The Relationship Between Social Media Sentiment and Bitcoin Price Volatility

September 2021

MSc in Economics and Business Administration

Master's Thesis

Authors:

Klaudia Byc

Stefania-Cornelia Ilinca

Supervisor:

Raghava Rao Mukkamala

Characters:

153,753

Number of Pages:

100

Abstract

This paper examines the relationship between the sentiment derived from social media and the cryptocurrency price movement. To achieve this objective, the study combines two natural language processing tools, specifically a document-level sentiment analysis and an aspect-based leved analysis, which is complemented by topic modelling. The dataset has been collected from Twitter using Python and it consists of tweets related to Bitcoin from 2017 to 2021.

The research reveals that the aggregate daily tweets conveyed preponderently positive sentiments. Moreover, the correlation coefficient between sentiment polarity and prices is close to 0, indicating the lack of a connection between the two variables.

The aspect-based sentiment analysis provides slightly better results than the document-level one, however, it is still not able to explain how does the prevailing sentiment of a topic explain the changes in Bitcoin price.

There was not enough evidence to confirm a significant relationship between the sentiment derived from Twitter and Bitcoin price movements. Several limitations could have affected the accuracy of the models: lack of multimedia content processing, use of slang and sarcasm.

Table of Contents

1. Introduction

The birth and development of blockchain technology created many opportunities for digital currencies to emerge and expand. Since the invention of the blockchain technology in 2008, thousands of cryptocurrencies of various sizes have been created, displaying different characteristics and processes. The most popular cryptocurrency to date is Bitcoin, which was first introduced in 2009. It is also the currency with one of the most spectacular evolutions, both in terms of price and volume. Bitcoin started catching attention at a global level at the end of 2013, when its price crossed the $1,000 mark. After that, its price continued to increase significantly at a fast pace, reaching its highest price of over $64,000 this year.

The price development of Bitcoin can be described as hectic, with a lot of up and down movements. Looking at the price evolution over the years, it can be noticed that the price has great fluctuations from day to day. For example, at the beginning of the year, Bitcoin price dropped more than 30% in one day. The high volatility exhibited by the Bitcoin price arises from its design, which involves a lack of a central authority and a limited supply of coins. The high volatility implies high risks, but at the same time it implies great opportunities for speculating investors.

Over the past years, we have observed an increasing popularity of cryptocurrencies and a growing interest for speculating activities. Since some investors were able to make huge profits based on the swings in the cryptocurrency prices, more and more individuals have started investing in these currencies. There can now be seen an abundance of information, discussions and advices related to cryptocurrencies all over the internet.

Social media platforms have also been used to share information and express opinions about cryptocurrencies. It appears that recently, content related to some of the well-known currencies has boomed, with even famous people expressing their thoughts on these. Some of the most popular cryptocurrencies on social media platforms such as Twitter are: Dogecoin, Ethereum and Bitcoin.

Given the high volatility exhibited by cryptocurrencies and the rise in the popularity on social media platforms, it would be interesting to study how does the mood of the social media content impact the price movements of cryptocurrencies.

1. 1 Motivation

The main motivation behind this study is grounded in multiple sources. First, the curiosity for the topic arose from personal experience and observations, where individuals use social media platforms to keep themselves informed and updated about cryptocurrencies. This, together with an increasing Bitcoin popularity and a surge in the volume of online content have provided a starting point for the interest in this topic.

Moreover, a recent event from Twitter that can be correlated to a sudden change in Bitcoin price raised some questions that could be potentially studied. There have been several occasions where Elon Musk, the founder of SpaceX and Tesla Motors, posted content related to Bitcoin and shortly after, the Bitcoin price suddenly increased or decreased. One of the most recent examples is from June 2021, when Elon Musk posted that Tesla will allow Bitcoin payments again once the sustainability issues will be addressed. In less than 24 hours after he posted this tweet, the price of Bitcoin increased by more than 10%. Similar examples can be observed when looking at some of his other Bitcoin related tweets and the price changes following these posts.

Therefore, a question can be asked about how important is the information from social media in explaining the sudden changes in prices of Bitcoin.

Additionally, the paper has a starting point in behavioral finance, which supports the concept of market participants acting less than fully rational when dealing with situations where there is either high risk involved or high uncertainty.

Observing the behavior of the cryptocurrency investors, especially their activity on social media and the price movements of Bitcoin, this can be associated with the concept of herding, since investors seem to follow the general opinion when making their investment decisions.

Therefore, taking into consideration the above mentioned information, this paper will attempt to research the relationship between the sentiment derived from social media platforms and the

volatility of cryptocurrency prices. The objective is to expand the knowledge on how cryptocurrency prices are influenced by social media content.

1.2 Research question

Consequently, this paper seeks to provide an answer to the following research question:

"What is the relationship between sentiment derived from social media content and Bitcoin price movements? "

The following sub-questions are formulated to provide a structure that would facilitate the process of answering the main research question:

1.3 Sub-questions

1) "Is there any significant correlation between tweets polarity and Bitcoin price?"

2) "Are there any dominant topics discussed among Twitter users that affect Bitcoin price?"

1.4 Delimitations

The topic of the study is fairly broad, so this section presents the study limitations, which will help narrow down the topic.

The first limitation to be mentioned is that this study will focus only on one cryptocurrency. The aim of this study is to identify whether the sentiment derived from social media can help explain the movements in the prices of cryptocurrencies, however, given that there is a multitude of cryptocurrencies trading at the moment of writing the thesis, only one will be studied: Bitcoin, since it's the most popular one.

Another limitation worth mentioning is the use of a single social media platform. Although there are several platforms to choose from, this study will include only data from Twitter. Additionally, the time frame chosen for this research is four years, from 2017 to 2021, so the dataset will contain tweets belonging to this time frame. Beside the time frame, there is an additional criteria which limited the dataset: when collecting the tweets, only a sample with a count of likes and retweets above a certain threshold has been selected.

Moreover, there is also a linguistic limitation. Twitter is a social media platform available worldwide, so individuals all over the world can express their opinions in their own language. Tweets are available in a variety of languages, but only the ones written in English have been selected for this study.

There is a wide range of statistical and natural language processing tools that can be used on the dataset in order to answer the research question, however, due to computational power and time constraints, a selection had to be made among these tools. Thus, to conduct the sentiment analysis, only lexicon-based models will be implemented.

1.5 Structure of the paper

This section will describe the entire process behind the research conducted in this paper and all the necessary stages to find answers to the research question.

First, the process will start by creating a theoretical background, with various terms and concepts relevant for the study, which will be useful for understanding the overall topic of the thesis. Then, a review of empirical studies related to the research topic will be presented.

Once the theoretical and empirical background has been established, we move on to identifying what type of data is needed to achieve the purpose of this thesis. Next, the data will be collected from Twitter using a tool from Python.

The raw dataset has to be cleaned and preprocessed using a variety of tools from Python packages. These processes will be described more thoroughly in the following sections.

After obtaining the cleaned dataset, the analysis process can start. This consists of three different models: sentiment analysis, topic modelling and aspect based sentiment analysis. Based on the results presented in this section, a discussion will be included to debate the outcomes of the tools implemented.

Finally, a conclusion will be developed to wrap up the paper and provide an answer to the research questions.

The elements of this master thesis have been summarized in Figure 1.

