financial or real assets. Managing the assets is not the only thing that is important for successful investing. What is really important is to understand what drives that value and what causes its volatility. So one will need different information for valuing different types of assets. For example information required to evaluate the real estate assets will be different than the information needed for evaluation of publicly traded stocks. There are a lot of different areas how to analyse market and how to determine the true value of the asset. Not everyone will agree about how successful different approaches are for determination of the true value of the asset . However, everyone will agree that asset price cannot only be based on belief that there will be other investors who will be willing to pay more in future [Dam12].

It is believed that one of the factors that impacts the prices of the assets is economic news. Also, stock prices are closely related to external forces. Although some other parameters and variables may influence their price [CRR86]. In finance, risk is considered as a return of investment which is different from the return we expect to have. Risk can go both ways and that is why we have upside risk and downside risk. Return can be higher than expected, which will mean that the outcome is good. On the other hand return can also be lower than expected and in that case outcome is bad for investors. However, when risk is estimated, both types of outcomes are taken into consideration. When an investor buys assets, they expect that investment will be successful and it will earn money over the period of time when the asset was in their possession. This is expected return. However, expected return can be different from an actual return. Difference between the expected and actual return is what defines risk. [Dam12].

Stock market prediction was always interesting for researchers. Possibly because of the difficulty and complexity behind how to model market movements and dynamics [SC09]. Stock market prediction comprises of different techniques which try to predict the value

[7] https://www.world-exchanges.org/our-work/statistics (Accessed March, 2019)

of a stock and also to maximise the return of investment. Successful prediction of the price can bring a significant profit. Many different approaches are used for forecasting the market movements. They can be divided into two areas: fundamental analysis and technical analysis or chartists [Lam04]. Fundamental analysis concentrates on how did the company perform in the past, as well as on its accounts' credibility. On the other hand, technical analysis tries to predict the future price by entirely concentrating on the past trends [Lam04]. The most used forms of technical analysis are trend analysis, cycle analysis, charting techniques and other types of historical data analysis. All of these tools help analysts to get necessary information in order to form the prediction on how the market will move. However, they can not just rely on one method or on a single tool. They have to combine different techniques and tools to get more precise approximations. Both qualitative and quantitative measures are used daily in analysis. [BC87].

Even though there are many different technical or chartists theories, all of them have the same basic assumption. The patterns that determined the price before will happen again. It is assumed that there is a lot of information in historical prices and by carefully analysing and understanding these historical patterns, future prices can be predicted. Those predictions will eventually cause profit to increase [Fam65]. Some of the information that is analysed are exchange rates and interest rates, specific information for an industry like growth rates and consumer prices, and specific information for a company like income statements and dividends [ET05]. As Enke et al. mention in their work, there are ways of predicting stock market returns by using publicly available information like time-series data on financial variables. Some of the variables that can influence stock returns include interest rates, monetary growth rates, inflation rates and also changes in industrial production. Furthermore, certain macro-economical factors can also influence in which direction the stock market will move. Some of them are political events, general economic conditions, firms' policies, bank rate and exchange rate, investors' expectations, psychology of investors, movements of other stock markets [PSTK15]. Previous studies were trying to find the relationship with these variables by using statistical modeling, such as linear regression. However, because there is no evidence that relationship is strictly linear, alternative non linear methods are introduced [ET05].

Over the past two decades, with the emergence of digital era stock market predictions improved. Also larger amounts of data can now be processed faster and therefore provide investors with precise results [ET05]. Some of the new alternative techniques that are used for analysing stock markets are artificial neural networks, machine learning, data mining, analysis of the public sentiment. [KAYT90, PSTK15, SC09, BMZ11, ET05].

As mentioned before, traditional stock market prediction techniques use linear statistical models. However, since the relationship between variables and stock market prices is not strictly linear, new ways of predictions should be considered. Some of the non linear models can explain this dependency in a more detailed way. New ways of prediction are done using Artificial Neural Networks (ANN), Genetic Algorithms (GA), fuzzy logic and also other techniques. Two of the most used machine learning algorithms for stock price prediction are Artificial Neural Networks (ANN) and Support Vector Regression (SVR).

Both of the mentioned algorithms have unique ways of learning patterns [PSTK15].

More precise prediction of future stock returns has motivated a lot of financial analysts and researchers to switch from linear to non-linear models in their analysis. Some of the studies that use neural networks for stock market prediction besides Enke and Thawornwong (2005) are Chenoweth and Obradovic (1996), Leung, Daouk, and Chen (2000), Motiwalla and Wahab (2000) [ET05, COL96, LCD00, MW00]. A neural network is a series of different algorithms and techniques that can mimic some parts of a human brain and nervous system. [ET05] Some neural network models, like feedforward neural network with supervised learning, learn from provided samples how to behave in the future. They are not programmed with specific rules. These neural network models work as a set of input and output units. Those units are called artificial neurons. Each neuron is connected to other neurons and that connection has a certain weight. Usually artificial neurons are positioned in layers. Signals go from the first (input) layer to the last (output) layer. During the first phase which is a learning phase, this type of neural network adjusts weights on each connection according to the input signals. In that way it learns how to correctly predict output for the given input samples. One of the new techniques that neural networks introduced, in comparison to the previously used linear techniques, is that they do not require any pre-specification during the modeling. This means that they learn independently the relationship between the variables and forecast the price according to what they learnt over time. This is very important for financial systems because it adds a new level of accuracy, since in those systems much was assumed and based on educated guesses. There are different types of neural network learning algorithms, architecture types and validation procedures which gives flexibility to the analysts. However, three types are most common in financial prediction systems: generalised regression, probabilistic and multi-layer feed-forward neural network. [ET05].

In the last three decades, significantly larger amounts of data are stored in digital form. Data mining can be described as a process of finding patterns in large data sets. Different methods are used for extracting useful information from the large amount of data. Some of them are database systems and query languages, statistics and machine learning. Ideally, all different types of data should give us answers in different areas. Data mining found its purpose in the stock prediction as well. A lot of trading nowadays is done using automated computer programs. Those programs use data mining and other predictive technologies in order to establish good forecasts. Data mining is especially used for technical analysis type of market predictions, where the focus is on analysis of historical data. In this way, analysts are now able to discover hidden patterns from large historical data sets, which was not possible before. They can use that information later for determining whether stock prices will increase or decrease [KSSA10].

Similar to data mining, text mining has also drawn the attention of the public in recent years. Text mining or text analytics refers to getting useful insights from text. With a combination of machine learning and text analytics, content from the Internet can be used as an input for prediction of price movements in financial markets. Public opinion from Twitter has also been linked to the stock market prices [SC09, BMZ11, MG12, ZFG11, RS12]. As stated in the efficient market hypothesis (EMH), stock market prices are the result of all available information, and they are especially driven by new information like news. That is different compared to theories which rely only on present and past prices. According to Tetlock, stock market movements can be predicted by content from news [Tet07]. Researchers have shown that news are also unpredictable. Hence, stock market prices can be predicted with only 50 percent accuracy [BMZ11]. However, there are very early indicators that can show various economic changes and trends. Those early indicators can be found on the Internet on blogs or Twitter where people, including investors, share their thoughts. So besides news, which for sure play a very important role in future stock market prices, public sentiment on the Internet can have influence as well. Especially because it has been shown that emotions and mood can have an important role in financial decision making. [KT13, BMZ11, Nof05]. Baker and Wurgler in their research have proven that investors sentiment plays a really significant part in future market movements [BW07].

2.3 Cryptocurrency markets

With use of digital currencies, there is a huge potential for improving payment systems by reducing fees and transaction costs. Technology improvements will allow businesses to develop even more from economic point of view. A lot of companies are adapting to cryptocurrencies which is revolutionizing the payment systems as we know them. Companies that might be affected include money transfer and credit card companies, as well as securities exchange companies and payment hardware companies. However, bitcoin still largely depends on the fiat currencies when it comes to buying it and trading. Since mining requires a lot of equipment, knowledge and effort, a lot of people would rather opt to buy it on a cryptocurrency exchanges than to mine it. Cryptocurrency exchanges serve as an intermediate between bitcoin and other cryptocurrencies and fiat currencies. With popularity of bitcoin and cryptocurrencies in general, there is an increase in interest for cryptocurrency exchanges [BC15].

People can buy bitcoin on cryptocurrency exchanges for fiat currency, or they can sell bitcoin there and earn money. This also applies to other cryptocurrencies. These exchanges serve as the primary source of access to the cryptocurrency network. Cryptocurrency prices are determined only by demand and supply, hence these exchanges have a major influence on their value. This means that value of the bitcoin is influenced by other people buying it or selling it on exchanges. At the cryptocurrency exchanges, people are able not only to buy or sell bitcoins or other cryptocurrencies, but they can trade it for other altcoins as well. Exchange rates for that particular exchange are applied. These rates can vary on different cryptocurrency exchanges. People can also store their bitcoins there. Clients can access exchanges via secure SSL connection on their website. Cryptocurrency exchanges are only accessible online. Since there are a lot of problems connected with cryptocurrencies such as money laundering as well as trading on the black markets, exchanges request client identification before they can do trading on it. Exchange has identity of the client as well as his or her account details. Every user on the exchange has their own digital wallet. However, exchange has all of the private keys for all of their clients' wallets. If client wants to sell bitcoin, they transfer desired amount from their wallet to the wallet of the exchange. Real transactions like bitcoin deposits and withdrawals are documented on the blockchain. However, trades of bitcoins are only recorded in the history of the cryptocurrency exchange itself. Identity of the client is not shown on the blockchain, rather they are identified as exchange trades. Details of the client who made transaction are only stored in the internal exchange's database. [BC15]. Some of the services that cryptocurrency exchanges provide are digital currency trading, digital currency exchange, currency tracking, mining, general information and gift cards. Currently there a couple of hundred cryptocurrency exchanges over the internet. However, only dozens of exchanges hold the majority of the market. According to Bloomberg, these are the largest cryptocurrency exchanges if we look estimated revenues and trading volume: Binance (Malta), Upbit (South Korea), Bittrex (USA), Coinbase (USA), Bitstamp (Great Britain), Kraken (USA). [8] How popular exchanges will be, depends on the diversity of services they provide, as well as on the fees that they charge. There are different strategies which exchanges follow in order to maximise their profits from fees. However, some of the exchanges, in order to be competitive on the market do not have fees at all [BC15].

For exchange to stay competitive on the market it needs to keep trading volume above a certain level. When it reaches that level, it becomes a point of interest for hackers. This is why cryptocurrency exchanges are highly exposed to cyber attacks and security threats. One of the biggest attacks on cryptocurrency exchange was one in 2011 on Mt. Gox. In 2013 and 2014, before it declared bankruptcy, Mt. Gox handled over 70% of all bitcoin transactions worldwide. It was the largest bitcoin exchange at the time. In 2011, a hacker transferred a large amount of bitcoins to his account, by illegally using credentials of the one of exchange's auditors. He manipulated the software and created massive 'ask' order for bitcoins at any price. The price temporarily went high, when he sold bitcoins. After a couple of minutes price returned to normal user-traded price. This incident affected $8.750.000 worth of accounts on the exchange. In order to gain confidence back, Mt. Gox transferred 424.242 bitcoins from one of its cold wallets to its address. The other incident happened just a couple of months later, in October 2011. In that incident 2609 bitcoins were sent to an invalid address. This would be the same as if those bitcoins were destroyed, because private key for that address does not exist, and it is almost impossible to restore it [DW14, BC15]. In February 2014, Mt.Gox stopped all the transactions on the exchange. For more than a couple of weeks, withdrawals from the exchange were suspended. This raised many concerns from its customers. Finally Mt. Gox declared bankruptcy, as well as that around 750.000 client's bitcoins were missing together with 100.000 bitcoins from exchange itself. At that time value of the missing bitcoins was evaluated to $473.000.000. The collapse of Mt. Gox was one of the biggest cryptocurrency crises. The fourth largest cryptocurrency exchange of 2013 also soon closed due to security breaches. Around 24.000 bitcoins were stolen from this exchange

[8] `https://www.bloomberg.com/crypto` (Accessed April, 2019)

in one attack. At that time, already 730.000 bitcoins were already missing before the collapse of the Mt. Gox. Together with the bitcoins lost on Mt.Gox it represented almost 6% of total bitcoin supply [BC15].

As mentioned before, if bitcoin is sent to an address for which private key is unknown, it is extremely hard to retrieve it. It can be considered as lost or destroyed. This can be done intentionally or accidentally. The private key ownership can easily be identified. However, it is not known if the private key has been lost or it is still by its owner. Thus, it is hard to determine if the bitcoin is saved or lost. Private key is almost impossible to retrieve from the address itself. It is really time-consuming and it requires a lot of hashing power to guess it [BC15].

According to blockchain, in April 2019, there is a bit more than 17.5 million bitcoins in circulation. [9] It is estimated that 3.79 million bitcoins are lost or destroyed. This represents a worth of more than $19 billion. [10] This implicates that around 18% of all bitcoin supply is lost forever [BC15].

There are more than a couple of hundred exchanges worldwide. However, approximately a bit less than a half will close eventually. Some of the most important factors that can influence the life of one exchange are transaction volume, security breach, financial strength, compliance capabilities and backroom and settlement support. Transaction volume is one of the most important factors which can determine if the exchange is profitable or not. Since there is no good methods to verify trading volume, exchanges often inflate that number. The reason behind is that exchanges with higher transaction volume are more attractive to traders and easier gain popularity. On the other side, exchanges with lower trading volume are more likely to get out of business and eventually shut down. Thus, bigger exchanges become better targets for attackers. Apart from the transaction volume, other factors also influence bitcoin exchanges' lifetime. Because of anonymity and ease of trading on black markets, cryptocurrencies become interesting for criminals. They use it for money laundering as well. On the other hand, with increase of the bitcoin value, many viruses are designed to attack and steal bitcoins from wallets. If bitcoin exchange experiences a security breach, it is more likely that it will close, since the breach can cause loss of its profits, decreasing the cash flow and losing trust from clients. What is interesting is that most of the malware softwares are designed to steal only bitcoins, and just 1% of all would target other cryptocurrencies [BC15].

Pressure on cryptocurrency exchanges could be reduced if bitcoin gets recognised as a common medium of exchange and gets widely accepted. In that case commodities could be traded directly for bitcoin, instead of having third party intermediary. Interest for bitcoin exchanges would be significantly reduced and security breaches would be minimised. However, exchanges would still exist in terms of trading among other cryptocurrencies. Almost all exchanges are not regulated. Even if there would be possibility to regulate, the majority of parties involved would not be interested. This is mostly because they

[9] `https://www.blockchain.com/about` (Accessed December, 2018)
[10] `http://fortune.com/2017/11/25/lost-bitcoins/` (Accessed April, 2019)

think that reluating the bitcoin exchanges would hamper growth and innovation. At the same time, even if they are unregulated, all transactions are stored in the blockchain and publicly announced. That is the main difference compared to how financial institutions are handling the fiat money and it gives more transparency. However, it is possible that the government will pressure Bitcoin community at some point to be regulated. And regulation, on the other hand, can bring positive effects such as gaining access to the mainstream of finance. If that happens, that moment can lead to division of the Bitcoin community [BC15].

2.4 Twitter

New form of blogging that usually consists of posts that have less than 200 characters is called microblogging. This is a new phenomenon emerged with new on-the-go lifestyle. It allows users to post more concise posts from different locations and from various devices including a mobile phone. With shorter posts, users can have less engagement for creating content and they will also spend a shorter period of time preparing a post. The other difference, when it comes to comparing to regular blogging, is the frequency of updates. Because users spend less time to create a post, users will share their thoughts more often. Hence, while regular bloggers would post once every couple of days, microbloggers can post multiple times during one single day. Popular topics on microblogging platforms are daily activities and information sharing [JSFT07].

Twitter is one of the most popular microblogging platforms nowadays [JSFT07, KLPM10]. According to Alexa Twitter is the 11th most visited site in the world. [11] It can be said that Twitter has influenced parts of modern communication. It allows people to share most recent events, even faster than news outlets are publishing them. Twitter was launched in July 2006 and since then it has experienced a fast growth. As per latest financial report from Twitter, it has around 330 million monthly active users. [12] On Twitter, users can share short messages up to 280 characters. Those short messages are called tweets. Before September 2017, tweets were limited to 140 characters only. However, twitter decided to increase that boundary in order for people to easier express themselves. [13] In tweets, people can share their thoughts and describe their current status. Twitter users can follow other users or be followed. This is not necessarily reciprocal, since users can choose who to follow and those users do not have to follow back. If user follows someone, he or she will receive all of their tweets [JSFT07, KLPM10]. Besides tweet and follow options there are three more things that are common for Twitter users. First one is called a hashtag. Hashtag or '#' is used to connect tweets with the same topic in one place. Those are the words or phrases that have hashtag symbol (#) in front of it. If user is interested in a particular topic, he or she can just click on the desired hashtag, and they will be able to see all posts that are about that subject. Hashtags

[11] https://www.alexa.com/topsites (Accessed December, 2018)
[12] https://investor.twitterinc.com/static-files/ (Accessed December, 2018)
[13] https://blog.twitter.com (Accessed May, 2019)

allow users to easily follow topics that they are interested in. This was first created on Twitter, and later it spread across other social networks. Second one is retweet. Users can retweet other users' tweets. This is the way how news and other interesting topics are easily shared across the platform and beyond followers from the original tweet's creator [KLPM10]. Kwak et al. in their research found out that retweeted tweet can reach on average 1000 users, regardless of the number of followers of the user who posted the tweet originally [KLPM10]. Third, there is an option to mention someone. This is done by putting '@' sign before other user's username. [14] This was also first developed by Twitter in a time when there were no private messages. This was the only way how users could communicate directly with each other [JSFT07].

Twitter users have profiles, where they have brief information about themselves. Users can choose between making their profile public or private. In case the profile is public, tweets of that user may appear in public timelines [JSFT07]. Usually public profile of the user includes a nickname, profile picture, number of followers, number of accounts that user follows and number of tweets. Some of the profiles can have in addition a full name, a location, a web page and a short biography [KLPM10]. Users that have the highest number of followers are mainly public figures and celebrities. Twitter also introduced verified accounts. Verified accounts have a blue badge next to the username. This is how Twitter verifies accounts of public interest. Accounts that get chance to be verified are usually accounts from musicians, actors, politicians, sportsmen, media etc. [15] Those top users mainly do not follow back their followers. Going further, according to Kwak et al. 67.6% of twitter users are not followed at all by people who they are following. It appears that those users use Twitter rather for getting information than as a social platform. [16]

According to some studies content of the tweets may vary between news, conversations, self-promotion, daily life and even spam [Ana09, JSFT07]. A lot of applications nowadays have integrated option to instantly share your posts to Twitter. According to Java et al. daily chatter, sharing information, reporting news and conversations are major drivers for users to post on Twitter. They also concluded that user can take different roles depending on what is their current purpose of visiting the social network. For example, the same user can at one point be information seeker, while at the other point they can be the source of the information. Most Twitter posts are related to what people are currently doing. Around 13% of users share information of some sort or URL. Likewise, many users share the latest news or they are commenting on current events [JSFT07]. Twitter has also introduced trends. Those are the topics that are mentioned more times compared to other topics [KLPM10]. Phrases, words and hashtags are the main drivers of the algorithm. However, they can also be personalized according to the user's interests, accounts that they follow and location. Furthermore, Twitter shows trends that are popular now, rather than in some past time period of the same day. If multiple tweets

[14] `https://help.twitter.com/en/twitter-guide` (Accessed May, 2019)

[15] `https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts` (Accessed May, 2019)

[16] `https://help.twitter.com/en/twitter-guide` (Accessed May, 2019)

or hashtags are related to the same topic, they are grouped together [17], which was not the case a couple of years ago [KLPM10]. Over 85% of trending topics are correlated with news, either headline news or persistent news in nature [KLPM10]. Among popular topics it is not rare to find spam tweets. Number of spam tweets has increased with the growing popularity of Twitter. However, Twitter tends to suspend all suspicious and reported accounts, thus making sure that only the meaningful tweets are reaching its users. [KLPM10].

According to Nofsinger and his research, mood and emotions of individual are notably influencing his or her financial decisions [Nof05]. However, the question is can public mood also affect common decision-making [BMZ11]. As mentioned in section 2.2 of this thesis, prediction of stock market has attracted attention from both researchers and businesses [SC09]. Many techniques have been used for these purposes. However, it was hard to determine in which direction the market will go, so need for new ways of stock market prediction emerged [Lam04, BMZ11]. According to Bollen et al. new information highly influences stock market prices, even more than comparison to past prices. However, when it comes to news, there is a high level of uncertainty, which means that stock market prices cannot be predicted with high accuracy. Even though news are unpredictable, stock market research argues that early signals of change can be found on social media like Twitter. Even though each individual tweet is restricted to 280 characters only, millions of tweets accumulated together can give a wider picture of general public sentiment on specific topic [BMZ11]. For example, Gruhl et al. in their research show how book sales can be predicted only by analysing online chat activity [GGK$^+$05].

2.5 Google Search

Google Search is a web search engine developed by Google LLC. [18] It is without any doubt the most visited, not only a search engine, but the website in the whole world. It is number one globally ranked site on Alexa. [19] In April 2019, Google had 92.81% of the worldwide market share among web search engines. [20] On average Google has 3.5 billion search queries per day. Furthermore, it has more than 4.5 billion of monthly active users. Google's main competitors are Bing (Microsoft), Yahoo, Baidu, Yandex and DuckDuckGo. However, when it comes to comparison, they all have negligible percentage of the market share. Google was developed in 1998 by Larry Page and Sergey Brin. [21] They also developed a PageRank algorithm, using which Google ranks search results. It functions in a way that it counts the number of links to the page and also the quality of those links. In that way it establishes how important is that particular website [PBMW99]. Google also provides data about frequency of search terms through

[17] https://help.twitter.com/en/using-twitter/twitter-trending-faqs (Accessed May, 2019)

[18] https://www.google.com (Accessed May, 2019)

[19] https://www.alexa.com/topsites (Accessed December, 2018)

[20] http://gs.statcounter.com/search-engine-market-share (Accessed May, 2019)

[21] https://www.forbes.com (Accessed May, 2019)

a service called Google Trends [22]. Some analysis have shown that search term frequency can suggest new economic and social trends. [GTMP14, Hub11, CV12].

Google Insights for Search was created by Google in 2008 and this was the first tool developed to get insights from Google search queries. However, after four years, they decided to merge it with more advanced tool for search term analysis. This tool is called Google Trends. [23] Google Trends is a website provided by Google and it shows frequency of searched terms on Google Search across various locations and among different languages. Search queries can be compared over time based on volume, popularity, region, country, etc. One can type term that they are interested in and they will get the requested data. Results are showing how interest changes over time, interest by subregion, related topics and related queries. Furthermore, the results can be filtered by country, time, category and type of web search (whether it is image search, news search, youtube search). Trending searches page shows what are current most popular searches around the world. Daily search trends and real time search trends are part of the trending searches page. The difference between the two is that first one highlights searches that popped out significantly comparing to other searches in the 24 hour period, while the second one highlights searches that popped out significantly comparing to all recent searches. Users can compare up to five different search terms. Google Trends data has two types of data that users can search for. First one is real time data and it represents random sample of searches in the last seven days. Second one is non-real time data, and this data represents historical data sample. User can retrieve data as far as from 2004. There is also data that is excluded from the sample. For example, that data can be searches made by very few people, duplicated searches (searches from the same person over the short period of time) and searches that contains special characters. One of the challenges that Google Trends was facing was how to adjust data results in order that regions with the highest search volume do not always be on the top. For example two regions may have the same interest over one topic, however, not both of the regions have the same volume of searches. They solved it in the following way: “Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity”. [24] Going further, there are different options in terms of what one might search for. There are search terms and there are topics. All results can be downloaded in the .CSV format, so it is easy to do further analysis if needed. [25]

One of the common limitations when it comes to analysing macroeconomic data variables is delay in obtaining data. Thus, traders are restricted to make a decision on the last available data, which is not necessarily the newest data. This is why nowcasting becomes more and more important. Nowcasting stands for predicting the present. Data availability would influence decision makers’ observations of economic trends almost in real time. This would help them to make more informed and precise decisions [CSL13]. Google

[22] https://trends.google.com (Accessed December, 2018)

[23] https://search.googleblog.com/2012/09/insights-into-what-world-is-searching.html (Accessed May, 2019)

[24] https://support.google.com/trends (Accessed May, 2019)

[25] https://trends.google.com (Accessed December, 2018)

Trends can provide almost the most recent data of public queries on Google. Various researches have shown that there is a correlation between economic trends and volume of search queries [GTMP14, Hub11, CV12, PMS13]. One of the researches shows how number of clicks on search results in one country is correlated to the amount of investment in that particular country [MWZ10]. Preis et al. in their research explain how different query volumes can determine the volume of stock market transactions [PRS10]. Bordino et al. in their paper indicate that the web search data can be used to predict stock market daily trading volume [BBC+12]. One of the research mention a relation between stock price changes and breaking financial news [BMZ11]. Furthermore, one of the research shows how Google Trends data is not only related to stock market changes, but it can also be a predictor for prices of digital currencies. For example, frequency of search of one digital currency can determine public interest in that particular currency. Kristoufek explains that there are certain disadvantages of only measuring interest using search queries on Google Trends because it is hard to make a difference if a search was driven by positive or negative events. Upturn in interest could indicate both an increase and decrease in the price of the digital currency, but even with this ambiguity, Google Trends data proves to be useful [Kri13].

2.6 Sentiment analysis

Natural language processing dates since the 1950s [NOMC11, MMS99]. In 1954 the Georgetown experiment took place in New York as collaboration of Georgetown University and IBM. The idea behind the experiment was to publicly demonstrate Russian - English machine translation system. It has attracted a lot of public interest [Hut04]. This is considered as one of the first non-numerical actions done by computers. For this experiment, six grammar rules and 250 words were used. Although the translation itself was not very successful, this experiment had set high expectations on what can be done in the field of translation and how it could develop in the future [Hut04]. With the emergence of the Internet and generation of large amounts of unstructured data, the need for computers to understand natural language was increasing. Natural language processing or NLP combines different areas like artificial intelligence, computer science and computational linguistics in order to discover how human language can be processed and analysed by computers [MMS99]. At first, NLP systems were designed to follow set of rules, like designing specific grammar for the system. However, this was often very complex, so new ways were introduced. In later years statistical and machine learning methods took over NLP systems. When compared with hand-produced rules, machine learning algorithms were better because they were automatically focusing on cases that are most common and since they had more input data, they could be more precise. Nowadays with the help of NLP it is possible for computers to process and analyse text and speech, also to determine which part of the sentence is most important, and to estimate sentiment. Some of the NLP tasks include parsing, identification of semantic relationships, lemmatization and stemming. In general natural language processing breaks sentences into smaller pieces and tries to understand the meaning of the pieces and what are the relationships between them. Use of NLP nowadays can be found in machine translation, document summarisation, content categorisation and topic discovery, speech to text and text to speech conversions and sentiment analysis [CWB$^+$11, NOMC11, MMS99].

Sentiment analysis is a field of study which analyses people's sentiment and opinion about products, services, individuals, organisations, different topics and events. It is also known as opinion mining, sentiment mining, subjectivity analysis, opinion extraction etc. Sentiment analysis evaluates positive or negative sentiment based on how people express their opinions. Although it is a subgroup of natural language processing, researchers did not pay much attention to it up until 2000s. The reason might be because this is the first time in history where there is a large amount of opinionated data available on the internet. Social networks are one of the biggest contributors to that large database. Without data available, much of the research work would not be possible. Different science fields which highly depend on people's opinions will be impacted by development of sentiment analysis. Some of them are economics, social science and political science. People's behaviour is often determined by their opinions. Others' opinions are very important to us when we want to make a decision. Large organisations invest a lot of money in trying to get consumer's feedback about their products or services. In that way they will know what and how to improve. Before spending money on a product or service, individual users also want to know reviews from already existing users. In this way they will not waste their money if the product is not good. In politics, a lot of research is done to determine public opinion about presidential candidates and who is most likely to win elections. In the past, when a company wanted to receive feedback, they were conducting surveys, focus groups and opinion polls. On the other hand, individuals would ask their friends and family if they would need recommendation for certain product or service. In contrast, nowadays, consumer and public opinions are becoming large business on their own. They are especially important in political campaigns, public relations and marketing. Large companies do not need to conduct surveys in order to get feedback on their product, because a lot of information about it is now available on the Internet. Individual consumers will focus more to find reviews of the product on the web rather than asking friends of their opinion or recommendations. Social media posts like reviews, blogs, micro-blog posts on Twitter, comments on various sites are largely used in decision making [Liu12, BMZ11]. However, so much data is generated that it might become overwhelming for someone to find and summarise all that information. Furthermore, it gets challenging to identify sites and blog posts where useful information can be gathered. Usually a lot of opinionated data can be found that is not very well structured. With these restrictions it is hard for someone to make an informed decision. This is why we need an automated process to help us summarise and analyse large amounts of data. Lately, it has been shown that a lot of opinionated data has helped in business decisions which eventually brought higher profits. Going further, analysis of public opinions and sentiments highly influenced political systems. And lastly, opinionated data cannot only be found online. Many organisations have their own internal data. This data is gathered by customer feedback, call centres or emails. Since information can be found for various domains like consumer products, financial services, social events, healthcare and political elections, interest in sentiment analysis is growing fast. Some of the big multinational companies like Google, Microsoft, HP and SAP have developed their own systems [Liu12, PL08].

When it comes to sentiment analysis, there are three main levels of analysis:

- Document level analysis
- Sentence level analysis
- Entity and Aspect level analysis

Document level analysis concentrates to determine whether the whole document has positive or negative sentiment. This level of analysis expects that the whole document expresses opinion on one single product. In case there are more products evaluated in the same document, this level of analysis will not be able to determine that. This is known as document-level sentiment classification [Liu12].

Sentence level analysis determines whether each sentence has a positive, negative or neutral opinion. Neutral opinion in this case represents that there is no opinion in the sentence at all. This level of classification is known as subjectivity classification [Liu12].

Entity and aspect level analysis is also known as feature level analysis. Compared to document and sentence level analysis, this kind is based on finer-grained analysis. Aspect level determines what precisely did people like or dislike. It tries to locate the part of the text that expresses opinion, rather than looking at the whole document or the whole sentence. It is considered that opinion consists of positive or negative sentiment and of its target. Target itself helps us to understand sentiment analysis better. On this level of analysis unstructured opinions can be summarised into structured data and can be used later for different quantitative and qualitative analysis. All three levels of analysis are highly challenging, with the last one being the most difficult [Liu12].

In addition to different levels of analysis, there are two types of opinions. Opinions can be regular and comparative. While regular opinion indicates sentiment of only specific aspect or entity, comparative opinion brings into comparison two or more entities which are based on their shared aspect. One example of comparative opinion would be “iPhone has better camera than Google Pixel phone”. Here iPhone and Pixel phones are compared based on their camera performance and preference goes to iPhone [Liu12].

Best way to determine sentiment is by analysing sentiment words. Sentiment words, also known as opinion words are words that are used to express positive or negative opinion. Some examples of positive sentiment words can be amazing, good, wonderful, while bad sentiment words are terrible, horrible, bad, poor. Besides individual words, phrases and idioms can also be used in order to determine sentiment. Sentiment lexicon represents a list of opinion words and phrases. Opinion words are very important indicators of sentiment, however they are not sufficient. There are different types of sentences where opinion words can be used but still give completely different context to the whole sentence. As Liu mentions in his book, there are four main points related to this issue [Liu12]:

1. The same sentiment word can have different meaning

   '*This product sucks!*' and '*This robot vacuum cleaner sucks perfectly.*'

2. A sentence that contains opinion words does not have to express any sentiment

   '*Can you tell me which laptop is good?*'

3. Sarcastic sentences with or without sentiment words

   '*What a great product! It stopped working immediately after the purchase.*'

4. Sentences that do not have opinion words can also express sentiment

   '*This kettle uses a lot of electricity.*'

With increase of unstructured opinionated data that could be used to better understand customer needs, a lot of sentiment analysis tools were developed. Even big companies like Google and Microsoft created their own systems for opinion mining [Liu12]. In this thesis we will use two sentiment analysis tools in order to determine public sentiment on Twitter towards bitcoin. The first one is the Stanford CoreNLP toolkit, set of tools specialized in the natural language processing with support for sentiment analysis [MSB+14]. The second algorithm that we will use was developed specifically to cope with short messages and informal spelling of words [PL08]. It is called SentiStrenght [26], and it showed good results when analysing short messages from a social network site [TBP11].

[26] `http://sentistrength.wlv.ac.uk` (Accessed May, 2019)

CHAPTER 3

Literature Review

3.1 Using Twitter to analyse public opinion

In the past, in order for researchers to find out what the public thinks about certain topics, they would have to conduct research on a random sample of the population. During the 20th century, different methodologies were developed for that purpose like surveys and polling [OBRS10]. However, nowadays with the high popularity of the Internet and social networks, people share their thoughts and opinions all the time. There is a lot of publicly available data that could be analysed in the same manner as polls. Such analysis would have a higher number of people involved in the survey which would give more precise results and it would be significantly cheaper in comparison of having standard polls or surveys. Moreover, time and effort used for analysis would be significantly reduced. However, topics that could be analysed in those types of research would be limited only to opinions that people share [OBRS10, PD11, BMP11].

There are multiple researches where Twitter is used for analysing public opinion [OBRS10, PD11, TSSW10, BMP11]. O'Connor et al. in their research investigate what is the connection between public opinion measured from traditional polls with sentiment measured from text collected from short messages from Twitter. They concentrated on two topics: consumer confidence and political opinion [OBRS10]. Consumer confidence represents how one feels about their personal finances, economy and health. Furthermore, if consumer confidence is higher, it will lead to more consumer spending. In regards to political opinion, researchers were tracking Barack Obama's presidential job approval rating and presidential elections between Barack Obama and John McCain. Researchers have collected messages from Twitter which were correlated to either of the topics mentioned above. They estimated whether message express positive or negative opinion about the topic. Results show that even with simple sentiment analysis of tweets, data from Twitter mirrors data collected from polls. In another research conducted by Michael Paul and Mark Dredze, authors are investigating whether messages on social media can analyse general public health [PD11]. They analysed around one and a half million of health related tweets by using Ailment Topic Aspect Model. Tweets that mentioned different illnesses like allergies, insomnia and obesity were tracked. Although somewhat intimate, this information was not hard to find, because users usually tweet about how they feel. Information that one person has flu might not be particularly interesting. But if you have a large amount of tweets mentioning the same thing this could track flu rate at particular geographical location [PD11, LDBC10]. In their research they concentrated on different aspects of public health like analyzing symptoms and medication usage, tracking illnesses over time and localizing illnesses by geographic region. That is how they found out in which parts of the world is more common for people to exercise. Furthermore, they discovered how people treat both milder and acute illnesses in comparison to more chronic ones. Findings included which medications are used, and that the majority of people do not seek medical help. Their quantitative results are in strong correlation with data provided by government. Results of their research suggest that public health research can be analysed through data collected from Twitter [PD11]. As a result of Barack Obama's campaign in 2008, Twitter became inevitable part of political sentiment prediction [OBRS10, TSSW10]. In the study done on Technical University of Munich by Tumasjan et al. it is analysed whether Twitter can forecast results of the German federal elections which took place in September, 2009 [TSSW10]. More than 100.000 tweets that contained either a politician or a political party were analysed. Among other things, researches studied whether political sentiment on Twitter is correlated to the current real world sentiment. And secondly they analysed if content of the short microblogging messages can be used to determine popularity of parties and coalitions. Their results show that only number of tweets mentioning the party can mirror results of elections. Besides, they show that joint mentions of the two parties are correlated to the offline political ties. They prove that political sentiment in tweets can reflect real world political situation. They conclude that Twitter messages can be used as a valid indicator of political sentiment [TSSW10]. In another study done by Bollen et al., researchers try to find the relation between public mood patterns obtained from Twitter by using sentiment analysis and macroscopic social and economic indicators during the same time period [BMP11]. As mentioned previously in section 2.4 of this theses, users tweet about different topics like daily chatter (what one is currently doing), conversations, information sharing (posting links to web pages) and news reporting (commenting on news and current affairs). Even though there is a high variety of topics, the majority of users' tweets are about either themselves or about sharing information. In both cases, mood of the author of the tweet can be collected and analysed [BMP11, JSFT07]. In the study conducted by Bollen et al., psychometric instrument was used in order to extract six mood states from microblogging messages - confusion, depression, anger, fatigue, vigor, tension. They compared their results collected from Twitter with the popular events from media in the same time period. Finally, the conclusion of their study was that social, political, cultural and economic events are significantly correlated to various dimensions of public mood. They also argue that large scale analysis of public mood can be used to determine social and economic indicators [BMP11].

3.2 Twitter sentiment analysis

Sentiment analysis has the ability to automatically measure emotions in online texts, which enables researchers to better analyse online communication. Numerous algorithms were developed in order to automatically detect positive or negative sentiment in the text. On one hand, some of the algorithms detect an object or a topic which is discussed and analyse polarity expressed about that. On the other hand, there are algorithms which identify polarity of overall text [PL08, TBP11]. Various studies analyse sentiment of online text on different platforms like blogs, Facebook, online reviews and online newspaper articles. Nevertheless, a lot of studies tackle sentiment analysis of microblogging messages on Twitter [PL08, OBRS10, BMP11, TBP11, KWM11, PP10, AXV^{+}11, LLG12].

Thelwall et al. in their research investigate whether sentiments in tweets is associated with popular events online. They came to the conclusion that an increase in negative sentiment in tweets is in correlation with important events on Twitter, like Oscars. They also argue that positive sentiment is significantly less correlated to the same type of events [TBP11]. Going further, Kouloumpis et al. in their research investigate how existing linguistic features can be used for detecting sentiment on Twitter. And if sentiment lexicon, that has been developed primarily for formal texts, can be used on short and informal microblogging messages. Conclusion of this research suggests that existing lexicon can be useful, but in combination with other microblogging features like abbreviations, intensifiers and emoticons [KWM11]. Pak and Paroubek went even further in their research [PP10]. They have shown how to automatically collect data from Twitter that can be analysed. Moreover, they show how collected data can be used to determine neutral, positive or negative sentiment in the text. They suggest that since the number of users that share their opinions about products and services is in constant increase, microblogging platforms have become an important source of people's sentiments and opinions. Data collected on microblogging platforms can be used for marketing, because corporations can be interested in for example what people think about their product or, how positive are people about their product, or what would people change about it. In their research they study how sentiment analysis can be used to identify people's opinion on Twitter. They opted for Twitter because of the following reasons [PP10]:

- "Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large."
- "Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interest groups."
- "Twitter's audience is represented by users from many countries."

During their research they collected three sets of tweets depending on emotion shown in the message - positive, negative and objective. Similar to the previous researches, Agarwale et al. introduces new models of Twitter sentiment analysis - tree kernel and feature based models. They found out that combining features with prior polarity words and their parts-of-speech tags give the most successful results. Moreover, they concluded that there is no difference between analysing sentiment analysis of tweets compared to sentiment analysis of other types of texts when using the system they developed [AXV^{+}11]. There are different models of how sentiment of tweets can be analysed. Those models use different types of training data. Some of them use manually labeled data, while others use noisy labels such as hashtags and emoticons. Both types of the models have positive and negative sides. For example, full supervised models which use manually labeled data, are more accurate, but are very time consuming and have a limited amount of data. On the other hand, with noisy label models it is easier to get large amounts of data. However, those models' performance is less accurate. Liu et al. in their research try to combine both models of training data in a new model called Emoticon Smoothed Language Model (ESLAM). They try to train their model using manually labeled data and then to add noisy data to improve performance [LLG12]. As per [PL08] most of the methods used for sentiment analysis use two-step strategy. First step is the subjectivity classification step, where text is classified either as subjective or objective (neutral). Second step is the polarity classification step, where text is further classified as positive or negative. Liu et al. argues that tweets are different from other forms of texts because they are shorter, use slang, contain misspelled words and acronyms. Hence, a lot of new methods of sentiment analysis which specifically concentrate on tweets have been developed. These methods can be divided into two categories: fully supervised methods and distantly supervised methods. Results of their research show that their newly developed method ESLAM can outperform methods which use only one type of training data [LLG12].

3.3 Stock market prediction using Twitter sentiment analysis

Nowadays it is no longer a question whether investor sentiment influences stock market, but it is more about how that sentiment can be measured [STY17]. In the first half of the 20th century it was first noticed that stock markets can be influenced by investors' sentiment. A few years ago that was also demonstrated by a couple of researchers [STY17]. Different research identified two different ways how sentiment can be measured. The first one shows that sentiment can be measured from surveys, while the second one concentrates on the objective variables that are in correlation with investor sentiment. However, because of different limitations, none of the measurements mentioned can directly measure sentiment [STY17]. In research done by See-To and Yang, authors try to examine how stock markets are influenced by individual sentiment dispersion and can it predict future stock market returns and volatility [STY17]. They used Twitter data to extract sentiment and they analysed the data using linear regression and support vector regression. They stated that sentiment dispersion is highly correlated with market volatility and in such a way that volatility will be influenced on the same day or the day after, while stock prices will change only after three days. Also they argue that number of tweets is related to the future stock returns and volatility. Lastly, they concluded that information that can be found on social media like Twitter can be of high importance when it comes to stock market evaluation. Twitter is a dominant platform where investors can share their opinions and this is why it should be considered for achieving more precise results when it comes to stock market predictions [STY17]. Many more scholars research influence of Twitter sentiment on stock markets [BMZ11, RS12, MG12, ZFG11]. Bollen et al. investigate whether collective mood is correlated or can it even predict Dow Jones Industrial Average (DJIA) over time [BMZ11]. DJIA is an index that represents value for 30 largest United States based companies and what was their stock market trading behavior during a certain period of time [OS07]. Prediction of the stock market has attracted a lot of attention from both researchers and investors. Before, it was believed that past and current prices influence future movement of the market. Some research papers of stock market prediction were established on the Efficient Market Hypothesis (EMH) and theory of random walk. EMH suggests that stock market movements are driven by new information. Hence, because one cannot predict the news, random walk theory will determine prices with 50% accuracy. However, different studies show that stock market prices can actually be predicted to a certain level. Going further, even though news cannot be predicted, there are certain indicators that can be extracted from social media, especially from blogging and microblogging platforms [BMZ11]. Beside the news, public mood can also influence stock market prices. From the previous research we know the importance of emotions when it comes to making financial decisions [BMZ11, RS12, MG12, ZFG11]. Results of Bollen et al. research indicate that if specific sentiment dimensions are included, DIJA values can be predicted with much higher accuracy than without those dimensions. Changes in the Dow Jones values will be noticed only after 3 to 4 days. They also found that daily changes of the DJIA closing values can be predicted with 87.6% accuracy [BMZ11]. Even further research was done by Mittal and Goel. Their paper was actually based on Bollen et al. research. They wanted to find relation between "public sentiment" and "market sentiment" by using sentiment analysis and machine learning [MG12]. As a public sentiment indicator they used data obtained from Twitter, and they compared it to the DJIA values from previous days. They divided public sentiment into four categories: Calm, Happy, Alert and Kind. Results from this research show that public sentiment can indeed be obtained from the microblogging platform like Twitter. Further, they found that public moods like calmness and happiness can have a large impact to the DJIA values with a delay of 3 to 4 days. What was also interesting, their portfolio, which was set only to follow their investment strategy indicated a profit over a range of 40 days [MG12]. Even though most of their results are consistent with Bollen et al. results, there are some differences. Firstly, Mittal et al. showed in their research that both happy and calm mood can influence DJIA values, while Bollen et al. states that only calm mood is correlated with DJIA values. Secondly, Mittal et al. could prove only 75.6% of prediction accuracy. On the other hand Bollen et al. in their research mention 87.6% accuracy in prediction DJIA values. [BMZ11, MG12]. Similar work was done by Rao and Srivastava [RS12]. In their research they try to find correlation between tweets and financial market instruments like stock prices, volatility and trading volume. The main purpose of their work was to improve the accuracy of stock market predictions. They derived bullishness and agreement terminologies from positive and negative tweets. Their results show that both positive and negative mood collected from Twitter is highly correlated to short-term price movements on the stock market [RS12]. Likewise, Zhang et al. try to predict stock market indicators like DJIA, NASDAQ and S&P 500 by analysing tweets [ZFG11]. They measure if collective emotions such as hope and fear can influence stock market indicators. According to them, when people are uncertain about the future or when they are pessimistic, they will be more cautious about investing. This is why public mood can be one of the predictors for stock market prediction. Unlike other similar research, they use positive and negative mood words for Twitter analysis. Words were divided into two groups - positive ("hope, "happy") and negative ("nervous", "worry", "fear", "anxious", "upset"). Collective mood was measured simply by counting the number of tweets containing some of the words mentioned above. In addition, they also wanted to see how long does it take for stock market to react on Twitter's public mood. Contrary to previous studies, they found out that when people express more, no matter whether it is positive or negative expression, DJIA values will go down the next day. When people express emotions less, DJIA values will go up. However, they still conclude that emotions collected from Twitter can be a predictor of the stock market movement on the next day [ZFG11].

3.4 Stock market prediction using data search queries

Stock market movements have large influence on both geopolitical events and individual fortunes [PMS13]. One of the articles suggests that Google Search queries can provide early indicators of consumer spending [BMZ11]. Mondria et al. argues that number of clicks on web search results in a particular country is correlated to the investments in that country [MWZ10]. Some further studies that were investigating influence of Google Trends on stock markets suggest that fluctuations in the volume of queries strongly correlate to changes in the transactions done on stock exchanges [PRS10]. Different ways of measuring online interaction can give a new perspective on traders behavior on stock exchanges, especially during the large market movements. Market participants are the ones whose decisions will shape financial data. However, their decision making process is largely driven by information gathering. And nowadays, information gathering is often done on the Internet [PMS13]. As the largest search engine, Google introduced Google Trends, which enables users to get information about volume of different search queries over time. [1] [2]

Preis et al. analyse if changes in volume in Google Search terms related to finance are correlated to stock market variations. They had analysed 98 different search terms connected to stock markets. All data was collected from Google Trends. Their results show that not only Google Trends data can reflect the current state of the market,

[1] `https://www.alexa.com/topsites` (Accessed December, 2018)

[2] `http://gs.statcounter.com/search-engine-market-share` (Accessed May, 2019)

but it can also suggest future trends. In addition, the researchers were able to detect increase in Google search volumes before stock market fell in 2008. Furthermore, they suggest that "early warning signs" collected from Google Trends can be used in creating profitable trading strategies. They also suggest that data collected from Google Trends can have a role in different stages of investor's decision making process. One of those stages is when investors tend to sell stocks due to low prices concern. During those periods of concern people tend to gather more information [PMS13]. Similarly, [BBC$^+$12] tries to prove if daily trading volumes of stocks is correlated to daily volumes of search queries corresponding to the same stocks. In their research, Bordino et al. looks at the stocks that are traded in NASDAQ100 and web search queries from Yahoo search engine. Furthermore, they investigate if individual users' search activity can provide us with insights about collective behaviour. Their results show high correlation between web search queries related to stocks to the trading volume of the same stock. They were able to see correlation on the market in the following days. Moreover, their individual user behavior analysis shows that users usually search for only one stock per month. This means that market can be predicted by uncoordinated activity of many users - "wisdom of the crowd". Finally their findings contribute to the previous research done by Preis et al. - volume of web search queries can be an indicator for early warning signs in financial markets [BBC$^+$12].