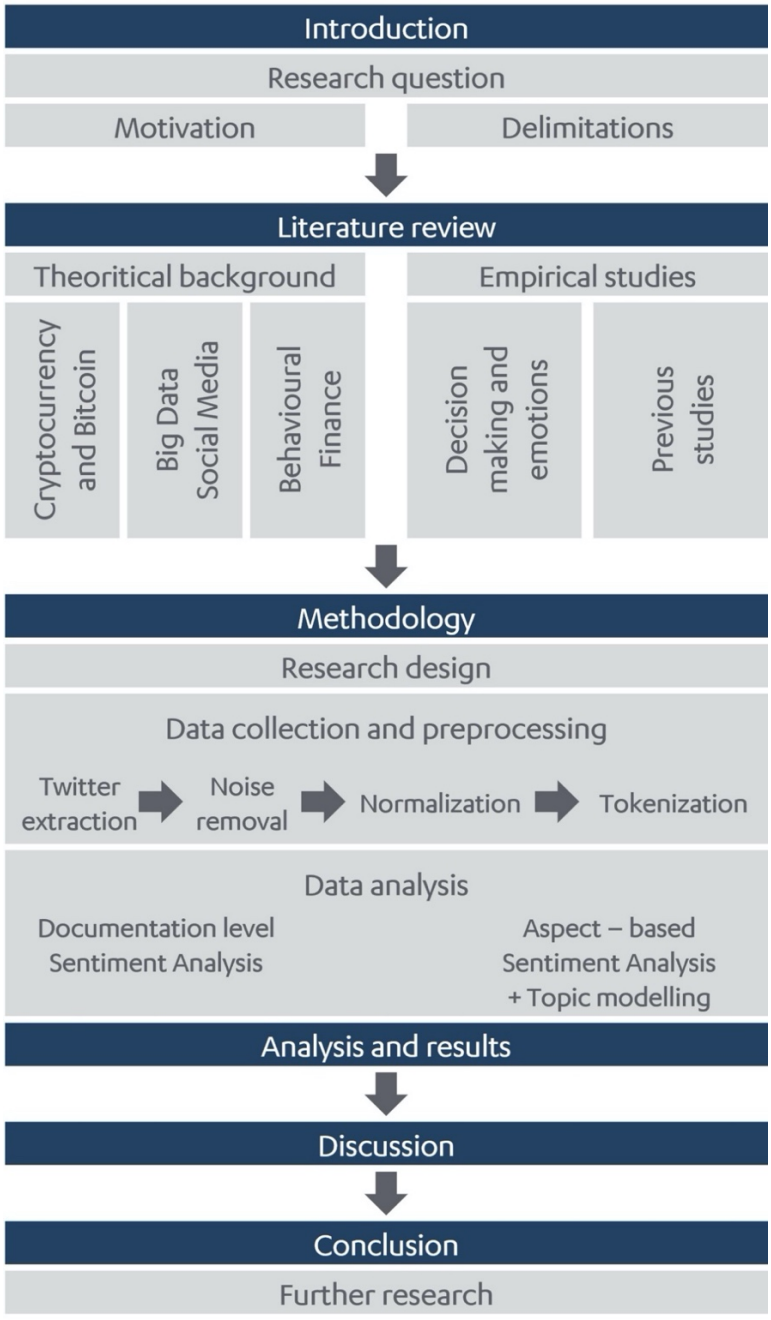

Figure 1 - Structure of the paper

2. Theoretical Background

This section introduces the concepts and the theories which form the foundation of this research paper. All the information presented in this section represents the theoretical starting point of this paper, as well as the essential knowledge related to the chosen topic, facilitating the understanding of the overall purpose of the project.

2.1 Cryptocurrencies

Cryptocurrencies can be generally defined as a digital asset used as medium of exchange, that depend on the use of blockchain technology to provide secure transactions. (Härdle et al, 2020) One of the most significant terms when discussing cryptocurrencies is cryptography. Cryptography represents the discipline of secure communication, taking into account the existence of rivals attempting to interfere with the communication channel. (Franco, 2015) Therefore the central interest of cryptography is encryption. The cryptography has multiple roles for cryptocurrencies, but mainly it is used to encrypt transactions, control the issuance and the supply of supplementary units and to monitor the transfer of assets. (Härdle et al, 2020)

One important characteristic of cryptocurrencies, i.e., a currency without an intrinsic value, is that they can only exist and operate if two conditions are fulfilled: market acceptance and the common belief that the cryptocurrency has value associated with it. (Härdle et al, 2020)

There is an abundance of cryptocurrencies available on the market, with a high level of differentiation among them, such as different rules, different driving mechanisms, latency or even cryptographic hashing algorithms. (Härdle et al, 2020)

One of the most popular cryptocurrencies worldwide is Bitcoin, which will be introduced and described in the following section.

2.2 Bitcoin

Bitcoin was first created in January 2009, based on the ideas proposed in a whitepaper by a person using the pseudonym Satoshi Nakamoto, with an unknown identity to date. Although Bitcoin is not legal tender in most parts of the world, it quickly gained a lot of popularity and it became the largest cryptocurrency traded globally, influencing the creation of many other cryptocurrencies.

To provide a generic definition of Bitcoin, it is a virtual currency, i.e., with no tangible form, created independently of a central institution, such as government or bank. Bitcoin, similar to a traditional currency, is also used as a medium of exchange, but differs greatly from traditional means of payment in almost every other way. (Segendorf, 2014) One Bitcoin unit can be divided into 100 million "Satoshis", which is the smallest fraction of a Bitcoin. (Berentsen&Schar , 2018)

Two of the most prevailing characteristics of Bitcoin are its decentralization and its peer-to peer trading nature, characteristics which also contribute to its desirability. These features are made possible through the use of a system called Proof-of-Work.

The Bitcoin network relies on a technology called Blockchain, which is a shared public ledger, consisting of a history of all the confirmed transactions of the cryptocurrency. As implied by the name, the blockchain comprises of numerous sequential blocks, linked using a cryptographic mechanism. Each block contains information about the newest Bitcoin transactions, and it is built on its predecessors.

The Bitcoin Blockchain is a public record, therefore any individual can access Bitcoin ownership information at any point. New Bitcoin blocks can be added to the Blockchain using Proof-of-Work algorithm. This process is being accomplished by individuals referred to as miners. In consequence, it is necessary to introduce these two terms to understand how Bitcoin is being created and its functionality.

One of the difficulties associated with maintaining a virtual currency such as Bitcoin is establishing the amount of monetary unit existent at every point in time, the amount of new units created with each block and reaching an agreement regarding the ownership rights. For

this purpose, Bitcoin is built with a consensus mechanism which allows its users to reach an agreement, despite its increasing size and the lack of personal interactions. To summarize, the objective of the consensus mechanism is to provide an environment where working together becomes possible even when the users lack reciprocal trust. A few consensus models are available, but Bitcoin uses the Proof of Work. (Berentsen&Schar , 2018)

The consensus among the Bitcoin miners that keeps Bitcoin alive is that all the block contenders with valid fingerprints will be added to each miner's copy of the Bitcoin Blockchain. (Berentsen&Schar , 2018) Thus through this consensus, Bitcoin keeps its decentralized form, as there is no central authority or institution to control the currency and to enforce the rules. Although this raises several issues, including questions about ensuring a secure virtual currency, the Proof of Work mechanism provides a powerful incentive to adhere to the consensus among the miners, because it can be a highly difficult task to unilaterally modify any feature of the blockchain, since an alteration would be reflected in all the following blocks. (Berentsen&Schar , 2018)

2.3 Bitcoin Mining

Bitcoin miners have an essential role in maintaining the Bitcoin. Their tasks consist of gathering the pending Bitcoin transactions, certifying their legitimacy and congregate them into a "block candidate". Following this, the miner has to persuade the other miners to add his block candidate to their own Bitcoin Blockchains. By doing this, the miner can obtain recently created Bitcoin units.

Generally, the process of Bitcoin mining is freely available, so any individual can become a miner, with only two prerequisites: downloading the required software and obtaining the most recent copy of the Bitcoin Blockchain. The reality, however, proves to be slightly different, as the majority of the newly accepted blocks are produced by a few large miners, as opposed to numerous small ones. The main explanation for the existence of only a few large miners is the fierce competition among the miners, which means that only miners with economies of scale can survive as they have access to cheaper electricity, therefore being able to gain some profits from their mining activity.

Mining is mostly being undertaken by large producers since, as explained above, it is an activity that involves high costs. The high costs arise due to the highly specialized hardware needed for mining, as well as the high electricity consumption from the complex computations. Another factors which contribute to the high costs of mining is the process of identifying valid block candidates, as this can only be done through trial and error.

The high costs of mining contribute to its secure feature, as it is increasingly difficult for a single user or a group of users to dominate the blockchain's computing power, given the expensive equipment and resources needed to achieve this.