3.5 Cryptocurrency market analysis

Research also focused on cryptocurrency markets. Gandal et al. in their paper analyse competition between different cryptocurrencies and also competition between various exchanges where those currencies are traded [GH14]. A different research done by Alessandretti et al. focuses rather on predicting cryptocurrency exchange rates using machine learning. And going further, it explains how analysing simple trading strategies using machine learning can improve investors' gain [AEAB18].

3.6 Bitcoin price prediction

The bitcoin has attracted a lot of attention in previous years [GTMP14, Nak08, Kri15]. It is not under control of any country or a government. Furthermore, it is completely decentralized and transparent [MLM15, Kri15]. The other reason for getting public interest is its highly volatile exchange rate [GTMP14]. This kind of behaviour is very hard to explain using standard economic and financial theories [Kri13]. Even other alternative cryptocurrencies, "altcoins", were not able to endanger its leading position in the cryptocurrency market [Kri15]. However, there are certain downsides of the bitcoin, as it has often been connected to the organised crime and money laundering, and also, it has often been a target for malicious hackers [Kri15]. Even taking those negative aspects into consideration, its high price volatility and how it can be predicted is still point of interest for both investors and scholars [GTMP14, Kri15, Kri13, MLM15, Bla17, GPB$^+$15, Kam14]. One of the research focuses on possible sources of price movement [Kri15].

Researcher investigates behaviour of different sources such as speculative, fundamental and technical sources. He points out that both time and frequency are important in order to determine bitcoin price. Firstly, his results show that over a long term standard fundamental factors like money supply, price level and usage in trade will play a significant role in bitcoin price. Secondly, he found out that users get motivated to become miners with increasing price of the bitcoin. Although, this effect disappears over time, because hash rates and difficulty are getting too high. Thirdly, Kristoufek argues that bitcoin price is highly driven by investors' interest [Kri15].

Besides concerns that bitcoin is used for criminal activities and money laundering, some economists were worried if bitcoin acts more as a speculative investment than a currency [Bla17]. Blau, in his research concentrates on two things. First, he tries to find out facts about the price dynamics of bitcoin. Second, he tries to see if speculative trading influences bitcoin exchange rate volatility. He also argues that fast price dynamics with high spikes can indicate the presence of a bubble. During his sample time, bitcoin volatility was 6%, which is double than average volatility compared to other currencies. Blau's study shows that even though there was a significant rise in the value of the bitcoin, speculative trading was not following that trend. Furthermore, he also proves that bitcoin price volatility is not directly correlated to speculative trading [Bla17].

Similarly to Kristoufek's research, [GPB$^+$15] investigate which factors can determine bitcoin price in shorter and longer time periods. They researched if Twitter sentiment analysis and frequency of Wikipedia views can have an influence on the bitcoin price. First, sentiment ration of Twitter users that were interested in bitcoin was measured by machine learning algorithm - Support Vector Machines. They came to the conclusion that Twitter sentiment ratio has positive influence on the bitcoin price in the short term. This means that collective mood can actually predict bitcoin price in the short time period. Second, they investigated if bitcoin price can be affected by the number of Wikipedia search queries. Basically they tried to see if the public has higher interest in the bitcoin, will market price go higher. Georgoula et al. came to the conclusion that the price of bitcoin is positively affected by higher frequency of Wikipedia views related to bitcoin. Third, they prove that increase in the hash rate, which indicates mining difficulty, will also have a positive effect on bitcoin price [GPB$^+$15].

Kaminski in his research analyses what are correlations and causalities between bitcoin price and Twitter [Kam14]. On the contrary to the previous research, instead of analysing public sentiment on Twitter, he analysed emotional signals. His data set was assembled from tweets that had term "bitcoin" together with one other positive, negative or uncertain related term. For instance some of the positive terms include: "happy", "fun", "good", "love". On the other hand, some of the negative terms include: "bad", "sad", "unhappy". Uncertainty was determined by the following terms: "fear", "hope", "worry". He found that static correlation and dynamic causality give different results of analysis. The static correlation (Pearson correlation) shows high correlation between emotions in tweets and bitcoin price and trading volume. On the contrary, a dynamic causality (Granger causality) does not confirm that sentiment of tweets predict price of a bitcoin.

It rather looks that emotional tweets are mirroring the bitcoin market [Kam14]. Garcia et al. examined if traces of online actions and interactions of individual users can influence growth of bitcoin price. And also how these online traces can be used to find the link between public interest and market dynamics [GTMP14]. They investigated relation among bitcoin price and social signals like volume of word-of-mouth social media communication, information search volume and user base growth. Vector autoregression helped them to define two positive feedback loops - social cycle and user adoption cycle. One is complementing the other. Authors argued that the social cycle has influence in the decisions of individuals to buy bitcoins. They define social cycle as follows: "Bitcoin's growing popularity leads to higher search volumes, which in turn result in increased social media activity on the topic of Bitcoin. More interest encourages the purchase of bitcoins by individual users, driving the prices up, which eventually feeds back on the search volumes." [GTMP14]. Furthermore, they define user adoption cycle as follows: "new Bitcoin users download the client and join the transaction network after acquiring information about the technology. This growth in the user base translates to a price increase, as the number of bitcoins available for trade does not depend on demand, but rather grows linearly with time." [GTMP14]. They concluded that higher user activity will actually correspond faster to negative events related to bitcoin network. This means that higher user activity can actually be an indicator for the price of bitcoin to drop [GTMP14]. These findings actually correspond to what Zhang et al. suggested in their research [ZFG11].

When it comes to analysing search queries in order to determine correlation to bitcoin price, Kristoufek went even further [Kri13]. In his research he combines search queries related to bitcoin from Google Search and from Wikipedia. In order to interpret values from Google Search he uses Google Trends. Similar as Garcia et al., Kristoufek claims that frequency of search terms related to cryptocurrency can be a good measure of public interest in that particular currency. The author argues that there is a strong correlation between bitcoin price and frequency of searched terms on both platforms. However, Garcia et al., based on their research, would not necessarily agree with that [GTMP14]. Furthermore, Kristoufek states that the relation between volume of searched terms and bitcoin price is causal and bidirectional. This means, smiliary to what Garcia et al. proposed in their paper, that search queries are influencing price, as well as price is influencing search queries. They concluded that when prices are high, increasing interest in the bitcoin market pushes prices even further up. On the opposite, if prices are low, increasing interest pushes prices even further down [Kri13, GTMP14].

Shen et al. argues that well-informed investors, who already have some previous knowledge about the cryptocurrencies will not search for it on Google, but they will rather tweet about it [SUW19]. They suggested that tweets may contain more informed content than just searching term on the Google Search. Investors may comment new posts related to bitcoin or make some sort of predictions which might have influence on the bitcoin price. Hence, the authors of this paper analysed volume of tweets connected to bitcoin rather then search terms on Google or Wikipedia like previously mentioned authors [GTMP14, Kri13]. They state that the volume of tweets is stronger measure than frequency of web search terms. Shen et al. found that the volume of previous day tweets is high indicator for trading volume and realized volatility. However, it doesn't have any influence on returns. All of the results are supported by linear and non linear Granger causality tests. They concluded that the volume of tweets can be used to predict realized volatility and trading volume of bitcoin [SUW19]. One other research analysed relation between bitcoin and US Dollar using some of the most popular parametric distributions in finance [CNC15]. Their findings show that generalized hyperbolic distribution gives the best result. They also gave predictions of the exchange rate for the upcoming three, six and nine years [CNC15].

CHAPTER 4

Methodology

This chapter describes the methodology used in order to successfully predict bitcoin exchange rate by using sentiment analysis on tweets and Google Trends data. In this work we will use bitcoin exchange rate and bitcoin price as two interchangeable terms. We describe data that was obtained by using Twitter and Google public APIs, and the format of that data. Retrieved data had to be pre-processed before it could be used for predictions, and that process is also described in this chapter. Two sentiment analysis algorithms were used in order to analyse tweets and compute data used for predicting the bitcoin price. Both are described in this chapter, and we also discuss how we used them. Finally, statistical methods used to evaluate bitcoin price and bitcoin trading volume are defined.

4.1 Data collection

4.1.1 Twitter data collection

All data from Twitter has been gathered from the Twitter Application Programming Interface (API) [1]. More specifically, Twitter Search REST (Representational State Transfer) API was used to obtain tweets that match our search criteria. Every HTTP request matched a unique set of tweets, and corresponding HTTP response contained detailed information about each tweet. Because this API imposes limits on how many tweets can be retrieved during some time period, we had to be careful not to fetch too many tweets and cause Twitter to blacklist our IP address. Additional limitation of this API is the limit of number of tweets that can be downloaded in a single API request i.e. the API supports paging and it expects clients to handle this behavior.

[1] https://developer.twitter.com/en/docs/api-reference-index (Accessed March, 2019)

Example of an HTTP request used to download tweets is `https://twitter.com/search?q=bitcoin%20until%3A2018-07-04%20since%3A2018-07-03&src=typed_query`. This is the first, and initial, request sent to Twitter API. Once response is received, it contains data in JSON (JavaScript Object Notation). This data is parsed and raw tweets data is stored to disk. Each request returns 20 tweets, which means that we had to make almost 46.000 API requests in order to fetch all tweets for this work. In addition to tweet data, each HTTP response also contains a unique key that is used to obtain the next page of results. For example, second API request contains additional parameter, *max_position = thGAVUV0VFVBaAgKOhnYGC0BwWgMC-qdDBgtAcEjUAFQAlAFUAFQAA*, which instructs that API to provide the client with the next page of results matching the search criteria of the first request.

One of the search criteria is that tweets must contain hashtag '#bitcoin'. Additionally, data is collected for the period between the 1st of July 2018 and the 31st of August 2018. In that time period we have collected a total of 918.354 tweets. Fetched data is stored in CSV files and total size of the files on disk is 320 MB. Each data record represents one tweet and it stores information about:

- Username of the user that has created the tweet
- Date of the tweet
- How many times particular tweet has been retweeted
- How many times particular tweet has been liked
- Text of the tweet itself
- Geographical information
- Mentions of other users that appear in the tweet
- Hashtags that are used in the tweet
- ID of the tweet
- Link to the original tweet on Twitter

Some examples of the data collected can be seen in the table 4.1. Due to space limitations, not all parameters have been shown in the table.

A number of validation checks has been ran in order to confirm the correctness of data. Firstly, manual checks were performed. This consisted of reading through the CSV file that stores all tweets, and visually confirming that everything looks good. Following that, 30 random tweets were selected from the file, and links to those tweets were visited in the Internet browser. Opening those links loaded the tweet data, and for all 30 tweets, data matched the one in the CSV files. Automated checks were also ran. One of them consisted of loading all tweet IDs, and making sure they are unique in the CSV file. This check found no duplicates. Second test consisted of trying to detect any invalid data by comparing tweet content. This analysis showed that the most common text appears in 2768 tweets (0.3% of the total number of tweets), and the second most common one in 2.016 tweets. Upon closer examination, we found that these tweets are all valid, and that they contain links to free giveaways. There is a possibility that these are created by an automated robot accounts, but this work does not try to detect that, and considers those tweets as part of the input data. Third automated test showed there are 90.236 different users in our data set, which means on average users create 5 tweets a month. However, top 10 users (by number of tweets) account for 62.037 tweets. Expanding this check to top 100 users, 223.881 tweets is generated by them, which is 25% of all tweets. This is in line with our expectations, as most of the social media sites have a handful of users that generate the majority of the content. [2]

Downloading tweets by using the Twitter Search API means that there are non-English tweets in the stored CSV file. As we are considering only tweets in English, firstly we had to separate English tweets from the rest of the original data set. For this purpose we developed filtering algorithm that should help us do that. The algorithm processed each tweet separately by analyzing the content of tweets. First step consisted of locating all URLs that are in the text, and removing them. A simple heuristic was used for this purpose and it finds all words that start with http or https, and removes them entirely. The second step of this algorithm simply removes any hash (#) signs from text, that are used as prefix for hashtags. Running these two steps transforms e.g. "Check https://bitcoin.org for #bitcoin data" is transformed to "Check for bitcoin data". Once tweets are processed, a heuristic is used to determine if tweets are in English. For all characters in the tweet, digits and punctuation is ignored. Remaining characters are analyzed and tweet is considered to be in English if at least 70% of the remaining (non-digit, non-punctuation) characters are from the English alphabet. After filtering data, we had a total of 870.778 tweets in English that we could process further.

4.1.2 Google Trends data collection

Another data source used in this work is information from Google Trends. Together with the collected tweets, this is used to predict the hourly trading volume. Data was obtained from the Google Trends REST (Representational State Transfer) API. This API allows clients to submit requests about a search term they are interested in, and Google Trends data about it is provided back. Similarly to Twitter Search API, number of requests during some period of time is limited. However, because we had to obtain 1.488 data points, for each hour of 62 days that we are analyzing, this was not a concern.

In order to download data, an HTTP request is sent to `https://trends.google.com/trends/api/widgetdata/multiline`, and it contains all parameters that spec-

[2] `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/387591/use-of-social-media-for-research-and-analysis.pdf` (Accessed August, 2019)

Username	Date	Retweets	Favorites	Text	Hashtags	Id	Permalink
Datavetaren	05/07/18	756	2108	Fiat money is essentially proof-of-stake where the ones with most stake run the full nodes (banks) and everybody else runs SPV. Whatever changes to consensus (money supply) we just have to accept. Our choice is to hard fork this system and start a new currency. #bitcoin	#bitcoin	1013375217094160000	https://twitter.com/Datavetaren/status/1014908808492175363
Crypto_Bitlord	02/07/18	18	204	#Bitcoin is not money. Money has inflation so you spend/invest that shit. Bitcoin is digital gold.	#Bitcoin	1013917422162590000	https://twitter.com/Crypto_Bitlord/status/1013917422162591744
LandM_Marius	06/07/18	28	186	Silence...before the storm. ALT coins will out perform #Bitcoin	#Bitcoin	1015147817973700000	https://twitter.com/LandM_Marius/status/1015147817973702656
SalihSarikaya	05/07/18	25	184	What are your most common regrets with #crypto and #bitcoin investing?	#crypto #bitcoin	1014657934301670000	https://twitter.com/SalihSarikaya/status/1014657934301675521
SJBMWGroup	05/07/18	101	180	Stephen James is now accepting Bitcoin for the purchase of your new BMW! Contact us today to find out more about how you can use Bitcoin to walk away with a brand new vehicle #bitcoin #bitpay pic.twitter.com/qN1Bq5Juc1	#bitcoin #bitpay	1014887028746570000	https://twitter.com/SJBMWGroup/status/1014887028746571777
cburniske	02/07/18	50	170	One of the most robust explorations I've read on the economic & game theoretic limits of #Bitcoin #PoW more broadly: http://www.nber.org/papers/w24717 pic.twitter.com/dUc7JdVrPl	#Bitcoin #PoW	1013795776860950000	https://twitter.com/cburniske/status/1013795776860958727
SimonDixonTwitt	04/07/18	31	160	If #Bitcoin was not on track to make $ USD and the USA less important in the world then I would pick Japan to be the next superpower of the world if I had to pick one country. pic.twitter.com/KqSKuyWpfT	#Bitcoin	1014564686791090000	https://twitter.com/SimonDixonTwitt/status/1014564686791094272

Table 4.1: Example of Twitter data

ify the data. In our case, parameters indicated that data should start on the 1st of July 2018, and end on the 31st of August 2018. In addition to that, time zone for those dates is specified (UTC timezone in this case). The API allows to specify region of the world to analyze the interest for (e.g. 'US' for the United States, or 'AT' for Austria), but because we are interested in the world wide interest, this option was not used. Because some search terms can be ambiguous, one can add a parameter describing the category, thus narrowing down the data. E.g. searching for "cat" may yield different results if category "animals" is selected, compared to selecting "shopping" where it most likely refers to the brand "Cat". In our case, no category was selected. Finally, the last two parameters used for the HTTP request were the ones to indicate that data should be per-hour, and of course, that we are interested in the term "bitcoin".

Every response from the API contains, at most, information about a full week. This means that we were able to obtain hourly information about global interest in term "bitcoin" in about 10 HTTP requests. Data is returned from the API in JSON (JavaScript Object Notation) format, and client parses that stream of information and stores it in a CSV file. Every data point contains the following information:

- Start time (in UTC) for the one-hour window the interest is computed for e.g. "2018-07-01 00:00:00".
- Interest, which is a value between 0 and 100 e.g. 79.
- If data is partial, which indicates if data fully known for the period we are interested in. In our case, no entry contains partial data.

Google Trends expresses interest in a particular term by using relative values that show how the interest evolves during time. E.g. for term "bitcoin" the figure 4.1 shows how interest was close to 0 for many years, and started to increase only recently.

As already mentioned, response contains per-hour information for one week. In more details, raw Google Trends data can be described as $x_{w,0}, x_{w,1}, \ldots, x_{w,168}, x_{w,169}$ where $1 <= w <= 8$, represents whole weeks, and $x_{9,1}, x_{9,2}, \ldots, x_{9,144}$ where $w = 9$ represent a partial week (week that has only six days). This raises an interesting problem. Each response can be thought of as a series of 169 integer values (there are 7 days, and 24 hours per day, and the first hour of the next week). These numbers depict how interest changes from the start of that week, until the end of that week. Across weeks, these numbers are not related in any way. However, the API is quite helpful here because at the week boundaries e.g. "2018-07-08 00:00:00" there are two data points. This means that $x_{w,169}$ and $x_{(w+1),1}$ are specifying trend data for the same time. The first data point is the last data point when fetching information for the 1st of July until the 7th of July. The second data point is for the first hour of the week between the 8th of July until the 15th of July. E.g. actual data obtained measures level of interest 73 for the first data point, and the 81 for the second one, where these two data points refer to the same time period. This means that interest 73 from the first week, corresponds to interest 81 in

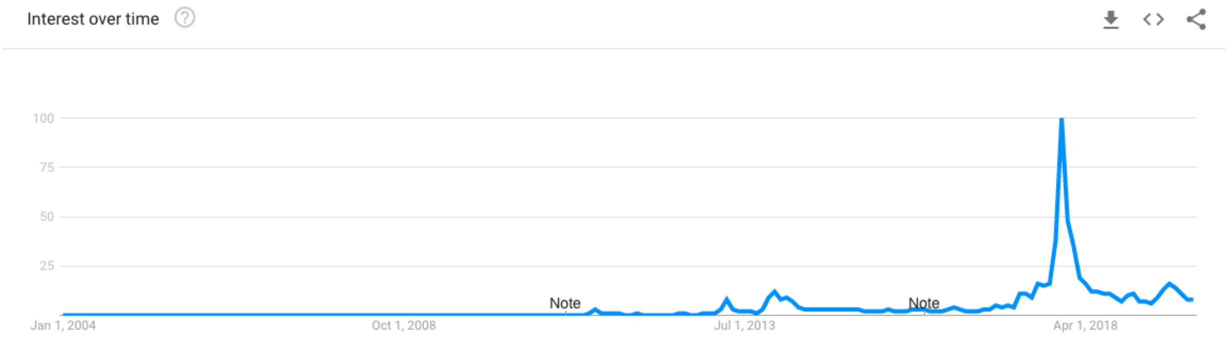


Figure 4.1: Interest in bitcoin over time (`https://trends.google.com`)

the second week. In order to correlate the values across the first and the second week, all values in the second week are multiplied by $^{73}/_{81}$. In order to get the interest during the entire time period, the same transformation is applied between every week boundary. There are 8 overlapping weeks to process in our data set. In general, transformed data looks like:

- First week, consisting of 168 data points:

$$x_{1,1}, \ldots x_{1,168} \tag{4.1}$$

- Second week:

$$z_2 = \frac{x_{2,1}}{x_{1,169}} \tag{4.2}$$

$$x_{2,1} * z_2, x_{2,2} * z_2, x_{2,3} * z_2, \ldots, x_{2,168} * z_2 \tag{4.3}$$

- Third week:

$$z_3 = z_2 * \frac{x_{3,1}}{x_{2,169}} \tag{4.4}$$

$$x_{3,1} * z_3, x_{3,2} * z_3, x_{3,3} * z_3, \ldots, x_{3,168} * z_3 \tag{4.5}$$

- N-th week:

$$z_n = z_{(n-1)} * \frac{x_{n,1}}{x_{(n-1),169}} \tag{4.6}$$

$$x_{n,1} * z_n, x_{n,2} * z_n, x_{n,3} * z_n, \ldots, x_{n,168} * z_n \tag{4.7}$$

After the data has been transformed, it has been saved in a CSV file that contains 1488 data points, corresponding to 1488 hours that we are analyzing. The interest for every hour is a positive double number. Mean interest is 69, with the median being 66.7. The lowest interest in term “bitcoin” is 37.33, that happened in the 515th hour which start on the 22nd of July, at 10AM. When it comes to peak interest, it happened in the 403rd hour, which corresponds to the 17th of July, at 6PM.

4.1.3 Bitcoin price and volume data collection

The last piece of information necessary for this work is price and volume data on bitcoin. Price refers to bitcoin (trading symbol BTC) to US Dollar exchange rate i.e. how much a single bitcoin is worth in USD. Volume is defined as number of bitcoins traded during a specific time window.

Raw data for this purpose was obtained from BitcoinCharts website [3], that is recommended as a source of information on the official Bitcoin website [4]. This website stores bitcoin related information, and it also provides data from multiple bitcoin exchanges. Individual transactions from many exchanges are stored in CSV files, and a single transaction corresponds to a single row in the file. Data about a transaction includes timestamp, price in USD, and volume of BTC traded. Timestamp is number of seconds since the 1st of January 1970, 00:00:00 UTC time zone.