In order for a block candidate to be valid and therefore accepted by all the users from the network, this must meet particular criteria, such as all the transactions registered in the block should be legitimate and the fingerprint should be below the desired threshold. The fingerprint is obtained by the miners when computing the block's candidate's hash value. (Berentsen&Schar , 2018)

To complicate this process even more, only the fingerprints with hash values below a specific threshold are generally accepted by all miners, which makes the valid fingerprints a rare feature to find. Therefore, there is a continuous process in trying to identify the rare block candidates fulfilling the threshold criterion. Each block has a data field, referred to as the "nonce", and this includes arbitrary data, which miners constantly adjust to get new fingerprints. Every alteration to the arbitrary data creates a new hash value and depending on where the hash value is positioned in relation to the threshold, the miners would either dispose the block candidate or share it immediately to the network, where the other users can check that the fingerprint fulfills the criterion. (Berentsen&Schar , 2018)

2.4 Bitcoin Price

Bitcoin is one of the cryptocurrencies with the highest growth since its origin in 2009. Following its price history and development since 2009, a tremendous surge in price can be observed. The price increase has been continuous since 2009, with periods of decline as well as slower development. As it can be observed on the graph below, the Bitcoin price started at $0 in 2009 and was constant until 2010, when the price increased to $0.1. Based on the graph, a small surge in price can be noticed in January 2014, when Bitcoin reached an unprecedented

price of $1014. The Bitcoin price has continued to develop, with highs and lows, reaching its peak of more than $64,000 in April 2021. At the time of writing this paper - the price of Bitcoin is around $33,500. Hence, the graph below (Figure 2) depicts how volatile Bitcoin price has been throughout the years.

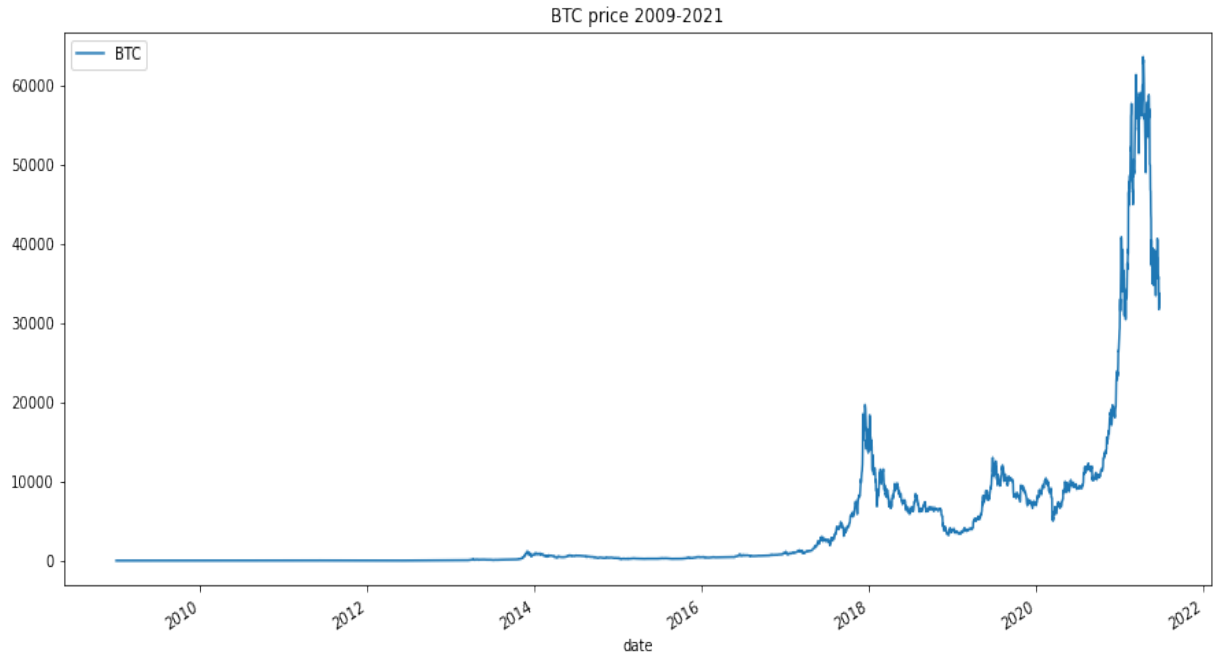

Figure 2 - BTC price 2009-2021, own creation

2.5 Use of Bitcoin

Based on the current set-up of the Bitcoin network, a valid block candidate is found on average every 10 minutes. The miner with the valid block candidate would then earn a predefined amount of newly created units of Bitcoin. The current Bitcoin system is scheduled in such way to achieve 21 million units of Bitcoin. (Berentsen & Schar, 2018)

Bitcoin is available for use to any individual, all that is required in order to start using the Bitcoin system is to obtain a Bitcoin wallet. The Bitcoin wallet is a software, through which users can store, receive and send Bitcoin units, and even fractions of it. Once a user has downloaded the Bitcoin wallet, the only thing left to do is to exchange a standard currency, such as Euros, for Bitcoin units. When a Bitcoin transaction takes place, the seller announces on the network the transfer of a specific amount of Bitcoin units, using the buyer's address.

The information regarding the ownership transfer will be shared from node to node, until all the network users learn about the transaction. (Berentsen & Schar, 2018) Of course, the transaction will follow the regular protocol where the transaction details, specifically the keys and the inputs, are being verified by the miners on the network. In order for a transaction to be successful, the inputs must be confirmed by the miners.

Besides the Bitcoin wallet, other important elements of a Bitcoin account are the keys. Every Bitcoin owner has two keys: a public key and a private one, which should not be shared. This system of using two different keys is called asymmetric cryptography and it is used in order to ensure the legitimacy of the transactions. The private key is used for message encryption, while the public key it's used to decrypt the corresponding message. This process is commonly known as "signature" because it mimics the role of handwritten signatures, to make sure that the transaction belongs to the correct person.

Once a user encrypts a message using their own private key, the other network users can check that in fact the encryption was done using that specific private key, if they can decrypt the message using the corresponding public key. Since only the specific user should have access to that private key, this process ensures the legitimacy of the transaction's origin. (Berentsen&Schar , 2018)

The use of this double key system is also useful in preventing manipulating transactions, because while everyone can decrypt the message and make changes to it, only the user with access to the private key can re-encrypt the message. Since the message will fail to be re-encrypted, the network participants will notice that, and they will reject the message. (Berentsen&Schar , 2018)

Bitcoin transactions can be considered anonymous, given that they do not reveal private information about one's identity if certain guidelines are being followed. However, anonymity should not be understood as complete privacy, since all the transactions are being published on the ledger, therefore being fully trackable and transparent to all the Bitcoin network participants. The anonymity is preserved through the use of pseudonyms in the form of Bitcoin addresses, which are used to hide the real identity of the users. Bitcoin owners are

recommended to use the addresses only once and to use a new address for each transaction, in order to prevent transactions from being associated to a common owner. Bitcoin addresses take the form of long strings of numbers and letters, which appear in a random order. (Franco, 2015) Therefore, to conclude on Bitcoin privacy, all the transactions are public and easily accessible, but the identity behind the transactions has been hidden, therefore protecting, in some part, the anonymity of the users.

2.6 Bitcoin Volatility

A significant aspect of Bitcoin, which has been widely discussed since its apparition is its volatility. Being a virtual currency, there is no intrinsic value associated with Bitcoin, which means that the expectations about the future price have a decisive role in determining its current price. This implies that Bitcoin is highly dependent on the expectations of the market participants, resulting in a very intense price volatility (Berentsen&Schar , 2018), as it can be seen in Figure 2 above.

Comparing Bitcoin with other fiat currency systems, it can be concluded that Bitcoin is generally a less desirable monetary policy because it doesn't meet the conditions to be a stable currency.

When referring to government-run currency systems, a crucial feature for the central bank is the possibility to stabilize the price level by making changes in the money supply. Unfortunately, given the limited constant supply of 21 million units of Bitcoin together with the price of Bitcoin being dependent on the aggregate demand, this makes a price stabilizing mechanism almost impossible, which results in higher price fluctuations for Bitcoin than for government run fiat currencies. (Berentsen&Schar , 2018)

2.7 Big Data

The concept of Big Data is one that has been increasingly popular over the past years, with applications in multiple disciplinary areas. Introducing the concept of Big Data is relevant since this is what makes the study of this research paper possible.

Despite the wide usage of the term Big Data, there is no consensus about a general definition. However, most data specialists agree that Big Data consists of a process where a large set of

data is being extracted, preprocessed and transformed. A well-known approach to describe Big Data is by using the 3Vs attributes: volume, velocity and variety. The volume refers to the aggregate volume of a dataset, the velocity refers to the speed at which the data can be generated and collected, and the variety refers to the diversity of the data, which can be both structured and unstructured and offer unlimited possibilities. (Buyya et al, 2016)

The reasoning behind Big Data is that these large amounts of data collected by companies, entities etc can become useful by being processed in such a way that useful insights and observations are being extracted. Moreover, Big Data can be combined with a multitude of methods and models from different academic fields in order to make some sense of the data and to adapt the analysis to the scope.