When talking about bitcoin exchange rate, in this work we refer to exchange rate at Bitstamp exchange [5]. This bitcoin exchange is one of the largest in the world, and it processed 18% of all BTC transactions in USD between the 15th of September 2019 and the 15 of October 2019. [6]

This exchange is one that has been identified as trustworthy, and to be providing actual transaction data. [7]

When it comes to the time period that we are investigating, Bitstamp is one of the top exchanges during that time as well. For example, using data available on Blockchain.org [8], we can find the total bitcoin trade volume in USD for major bitcoin exchanges per day for the period we are analyzing. On the other hand, by using raw data that we have about Bitstamp transactions, we can compute trading volume in USD for every day. A single transaction volume is computed by multiplying the number of traded bitcoins by exchange rate in that transaction (both of these are recorded in a single row of the CSV file), and trading volume is computed by adding up all of the transaction volumes. We can see that Bitstamps accounts for 20% of USD exchange trade volume among major bitcoin exchanges on the 2nd of July 2018. For the 23rd of July, this share is 27%. This confirms that Bitstamp is a credible source for bitcoin to USD exchange rate, and that

[3] http://api.bitcoincharts.com/v1/csv (Accessed September, 2019)

[4] https://bitcoin.org/en/resources#charts (Accessed September, 2019)

[5] https://www.bitstamp.net (Accessed September, 2019)

[6] https://bitcoincharts.com/charts/volumepie/ (Accessed Ocotber, 2019)

[7] https://www.sec.gov/comments/sr-nysearca-2019-01/srnysearca201901-5164833-183434.pdf (Accessed October, 2019)

[8] https://www.blockchain.com/charts/trade-volume?timespan=2years(Accessed Ocotober, 2019)

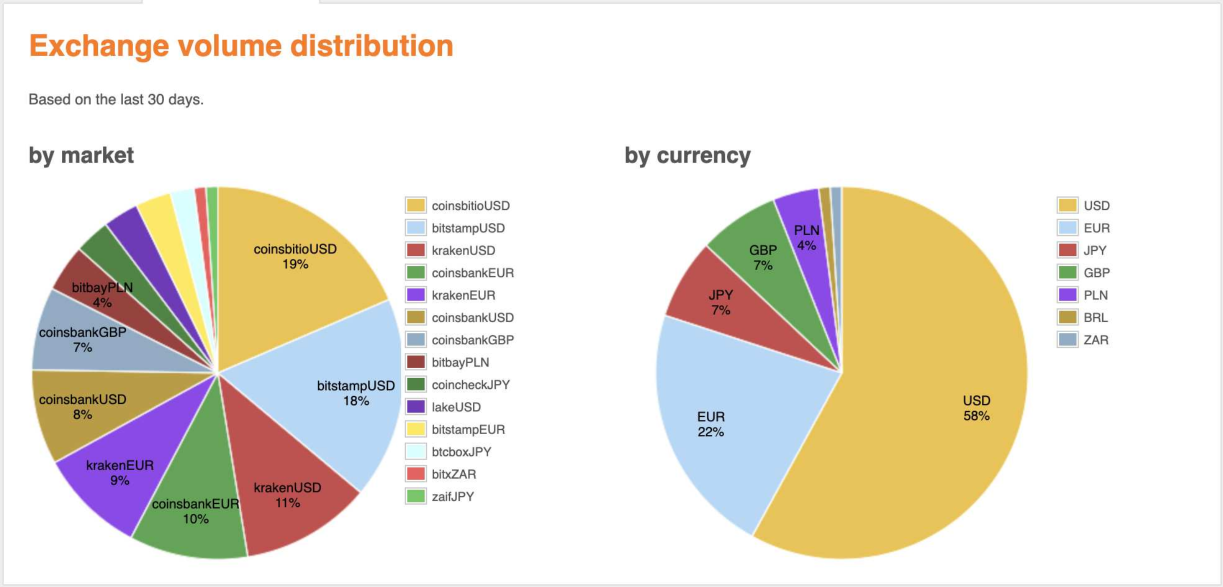


Figure 4.2: Exchange Volume Distribution (`https://bitcoincharts.com/charts/volumepie`)

we can rely on it in this work. The same holds for bitcoin volume. Also, for this exchange we are able to obtain comprehensive transaction data that we can use in the thesis.

Raw data downloaded from BitcoinCharts contains 1.4 GB of Bitstamp transactions data, including transactions as early as the 13th September 2011. Total number of transactions is 34.675.070. This data has been filtered to include only transactions between the 1st of July 2018 (UTC time), and the 1st of September. The number of transactions for that period is 1.538.551, with total size of 41 MB. An example transaction is "1530403241, 6392.09, 0.05479034", which means that on the 1 July 2018 00:00:41 UTC, 0.05479034 BTC was traded, with USD price of 6.392,09. This means that a total of 350,22 USD was traded in this transaction.

In order to determine bitcoin exchange rate for every hour of the two month period that we are analyzing, we choose the latest transaction during that hour and exchange rate for that transaction is considered to be the exchange rate for that hour. For example, exchange rate for 1st July 2018 00:00 UTC is the exchange rate of the latest transaction between 00:00 UTC (inclusive) and 01:00 UTC (exclusive) during that day. After this processing, we are left with 1.488 data points, each one representing the bitcoin price. This data has been validated by comparing minimum price, maximum price, and 20 random data points against price information available at `https://www.bitstamp.net/market/tradeview/` the official data source for Bitstamp exchange. Validation confirmed that data obtained from BitcoinCharts is correct. Minimum price is 5.941,33 USD for a single BTC, and maximum price is 8.427,36 USD for BTC. For the period we are analyzing, average price is 6.893,86 USD.

Secondly, in order to determine bitcoin hourly trading volume, we used the same raw data. Every row in the CSV file represents a single transaction, and the third column represents the number of bitcoins in that transaction. Bitcoin volume for a single hour is defined as the total amount of bitcoins traded in all transactions with timestamp greater or equal to that hour and strictly less than the next hour. For example, in order to compute bitcoin volume for the 1st of July 2018 00:00 UTC, we add up the number of bitcoins traded in all transactions between the 00:00 UTC (inclusive) and 01:00 (UTC) exclusive during that day. Data obtained in this way has been validated against the official Bitstamp data available at `https://www.bitstamp.net/market/tradeview/` by choosing 20 random data points. There were no discrepancies found between these two data sources. Total number of data points is 1.488, where the first hour is 1st of July 2018 00:00 UTC, and the last hour is 31st August 2018 23:00 UTC. The lowest trading volume was recorded during the 1011th hour, 12th of August 2018 02:00 UTC and 12th of August 2018 03:00 UTC, when it was 15,231848 BTC. When looking for the hour that has the highest trading volume we find that it was during the 37th hour, on the 2nd of July 2018, between 12:00 UTC and 13:00 UTC. Total of 2.760,205333 bitcoins was exchanged during this hour. Average hourly trading volume was 324,555127 bitcoins.

4.2 Sentiment analysis algorithms used

A lot of messages are posted daily on different social network sites, forums and blogs. Researchers have shown that emotions play an important part in those informal messages, especially because people can feel and share emotions through computer mediated communication (CMC). [WP02] Hence, new algorithms were developed in order to automatically identify emotions from short informal texts and to help understand their role in everyday online communication. Furthermore, algorithms for sentiment detection can potentially identify hater speech, inappropriate or offensive posts and threatening behaviour. However, most of the algorithms for opinion mining are not designed to detect user behaviour, but to detect opinions about products. [TBP+10]

There are several different forms how algorithms detect sentiment. Majority of algorithms identify whether text is positive, negative or neutral (trinary form). Some algorithms use binary form where they only detect positive or negative sentiment. Other algorithms use scales in order to identify the strength of the sentiment, e.g. 5 (strongly positive) to -5 (strongly negative). [The] However, this is not always sufficient, since mix of sentiments can be present in one text. Hence, some of the algorithms concentrate to identify different sentiments in parallel by using dual scales. One of the other challenges, when it comes to online texts, is that they are often informal and they do not follow rules of spelling or grammar. Furthermore, many times they can include words abbreviations (e.g. “u” - you; “lol” - laughing out loud) and emoticons (e.g. “:)”). [The, TBP+10]

Algorithms for sentiment analysis use one of the two approaches when it comes to determining sentiment: lexical or machine learning approach. A lexical approach will often have a list of sentiment words and their polarities. It will combine these with grammar rules. In contrast, machine learning will convert particular text to the list of words, and based on some previous rules which were human coded, it will "learn" how to associate sentiment score to each word or pair of words. Both of the mentioned approaches have similar level of accuracy. [The]

The first sentiment algorithm that will be used in this thesis is the Stanford CoreNLP toolkit, set of tools specialized in the natural language processing with support for sentiment analysis [MSB+14]. The second algorithm that will be used was developed specifically to cope with short messages and informal spelling of words. [PL08, Sen19] It is called SentiStrenght, and it showed good results when analysing short messages from a social network site. [TBP11]

4.2.1 Stanford CoreNLP algorithm

The majority of algorithms for sentiment prediction only isolate single words for the analysis, meaning that positive points will be given to positive words and negative points will be given to negative words, in the end summing up those points. In that way, the order of words is disregarded and some important insights may be lost. For example, in the following sentence "That film was neither funny nor good." even though both of the attributes are positive ("funny", "good") the whole sentence has a negative overall sentiment. On the opposite of the majority of algorithms, Stanford CoreNLP algorithm defines sentiment based on the whole sentence structure. It's sentiment determination is based on the meaning of the words composed in longer phrases. In that way, this particular model is not easily misled compared to other models. [SPW+13, MSB+14] In order to accomplish that, the Stanford CoreNLP natural language processing toolkit uses the Stanford Sentiment Treebank. The sentiment Treebank has fully labeled parse tree which enables analysis of the effects of compositional models on sentiment in language. This corpus is based on the set extracted from the movie reviews. It has 11,855 single sentences and it includes 215,154 unique phrases, which were annotated by three human judges. This data set actually enables to capture sentiment in more complex sentence structures. When this model is used on the new data set it outperforms previous models with 5.4% more accuracy in determining sentiment in a single sentence. Furthermore, this model is the only model that can capture negation in both positive and negative phrases. [SPW+13]

In this thesis we use the Stanford CoreNLP algorithm in the following way. All tweets written in English are analyzed using this algorithm. Although Twitter API did not allow us to download tweets written only in English, we developed an algorithm to filter out non-English tweets (using the method described in "Twitter data collection" subsection). This step is important, because the sentiment analysis relies on language-specific pre-computed model. More specifically, in order to compute sentiment for some text, algorithm needs to be able to understand what is a single word, and what is a sentence. This step is called tokenization. Further on, the Stanford CoreNLP must know how words are used. It must differentiate verbs and nouns, plural and singular forms etc.

This step is known as part-of-speech tagging. The final pre-requirement for sentiment analysis is full syntactic analysis of the sentence, which determines how to group words, what is a subject, what is an object etc. In Stanford CoreNLP, this is performed by ParserAnnotator. All of these steps are using pre-trained models, and all of them are specialized for the English language. There is support for additional languages, such as Chineese, German, French, but we are not relying on it in this work.

Now that we have some insights into how the sentiment analysis works, let's see how it is applied for tweet processing. The library we are using consists of multiple annotators that are composable, and those annotators add additional metadata to processed text. As described in the previous paragraph, there are also some dependencies between annotators. In order to compute sentiment for a sentence, we first need to build a tree that represents how different words in the sentence interact with one another, and only once this tree is available we are able to compute sentiment. First we build a processing pipeline that consists of all annotators we require for sentiment analysis, including the annotator that computes sentiment. Every tweet is processed independently, because that corresponds to how users are reading them. For each tweet, Stanford CoreNLP pipeline runs, it splits the tweet into sentences, and it assigns sentiment to each of those sentences. Sentiment values are "very negative", "negative", "neutral", "positive", "very positive". For a single tweet, sentiment consists of one or more of these values. E.g. tweet with ID 1017377774347186177 has sentiment "Neutral, Neutral, Negative, Negative", which means that the first and the second sentence are neutral, while the third and the fourth are negative.

We define sentiment of tweets by computing the average sentiment across all sentences in a single tweet. First, we assign values to each of the sentiment outputs. Very negative sentiment has value 0, negative has value 0,25, neutral 0,5, positive 0,75, and finally, very positive has value 1. This means that a tweet with a very negative sentence and a very positive sentence will have sentiment 0,5, and tweet with two neutral sentences will have sentiment 0,5. Using the input data, there were 5.229 "Very positive", 161.047 "positive", 345.038 "neutral", 1.014.612 "negative", and 5.317 "very negative" sentences. Now that we are able to compute sentiment for a single tweet, we can also define hourly tweet sentiment. For a specific hour, we compute the sentiment as average sentiment for all tweets with timestamp between start of that hour (inclusive), and the next hour (exclusive). Data obtained this way shows that the minimum recorded sentiment is 0,295873, and this happened during the 655th hour, on the 28th of July 2018, between 6AM UTC and 7AM UTC. The most favorable sentiment was 0,373809, during the 402nd hour, on the 17th of July 2018, between 5PM UTC and 6PM UTC. On average, hourly sentiment is 0,328664, which shows that Stanford CoreNLP algorithm classifies bitcoin related tweets as mostly negative.

4.2.2 SentiStrength algorithm

SentiStrength uses a lexical approach to determine sentiment in short informal texts. Besides a dictionary of sentiment words and their strength, it has rules to detect sentiment in online social communication, such as emoticon list, deliberate misspellings, idioms and repeated punctuation. Its lexicon is based on 2.310 sentiment words which were obtained from Linguistic Inquiry and Word Count (LIWC) program and the General Inquirer list of sentiment terms. Some of ad-hoc additions, particularly for CMC words, were made during development and testing. [The] SentiStrength was developed in 2010 based on 2.600 MySpace comments for which sentiment and strength were initially assigned by humans. Afterwards, it was assessed on a random sample of 1.041 MySpace comments. [TBP+10] SentiStrength first reads a text and then it splits it into words. It also separates punctuations and emoticons. After that, words are matched against the sentiment lexicon. If a match is found, SentiStrength assigns sentiment score to that word. The scores can be positive (from 1 to 5) and negative (from -1 to -5). The overall sentiment for that sentence is the highest positive and the highest negative score. If there are multiple sentences, maximum scores of individual sentences will be taken. [The]

As mentioned previously, beside the lexicon, SentiStrength has a list of emoticons that also have human-assigned sentiment scores. However, sometimes automatic extraction of emoticons can be challenging, for example if they are followed by some other punctuation which is not part of the emoticon. Going further, SentiStrength has a list of idioms. Each of the idioms has a sentiment score which can override the lexicon score. This is because the meaning of the idiom can be different from the meaning of the original words. [The]

In addition, SentiStrength has a list of rules which also help when it comes to informal texts and special cases. As per Thelwall [The], some of the rules are as follows:

- “The word “miss” is a special case with a positive strength of 2 and a negative strength of -2. It is frequently used to express sadness and love simultaneously, as in the common phrase, "I miss you".”
- “A spelling correction algorithm deletes repeated letters in a word when the letters are more frequently repeated than normal for English.”
- “At least two repeated letters added to words give a strength boost sentiment words by 1.”
- “A booster word list is used to strengthen (e.g., very +1) or weaken (e.g., somewhat -1) the sentiment of any immediately following sentiment words.”
- “A negating word list is used to neutralise any following sentiment words (skipping any intervening booster words). (e.g., “I do not hate him”, is not negative).”
- “Sentences with exclamation marks have a minimum positive strength of 2, unless negative.”
- “Repeated punctuation with one or more exclamation marks boost the strength of the immediately preceding sentiment word by 1.”
- “Sentiment terms in capital letters receive a strength increase bonus of 1.”

- “Two consecutive moderate or strong positive terms with strength at least 3 increase the strength of the second word by 1.”

On the regular PC, SentiStrength can process around 16.000 tweets per second. During the 2012 Olympic Games held in London, SentiStrength was used to power the light on the London Eye by monitoring average sentiment of tweets which were Olympic related. Among other commercial users is also Yahoo. [The]

In this thesis we use SentiStrength algorithm in the following way. All tweets in English are first stored in a single CSV file, with each row containing entire tweet information as retrieved from the Twitter API. Processing these tweets one by one produces a CSV file that contains only tweet text. Every word has associated sentiment, all words comprising a sentence define the sentence sentiment, and all sentences within a single tweet define the tweet sentiment which is expressed as two numbers, with the first one being between -5 and -1, and the second one between 1 and 5. The former represents negative sentiment, with -5 being the strongest negative sentiment, and -1 being the weakest negative sentiment. The later represents positive sentiment, with 1 being the weakest positive sentiment, and 5 being the strongest. In order to compute tweet sentiment, we add up those two numbers. This means that if tweet has SentiStrength computed sentiment [-2, 2], we define it to have sentiment 0. In case the sentiment is [-1, 3], the sentiment is 2. This means that sentiment is defined as a number between -4 (when negative sentiment is -5 and positive 1) and 4 (when negative sentiment is -1 and positive is 5).

In order to compute hourly tweet sentiment we compute average sentiment for all tweets with timestamp during that hour. Analyzing all tweets we can see that the maximum sentiment value of 0.259076 is recorded during the 856th hour, on the 5th of August 2018, between 3PM UTC and 4PM UTC. When looking for the minimum sentiment value, it is -0,143448, on the 1049th hour, on the 13th of August 2018, between 4PM UTC and 5PM UTC. Average sentiment is 0.065638, which shows that SentiStrength algorithm considers tweets to have slightly positive sentiment. This contrasts the output we got from Stanford CoreNLP sentiment analysis, that classified tweets as having almost negative sentiment on average.

4.3 Predicting bitcoin exchange rate

To predict bitcoin exchange rate in this work we use multiple input signals. However, in order to show that these signals are indeed useful when trying to predict the bitcoin exchange rate, we perform multiple tests. To forecast future results, we are using vector autoregressive model (VAR). Also, we make sure all signals are stationary i.e. that properties of the time series that we are analyzing do not change over time, and that the signals are correlated. This greatly improves our ability to make predictions. [BD16]

Vector autoregression is mainly used in macroeconomics for forecasting and analysing. Even though it is simpler compared to large-scale structural equation systems, it gives good or even better results than them. Besides its primary functions it is also used for testing Granger causality. Vector autoregression can be presented through the following formula:

$$y_t = \mu + \Gamma_1 y_{t-1} + ... + \Gamma_p y_{t-p} + \epsilon_t \tag{4.8}$$

Where y_t is a vector which represents the state of the system at hour t, and ϵ_t is vector of non autocorrelated disturbances with zero means. [Gre03]

Let us now define how our model looks like. In the previous sections of this thesis, we have seen how raw input data was obtained. When building the prediction model, we have further processed the data in order to enhance our ability to predict the bitcoin exchange rate. We define bitcoin exchange rate as time series consisting of 1488 data points, $price_0, \ldots, price_{1487}$, where $price_0 = 0$, and $price_t = rawPrice_t - rawPrice_{t-1}$ i.e for specific hour we are keeping track of the change in the price from the previous hour. We define sentiment as $sentiment_t = rawSentiment_t$ as obtained from the Stanford CoreNLP and SentiStrength algorithms. The hourly number of tweets for hour t is defined as $count_t = rawCount_t$. When it comes to the hourly Google Trends Interest for hour t, we define it as $interest_t = rawInterest_t$.

To check if the time series we are analyzing are stationary, we are performing the augmented Dickey–Fuller (ADF) test that checks if a unit root exists for the time series under test. Absence of unit root means that the time series is suitable for the prediction model without additional processing. The output of the ADF test that we run is p value. For all signals that we are analyzing, the exchange rate, Stanford CoreNLP sentiment, SentiStrength sentiment, tweet count, and Google Trends data, p values were below 0.05 which means that we can reject the hypothesis that there is a unit root. This implies that we can use all of input signals as-is. Obtained values for the time series data used in this model can be found in the table 4.2.

	ADF	$p - value$
Bitcoin price	−1.48732	0.015390
Stanford CoreNLP sentiment	−10.17964	0.006740
SentiStrenght sentiment	−4.96471	0.000260
Google Trends interest	−4.49897	0.000197
Number of tweets	−6.56461	0.000082

Table 4.2: Time series data values

One of the reasons why we are using differenced bitcoin price series is because using the raw data failed to pass the ADF test, and we could not discard the hypothesis that there is a unit root i.e. that the series is not stationary. By differencing the input series, ADF test was successful.

By using F-tests on lagged values of input signals, we show that lagged hourly tweet sentiment does improve our ability to predict the exchange rate. Using the same technique we run F-tests on lagged values of hourly tweet count and hourly interest in bitcoin. For these two input signals, the results we obtained show that both of these variables impact exchange rate in a statistically significant way. [Gre03] This means that we should consider them when building a model to predict the bitcoin price. More specifically when running Granger causality test to check if bitcoin price is caused by sentiment computed using Stanford CoreNLP showed that lags exist that yield p value below 0.05 which is statistically significant. The same process is repeated for the sentiment computed using SentiStrenght algorithm. When examining if tweet count is Granger-causing the price, we also examined all lags between 1 and 168, and found lags that produce p value below 0.05. The same test checked Google Trends interest data.

In order to predict bitcoin exchange rate which is defined as the cost of a single bitcoin in US dollars, we are building a vector auto-regression model. [Gre03] This model predicts the bitcoin exchange rate from tweet sentiment, number of tweets, and search interest in bitcoin, where all of these inputs are per-hour. Also, it captures the fact that all of these inputs interact with one another, while leveraging the past sentiment, tweet count and Google Trends interest data to make successful bitcoin price predictions. This means there are in total 1488 data points to analyze for each of these inputs, and for the value we are trying to predict. The model can be expressed as:

$$y_t = [price_t, sentiment_t, count_t, interest_t] \tag{4.9}$$

$$y_t = \sum_{i=1}^{maxLag} (k_i * y_{t-i}) + \epsilon \tag{4.10}$$

where t is a value between $maxLag$ and 1487 representing every hour between the 1st of July 2018 (inclusive) and the 1st of September 2018 (exclusive). Variable lag represents the number of hours input information lags behind the hourly price we are trying to predict. E.g. if lag has value 4, that means that inputs at 3PM, 4PM, 5PM and 6PM can be used to predict exchange rate for 7PM on the same day. By exploring the search space consisting of values for lags, we are able to find one that minimizes the root-mean-square error (rmse). In more details, we are finding lag that minimizes:

$$rmse = \sqrt{\sum (actual_i - prediction_i)^2} \tag{4.11}$$

where $actual_i$ is the actual price for hour i, and $prediction_i$ is predicted price for hour i.

Once we have identified the model parameters, we are using it to build the model using the training data set. The training data set is the first 1392 data points in our entire data set. With trained model, we are making predictions for the first hour outside the training data. This means that the model is going to forecast price for the 1393rd hour as the first step. After that, we add the actual price to the training data set, re-train the model, and repeat the process for the next hour. Using this process, we complete the prediction for the last 96 hours i.e. the 28th, 29th, 30th and 31st of August 2018.

4.4 Predicting bitcoin volume

In this thesis another time series that we are trying to predict is hourly bitcoin trading volume i.e. how many bitcoins are traded during a single hour. For this purpose we have developed a vector autoregressive model (VAR) using which we are forecasting the hourly bitcoin trading volume. [Gre03] This model is used because it correctly captures the fact that volume, hourly tweet count and Google Trends interest are all impacting each other. Using this model, we are able to leverage historical information as auto-regressor in order to predict future volume.

When looking vector autoregressive formula, in this case y_t represents trading volume of bitcoin, number of tweets and total interest on Google Trends at hour t that we are analyzing.

Time series that describes trading volume is described in details in the “Bitcoin price and volume data collection” section of this thesis. It can be described as $volume_0, ..., volume_{1487}$ i.e. it consists of 1488 data points, each representing the trading volume during that hour. In order to confirm that volume time series does not have a unit root, we ran augmented Dickey-Fuller test, which resulted in ADF Statistic = -5.22159 with p-value = 0.01794216. Value for p is significantly below 0.05, allowing us to reject the hypothesis that unit root exists. This means that the time series is not stationary. As already mentioned, tweet counts and interest information time series also do not have unit roots.

Before fully defining the model, let us examine the volume data a bit more. It is interesting to examine auto-correlation i.e. how much past values are impacting the future ones. In figure 4.3 we can see generated auto-correlation function (ACF) plot for volume. X-axis describes the lag i.e. how far in the past we are looking. In this case we are examining the impact of previous 32 hours on the current hour. From the graph, we can see that many lags are significant (they are outside of the shaded area). However, the partial auto-correlation graph shown on the figure 4.4 illustrates that lags 1, 2 and 5 are off significance, and that we should consider them when creating the model.

The signals used as input for this model are hourly number of tweets and hourly Google Search interest in term “bitcoin” on Google Search. We define number of tweets for hour t as $count_t$, where $0 <= t <= 1487$. The interest in term bitcoin is defined for the hour t as $interest_t$, with $0 <= t <= 1487$. In order to examine if these two time series are causing changes in bitcoin volume we run Granger causality test. This test will examine different lagged values of count and interest. [Gre03] By using F-test to check if count does cause volume, and if interest does cause volume, the results we get show that for different lags these two time series do improve our ability to predict volume in a statistically significant way. When examining if count causes volume, p value is below 0.05 for lags 1 to 10. When testing the interest time series, p is below 0.05 for lags 2

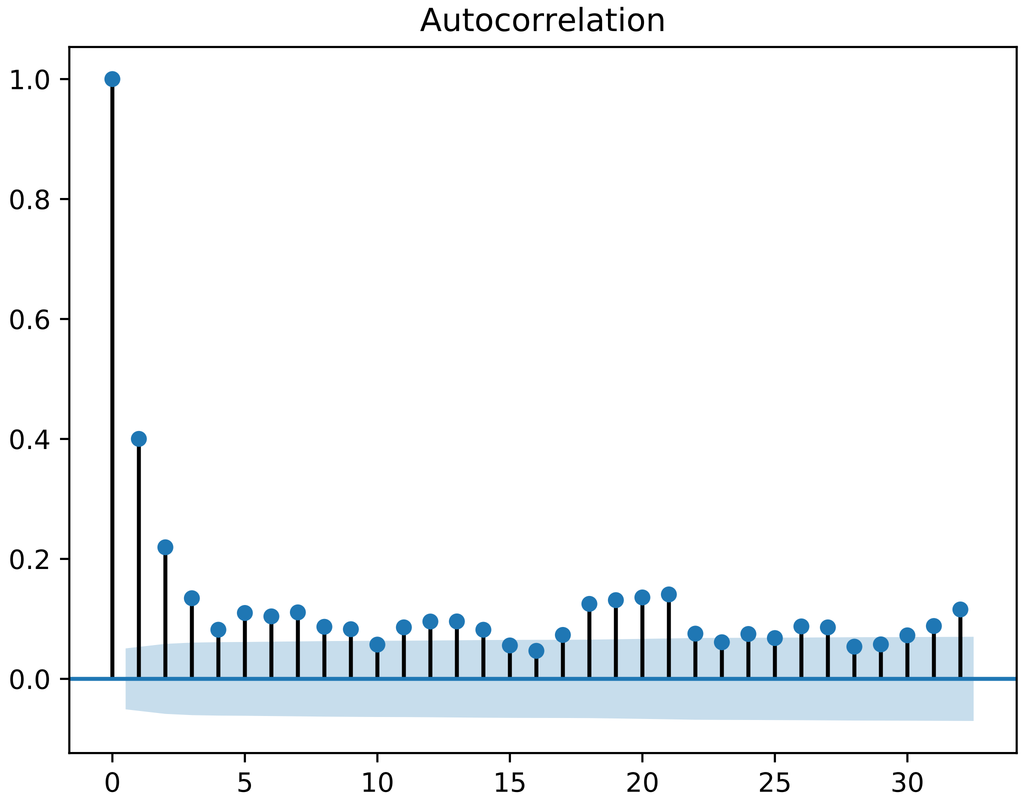


Figure 4.3: ACF Volume

to 10. This shows that we can reject the null hypothesis that count and interest do not cause volume, and that these two series can be used in the auto-regression model. Based on this information, and on ACF and PACF graphs, we choose lag 5 for the VAR model we are building.

The model we have developed is used to predict the last 96 hours of the observed period. Data is split into training and validation sets, and the first one is used to train the model. Because VAR model forecast does exhibit increased error as we are using it to predict further in the future, we are using a rolling window approach. This means that we first build the model on the training set, and that model is used to predict the value for the next hour. After this, we add one more data point to the model, and repeat the forecast for the next hour. In total, we build the model 96 times, and every time a single value is obtained. In this way, we obtain the most accurate information.

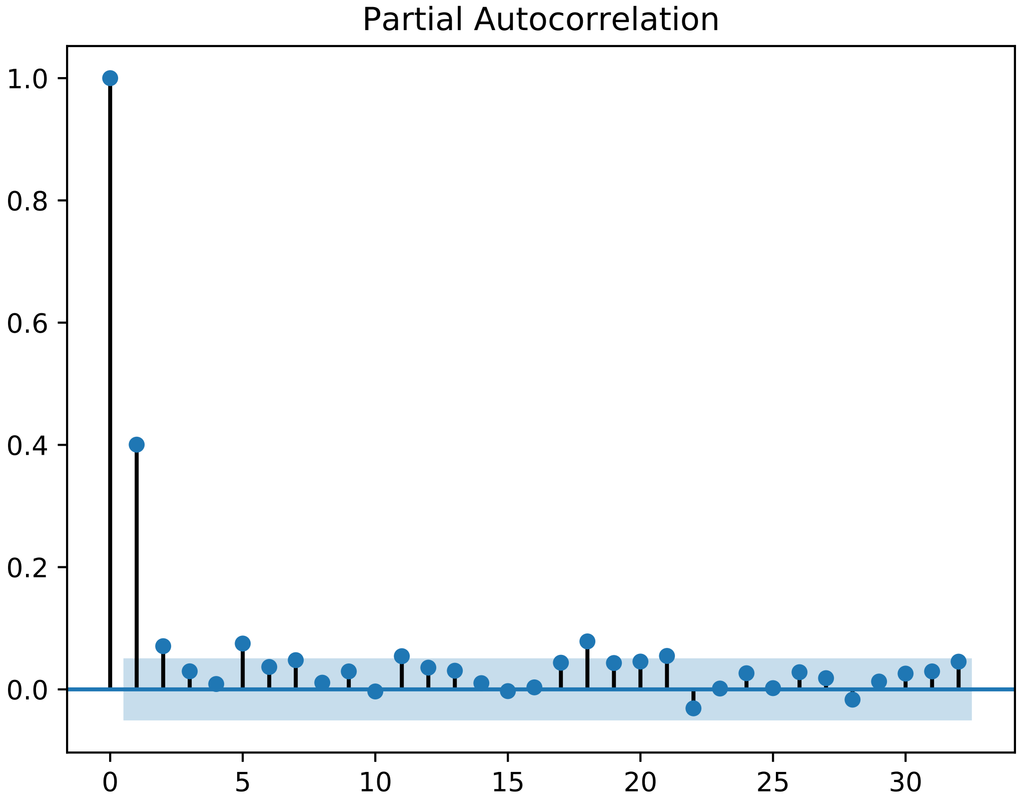


Figure 4.4: PACF Volume

CHAPTER 5

Implementation

This chapter focuses on describing implementation details of the system we have developed. It describes all components we have built in order to obtain the necessary data by creating connectors for different APIs we are using. Also, it describes how data cleanup was performed, both automated and manual one. Finally, it depicts how sentiment analysis algorithms were applied and how all of the input data was used to build the prediction model that allowed us to successfully forecast bitcoin exchange rate and trading volume.

5.1 System components

There are three main components of the system we have developed. These components are:

- Data retrieval components which further are split into:
  - TwitterAPI retrieval module
  - GoogleTrendsAPI retrieval module
  - Bitcoin price and volume retrieval module
- Raw data cleanup and processing component:
  - Twitter data processing
  - Sentiment analysis
  - GoogleTrends data processing
  - Bitcoin price and volume data processing
- Statistical model component:

- Final input signal processing
- Granger causality test
- Model building
- Model validation

There is a linear dependency between these three components. Data cleanup and processing depends on data retrieval component, and statistical model component depends on data that is output by the data cleanup and processing component.

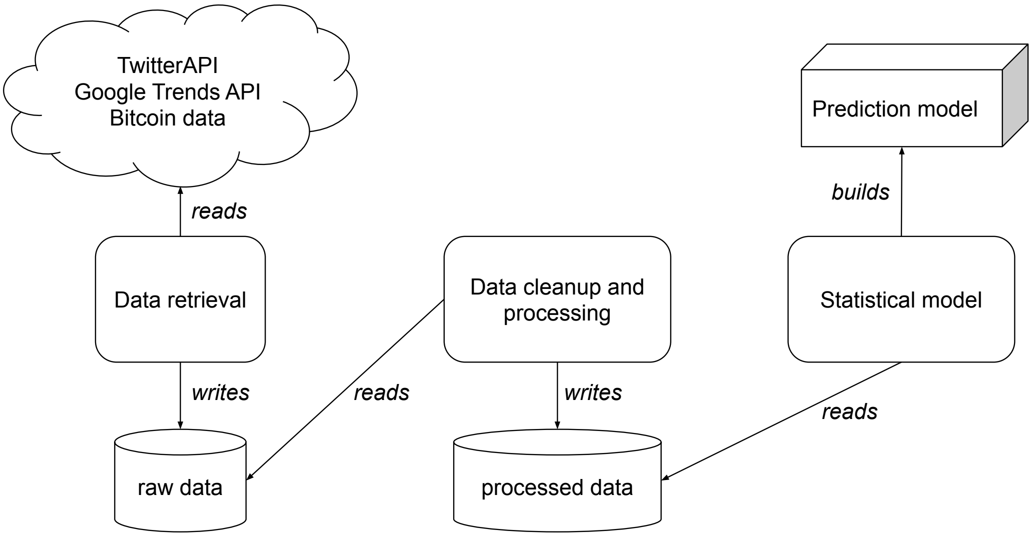


Figure 5.1: System Components

Once a component is done processing, it persists the processed information on disk. Data format used to store information in our system are CSV files. This data format is efficient for this purpose because it can easily store millions of data entries in a single file. Also, this system is not meant to be used for real time data processing and it is doing batch processing. However, adding a real time processing features is feasible, and it would only require setting up a stable API between components.

5.2 Data retrieval component

Part of the system that fetches the data is responsible for obtaining entire data used in this thesis. As already outlined in the "Data collection" section, there are three different data sources we had to use in order to download data: Twitter API, GoogleTrends API, and Bitcoin Charts website. In total, we have obtained around 2 GB of data, where most of it is bitcoin data, followed by tweets, and finally Google Trends data. Retrieval components were built in a way to handle rate-limiting APIs and to be able to download data partially. This means that if at some point access to the API was prevented (e.g. because limit was exceeded) we did not have to start from the beginning, but we could continue from the last downloaded timestamp. Because of the way we are processing the data and building the prediction model, we had to fetch entire data set before invoking the next component. While retrieval process was automated for a single source, it was not between data sources. This means that once e.g Twitter data was downloaded, we invoked sub-component that downloaded Google Trends data, followed by bitcoin data download.

Firstly, we downloaded tweets using the Twitter REST API. Through this API, we are able to specify search criteria and to fetch only tweets that we are interested in. Algorithm, looks like this:

Listing 5.1: Twitter data retrieval

```
latestDownloaded = loadLatestDownloadedTweet()
if (latestDownloaded > END) {
   // we are done with downloading
   return
}
client = TwitterClient(hashTag = 'bitcoin',
startDate = latestDownloaded, endDate = END)
while(client.hasMoreData()) {
  response = client.getResponse();
  if (response.status() != SUCCESS) {
    restartJobIn(hours = 2)
    return
  }
  nextRequestToken = response.nextPageToken()
  storeToDisk(response.getData())
  client.setNextPage(token = nextRequestToken)
}
```

With this algorithm implemented in Python, we started fetching the data from the 1st of July 2018, 00:00 UTC. If some data was previously persisted, we are processing it in order to find the tweet with the highest timestamp. This timestamp is used to specify the "since" search parameter to the REST API. In case this timestamp is already past the end of the time period we are analyzing, all tweets have been downloaded. It is important to specify that the API returns tweets in ascending chronological order. This means that if the latest retrieved tweet has timestamp t, all tweets with timestamp less than t have already been downloaded. Once the start date is known, we can fully specify our search criteria i.e the hashtags we are interested in and the end date. This API supports paging, which seems very suitable as there is almost 330 MB of data we have downloaded. Also, this means that Twitter servers do not have to do unnecessary work and send data in its entirety unless it is actually needed and used. The client that is created contains all search criteria, and it is able to provide us with information if there is more data to fetch. If all data is retrieved, we are done with data retrieval. However, if there are more tweets to download, we issue a request to the endpoint. As already mentioned this API has rate limiting, and during our work we have frequently encountered rate limiting by the Twitter API. For any response code other than success, we will stop processing, and we will restart the download job in two hours. This downtime proved to be enough in order to allow us to regain access to the API. Upon successful response, there are two important steps. The first one is fetching the token that is used to fetch the next page of results, where a single page has 20 tweets. This protects against malicious users as it forces them to wait for the request to be completely done, before issuing another one. Another important step is persisting tweets to disk, where they are permanently kept. In order to do that the API response is completely loaded in memory. Once loaded, this JSON data is parsed in an automated way, each tweet is pre-processed and data (username, text, mentiones, hashtags etc.) is extracted. Finally, this data is stored in a CSV file with all previously downloaded tweets.

In order to estimate global interest in bitcoin, data from Google Trends was used. This service provides an API that allow its users to fetch information about how interest in certain search terms during some period. Similarly to how we retrieved tweets, this algorithm was also written in Python. Retrieval mechanism can be described using the following pseudocode:

Listing 5.2: Google Trends data retrieval

```
currentDate='2018-07-01_00:00_UTC'
endDate='2018-09-01_00:00_UTC'
while (currentDate < endDate) {
  duration = min(7days, (endDate - currentDate))
  request = GoogleTrendsApi.request(start = currentDate,
                                    duration = duration,
                                    term = 'bitcoin')
  response = request.execute()
  if (response.statue() == LIMIT_EXCEDEED) {
    sleep(1hour)
  } else {
    storeToDisk(response.data())
    currentDate = currentDate + 7days
  }
}
```

GoogleTrends data is downloaded from the start of the period we are analyzing. Requests are limited to time period of seven days, if hourly information is requested (which was the case here). That is why we had to split data retrieval in multiple stages. In case the rate limit was exceeded, data fetching is suspended for 1 hour, until the ban is lifted. With every request, duration of seven days is specified, except for the last request which contains six days. Upon successful API response, data is stored to disk. Current date is increased by *duration* days, and it now reflects a new start date for which to retrieve the interest for. In order to obtain Google Trends data, we had to submit nine API requests, and data was stored in a CSV file with size 44 KB on disk.

The last data source used is Bitcoin Charts website, that provided use with raw information on Bitstamp exchange bitcoin to USD information. This website contains historical information about all transactions from bitcoin to and from USD that happened on Bitstamp, which we have additionally verified for validity. In order to download this information, we have used the Python request library. This library allows us to specify the URL where the file containing information is stored. Because it is quite large in size, in order not to exceed the machine physical memory limits, file is downloaded in pieces, where each piece is just a couple of MB (see listing 5.3).

Listing 5.3: Bitcoin data retrieval

```
req = request(url=url, chunk=8MB)
file = createForWriting()
while(req.hasMoreChunks()) {
    data = req.getNextChunk()
    file.writeToDisk(data)
}
file.finish()
```

File that was downloaded using the script above is compressed, and before we can access the transaction data, we have to uncompress it. In order to do that, we are using gzip Python library, that allows reading files in pieces. With this, we are able to read a couple of MBs from the compressed file, examine the transaction data, and decide if we want to keep it in the final output (see listing 5.4). Result of this phase is a single CSV file that has three columns. The first column is the transaction timestamp, the second one is the price of bitcoin in USD, and the third one is the number of bitcoins in the transaction. Only transactions with timestamp within the period we are analyzing are kept. This is to make the data loading faster, as the original data set has 1.5 GB of data and 34.675.070 transactions. Reduced data set, consisting only of transactions we are interested in has 1.538.551 transactions and its size is only 41 MB which makes it much easier to work with.

Listing 5.4: Uncompress data

```
fileReader = readGzFile()
output = createForWriting
while(fileReader.hasMoreChunks()) {
  data = fileReader.getNextChunk().decompress()
  for (line in data) {
    timestamp, price, volume = split(line)
    if (timestamp >= START_DATE && timestamp < END_DATE) {
      output.addEntry(timestamp, price, volume)
    }
  }
}
```

5.3 Data cleanup and processing component

Data cleanup and processing component is the next in our processing pipeline. Goal of this part of the system is to cleanup, validate and prepare data for the final phase, which is model building. In order to do that, it consumes raw data output by the data retrieval component. It also needs to wait for that component to finish its run, before it can start any work. This can be improved in the future versions of the system that can process data in a more streamed way, rather than having it batch processed. Input to this phase are three CSV files, one containing downloaded tweets, another one containing the Google Trends data, and the final one containing bitcoin transaction information.

Tweets were downloaded using the methods described in the previous chapter. In order to use this as input to our prediction model, we preprocess and validate the information. Considering there are almost 330 MB of tweets, pipeline for tweet processing was developed to handle them in a streamed way. This avoids loading all of them in memory, potentially straining the machine used to run this operation. The system consists of data loader, that is connected to the data source, a CSV file in our case. The data loader is able to load tweets one by one from the data source, and to pass it to a chain of processors. Every processor is able to transform and examine the tweet, storing any information it cares about. Also, because these processors are chained, if a tweet is not matching some criteria (e.g. it is a duplicate), processor can remove it from the pipeline, thus preventing downstream processors from being polluted by this tweet. This is shown in the figure 5.2.

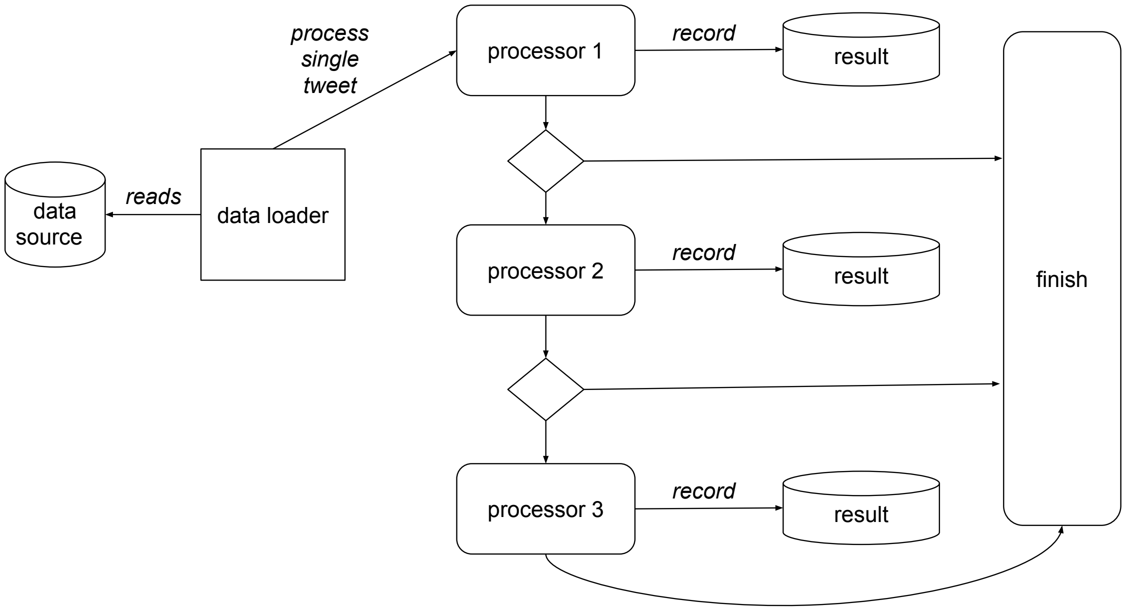


Figure 5.2: Data Processing

A single processor has the following structure, as depicted in the following Java class:

Listing 5.5: Structure of the processor

```java
public abstract class AbstractProcessor {
  public AbstractProcessor(AbstractProcessor delegate) { ... }
  public void doProcess(Tweet tweet) {
    boolean shouldPassToDelegate = processingImpl();
    if (shouldPassToDelegate) {
      delegate.doProcess(tweet);
    }
  }
  public abstract Result getProcessingResult();
  /* Does actual processing, and stores
  information in Result. */
  protected abstract boolean processingImpl(Tweet tweet);
}
```

Because every processor has a delegate i.e. a processor it delegates more processing to, we can chain as many processors as we like. For example, in our implementation we are adding transforming processors, validating processors, and filtering processors. Transforming processors are used to process the tweet text and to remove any URLs or hashtags from the text. This is needed in order to cleanup data before invoking Stanford CoreNLP sentiment analysis as it is not optimized for such input. There are multiple validating processors we have implemented. The first one is the one requiring tweet uniqueness and we accomplish that by keeping track of tweet ID that must be unique in our entire data set. This processor captures tweet ID whenever it processes a tweet, and it asserts that no two tweets have the same ID. Another validating processor is the one analyzing unique tweet texts. This processor examines tweet messages and it counts how many times some text appears in our data set. The last validating processor we have implemented is the one that randomly picks tweets, and for those tweets we are checking the URL information by downloading the tweets using a Java library and asserting that the text we have matches the one we have just downloaded. The last category of processors we have implemented is the filtering processor. This processor decides if tweet should be kept in the data set or discarded. Because the sentiment algorithms we are using support English, we had to make sure that the input data is also in English. For that purpose, we developed an algorithm that determines if tweet is in English or not. It relies on heuristic to make that decision. Once tweet is processed, and URLs and hashtags are removed, we count the number of non-space, non-digit, non-interpunction characters. If 70% of the remaining characters are letters in the English alphabet, we consider that message to be in English. This simple heuristic gives very good results, and from our initial data set we filtered out around 5% of the tweets. Algorithm below (listing 5.6) depicts the method that implements this logic. At the end of the processor chain, the last processor is used to store validated and filtered tweets to disk. They are stored in a format that can be used by the statistical model component.

Listing 5.6: Algorithm for detecting tweets in English

```
public boolean isInEnglish(String text) {
  int totalLength = text.length();
  int englishLetters = 0;
  int ignore = 0;
  for (int i = 0; i < text.length(); i++) {
    char letter = text.charAt(i);
    if (letter == '␣' || letter == '.' || letter == '#'
    || letter == '!' || letter == ',' || letter == '?'
    || letter == '\'') {
      ignore++;
    } else if (Character.isDigit(letter)) {
      ignore++;
    } else if (letter >= 'a' && letter <= 'z') {
      englishLetters++;
    }
  }
  totalLength -= ignore;
  return (englishLetters / (double) totalLength) > 0.7;
}
```

There is one processor that was not yet mentioned, and that is the one that runs Stanford CoreNLP sentiment analysis. Because there is an implementation of this library in Java, it was a natural fit to do sentiment analysis in this component written in Java programming language. Similarly to processors that were already mentioned, sentiment analysis processor analyzes tweets one by one. Each tweet is processed as an independent document, it is split into sentences, and each sentence is individually scored. The sentence sentiment is associated with the tweet ID, which is unique for the whole data set. Because the sentiment analysis is very CPU intensive, we are leveraging feature of Stanford CoreNLP sentiment analysis that allows specifying number of threads to use for processing. Because we are using a computer with 8 CPU cores, this has greatly decreased processing time. How we integrated this library in our project is shown in the listing 5.7.