Big Data application is not limited only for business purposes but can be observed in almost any field where technology is being used, ranging from IT, media, entertainment sector to biology, biochemistry and healthcare sectors.

Recently, Big Data has been associated with Machine Learning, a term that has also received a large amount of attention in the recent years. The purpose of Machine Learning is to create systems which can handle complex issues and tasks better and faster than humans. It comprises of different techniques and models, depending on the task to be fulfilled. (Buyya et al, 2016) Therefore it is logical to combine these two terms, as they are intertwining frequently, given that Big Data facilitates and enhances the use of Machine Learning.

One of the subfields of Machine Learning, highly relevant for achieving the goal of the study is Text Mining, which will be introduced later in the paper.

2.8 Social Media

It might be a challenge to precisely define social media, however the general principle is to create, exchange and share information, ideas, opinions or other content and participate in social networking. (Kietzmann & Hermkens, 2011) There has been a broad range of applications found for social media with different stand-alone or built-in services. Social media major characteristics are considered to be interactive web-based applications, ability to generate content by users, dedicated profiles for users in each service, ability to develop and

maintain social networks by creating a platforms to exchange thoughts and interests. (Obar & Wildman, 2015) The prime access point for social media services is through web-based apps or mobile apps which can be used on desktops, laptops or convenient mobile devices – tablets and smartphones. First developments on social media happened in the second half of twentieth century, however the real breakthrough in social media industry is dated for the first decade of current century were companies like Facebook, MySpace and Twitter were founded. It is aligned with rapid innovation within mobile industry where smartphones started to gain popularity due to its ease of use and interactive approach with user. Today, multiple social media platforms like Facebook, Twitter, LinkedIn, Instagram, TikTok and Snapchat are international corporates who dominate technology industry. (Obar & Wildman, 2015)

The increased adaption of social media platforms has led to enormous growth of usage. The usage has grown two-fold, in terms of amount of users but also time spent on a platform by a single user. (Bello-Orgaza et al., 2016) The gigantic success of social media is due to its ability to connect people and creating new societal groups which have positive outlook on being engaged in social media by interacting, sharing and collaborating with others. Engagement in social media quickly became a daily standard for teenagers and young adults. (Bello-Orgaza et al., 2016) Data analysis experts quickly realised the scale of data insights produced by each post and interaction which happens on social media platforms like Twitter and Facebook. Such big data is often used for pattern mining, analysing user behaviors as well as visualizing and tracking the data. (Bello-Orgaza et al., 2016) The main challenge with big data acquired from social media is to critically filter through datasets in order to identify meaningful data records. Such data is later analysed with consideration of decision making processes.

The concept of social media mining is emerging among literature publications in the recent years. It can be categorised under data mining and is understood as “the process of representing, analysing and extracting actionable patterns from social media data”. (Zafarani et al., 2014) The approach proposes simple methods and algorithms which are applicable to social media data. The foundation for social media mining lies in traditional disciplines such as sociology but also computer science, machine learning, social network analysis, mathematics.

(Zafarani et al., 2014) The most popular use cases of social media mining are trend analysis, event detection, social spam detection

2.9 Twitter

Twitter is a social networking service which was founded by Jack Dorsey in 2006. It is an interactive platform which enables the users to post "tweets" – short messages which usually encourage others to interact either by liking, replying or retweeting such post. Twitter is primarily available as a mobile app but it also has a web based application for desktop users. (twitter.com) Initially, the service allowed to post messages no longer than 140 characters, however the amount of characters was increased to 280 in the late 2017. Twitter is accessible globally with 330 milion monthly active users dated for first quarter of 2019. (statista.com) Apart from its prime function, Twitter has been used for multiple purposes among societies. There are plenty examples where Twitter was used as a mean to organise protests such as 2009 student protests in Austria, the 2011 Egyptian Revolution or the 2012 Gaza-Israel conflict. On the other hand, Twitter is widely used by the governments and other entities to gauge the emotion sentiment of citizens, manipulate the users by spreading misleading news or to censor the spread of information. Twitter is trying to actively confront such campaigns by detecting suspicious behaviors. For instance, the service has suspended over 1000 accounts during 2019 and 2020 Hong-Kong protests due spreading the misleading information and apparent connections to Chinese government. (Makena, 2019)

One can argue that Twitter interactions among users have high degree of emotions which makes the obtained big data fascinating to analyse for any data scientist.

2.10 Behavioral Finance

One of the most important assumptions in finance theories is that individuals are perfectly rational, and they are able to apply this rationality when valuing securities. This assumption is at the foundation of many theories proposed within finance, such as efficient market hypothesis. However, the premise that market participants are fully rational has been challenged based on observations throughout time that proved the contrary. (Tseng, 2006)

Consequently, alternative concepts have been proposed which are more grounded in reality. For example, the concept of bounded rationality takes as starting point the idea that market

participants are rational, however, it relaxes the assumption that they are perfectly informed at all times. According to bounded rationality, the behavior of market participants deviates from rationality when other elements are included into the equation, such as risk, uncertainty or incomplete information. (Tseng, 2006)

As defined by Simon (1997), bounded rationality implies a rational choice when individuals are bounded by cognitive limitations. This concept is an advancement from the perfectly rational assumption, however, it's still not able to fully describe the behavior of market participants in some situations.

These challenges led to the establishment of behavioral finance as a subdiscipline. Behavioral finance seeks to explain the actual behavior of the market participants against the behavior suggested in financial theories. The objective of behavioral finance is to gain insight into individual's process of making financial decisions in the presence of risk, unpredictability and insufficient information. (Pompian& Pompian, 2012)

Behavioral finance builds on traditional finance, which proposes perfect rationality, but it incorporates components such as emotions and biases to explain the deviations from rationality. It also includes social elements and draws from psychology, therefore being better at explaining human behavior. (Pompian& Pompian, 2012)

An indirect link has been made between emotions and behavioral finance, because emotions are connected to psychology and physiology, which, in turn, is connected to behavioral finance. Therefore, emotions can be highly influential in the decision-making process of humans. To exemplify the importance of emotions during decision making, the notion of emotional states is introduced. Generally, they have been split into two categories: hot states and cold states. The hot states are a result of the human mental state being affected by physiological elements, such as tiredness, hunger or other strong emotions. On the contrary, the cold states are characterized by homeostasis, where emotions are not clouding the judgement, so humans are calm and more rational.

The distance between behavior in hot states and in cold states is called empathy gap. Conventionally, people are not able to recognize the temporary aspect of these states and they

can over- or underpredict their reactions when they will be in the opposite state. The existence of the empathy gap implies flawed expectations about future behavior. As indicated by academic literature, investors have higher chances of misjudging during hot states, which can result in losses. (Tseng, 2006) Therefore understanding individual emotions and how they affect our decisions is highly important.

A brief conclusion of this section is that through behavioral finance, the assumption of fully rational individuals has been relaxed and the role of emotions and psychology in the decision-making process has been emphasized. Moving further, in the next subsection the investor behavior will be described more thoroughly,

2.10.1 Investor behavior

The investor behavior has been thoroughly studied within finance, with contributions from a multitude of researchers shedding some light into this subject. Investor behavior attempts to explain the decision-making process of an investor by combining aspects from psychology and investing. As expected, the behavior can variate depending on the context and the risk involved, while the decision-making process can be affected by a variety of factors, such as cognitive biases and past experiences.

Some of the most common biases which can affect the investor's decision-making process are representativeness, loss aversion, anchoring, trend-chasing bias and overconfidence. The representativeness bias consists of investors classifying an investment based on its historical performance. For example, given a stock with abnormal returns during the past year, this impressive performance would be used as an indicator of the general performance of the stock and used to make decisions further on.

Loss aversion can be described by the phenomenon of attributing more importance to a potential loss than its equivalent gain. In finance, this bias explains why generally, individuals need higher incentives when the threat of loss exists. Anchoring is another common bias among investors. It represents the habit of "anchoring" to the first piece of information learned and using this as a resource for further decisions. This happens when people allow a certain bit of information affect their whole decision-making process, taking a subjective approach rather than an objective perspective. (Baker& Ricciardi, 2014)

The trend-chasing bias, as the name implies, occurs when investors follow the historical performance of securities and mistakenly rely on these historical results to predict the future performance of the asset. (Baker& Ricciardi, 2014) Lastly, the overconfidence bias refers to the overestimation of own's abilities. The overconfidence bias can affect the decision making of investors, by making them overestimate their own expertise and capabilities, therefore encouraging riskier or unjustified investments. (Said et al, 2011)

There is a great amount of studies corroborating the existence of these biases and their effects on the investor's behavior. Daniel et al (1998) provide evidence for the overconfidence bias: their results suggest that investors are more responsive to private information indicators compared to the public information signals. They have discovered that when the public information matches their own private facts, their confidence level increases substantially, however, when faced with information that's contrary to their own facts, their confidence level only slightly decreases. (Tseng, 2006)

Moreover, Lim (2012) also confirms that the overconfidence and the loss biases have a compelling influence on the investor's decisions. Quareshi et al. (2012) studied multiple biases, including representative bias, anchoring and overconfidence. He conducted the research on investors from Pakistan and he was able to determine that all three biases mentioned above are meaningful factors that can affect the decision-making process of an investor.

Another concept that has been recently adapted to financial markets is herd behavior. The concept emerged by observing animal behavior and how at times, they tend to imitate the actions of the others or of a leader, in a manner that seems thoughtless. Adapted to financial markets, the herd behavior is illustrated by situations where market participants behave less rationally, by making similar decisions, usually based on public information at the detriment of private information. It is important to understand this behavior and its implications: it can affect not only the investor's results and gains, but also the economic efficiency. (Darity, 2008)

Herding is not always considered an irrational behavior; this depends on the context and the factors affecting the decision-making process. When there is either high uncertainty, lack of information or potential external threats, herding can be considered the optimal choice,

assuming that the group is more experienced or knowledgeable than a single person. However, if the main motivation behind herding is the social pressure, then the behavior can be described as irrational because the driving factor is conformity with the general opinion, rather than forming a logical conclusion based on available information. (Darity, 2008)

There is extensive empirical research conducted on herd behavior and its implications for the decision-making process, most of the studies confirming the existence of herding among investors.

Kumar (2009) identified a strong presence of herd behavior among investors in the retail segment. Kallinterakis and Gregoriou (2017) reach the same conclusion that confirms the existence of herding, not only at an individual level, but also at a macro level.

A more recent study, conducted in 2019 by Bouri, Gupta and Roubaud investigates the herding behavior among cryptocurrency investors. First, they attempted to study the herding behavior using a static model, which provided no significant results. They also tried a more dynamic approach and through a rolling-window analysis, they were able to indicate the presence of herding among cryptocurrency investors. Lastly, they also conducted a logistic regression which again confirmed the existence of herd behavior, especially when the uncertainty intensifies.

Behavioral finance suggests that investors are not always perfectly rational as expected by standard finance theory and they can be influenced by cognitive biases and also by emotions. The empathy gap becomes relevant for investors, who should aim to be aware of the differences in the behavior between hot state and cold state. Moreover, the herd behavior among investors implies that they behave in similar ways to other investors. Therefore, given the theoretical concepts introduced in this chapter, some pertinent questions can be formulated: is social media capable of reflecting this behavior? If so, can the emotions derived from social media affect the decision-making process of investors, more specifically cryptocurrency investors?

3. Empirical evidence

This section will introduce previous studies following the same direction as this paper, as well as other relevant studies which support the reasoning behind the analysis.

The discussion of the related studies will be divided into two complementary parts: the first part will focus on presenting evidence towards the usefulness of sentiment analysis and the value of social media. The second part focuses on studies that have a similar topic to this master thesis: the relationship between cryptocurrency price movements and the users' sentiment derived from social media.

There is empirical evidence that supports the relevance of implementing sentiment analysis as a tool to answer the research question. Lerner & Keltner (2001) find a significant influence of emotions on the decision-making process. They determine that especially negative emotions can have a direct interference with decision making, by affecting the perceptions of risk.

Fenton-O'Creevy et al.'s (2011) findings support the importance of emotions in the decision making process too. Their study is limited to investors from London, and it concludes that emotions affect the judgement of traders. Since emotions affect decision making, and sentiment analysis is a tool which helps identify emotions, implementing the sentiment analysis is useful in studying the relationship between cryptocurrency prices and the emotions on social media.

Additionally, there is also evidence supporting the value of social media analytics. Oumayma (2020) investigate the impact of social media on the purchasing behavior of consumers using a sample of 828 social media users. He determines that the social media platforms have a high impact on the user's decision to purchase a product/service. The two stages that are the most influenced by social media in the purchase decision making process are the information and evaluation of alternatives. Therefore, information from social media can influence the decisions that individuals make at any given point in time, so it is relevant to include social media platforms as proxy for investor sentiment.

Reviewing the academic literature, numerous studies trying to establish the nature of the relationship between cryptocurrencies and social media content can be found. A recent study conducted by Mirtaheri et al. (2021) investigates cryptocurrency manipulations using social media platforms. They are using data from Twitter and Telegram in an attempt to discover and predict the existence of advertising campaigns and whether these actions will succeed. They are able to identify successful operations of manipulation, especially for the data extracted from Twitter. Their study proves that the activity on social media platforms such as Twitter can affect the decision-making process of the investors.

Another interesting finding is the relationship between the cryptocurrency price behavior and the content from Google Trends and Telegram identified by Smuts (2019). Through his extensive analysis, he discovered a negative correlation between Bitcoin price movements and the search volumes extracted from Google Trends in June 2018. Smuts (2019) also derives sentiments from investment groups on Telegram and finds a positive correlation between these sentiments and the Bitcoin prices. Moreover, he also finds that the number of messages posted in Bitcoin Telegram groups are helpful in predicting the Bitcoin price movements in the following week.

Rahman et al. (2018) study the correlation between Bitcoin price change and the user's sentiment using Twitter data. They implement a more complex classification algorithm, where the sentiments are divided into 10 different groups of emotions, so instead of the classical labels such as positive, negative and neutral, they use emotions such as anger, sadness, surprise etc. The authors apply a variety of machine learning models and are able to discover a positive correlation between change in price of Bitcoin and the sentiment from Twitter.

An abundance of articles use sentiment analysis tools on data extracted from social media and try to build models that can predict prices of cryptocurrency. Wołk (2020) proposes a multimodel approach to evaluate the effects of social media content on Bitcoin prices. He recommends the use of a hybrid model which combines Google Trends data with tweets and the sentiments extracted. Using the hybrid model, the author was able to achieve a significant predictor for fluctuations in Bitcoin. The results shown in his paper indicate a negative correlation between the two variables: Bitcoin prices and sentiments.

Philips and Gorse (2017) have similar results as presented previously. They conclude that social media can contain useful information related to cryptocurrency price movements. The same conclusion is presented in Kim et al.'s paper (2017), where they discuss the usefulness of social media analytics and how several topics can be linked to price movements.

Lastly, there is evidence on the use of topic modelling on Twitter content. The study conducted by Lim and Buntine (2014) and uses over 9 million tweets for electronic products to derive the main topics of discussion and the opinions for these topics. The results show that the aspect-based sentiment analysis conducted on a large amount of tweets generates valuable opinions about the products.

3.1 Research Gap

Based on the information presented in the previous sections, it has been established that there is a research gap which this study can help fill by bringing some knowledge into.

According to the findings from the empirical evidence section, there have been recent studies with similar topics, that is conducting a sentiment analysis on cryptocurrencies. However, the different techniques used, preprocessing elements and the dataset provide a novelty to the already existing academia.

Furthermore, this paper presents a unique approach by combining three different tools to extract information from social media: sentiment analysis, topic modelling and aspect-based sentiment analysis. There was also no indication of other studies using topic modelling on cryptocurrency-related tweets.

This paper brings information and insights from two perspectives: from an individual level and from an aggregate level, using topics as proxies for the sentiments.

4. Methodology

This chapter will intrduce the methodology used throughout the paper, that provides a systematic, planned approach to examine the proposed topic and address the outlined research question. Additionally, it ensures the consistency among all aspects in the study. (Teresa & William, 1997)

In order to construct a well-structured methodology, the researchers have incorporated the Research Onion model created by Saunders et al. (2007). Hence, the study will go through different layers of the research onion as presented in Figure 3.