Listing 5.7: Stanford CoreNLP data processing

```
@Override
protected boolean processImpl(Tweet tweet) {
   activeProcessing.increment();
   Annotation document=new Annotation(tweet.getProcessedText());
   pipeline.annotate(document, r -> {
      for (CoreMap sentence :
 document.get(CoreAnnotations.SentencesAnnotation.class)) {
         String sentiment =
 sentence.get(SentimentCoreAnnotations.SentimentClass.class);
         List<String> sentiments =
 result.idToSentiment.getOrDefault(tweet.getId(),
 new ArrayList<>());
         sentiments.add(sentiment);
         result.idToSentiment.put(tweet.getId(), sentiments);
      }
      activeProcessing.decrementAndGet();
   });
    return false;
}
```

Before starting the processing of the current tweet, we are increasing the number of actively processed tweets. This is important, because it allows us to find one when we are done processing. Tweet text with URLs and hashtags removed is used to build a new document, which is the basic unit of processing in the sentiment analysis algorithm. Next part of the method is leveraging the parallel processing. It triggers the document processing, but it does not wait for the document to be fully processed. Instead, it specifies a callback (passed as a second parameter), that will be invoked once the document is annotated and all metadata is added to it. Once done, the sentiment analysis will invoke the callback which examines the algorithm output. As previously explained, Stanford CoreNLP relies on annotations (e.g. SentencesAnnotation, SentimentCoreAnnotations) to store information it has computed. As a first step, we are fetching the annotation that contains information about sentences in our text. The list of sentences is processed one by one, and sentiment information about a sentence is obtained by querying the SentimentCoreAnnotations.SentimentClass. The result of that query is a sentiment category that can be “very negative”, “negative”, “neutral”, “positive”, “very positive”, and it is associated with the tweet ID. There is a separate part of the system that processes this data, and it persists a CSV file where each row starts with a tweet ID, followed by sentiment computed for each sentence in that tweet. It is important to notice that we are also decrementing the number of tweets that we are actively processing. This is to make sure that the part of the system that handles this CSV file generation is able to start its work once there are no more tweets being processed.

We have already mentioned that one of the outputs of tweet data cleanup and validation was a CSV containing all tweets in English. This file is created by identifying all tweets written in English, but when generating the file, original tweet text was written out. This means that URLs and hashtags are stored unchanged. One of the main features of SentiStrength algorithm is that it is optimized for short, tweet-like, texts and that is why we are not preprocessing tweets. The SentiStrength executable was invoked by specifying additional configuration parameters and the input file. Because each tweet is processed independently, we split the input data into multiple files. This allowed us to have multiple invocations of the sentiment analysis running in parallel. Each of these invocations produced an output file that contains tweet text and computed sentiment for individual tweets in the input files. Data in the output file is very granular and it allows for per-word sentiment information. However, because we are interested in the sentiment of entire tweet, we transform the output files further. Notably, we load each output file using Python pandas library, and we discard all information except for the recorded sentiment. In SentiStrength algorithm, this is represented as two numbers, the first one -5 to -1 for negative sentiment, and the second on 1 to 5 for positive sentiment. Additional challenge was associating this information to tweet IDs. We accomplished this by leveraging the fact that the order of extracted sentiment information in the file is the same as in the original input files (which contain tweet ID information). Using this information, as a final step in recording the SentiStrength sentiment, we generate a CSV file which rows contain tweet ID followed by two numbers that represent computed negative and positive sentiment, respectively. This process is shown in the listing 5.8:

Listing 5.8: SentiStrength data processing

```
chunks = 8
inputs = getTweetsInEnglish(chunks=8)
for (input in inputs {
  outputs[i++] = sentiStrength(input).launch()
}
for (output in outputs) {
  output.waitToFinish()
}
finalOutput = createForWriting()
for (i=0 to inputs.size) {
  tweetIdWithSentiment = merge(inputs[i], outputs[i])
  finalOutput.write(tweetIdWithSentiment)
}
finalOutput.finish()
```

Within this system component there is a module that processes Google Trends information. Input to this module are multiple CSV files that contain raw information retrieved from the Google Trends REST API. These files contain two columns, where the first one is the timestamp representing an hour of the time period that we are analyzing. The second column is an integer representing the level of interest in search term “bitcoin” during that hour. Because the API allows fetching hourly data for a maximum of seven days, this raw data contains continuous information only for that period. Across weeks, this raw data is not comparable, and we have to normalize it. In order to do that, data retrieved across two API responses is overlapping for one hour. This allows us to transform the interest rate from the second response to the same scale that is used in the first response. Python code for this transformation is shown in the listing 5.9:

Listing 5.9: Google Trends data processing

```
start_time = np.datetime64("2018-07-01␣00:00")
hour_time_period = np.timedelta64(1, 'h')
data["Hour"] = pd.to_datetime(data["Hour"])
data["Hour"] = data["Hour"].apply(lambda x:
int((x - start_time) / hour_time_period))
transitions_between_weeks =
data[data["Hour"].duplicated()].index.values

current_coefficient = 1
adjusted_interest = []
for i in range(len(data)):
  if i in transitions_between_weeks:
    current_coefficient =
data["Interest"][i - 1] / data["Interest"][i]
  else:
adjusted_interest.append(current_coefficient * data["Interest"]
[i])

interest_data =
        pd.DataFrame(adjusted_interest, columns=["Interest"])
interest_data.to_csv(generated_file, index=False)
return interest_data
```

To apply this transformation, we are using pandas Python library. First we load the CSV file, and we convert timestamps from the raw data to number of hours since the beginning of the period we are analyzing. Values for the column "hour" are between 0 and 1487. Variable current_coefficient is used to store the coefficient that should be used to multiply the raw data with. Because we are using data from the first seven days as-is, coefficient has value 1 in the beginning. Another variable, transitions_between_weeks, is storing information about indices of rows that are overlapping between two consecutive API requests. If we are processing a row that is not an overlapping one, we just multiply the current coefficient and the raw interest value. However, if we are processing an overlapping row, this interest should not be stored, as it was stored in the previous iteration of the loop. This row should only be used to compute the new coefficient that will be used to multiply the raw interest rates of the next seven days we are processing. Once we are done processing all downloaded information, we store transformed interest data to a CSV file. This file has two columns. Data in the file is sorted chronologically, and the first column is the number of hours since the 1st of July 2018, and the second column is adjusted level of interest in search term “bitcoin” for 2 months.

There is one more part of this subsystem to describe, and it is the one that deals with processing bitcoin exchange rate and volume information. The input to this module is a CSV file that contains bitcoin to USD transactions that were executed on Bitstamp exchange. A single row contains information when the transaction happened, what was the exchange rate, and how many bitcoins were traded in the transaction. Because there are many transactions during a single hour, we decided to pick the last transaction during an hour as the one defining the price for that hour. In order to do that, all transactions are sorted, they are grouped by hour they belong to (a number between 0 and 1487), and the last price in that hour is kept. Python code in listing 5.10 does this processing. Once this code finishes, there are 1488 data points, where data is sorted chronologically starting from the 1st hour. These values are used in the component that creates the prediction model. In order to validate the data, pandas library was used to find the lowest and the highest prices. Validation checks also picked a number of random data points for which we performed manual checks. This information matches the one available on the Bitstamp website, confirming the validity of the data.

Listing 5.10: Bitcoin exchange rate processing

```
raw_prices = pd.read_csv(path, header=None, names=["Timestamp",
"Price", "Volume"])

starting_hour = int(raw_prices["Timestamp"][0] / 3600)
raw_prices["Timestamp"] = raw_prices["Timestamp"].apply
(lambda x: int(x/3600) - starting_hour)
raw_prices.drop_duplicates(subset="Timestamp",
keep="last", inplace=True)

data = pd.DataFrame()
data["Price"] = raw_prices["Price"]
data.to_csv(generated_file, index=False)
```

Another important information to extract from the input records is bitcoin hourly trading volume. It represents the total number of bitcoins exchanged between start until the end of an hour. In order to obtain this data, we are leveraging pandas Python library (listing 5.11). After the raw data is loaded in memory, it is available in a table-like data structure. This allows us to access the entire column just by using the column name. In order to find volume for a single hour, we first subtract the number of hours of the first timestamp and divide the timestamp (which is in seconds) with 3600 (which is the number of seconds in a single hour). Outcome of that division is a number that we round down to the first integer. After this, “timestamp” column contains values between 0 and 1487, where multiple rows share the same timestamp value. It is exactly this property that we leverage when we invoke the method to group the rows by timestamp value, and to compute sum for all groups. This method produces 1488 records, where “volume” column contains hourly bitcoin trading volume. We store this information in a CSV file, and this is used as input for the prediction model. There are also validation steps that we perform for this data. We explore weekly and monthly trading volume and compare that with information available on the Bitstamp website. In addition to that, this module picked a number of random data points, and we have checked that against hourly trading volume recorded at the official source. There were no differences found, which confirmed the correctness of obtained data.

Listing 5.11: Bitcoin trading volume processing

```
raw_prices = pd.read_csv(path, header=None, names=
["Timestamp", "Price", "Volume"])
starting_hour = int(raw_prices["Timestamp"][0] / 3600)
raw_prices["Timestamp"] =
raw_prices["Timestamp"].apply(lambda x: int(x /
3600) - starting_hour)
raw_prices = raw_prices.groupby(["Timestamp"]).sum()

data = pd.DataFrame()
data["Volume"] = raw_prices["Volume"]
data.to_csv(generated_file, index=False)
```

5.4 Statistical model component

Component of the system that creates the prediction model is the last one to run. The reason for that is because it depends on processed data that is the output of “data cleanup and processing” component which is described previously. Inputs for this subsystem include sentiment from the Stanford CoreNLP algorithm, sentiment from the SentiStrength algorithm, tweets, bitcoin price and volume information. These time series data are further analyzed in order to learn about their properties and to discover which prediction model works the best.

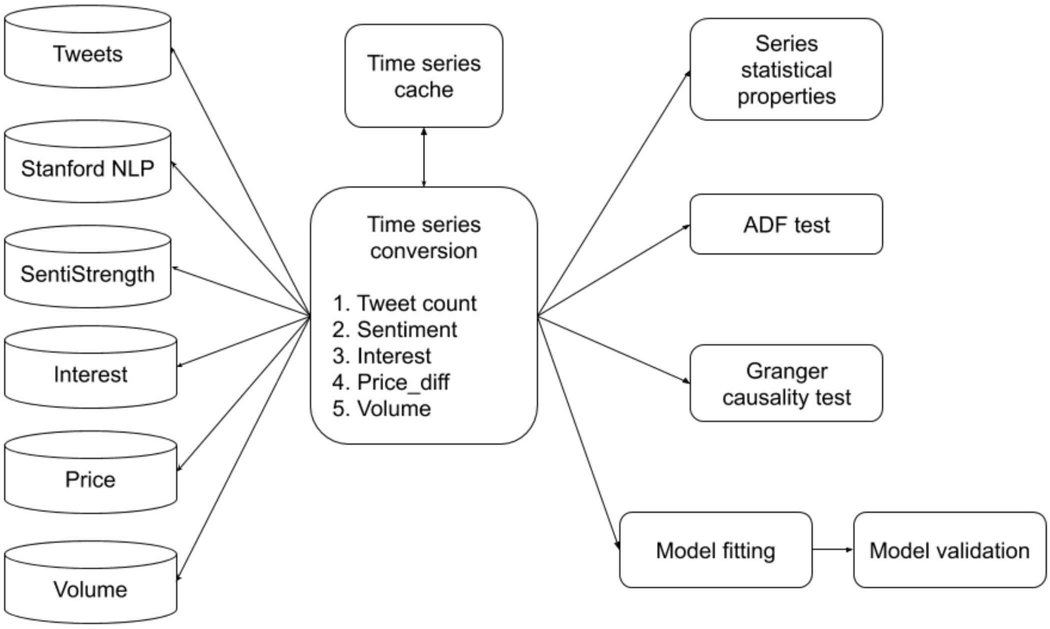


Figure 5.3: Statistical Model

This subsystem consists of multiple parts that form a processing pipeline. Initial data is first transformed to time series based data. Following that, a series of statistical tests are run in order to learn about the properties of every time series, and to visualize data we are processing. Additional tests are run to confirm that one time series does cause other, and that data is suitable for model building. Finally, we build multiple models, two for price prediction with Stanford CoreNLP data and SentiStrength data, and a third one that forecasts trading volume. The figure 5.3 depicts different parts of this system.

The first part of this component converts data from a CSV file to time series data. This makes it more suitable to work with, and we are able to leverage Python statistics library. Because we were running this component a lot, we optimized for fast startup time and time series conversion implemented a simple caching mechanism. It consisted of processing data once, producing time series of 1488 data points, and storing that information to disk. Next time a conversion for a data source runs, we are able to load a small CSV file and data is loaded in memory very fast. In our case, all data sets were loaded in memory in less than 500 milliseconds. This was especially helpful when processing tweets and sentiment information. Tweets are specified in a CSV file, where every row consists of tweet timestamp, tweet ID, and other metadata. However, in this thesis we are not using this information, and the only information we require is the tweet ID and the timestamp. This data is loaded in memory in a table-like structure. Then, we process it and keep tweet ID and timestamp columns. This information is used in two ways. Firstly, we round down the timestamp to the full hour, and we count the number of rows within the same hour. This represents hourly number of tweets time series, and we store that information on disk. Secondly, we are correlating tweet IDs from this data set with tweet IDs from sentiment data sets (this is shown in the listing 5.12). By generating a merged table that contains tweet ID, timestamp and sentiment, we are able to get sentiment information per-timestamp. Similarly to computing tweet count, we round down the timestamp to the full hour, and we compute the average sentiment during that hour. This process is repeated twice, once for Stanford CoreNLP results and once for the SentiStrength results. Two time series are stored to disk. In this way we generate six time series that we use in our statistical component.

Listing 5.12: Twitter time series data

```
tweets = get_tweets().get_tweets_per_hour()
dataFrame = []
with open(raw_file, 'r') as f:
  for line in f.readlines():
    tokens = line[:-1].split(",")
    tweetId = tokens[0]
    avgSentiment = sum(map(lambda text:
    _sentiment_values[text], tokens[1:])) / (len(tokens) - 1)
    dataFrame.append([tweetId, avgSentiment])

sentiment = pd.DataFrame(dataFrame, columns=["id", "sentiment"])
sentiment["id"] = sentiment["id"].apply(lambda x: x[1:-1])
joined = pd.merge(tweets, sentiment, on="id")
[["id", "sentiment", "date"]]
hourly_sentiment = joined.groupby("date").mean()["sentiment"]
```

Once data is available in time series format we use pandas Python library. This library is very helpful when it comes to generating graphs and examining statistical properties for a particular time series. More specifically, one is able to find minimum, maximum, average value, or to plot the autocorrelation function of a time series. All of these features provided us with invaluable insights which allowed us to make a better prediction model. When predicting bitcoin price by using the Stanford CoreNLP algorithm, tweet count and level of interest, our initial assumption was that sentiment, tweet count and level of interest are indeed driving the price. In addition to that, we assume that there is also a dependency between these time series, where the past is impacting the future data points.

This encouraged us to pursue vector autoregressive prediction model (VAR), but before doing that, we ran multiple tests. The first test we run is augmented Dickey-Fuller test, which analyzes a single time series, and checks if there is a unit root to check if series is stationary or not. The first test produces p value that is above 0.05, and that is why we decided to use hourly price difference as input to our model (denoted as price_diff series). One of the parameters used for the ADF test is maximum lag. We intentionally choose a very high value to make sure data is not stationary. In this work, we check all six time series that we are using. The code in the listing 5.13 shows this check for the price time series.

Listing 5.13: Price time series data

```
price = statsmodels.tsa.stattools.adfuller(price, maxlag=168)
print(f'ADF␣Statistic␣=␣{price[0]}')
print(f'p-value␣=␣{price[1]}')
price_diff = adfuller(price_diff, maxlag=168)
print(f'ADF␣Statistic␣=␣{price_diff[0]}')
print(f'p-value␣=␣{price_diff[1]}')
```

Having confirmed that data is not stationary, we proceed to examine time series autocorrelation. This is available as a method in the statsmodels library. In order to check causality between time series, we are running Granger causality test, which is available in the statsmodel Python library that we are using as statsmodels.tsa.stattools.grangercausalitytests (see the listing 5.14). This method accepts two time series and it checks if the second time series is impacting the first one. It will examine if historical values of one series are improving our ability to predict the second time series. In our work we check lag up to 168 data points (7 days lag). Four different statistical tests run in order to determine if there is causality, and they run for every lag between 1 and 168. In case one of the tests has p value that is statistically significant, we record that lag. All lags with p values below 0.05 are reported, and that confirms that having predictor series does improve our ability to predict to_predict time series. This test reported multiple lags that tweet count and level of interest cause volume, and it also confirmed that both sentiment time series, tweet count and level of interest do cause bitcoin price.

Listing 5.14: Granger causality test

```
joined = [[to_predict[i], predictor[i]] for i
in range(0, len(to_predict))]
result = grangercausalitytests(joined, _MAX_LAG, verbose=False)

res = {}
for lag in result.keys():
  stats, _ = result[lag]

  for test_name in ['ssr_ftest', 'params_ftest',
  'ssr_chi2test', 'lrtest']:
    if test_name == 'ssr_ftest' or test_name == 'params_ftest':
      _, p_value, _, _ = stats[test_name]
    else:
      _, p_value, _ = stats[test_name]
  if p_value < 0.05:
    if lag not in res or res[lag] > p_value:
      res[lag] = p_value
return res
```

The final part of this component is used to build the prediction model and to validate the prediction against the actual data. The Python statistics library provides implementation of VAR model that we are using in our project. Because we confirmed that the time series we are using are suitable for that model, we proceed to build it. In the listing 5.15 it is shown how we build the model and obtain predictions when using Stanford CoreNLP sentiment data. Firstly, we obtain the time series data, and create input matrix that has hourly price difference, sentiment, interest, and count as inputs. In order to evaluate the model, we split data set into training data set and validation data set. The former is used to fit the model, and the latter is used to evaluate the model correctness. Some of the parameters that we are able to configure are number of predictions we make before we fit the model again and maximum number of lags. The first parameter shows how far in the future we can make predictions, and the second one specifies the maximum of historical data points model can use in order to predict the next data point.

Listing 5.15: Prediction model

```
data = pd.DataFrame()
_, price_diff, _, stanford_sentiment, senti_strength,
interest, tweet_count = get_raw_inputs()

data['Price'] = price_diff
data['Sentiment'] = stanford_sentiment
data['Interest'] = interest
data['Count'] = tweet_count

TO_PREDICT = 96
data_train = data[:-TO_PREDICT]
data_valid = data[-TO_PREDICT:]
data_valid.reset_index(inplace=True, drop=True)

from statsmodels.tsa.vector_ar.var_model import VAR

prediction = list()
for i in range(0, TO_PREDICT, STEP):
  model = VAR(data_train)
  model_fit = model.fit(maxlags=MAX_LAGS, trend='nc')
  actual_lag = model_fit.k_ar

  forecast = model_fit.forecast(model_fit.y[-actual_lag:],
  steps=STEP)
  for j in range(0, STEP):
    prediction.append(forecast[j][0])
    data_train = data_train.append(data_valid[i:(i +
    STEP)]).reset_index(drop=True)

rmse = sm.tools.eval_measures.rmse(data_valid["Price"],
prediction)
```

This model is using a rolling window in order to build the model. After model is fitted, and predictions are obtained, we expand the training data to include the next STEP data points. In this way we retrain the model with fresh data and we are able to make more precise future decisions. Forecast data is added to prediction time series until we compute all TO_PREDICT data points. In this way, we will retrain model TO_PREDICT/STEP number of times. Because STEP and MAX_LAGS are configurable parameters, we explore the search space in order to find the one that minimizes the root mean square error between actual data and prediction data. Values that we examine for STEP are 1, 2, 3, 4, 6, 8, 16, 24, while for MAX_LAGS we are examining all values between 5 and 168.

Because there are 1312 combinations to examine we are building individual models in parallel. In this way we are able to speed up the process substantially. Each built model reports back the computed root mean square error, together with configuration it used, and in that way we are able to find the most optimal model. This process is repeated three times: when predicting price using Stanford CoreNLP sentiment, when predicting price using SentiStrength sentiment, and when predicting trading volume. Once the best model is found, we are using pandas library to plot the prediction data and to examine forecast data further.

CHAPTER 6

Results

In the previous chapters we have described data, statistical methods, and implementation of our prediction system. In this chapter we will examine results we obtained, provide explanations for those results, and also look deeper into properties of data we have used as inputs. We will examine how bitcoin price predictions change depending on the sentiment analysis algorithm used, and how we predict hourly trading volume.

6.1 Bitcoin exchange rate prediction

In the figure 6.1 we can see the results obtained from running our prediction model that relies on the SentiStrength algorithm for sentiment analysis. Forecast runs for the last 96 hours of the analyzed period by using the VAR prediction model with rolling window. Prediction results show root mean square error from the actual data of 36.77. We are able to successfully predict price movements, and this is visible from the graph.

The comparison of the prediction and the actual data can be seen in the table 6.1.

Hour	Predicted value	Actual value
2nd	−26.21	−28.23
19th	28.02	32.46
52nd	−34.27	−42.26
82nd	−8.56	−0.79

Table 6.1: Comparison of actual and predicted values

To get even better idea about the prediction ability of this model, we have measured if it is able to predict price movement correctly. Because we are forecasting hourly price difference, we are measuring how many times our model successfully predicted rise or fall

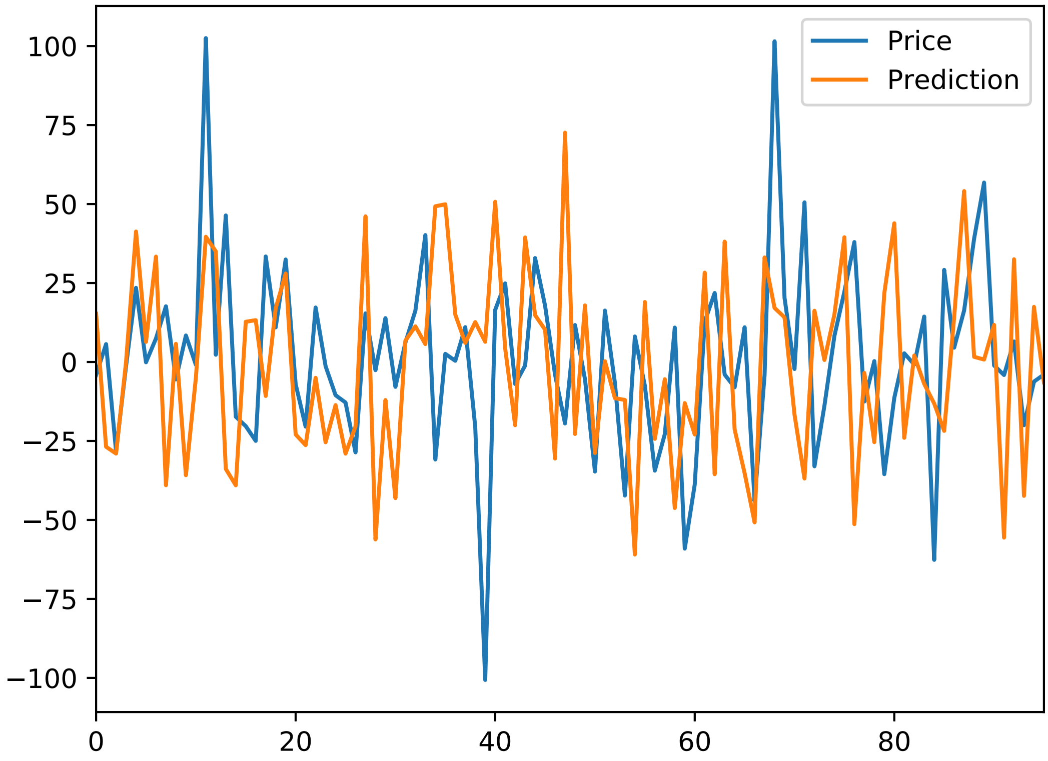


Figure 6.1: Bitcoin price prediction using the SentiStrength algorithm

of the price. This means that both actual price difference and prediction are negative or both are positive. In 63% of cases we are able to successfully predict whether the price will be higher or lower from the current one. The VAR model we have built is using maximum lag 168, and for the first five lags, the coefficients are described in the table 6.2. One may notice that the coefficients for the price and sentiment vector components are the highest for lag 5. When analyzing the impact of level of interest on price, lag 2 has the highest positive correlation, but lag 3 has negative correlation which has the highest intensity. Observing coefficients for count, lag 3 and 4 are the most prominent, and they are positively and negatively correlated with price, respectively. In total, our model has analyzed coefficients for all four components across 168 historical values.