The section will start with the choice of general philosophy of the research that will be followed by research approach. Next phase will include a discussion of the research design that consist of methodological choice, research purpose, research strategy and time horizon. Even though research design also includes data collection as well as data analysis together and research quality, those elements will be covered in the separate subsections.

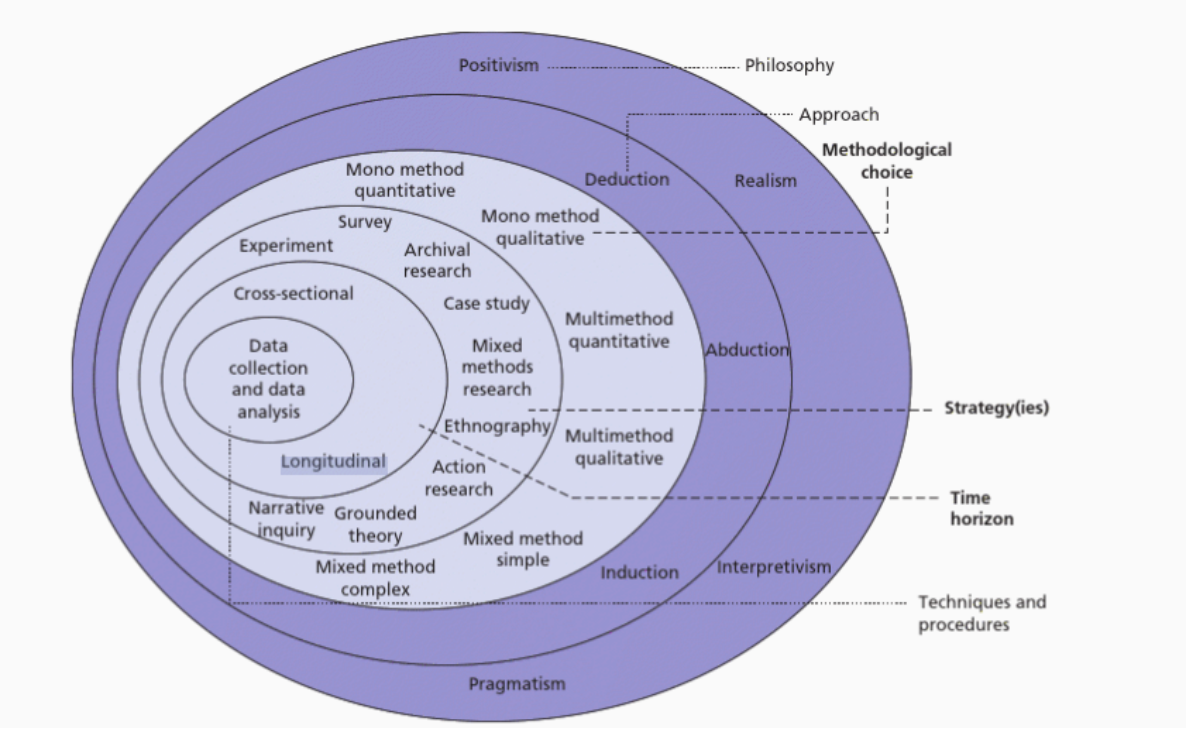

Figure 3 - The research onion, (Saunders et al., 2015, p.164)

4.1 Research philosophy

The first layer of the research onion, introduced by Saunders et al. (2007), refers to the research philosophy, which is defined as "a system of beliefs and assumptions about the development of knowledge". Knowledge development is not only depicted as a new theory invention but it could be also an answer to a specific problem, which is a case in this study. There are four major research philosophies: pragmatism, positivism, realism and interpretivism (Saunders et al., 2007). In order to distinguish between aforementioned philosophies, one need to consider the differences in the assumptions among each of them.

In each phase of the study, there is a continuous need to form numerous assumptions, especially regarding ontology, epistemology and axiology (Burrel & Morgan, 1979).

Ontology is a system of believes about nature of reality. It reflects an interpretation by an individual about what constitutes a fact (Blaikie, 2010). Therefore, ontological assumptions shape the way in which study objects are seen and researched. These study objects can include social groups of people, institutions, organizations as well as social situations, events and social behaviour (Matthews & Ross, 2010).

Epistemology refers to the relationship between the researcher and the phenomenon being studied (Veal, 1997). It focus on the assumptions regarding knowledge, more specifically it target the question what constitute acceptable, valid and legitimate knowledge, and how this knowledge can be communicated to others (Burrel & Morgan, 1979). It is vital to understand the fundamentals of epistemological assumptions since they play a crucial role in the further research design phases, such as methods choice.

Axiology relates to the assumptions about values and ethics within the research process. In particular, axiology is engaged in assessment of the role of researcher's own value throughout all stages of the study (Saunders et al., 2007). For example, knowing the researchers value, one is able to conclude that chosen topic was more important for the researchers than any other. Moreover, Heron (1996) is pointing out that researchers values are being articulated by the way in which the study is conducted, for instance by depicting data collection method.

Additionally, Lee & Lings (2008) argue that axiology also guides the aim of the study as it determine whether the researchers are seeking to explain, predict or understand the world.

Presented three types of philosophical assumptions can undertake two extreme positions, namely objectivism and subjectivism. Objectivism is based on the believes, which states that social reality is external to the researchers and others. It portrays the position that social phenomena exist independent of social actors. Therefore, the perception of social actors do not affect social word (Bryman, 2012). According to Saunders et al. (2007), the most extreme form of objectivism incorporate the assumption that there is only one true social reality encountered by all social actors.

On the other hand, subjectivism focus on believes declaring that social reality is made from perceptions and consequent activity of social actors (Saunders et al., 2007). The most extreme form of subjectivism asserts that the social phenomena is only real when it is created by social actors that are continually reviewing and reworking it (Matthews & Ross, 2010). Therefore, there is no underlying reality to the social world beyond what people constitute to it.

After acknowledging the differences between four major philosophies, namely, pragmatism, positivism, realism and interpretivism by understanding various assumptions and their extreme positions, researchers are able to define research philosophy that will further determine methodological choice, research strategy and data collection techniques and analysis process.

Researchers of this study have developed reflexivity on the philosophical position that led to undertaking positivism as a main stance. Positivism portrays that the researchers see the studied phenomena from the outside and explain its behavior on the basis of objectively gathered data (Veal, 1997). Therefore, the researchers are independent of the data and have no impact on it. Additionally, as argued by Crotty (1998), positivists aim at discovering observable and measurable facts as only the phenomena that could be observed and measured will produce sufficient and valid results.

Following proposed positivist perspective, it is believed that relationship between Bitcoin price volatility and Bitcoin related Tweets is universal and it does exist independent of whether or not it is examined. Researchers aim at investigating the effect of Twitter posts on Bitcoin price

through observable and recorded data set rather than subjective understanding. Lastly, to achieve unbiased findings, researchers detach from their own values and belief during the whole study (Saunders et al., 2007).

4.2 Research approach

According to Malhotra, researchers should endeavor to base study investigations upon objective evidence, supplemented by theory identified by reviewing academic literature (Malthora, et al., 2017). Theory is a set of ideas or related concepts that help to understand, explain and investigate studied event or social phenomena. Depending on the epistemological stance, theories can be explicit or implicit to the chosen research study (Matthews & Ross, 2010).

There are two main research approaches that could be employed in the study, namely, deduction and induction. Deduction undertakes the approach to test the hypothesis that is based on existing theory. The concepts in the hypothesis need to be operationalized, which allows for quantitatively measures (Saunders et al., 2007). Afterwards the hypothesis is either rejected or confirmed. However, researchers may modify the hypothesis throughout the study in order to test again.

On the other hand, induction starts with the data collection process in order to explore studied phenomenon that further leads to the theory formulation. Hence, it is a theory that follows the data. Induction approach is also associated with collection of small sample in contrary to the large number in deduction approach.

Considering the positivist stance of the researchers, it was decided to adopt deduction approach for this thesis. That is, the research question has been formulated based on existing theory regarding investors behavior and decision making that establish a framework and set the basis for the application of our chosen model. The deductive reasoning approach is tied to the mostly quantitative nature of the paper, which is also a case in this study.

Research design

The following section presents the research design, which serves as a guideline to methodological factors on which the thesis is built. The research design refers to the overall strategy chosen to integrate the different components of the study in a coherent and logical way, thereby, ensuring that the research question will be addressed effectively. Although, there is a wide choice of alternative research designs that can meet study objectives, the researchers should aim at creating a design that enhances the value of the information obtained, while decreasing the cost of obtaining it (Malhotra et al. 2017).