When determining the number of data points to predict between two model fittings, multiple values were examined. Figure 6.2 shows how our optimizing function changes as we explore different values for step. It shows that as we increase the number of

Lag and Input	Coefficient	Std. error
L1.Price	-0.005383	0.044147
L1.Sentiment	9.634089	41.864154
L1.Interest	-0.047750	0.507066
L1.Count	-0.008614	0.039420
L2.Price	-0.000543	0.044236
L2.Sentiment	-3.007671	42.819884
L2.Interest	0.999368	0.678856
L2.Count	-0.004521	0.041695
L3.Price	0.012625	0.044178
L3.Sentiment	7.262058	42.793783
L3.Interest	-1.399117	0.689621
L3.Count	0.054106	0.042698
L4.Price	0.044963	0.044031
L4.Sentiment	-16.581462	42.889945
L4.Interest	0.017984	0.693527
L4.Count	-0.051997	0.042811
L5.Price	0.074992	0.044040
L5.Sentiment	25.780452	43.108836
L5.Interest	0.691057	0.691881
L5.Count	-0.009263	0.042893

Table 6.2: SentiStrength Model Coefficients

forecast steps, the error increases. This is in-line with VAR model properties, which may demonstrate increased error as we are trying to predict more data points in the future.

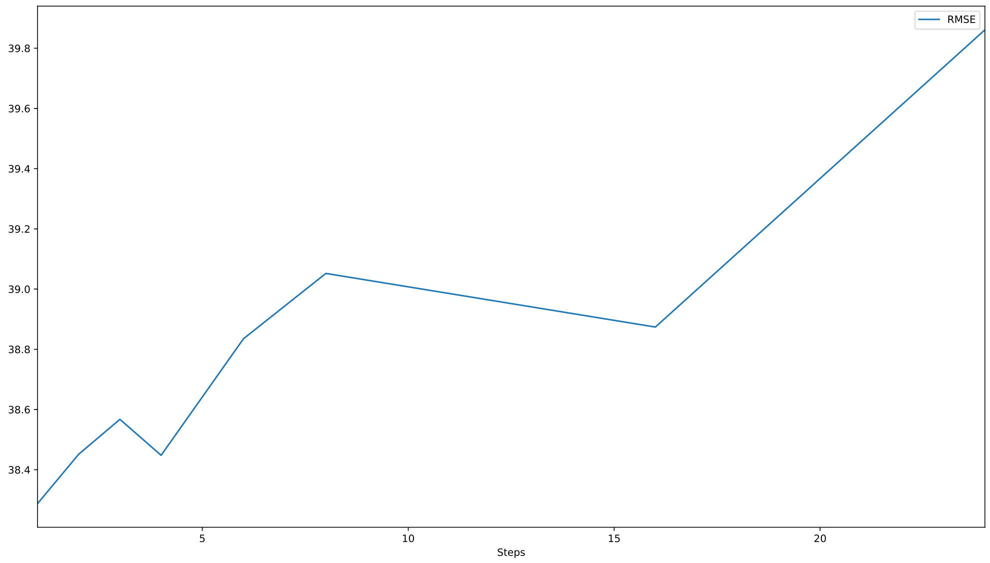
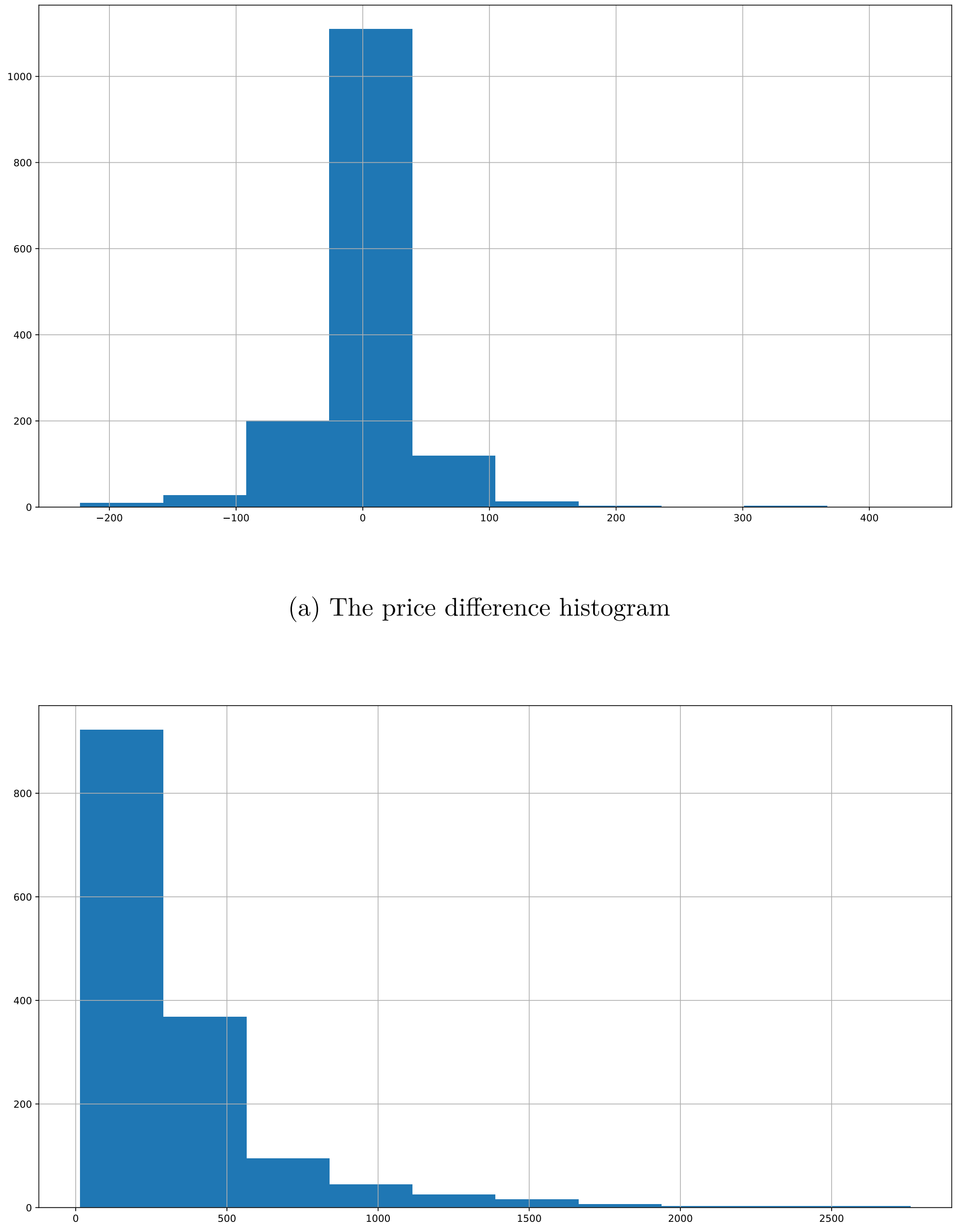


Figure 6.2: SentiStrength step Root Mean Square Error

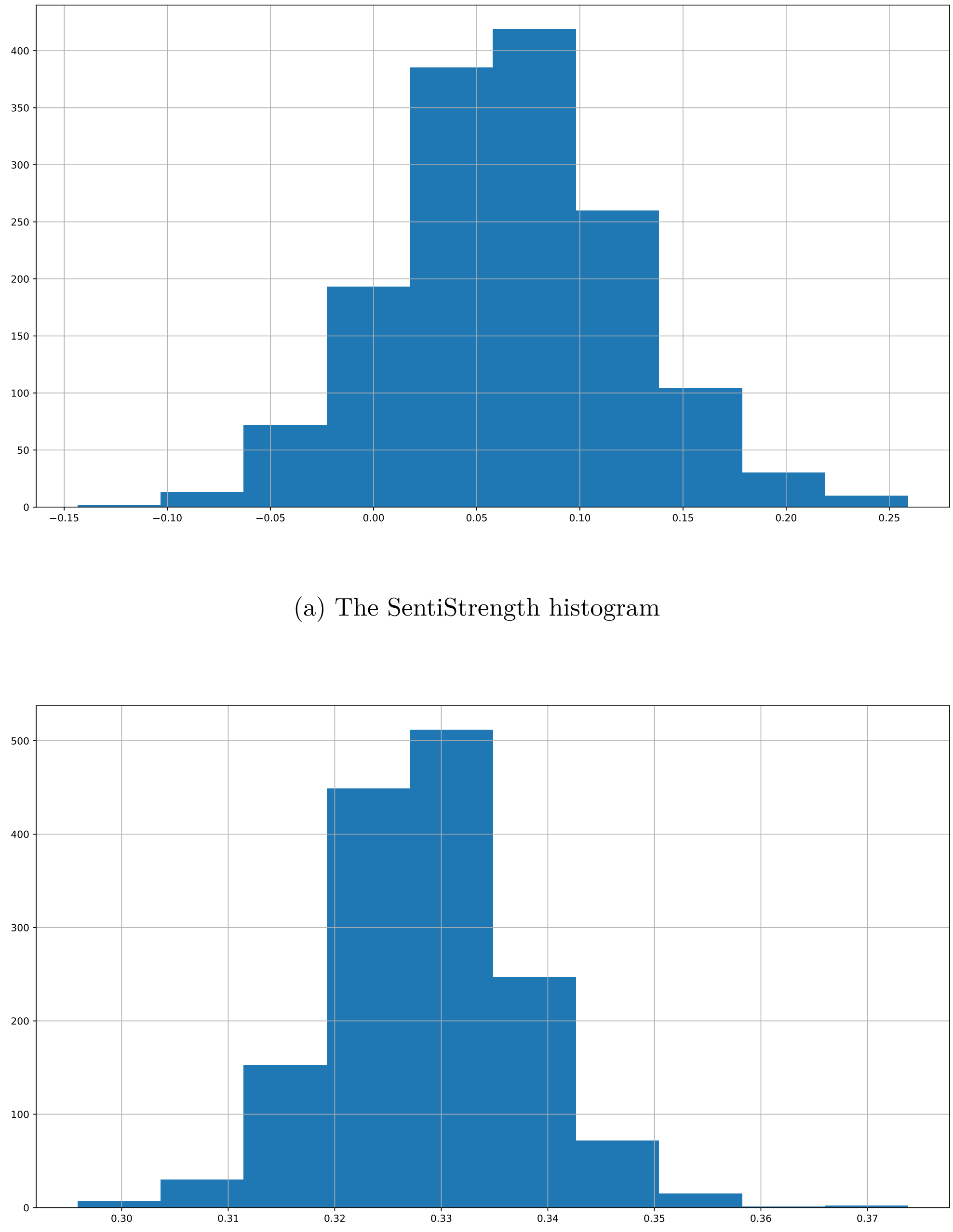
Examining the time series reveals interesting properties. The price difference histogram graph (Figure 6.3 (a)) shows that the majority of hourly differences are between -20 and 30 USD. Going further, analyzing volume histogram (Figure 6.3 (b)) shows a steep decline as the volume increases. This is in line with expectations, as there are a few events that cause large bitcoin trading volumes. Sentiment that is computed with SentiStrenght algorithm (Figure 6.4 (a)) does exhibit lean to positive values, confirming the fact that this algorithm evaluates tweets to have slightly positive sentiment on average. Looking into value distribution of hourly tweet count (Figure 6.5 (a)) shows majority of values between 400 and 600, and the level of interest (Figure 6.5 (b)) looks very similar to it, with most of the values between 50 and 80.



(a) The price difference histogram

(b) The volume histogram

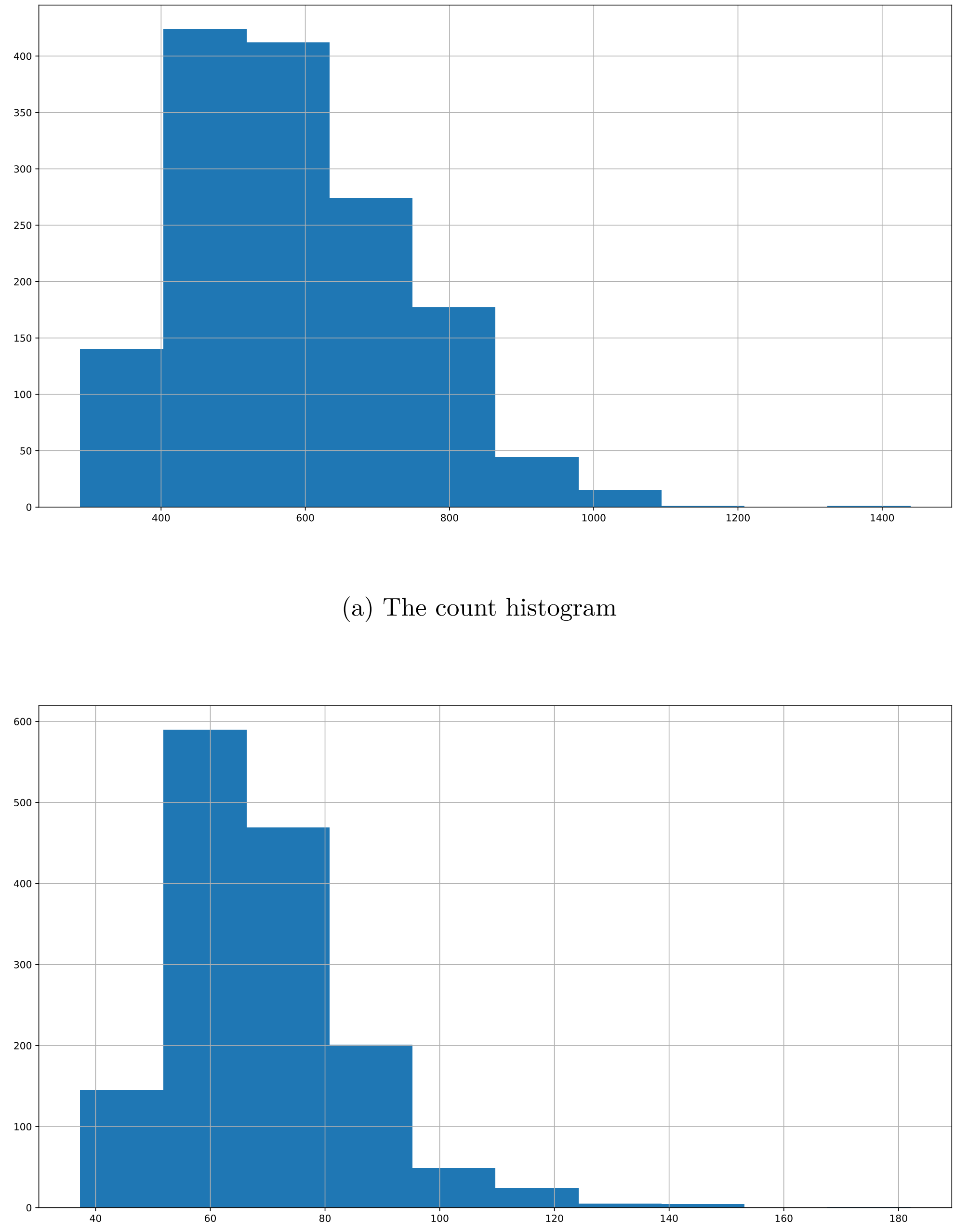
Figure 6.3: The price difference histogram and volume histogram



(a) The SentiStrength histogram

(b) The Stanford CoreNLP histogram

Figure 6.4: The SentiStrength and Stanford CoreNLP histograms



(a) The count histogram

(b) The interest histogram

Figure 6.5: The count and interest histograms

When building the prediction model that is using Stanford CoreNLP library, forecast results are depicted in the figure 6.6.

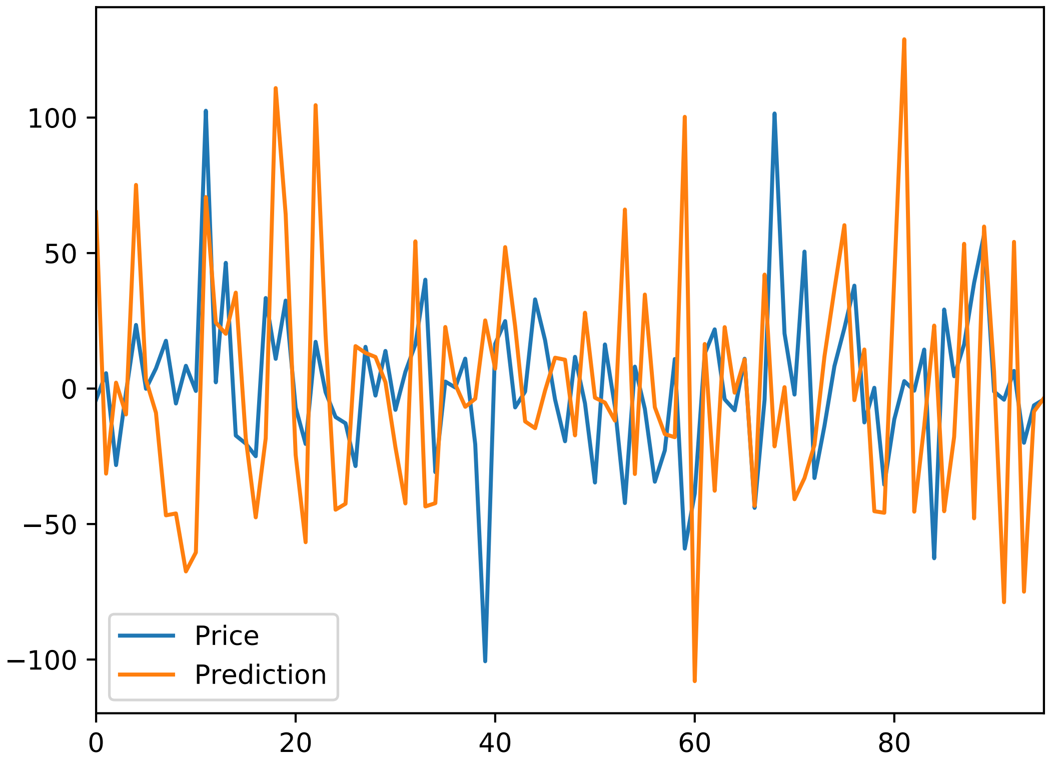


Figure 6.6: Bitcoin price prediction using the Stanford CoreNLP algorithm

Visually comparing the actual and predicted time series, we can notice that general trends do match, but that we are unable to fully match some of the outlier values like big decrease in price just before the 40th hour, or to fully match the intensity of increase around the 70th hour. The same process as for the SentiStrength algorithm was applied, and we obtain predictions for the last 96 hours. Comparing the prediction results with the actual data shows root mean square error of 49.49, which is higher than the value we got by using SentiStrength. This is in line with our expectations, as we anticipated SentiStrength to outperform Stanford CoreNLP algorithm when providing sentiment analysis information for prediction of the bitcoin price. The former algorithm is optimized for short text and in particular, it is optimized for tweets which contain abbreviations and slang. Using the same evaluation technique as for the SentiStrength algorithm model, we are evaluating how many times we successfully predict increase or decrease in the bitcoin price. This model is able to successfully predict price movements in 59% of the cases, which demonstrates that it is valuable when trying to forecast bitcoin price movement. For this model, the sentiment analysis yields values that are grouped around 0.33 (figure 6.4 (b)) which indicates overall negative sentiment value for the data we analyze. Maximum lag of 168 is used for this mode, and the table 6.3 shows coefficients for the first 5 lags.

Lag and Input	Coefficient	Std. error
L1.Price	-0.000831	0.043546
L1.Sentiment	-108.407258	262.421305
L1.Interest	-0.445105	0.517313
L1.Count	0.017434	0.037422
L2.Price	0.022302	0.043373
L2.Sentiment	205.190987	269.682228
L2.Interest	0.833543	0.688649
L2.Count	-0.015741	0.039426
L3.Price	0.011825	0.043378
L3.Sentiment	261.260262	271.933925
L3.Interest	-0.750986	0.691720
L3.Count	0.023555	0.040033
L4.Price	0.042282	0.043423
L4.Sentiment	188.675419	269.413742
L4.Interest	-0.440258	0.691504
L4.Count	-0.037775	0.040114
L5.Price	0.064005	0.043289
L5.Sentiment	348.764103	269.623575
L5.Interest	0.849277	0.691251
L5.Count	-0.010657	0.040128

Table 6.3: Stanford CoreNLP Model Coefficients

As we have already seen in the VAR model that is using SentiStrength, the highest coefficient for the price and sentiment vector components are for lag 5. However, this VAR model found level of interest at lag 5 to have the highest positive correlation, and count has the highest negative correlation at lag 4.

Two prediction models that we developed in order to predict the bitcoin price differ only in the algorithm that is used to compute the sentiment. One is using a general purpose natural language processing algorithm, while the other one is specialized for short texts such as tweets and has a proven track record in processing those. The first difference in data obtained from these two algorithms is that the SentiStrength sentiment is slightly positive on average, while the Stanford CoreNLP one is negative on average. In order to understand better why is that we analyze a few sample tweets in the table 6.4. This table shows three randomly chosen tweets, and how they were evaluated by these two algorithms. Scale is from 0 to 1, where 0.5 is neutral, and 0 to 0.5 is negative sentiment (0 being the most negative), and 0.5 to 1 is positive sentiment (1 being the most positive). In all three cases, Stanford CoreNLP rates tweets as having lower sentiment value. The first example "To help accomplish a smooth and successful ICO. . . " is clearly slightly positive as it encourages bitcoin usage. SentiStrength correctly assigns value that is a bit positive, but the other algorithm classifies this as having negative sentiment. The third tweet in the table, "... The ultimate Financial reward will be amazing. . . " is very positive and SentiStrength classifies it as very positive, but Stanford CoreNLP assigns it a slightly positive sentiment. These slight differences illustrate that SentiStrength is more suitable for obtaining sentiment information about tweets, and it explains why we achieve better accuracy when that algorithm is used.

Text of a tweet	Stanford sentiment	SentiStrength sentiment
To help accomplish a smooth and successful ICO, Spuul has engaged Deloitte & Touche LLP for Risk and Tax Advisory Services and Pinsent Masons MPillay LLP as its Legal Advisors. #Blockchain #Crypto #Bitcoin #BTC #ETH #Ethereum $ BTC #ETH #Cryptocurrency pic.twitter.comqyW9NGVqBe	0.25	0.625
Yale Researchers Develop System for Predicting Bitcoin Price Trends - UNHASHED http:/bit.ly2npuGNa #yale #bitcoin #bitcoinprice #crypto #cryptocurrency #cryptocurrencies #cryptonews #cryptocurrencynews	0.625	0.75
#Bitcoin will not jump out and tell you why you need to own it. You have to do your own work. Listen to these. Your education will be Free. The ultimate Financial reward will be amazing. Being part of what Changes the world in a HUGELY positive way, PRICELESS!!	0.535	1

Table 6.4: Stanford CoreNLP Model Coefficients

6.2 Bitcoin volume prediction

The final time series that we try to forecast is bitcoin hourly trading volume. The period that we are running the predictions for are the last 96 hours of the two months period that we are analyzing. Figure 6.7 shows comparison between the actual and predicted data.

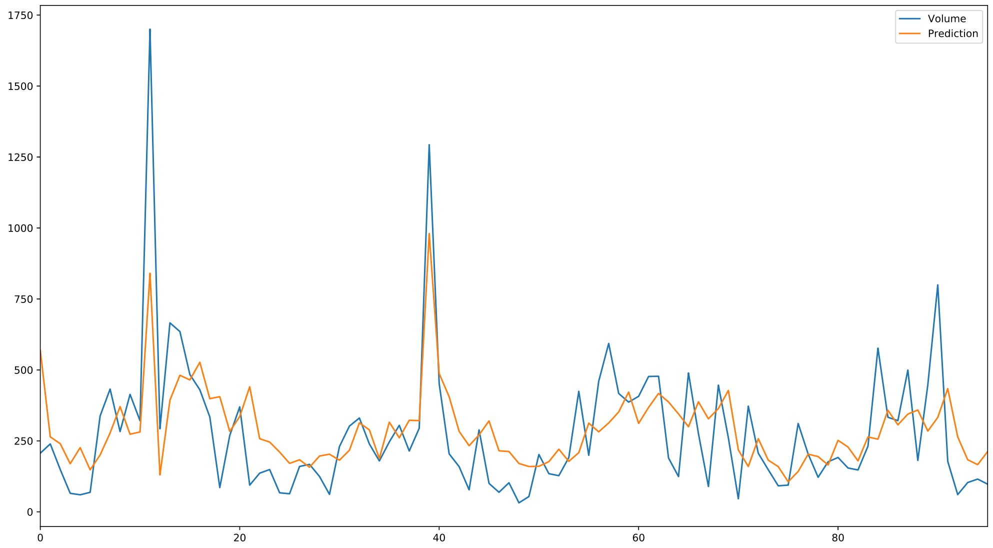


Figure 6.7: Bitcoin volume prediction

As in the two previous models, vector autoregression (VAR) is used to build the prediction model. Optimizing function used is root mean square error, and it is 172.76 when comparing actual and predicted data. Visually examining the graph shows that we are able to predict the hourly volume movement. Also, we are able to predict two large increase in hourly volume around the 10th hour and a bit before the 40th hour. Figure 6.8 provides more insights how volume, count and interest time series interact with one another.

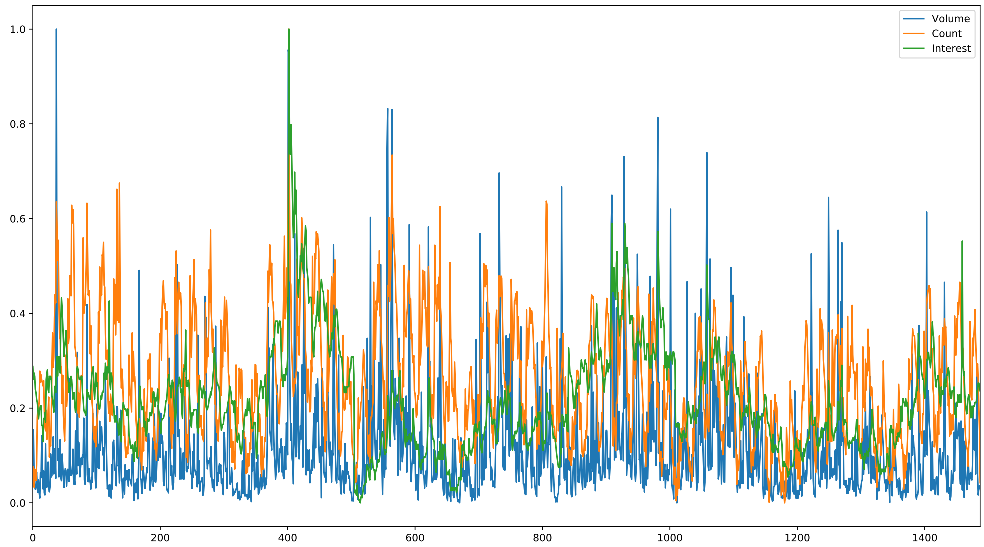


Figure 6.8: Volume, Count and Interest interaction

This graph shows normalized values for these three time series (values are from 0 to 1), and visually we can confirm that there is a correlation between them. Also, this graph shows how all of them surge around the 400th hour. The VAR model we have built in order to predict volume information is using max lag 5. This means that we can predict hourly volume based on 5 historical data points. Table 6.5 shows how we compute current value based on the historical ones. Volume in hour t is highest correlated with volume in the previous hour, with coefficient 0.3. However Google Trends interest has highest positive correlation at lag 2. It is interesting to notice that hourly count of tweets has the highest positive correlation at lag 1 and lag 5. This means that increase in tweet count will result in an increase in trading volume, but it also means that tweet counts five hours ago does impact volume as well. These coefficients allow us observe how these time series are impacting each other and how far in the future their impact lasts. From this data, and all statistical tests we run, we conclude that hourly tweet count and Google Trends data do allow us to make successful predictions about hourly bitcoin trading volume.

Lag and Input	Coefficient	Std. error
L1.Volume	0.305965	0.030002
L1.Interest	1.308062	1.991426
L1.Count	0.351861	0.108698
L2.Volume	-0.002491	0.031093
L2.Interest	2.813886	2.759127
L2.Count	-0.019148	0.130930
L3.Volume	-0.018242	0.031203
L3.Interest	-1.312551	2.741456
L3.Count	-0.056313	0.133417
L4.Volume	-0.054292	0.031079
L4.Interest	-1.470064	2.706341
L4.Count	0.265324	0.132240
L5.Volume	0.044568	0.029121
L5.Interest	0.332921	1.902425
L5.Count	-0.142892	0.109389

Table 6.5: Volume Model Coefficients

CHAPTER 7

Critical Reflection

In this chapter we will show what are the limitations of our system, how system components in charge of data collection and processing can be enhanced, and what are the opportunities for improvements and extensions of our prediction model in the future.

7.1 Open issues and limitations

When it comes to limitations, Twitter API rate limitations were one of the major challenges we had to overcome. This caused us to make a decision to collect data for a period of two months. Not having this constraint would enable us to gather data over a longer period of time which will potentially result in more accurate results. In case too many tweets were fetched, Twitter would blacklist our IP address, and we would have to stop data fetching for a couple of hours. Additional limitation of this API is the limit of number of tweets that can be downloaded in a single API request i.e. the API supports paging and it expects clients to handle this behavior. To the best of our knowledge, we have implemented the API integration correctly. Nonetheless, there is no way for us to verify that 918.354 tweets that we have downloaded are indeed all available tweets with the hashtag bitcoin.

In regards to language of tweets, our research focused only on English tweets. Even though English is the most used language on Twitter [1], it still excludes tweets about the topic in other languages which are as important as ones in English. Especially because we do not know their exact number. Other language related constraint was a heuristic that was developed in order to determine if tweets are in English. Our model determines whether text is written in Englsih based solely on if more than 70% of letters belong to the Englsih alphabet. However, in case some other languages use similar alphabet, we

[1] https://www.statista.com/statistics/267129/most-used-languages-on-twitter/ (Accessed November, 2019)

would not be able to detect them. Consequently sentiment algorithms used in this thesis would not have been able to detect sentiment from languages other than English.

In terms of other Twitter related constraints, it is worth to mention that we were not able to process tweets containing only an image or only a link which could provide valuable information. Our data cleanup and processing system component also detected many occurences of the same text in the tweets, which could imply that robot account were used for creating them. Being able to separate these accounts from actual users could be an interesting way to improve the accuracy of our prediction model.

When it comes to Google Trends, search frequency is expressed as integers between 0 and 100. This means that a minimum change that this system can detect is value of 1. Having access to actual number of search queries about bitcoin would provide us with a more precise data that we can use in our prediction model. For example, websites like Wikipedia, provide the actual number of visits for a given day.

7.2 Future work

The model developed in this thesis can be improved and extended in multiple ways. That is especially true for the rich data available from Twitter that can be processed and analyzed further.

As mentioned earlier, because of the API restrictions, current data set was limited to only two months. In the future, this research can be expanded to include larger data set which could provide better quality of scientific results. In order to do that, collaboration with Twitter can be accomplished.

When it comes to better accuracy, as bitcoin trading is happening globally, tweets from other languages should be included as well. This implies that method for language detection should be enhanced, and the sentiment algorithms should support different languages. Also, other sentiment analysis methods should be considered, in case they are able to provide better results.

There are reasons to believe that this research can be applied to other cryptocurrencies, and conducting the work to examine that hypothesis would be valuable. Examining five the most widely used cryptocurrencies can be a way to test this. Even further, bitcoin exchange rate could also be compared to other fiat currencies like EUR, GBP, CHF.

Further enhancements to the prediction model should be considered. We have not considered the number of “likes”, “retweets” and “comments” for one tweet. This can imply which tweet is more popular and which tweet will have greater reach. A special metric could be implemented to reflect these values and to include them in the final prediction model. Furthermore, geographic location of the Twitter users could also be taken into consideration. In that way, we could have localised predictions, which could be more accurate in some regions than in others.

Going further, tweets that contain only images, could be analysed with image recognition tools. These tools can provide us with information about the image content, and we could also leverage image sentiment analysis. In that way the scope of the available tweets for the analysis would be greater.

CHAPTER 8

Conclusion

Bitcoin is a cryptocurrency that is not backed by a single entity and it is highly decentralised. This means that the dynamics of its market parameters is highly dependant on the behavior of its users. These users can mine new bitcoins and maintain the bitcoin blockchain system or they influence the price by trading bitcoins to other cryptocurrencies or fiat currencies. Furthermore, emotions can influence investors' decision-making, and also interest in particular topic on web search can be used to predict market movements.

In this thesis we investigated whether data from Twitter and Google Trends can successfully predict two important bitcoin financial parameters, exchange rate and hourly trading volume. A period of two months for July 2018 and August 2018 is analyzed. When it comes to predicting exchange rate, we used sentiment obtained from tweets containing reference to bitcoin. Similarly, in order to predict bitcoin hourly trading volume we used number of tweets related to bitcoin and level of interest shown on Google Trends.

Firstly, we used two different algorithms for analysing sentiment from the obtained tweets, SentiStrength and Stanford CoreNLP. Our results have shown that with both sentiment algorithms there is a correlation between bitcoin price and users sentiment. In order to forecast the price, we used vector autoregressive model (VAR). First prediction model was able to successfully predict whether the price will increase or decrease in the next hour in 63% of cases. Nonetheless, second prediction model was able to successfully predict price movements in 59% of the cases. This demonstrates that both models are valuable when trying to forecast bitcoin price movement. However, the prediction model built using SentiStrength algorithm gave us results with better precision. We concluded that this was expected, since SentiStrength is more suitable for obtaining sentiment information from short informal texts, compared to the Stanford CoreNLP algorithm.

Secondly, we have predicted hourly bitcoin trading volume, measuring how many bitcoins are traded during a single hour. For this purpose another vector autoregressive prediction model was developed. By using this model, we were able to leverage historical information as auto-regressor in order to predict future volume. Results from the model have shown that hourly tweet count and Google Trends data do allow us to make successful predictions about bitcoin trading volume.

Furthermore, our findings can be summarised as follows. By analysing data available through Twitter API and Google Trends API, we have shown that sentiment analysis, frequency of tweets and Google search frequency can be used to successfully predict the bitcoin exchange rate. Analyzed data is processed as time series data, and we describe causality between the time series that we are trying to predict and this data. We have also shown that the sentiment analysis algorithm specialized for tweets does provide us with a more precise prediction model. In addition to that, we have proven that examining popularity of web searches that refer to bitcoin and the number of tweets that contain the reference to bitcoin can predict the bitcoin hourly trading volume.

In summary, this thesis has shown that it is possible to predict two main bitcoin parameters by using publicly available data gained from Twitter and Google Trends.

List of Figures

List of Tables

Listings

Bibliography

[AEAB18] Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, and Andrea Baronchelli. Anticipating cryptocurrency prices using machine learning. *Complexity*, 2018.

[AG17] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[Ana09] Pear Analytics. Twitter study–august 2009. *San Antonio, TX: Pear Analytics. Available at: www.pearanalytics. com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf*, 2009.

[AXV$^+$11] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

[Bac02] Adam Back. Hashcash - a denial of service counter-measure. *Tech Report*, 2002.

[BBC$^+$12] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. Web search queries can predict stock market volumes. *PloS one*, 7(7), 2012.

[BC87] Helmut Braun and John S Chandler. Predicting stock market behavior through rule induction: an application of the learning-from-example approach. *Decision Sciences*, 18(3):415–429, 1987.

[BC15] Nirupama Devi Bhaskar and David Lee Kuo Chuen. Bitcoin exchanges. In David Lee Kuo Chuen, editor, *Handbook of Digital Currency*, pages 559–573. Elsevier, 2015.

[BD16] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. springer, 2016.

[BHM12] John Black, Nigar Hashimzade, and Gareth Myles. *A dictionary of economics*. OUP Oxford, 2012.

[Bla17] Benjamin M Blau. Price dynamics and speculative trading in bitcoin. *Research in International Business and Finance*, 41:493–499, 2017.

[BMP11] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona, Spain, 2011.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[BW07] Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. *Journal of economic perspectives*, 21(2):129–152, 2007.

[CFN88] David Chaum, Amos Fiat, and Moni Naor. Untraceable electronic cash. In *Conference on the Theory and Application of Cryptography, Davos, Switzerland*, pages 319–327. Springer, 1988.

[Cha83] David Chaum. Blind signatures for untraceable payments. In Sherman A.T. Chaum D., Rivest R.L., editor, *Advances in cryptology*, pages 199–203. Springer, Boston, MA, 1983.

[Chu15] David Lee Kuo Chuen. *Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data*. Academic Press, 2015.

[CNC15] Jeffrey Chu, Saralees Nadarajah, and Stephen Chan. Statistical analysis of the exchange rate of bitcoin. *PloS one*, 10(7), 2015.

[COL96] Tim Chenoweth, Zoran Obradovic, and Sauchi Stephen Lee. Embedding technical analysis into neural network based trading systems. *Applied Artificial Intelligence*, 10(6):523–542, 1996.

[CPV$^+$16] Michael Crosby, Pradan Pattanayak, Sanjeev Verma, Vignesh Kalyanaraman, et al. Blockchain technology: Beyond bitcoin. *Applied Innovation*, 2(6-10):71, 2016.

[CRR86] Nai-Fu Chen, Richard Roll, and Stephen A Ross. Economic forces and the stock market. *Journal of business*, 59(3):383–403, 1986.

[CSL13] Yan Carrière-Swallow and Felipe Labbé. Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298, 2013.

[CV12] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88:2–9, 2012.

[CWB$^+$11] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[Dai98] Wei Dai. B-money. *Consulted*, 1:2012, 1998.

[Dam12] Aswath Damodaran. *Investment valuation: Tools and techniques for determining the value of any asset*, volume 666. John Wiley & Sons, 2012.

[DW14] Christian Decker and Roger Wattenhofer. Bitcoin transaction malleability and mtgox. In *European Symposium on Research in Computer Security, Wroclaw, Poland*, pages 313–326. Springer, 2014.

[ET05] David Enke and Suraphan Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, 29(4):927–940, 2005.

[Fam65] Eugene F Fama. The behavior of stock-market prices. *The journal of Business*, 38(1):34–105, 1965.

[Fra14] Pedro Franco. *Understanding Bitcoin: Cryptography, engineering and economics*. John Wiley & Sons, 2014.

[FSS93] Kenneth A Froot, David S Scharfstein, and Jeremy C Stein. Risk management: Coordinating corporate investment and financing policies. *the Journal of Finance*, 48(5):1629–1658, 1993.

[GGK$^+$05] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Chicago, Illinois, USA*, pages 78–87. ACM, 2005.

[GH14] Neil Gandal and Hanna Halaburda. Competition in the cryptocurrency market. *Centre for Economic Policy Research*, 2014.

[Gor62] Myron J Gordon. The savings investment and valuation of a corporation. *The Review of Economics and Statistics*, 44(1):37–51, 1962.

[GPB$^+$15] Ifigeneia Georgoula, Demitrios Pournarakis, Christos Bilanakos, Dionisios Sotiropoulos, and George M Giaglis. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. 2015.

[Gre03] William H Greene. *Econometric analysis*. Pearson Education India, 2003.

[GTMP14] David Garcia, Claudio J Tessone, Pavlin Mavrodiev, and Nicolas Perony. The digital traces of bubbles: feedback cycles between socio-economic signals in the bitcoin economy. *Journal of the Royal Society Interface*, 11(99):20140623, 2014.

[Hil14] Garrick Hileman. From bitcoin to the brixton pound: history and prospects for alternative currencies. In *International Conference on Financial Cryptography and Data Security, Christ Church, Barbados*, pages 163–165. Springer, 2014.

[HR17] Garrick Hileman and Michel Rauchs. Global cryptocurrency benchmarking study. *Cambridge Centre for Alternative Finance*, 33(1), 2017.

[HS90] Stuart Haber and W Scott Stornetta. How to time-stamp a digital document. In *Conference on the Theory and Application of Cryptography, Sydney, NSW, Australia*, pages 437–455. Springer, 1990.

[Hub11] Douglas W Hubbard. *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities*. John wiley & sons, 2011.

[Hut04] W John Hutchins. The georgetown - ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas, Washington, DC, USA*, pages 102–114. Springer, 2004.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, San Jose, California*, pages 56–65. ACM, 2007.

[Kam14] Jermain Kaminski. Nowcasting the bitcoin market with twitter signals. *arXiv preprint arXiv:1406.7577*, 2014.

[KAYT90] Takashi Kimoto, Kazuo Asakawa, Morio Yoda, and Masakazu Takeoka. Stock market prediction system with modular neural networks. In *1990 IJCNN international joint conference on neural networks, San Diego, CA, USA*, pages 1–6. IEEE, 1990.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web, North Carolina, USA*, pages 591–600, 2010.

[Kri13] Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. 3:3415, 2013.

[Kri15] Ladislav Kristoufek. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PloS one*, 10(4), 2015.

[KSSA10] K Senthamarai Kannan, P Sailapathi Sekar, M Mohamed Sathik, and P Arumugam. Financial stock market forecast using data mining techniques. In *Proceedings of the International Multiconference of Engineers and computer scientists, Hong Kong*, volume 1, page 4, 2010.

[KT13] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[KW89] Nobuhiro Kiyotaki and Randall Wright. On money as a medium of exchange. *Journal of political Economy*, 97(4):927–954, 1989.

[KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! 11(538-541):164, 2011.

[Lam04] Monica Lam. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision support systems*, 37(4):567–581, 2004.

[LCD00] Mark T Leung, An-Sing Chen, and Hazem Daouk. Forecasting exchange rates using general regression neural networks. *Computers & Operations Research*, 27(11-12):1093–1110, 2000.

[LDBC10] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector-tracking epidemics on twitter. In *Joint European conference on machine learning and knowledge discovery in databases, Barcelona, Spain*, pages 599–602. Springer, 2010.

[Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[LLG12] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *Twenty-sixth AAAI conference on artificial intelligence, Toronto, Ontario, Canada*, 2012.

[MG12] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. 15, 2012.

[MLM15] Martina Matta, Ilaria Lunesu, and Michele Marchesi. Bitcoin spread prediction using social and web search media. In *UMAP Workshops, Dublin, Ireland*, 2015.

[MM58] Franco Modigliani and Merton H Miller. The cost of capital, corporation finance and the theory of investment. *The American economic review*, 48(3):261–297, 1958.

[MMS99] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[MSB$^+$14] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, Baltimore, Maryland, USA*, pages 55–60, 2014.

[MW00] Luvai Motiwalla and Mahmoud Wahab. Predictable variation and profitable trading of us equities: a trading simulation using neural networks. *Computers & Operations Research*, 27(11-12):1111–1129, 2000.

[MWZ10] Jordi Mondria, Thomas Wu, and Yi Zhang. The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics*, 82(1):85–95, 2010.

[Nak08] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Working Paper*, 2008.

[NBF$^+$16] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. *Bitcoin and cryptocurrency technologies: A comprehensive introduction*. Princeton University Press, 2016.

[NC15] Lam Pak Nian and David LEE Kuo Chuen. Introduction to bitcoin. In David LEE Kuo Chuen, editor, *Handbook of Digital Currency*, pages 5–30. Elsevier, 2015.

[Nof05] John R Nofsinger. Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3):144–160, 2005.

[NOMC11] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[OBRS10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, USA*, 2010.

[OS07] Arthur O'sullivan and Steven M Sheffrin. *Prentice Hall Economics: Principles in Action*. Pearson/Prentice Hall, 2007.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[PD11] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain*, 2011.

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. 2(1–2):1–135, 2008.

[PMS13] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using google trends. 3:1684, 2013.

[PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Seventh International Conference on Language Resources and Evaluation, Malta*, volume 10, pages 1320–1326, 2010.

[PRS10] Tobias Preis, Daniel Reith, and H Eugene Stanley. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719, 2010.

[PSTK15] Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4):2162–2172, 2015.

[PW97] Dean Paxson and Douglas Wood. *The Blackwell encyclopedic dictionary of finance*. Blackwell Business, 1997.

[RS12] Tushar Rao and Saket Srivastava. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), Istanbul, Turkey*, pages 119–123. IEEE Computer Society, 2012.

[RS13] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph. In *International Conference on Financial Cryptography and Data Security, Okinawa, Japan*, pages 6–24. Springer, 2013.

[SB17] Ritesh Srivastava and MPS Bhatia. Challenges with sentiment analysis of on-line micro-texts. *International Journal of Intelligent Systems and Applications*, 9(7):31, 2017.

[SC09] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[Sch97] Berry Schoenmakers. Basic security of the ecashtm payment system. 1528:338–352, 1997.

[Sen19] SentiStrength. http://sentistrength.wlv.ac.uk/. Accessed May, 2019.

[SPW$^+$13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[STY17] Eric WK See-To and Yang Yang. Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, 27(3):283–296, 2017.

[SUW19] Dehua Shen, Andrew Urquhart, and Pengfei Wang. Does twitter predict bitcoin? *Economics Letters*, 174:118–122, 2019.

[Swa15] Melanie Swan. *Blockchain: Blueprint for a new economy.* " O'Reilly Media, Inc.", 2015.

[TBP$^+$10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[TBP11] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[Tet07] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168, 2007.

[The] M Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength (summary book chapter). *Cyberemotions: Collective emotions in cyberspace. Berlin, Germany: Springer*, pages 119–134.

[TSSW10] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media, Washington, DC*, 2010.

[TT16] Don Tapscott and Alex Tapscott. *Blockchain revolution: how the technology behind bitcoin is changing money, business, and the world.* Penguin, 2016.

[Wal11] Benjamin Wallace. The rise and fall of bitcoin. *Wired*, 19(12), 2011.

[WP02] Joseph B Walther and Malcolm R Parks. Cues filtered out, cues filtered in. *Handbook of interpersonal communication*, 3:529–563, 2002.

[ZFG11] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter “i hope it is not as bad as i fear”. *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.

[ZXD$^+$17] Zibin Zheng, Shaoan Xie, Hongning Dai, Xiangping Chen, and Huaimin Wang. An overview of blockchain technology: Architecture, consensus, and future trends. In *2017 IEEE international congress on big data (BigData congress), Honolulu, HI, USA*, pages 557–564. IEEE, 2017.