As depicted by Saunders et at. (2007), research design incorporates methodological choice, research purpose, research strategy, time horizon as well as data collection, data analysis and research quality. However, it is well worth to mention that the data collection and data analysis together with research quality will be introduced and discussed in two separate subsections due to the high level of information associated with those aspects.

4.3 Methodological choice

Methodology stands for the framework within which research should be undertaken in order to precisely achieve study objectives (Saunders et al., 2007). Methodological choice, that constitutes a third layer of Saunders research onion, will be inevitably influenced by two first outer layers – research philosophy and approach.

The first methodological choice concerns whether researchers follow quantitative, qualitative or mixed methods research design. Quantitative research methods focus on collecting structured data that can be represented numerically (Matthews & Ross, 2010).

Additionally, it is believed that quantitative data is highly associated with positivist philosophy, where collected data is gathered and statistically analyzed.

Contrarily, qualitative research methods concentrate on collecting a large amount of information about a relatively small number of people regarding social or behavioral aspects (Veal, 1997). Qualitative data is frequently gathered by interpretivists and it commonly consist of the words and expressions of the research participants (Matthews & Ross, 2010).

Lastly, mixed methods combine qualitative and quantitative methods in a way that is best for the given research.

Since Bitcoin price as well as Twitter text could be represented as numerically measured variables, quantitative methods will be applied to examine the relationship between those two variables. The choice of quantitative methods was also highly driven by undertaking positivist perspective as well as deductive approach.

4.3.1 Research purpose

The second component of methodological choice refers to determination of research purpose. According to Marshall & Rossman (1995) there are three types of research purpose: exploratory, that focus on investigating a new or an understudied idea, explanatory or causal, whose primary role is to explain factors or relationships related to the phenomenon being studied and the descriptive research, that seeks to describe a certain phenomenon. However, as noticed by Saunders et al. (2007), the purpose of the research may change throughout the study.

The primary purpose of this thesis is explanatory, since the aim of the study is to identify cause-and-effect relationship between Bitcoin price and Bitcoin related Tweets. Therefore, it seeks to determine whether one variable, the sentiment from Twitter affects another outcome variable, which is Bitcoin price. The research is conducted for a topic that has not been clearly studied and the result is inconclusive and so tries to identify a specific answer to the topic.

Moreover, Veal (1997) depicted that depending on the purpose of the research, it could be also categorized as either empirical or non-empirical/theoretical. An empirical study includes data collection, whether quantitative or qualitative, primary or secondary, followed by the analysis of the collected data and has a primary objective expanding the knowledge from actual experience, compared to just theory, as is the case of theoretical research (Veal, 1997).

Presented study is perceived as an empirical research, since its main objective is investigating the relationship between Bitcoin price and Twitter activity. Nevertheless, as Swartz et al. (1998) mention in their paper, “theory cannot be generated without data, and data cannot be collected without a theoretical framework.” Therefore, this research is not purely empirical as

it implicitly contains some theory necessary to explain different concepts related to the object of study. The theoretical background constructed supports the empirical research conducted in this project.

4.4 Research strategy

The fourth layer of the research onion concerns the determination of research strategy, which is defined as a general plan of how the research question should be approached and answered by the researcher (Saunders et al., 2007). It is known as a link between research philosophy and upcoming choice of methods to collect and analyse data (Denzin & Lincoln, 2017). Typically, research strategies are the consequence of philosophical stance and undertaken research approach.

There are a number of frequently used strategies available, such as experiments, surveys, documentary research, ethnography, action research, grounded theory or narrative inquiry, but for this research, case study has been chosen as a primary strategy. According to Robson (2002), a case study is a strategy that aims at conducting research that involves an empirical investigation of certain real-life contemporary circumstances using several sources of evidence. Additionally, Gerring (2008) defined a case study as an intensive study of a single spatially delimited phenomenon observed at a single point in time or over some period of time.

Above definitions seems to depict the strategy followed throughout this research as the aim is to investigate the relatively young Bitcoin phenomenon employing various sources of information and empirical data on Bitcoin price. Considering that this study only examines the price volatility of one cryptocurrency, it points in the direction of a single case study. Moreover, Yin (2009) argued that single case studies should be preferably either critical, unique or phenomenon revealing, which matches the case at hand given that sentiment analysis of social media data with respect to Bitcoin price volatility is still a relatively novel and unexplored area.

It is believed that selected strategy will provide coherence throughout research design, which lead to answering research question (Saunders et al., 2007).

4.5 Time horizon

The fifth onion layer represents the time horizon of the study. According to Saunders et al. (2007), the choice of time horizon is independent of the applied research methodology.

There are two types of time horizons, namely cross-sectional and longitudinal.

Cross-sectional studies concerns the collection of information only once from any given sample from population (Malthora, et al., 2017). It is suggested that cross-sectional research typically focus on investigating possible relationships between the variables about which data is gathered (Matthews & Ross, 2010).

On the contrary, the aim of longitudinal studies is to examine the same sample at key points in time, where data is gathered on at least two separated occasions (Matthews & Ross, 2010). The benefit of longitudinal studies is that it can account for change and development over time.

In line with a cross-sectional study type, the purpose of this study is to investigate whether Bitcoin price volatility can be explained by social media activity captured over 4 years.

Despite that the study covers significant number of days, it is still perceived as a snapshot from a single period of time. Moreover, particular sample was collected only once, thus no changes over time were considered.

The remaining layer of the research onion refers to data collection and analysis, which will be discussed in the subsequent section.

4.6 Data Collection and Preprocessing

The following chapter will provide some relevant insights regarding the data used throughout the study. At the beginning, the researchers will specify data type chosen for this paper, which will be followed by the description of data extraction process. Subsequently, detailed sample selection approach will be introduced, which results in the presentation of the final dataset. The final stages will refer to data cleaning and pre-processing.

4.6.1 Data type

There are two types of data that could be gathered in order to answer the research question. The first one is primary data and the second type is secondary data. Primary data is the new information that the researchers are gathering for the specific purpose of the project, meaning that they are the first users of the data (Veal, 1997).

On the other hand, secondary data composed of already existing information that have been collected for purposes other than the problem at hand (Malthora, et al., 2017). The secondary data can be extracted either from other researches on the same or related topics or from other sources, such as organizations' databases, books, journals, big data sets, industry statistics and reports etc. Therefore, secondary data stands for an economical and quick source of background information.

It is believed that prior to collecting primary data, researchers should locate and analyse relevant secondary data (Malthora, et al., 2017). Consequently, it has been established that there is no need for primary data for the purpose of this paper as the available secondary data would be sufficient. Despite the fact that the researchers constituted the dataset regarding the Bitcoin price and Bitcoin Tweets themselves, the primary raw data was obtained from publicly available and already existing databases. Furthermore, in order to portray the studied phenomenon and outline the theoretical background, literature review was conducted based on relevant books, journals and articles.

4.6.2 Data extraction

The process of data extraction consist of two different phases. The first phase focus on gathering relevant tweets from Twitter, while the second stage consist of collecting historical BTC/USD exchange rate data. The dataset was compiled using a dedicated Python Programming Language.

4.6.2.1 Collecting Tweets from Twitter

The initial process of Tweets extraction was based on accessing Twitter API, which enables programmatic access to Twitter data in unique and advanced ways (https://developer.twitter.com/en/docs). Nevertheless, this approach only allows for collecting tweets that were posted in real-time and the access to historical data was restricted to 7 days.

Therefore, considering several limitations of Twitter API, it was decided to gather relevant Tweets by scraping the Twitter and building own model using snscrape. Snscrape is a scraper for social networking services (SNS) that allows for efficient data collection by filtering based on hashtags or key words (https://github.com/JustAnotherArchivist/snscrape). In this way a significant sample was obtained that is discussed further in this section.

4.6.2.2 Collecting Bitcoin price

This section will present the process of collecting the Bitcoin price over four years, from 2017 to 2021 using Python.

The retrieval of Bitcoin prices has been done using the pycoingecko package, which is linked to CoinGecko's API. CoinGecko is a website that tracks prices and other relevant data for a multitude of cryptocurrencies. First, the package was installed in Python and the library was imported, before setting up the client.

Upon setting the client up, several inputs were needed to create an API call: the time period, which at the time of collecting the data was 1481 days, and the name of the currency as it appears on their website. These inputs were then used in a loop to collect the Bitcoin prices and the dates over 1481 days and save these in a list, later to be converted into a pandas data frame.

The date column in the price data frame is presented into a different format, a Unix timestamp. An example of how this looks like can be seen in Figure 4.

```
In [45]: crypto

Out[45]: [                    bitcoin
          1498262400000   2618.416736
          1498348800000   2606.863235
          1498435200000   2488.194163
          1498521600000   2531.348999
          1498608000000   2575.437770
          ...                     ...
          1631059200000  46995.164171
          1631145600000  46085.028616
          1631232000000  46518.941187
          1631318400000  44802.606402
          1631401972000  45200.119623
```

Figure 4 - Data with Unix timestamp

For ease of interpretation and uniformity throughout the paper, the Unix timestamp is converted into the standard date time format. Additionally, the unnecessary columns have been dropped, retaining only the date column, the variable name and the price. Lastly, the NA values have been eliminated, since these would later affect the NLP procedures. The code for retrieving Bitcoin prices and formatting them can be checked in Appendix 1.

```
In [43]: btc_price.head(10)
```

Out[43]:

	date	variable	value
0	2017-06-24	bitcoin	2618.416736
1	2017-06-25	bitcoin	2606.863235
2	2017-06-26	bitcoin	2488.194163
3	2017-06-27	bitcoin	2531.348999
4	2017-06-28	bitcoin	2575.437770
5	2017-06-29	bitcoin	2561.681435
6	2017-06-30	bitcoin	2490.753488
7	2017-07-01	bitcoin	2439.820560
8	2017-07-02	bitcoin	2518.545694
9	2017-07-03	bitcoin	2576.714657

Figure 5 - Bitcoin price dataset

4.6.3 Sample Selection

As indicated by Saunders et al. (2007), for many research questions and objectives it is impossible to gather and analyse all the potential data available owing to restrictions of time, money or access. Therefore, to reduce the amount of collected data a sampling method should be introduced/implemented. According to Becker (1998), a selected sample should represent the full set of cases in a way that is meaningful, and which could be properly justify.

The sampling design process begins with specifying target population, which is “the collection of elements or objects that possess the information sought by the researcher and about which inferences are to be made” (Malhotra et al. 2017). For the purpose of this study, a target population is represented by all the Tweets that relates to Bitcoin topic.

The next step refers to identifying the sample frame, which is a representation of the elements of the target population that consist of a list of certain directions for determining the target population (Saunders et al., 2007). The sampling frame for this research comprises of all posts on Twitter that includes Bitcoin related key words.

There are various ways in which Twitter users refers to the bitcoin topics. The most direct way is by using a hashtag (”#”) followed by ”bitcoin”. Other likely possibility is typing a hashtag and cryptocurrency abbreviation, which is ”#btc” for Bitcoin. Therefore, the filter keywords were chosen by selecting the most definitive Bitcoin context words that include: Bitcoin, bitcoin, BTC, btc, #btc, #bitcoin, #Bitcoin.

After determining the sample frame, the following step covers a selection of sampling technique. There are multiple sampling techniques, which are grouped into two main categories as probability sampling and non-probability sampling. The difference between aforementioned methods relies on whether the sample selection is based on randomization or not. The probability sampling applies randomization, which guarantees that every element gets equal chance to be chosen as a part of a sample (Riffe, Lacy, & Fico, 2005). On the other hand, non-probability sampling is based on the researcher’s ability to select elements for a sample. Thus, it relies on the personal judgment of the researcher (Malhotra et al. 2017).

In case of exploratory research the judgement of the researcher in choosing participants with particular features may be far more effective than any form of probability sampling (Malhotra et al. 2017). Therefore, a non-probability sampling will be applied to this study as the sample was determined by the researchers based on their topic knowledge. Moreover, a purposive sampling, which is a sub-category of non-probability sampling could be also identified in the research. It is based on the intention of the study, where the elements selected from the population suit the best for the purpose of the research (Riffe, Lacy, & Fico, 2005).

Finally, a sample size need to be determined, which is a crucial step in conducting an empirical research. Sample size refers to the number of elements to be included in the study. According to Saunders et al. (2007), for non-probability sampling the issue of sample size is ambiguous and, in contrast to probability sampling, there are no rules. Saunders further explains that the importance lies in logical relationship between the sample selection technique and the purpose of the study since the generalisations are made to theory rather than about the population (Saunders et al., 2007). As a consequence, sample size depends on the research question and objectives, which includes what need to be discovered, what will have credibility and what could be accomplished given available resources (Patton, 2002).

Despite the fact that validity, understating and insights from data will be more tightened to data collection and analysis skills rather than sample size, it is still important to consider several factors while determining the size of the sample (Patton, 2002). Few of the most significant aspects are the importance of the decision, the nature of the research and analysis, resource constrains and sample sizes used in similar studies (Nunan et al., 2020). All of the depicted features have been thoroughly taken into account by researchers.

Given that this study focus on daily Bitcoin price volatility with respect to Bitcoin related Tweets, the time frame was the first factor that limits the sample size. It was determined to conduct a research within the period from 24th June 2017 to 24th June 2021, which sum up to 1481 days. The decision was made based on the historic Bitcoin price variation that intensified in the mid June 2017 as presented in Figure 2. Therefore, it was assumed that the higher price increases and decreases, the higher the interest on social media, which will result in sufficient number of tweets.

Another crucial aspect that was taken into consideration was time and resource constrain. Given that this study was conducted using the Python Programming Language, sample collection was depended on the researchers computer power. Therefore, the researchers decided to limit the amount of requested tweets so that it could be manageable by available computers. It was achieved by setting by setting the minimum like count to 5 and minimum retweet number to 1 in the snscrape code. In this way, it was assured that tweets with at least

minimum engagement will be collected. Additionally, English language has been chosen as the last search operator.

Finally, the sample size of the previous research, regarding the similar topic, have been thoroughly examined. It was observed that the average sample size varies from 300,000 to 9,000,000 tweets. The large spread was mainly driven by the time span of the studies.

The volume of collected tweets vary depend on the range of active session during the given days. The final dataset contains of 2,195,405 tweets. It is considered as a large sample size, which is aligned with the statement that researches based on quantitative methods should have larger sample than studies with qualitative methods (Malhotra et al. 2017).

4.6.4 Dataset description

To obtain the abovementioned final dataset of 2,195,405 tweets, the researchers have decided to run the snscrape code for each year separately due to the computer power constrain. In this way 4 different datasets have been gathered that were further merged using the *pd.concat* function from pandas package. Figure X presents the final snscrape code for 2018.

```
df_BTC_2018 = pd.DataFrame(itertools.islice(sntwitter.TwitterSearchScraper(
    '#Bitcoin OR #btc OR #bitcoin OR Bitcoin OR btc OR bitcoin\
    lang:en\
    since:2018-06-24 until:2019-06-24\
    include:nativeretweets\
    min_faves:5\
    min_retweets:1').get_items(), 1000000))
```

Figure 6 - Snscrape code 2018

The output of the snscrape code not only consists of the user's text but also numerous variables associated to the single tweet, which are known as properties of the individual tweets. Consequently, each tweet includes 27 variables, where the most significant contain the information regarding: date, tweet content, user, number of replies, number of retweets, number of likes, language or hashtag, Nevertheless, not all of the features were important for the aim of this paper. Therefore, unnecessary columns have been removed from the dataset are the final variables are depicted and described in Table 1.

Variable Name	Variable Type	Variable Description
date	datetime64[ns]	The timestamp of the creation of the tweet.
content	object	The actual tweet text.
retweetCount	Int64	Count of retweets.
likeCount	int64	Count of likes.

Table 1 - Twitter Dataset Description

As evident from the table above, the first variable refers to the exact time when the tweet was posted. This feature is inevitable in order to group tweets by day, later in the pre-processing stage. It also plays a critical role while compering the sentiment of a given tweet with regards to the Bitcoin price from a particular day.

Second feature presented in the table includes the actual text of the requested tweet, which is crucial for the sentiment analysis process as it contains users opinions, believes and comments on a various topics. Twitter streams are known for its lack of structure and high level of noise. Therefore, they requires extensive pre-processing that will be described in the following section.

Lastly, the researchers have decided to keep retweetCount and likeCount columns in order to verify Tweets during the analysis process.

4.6.5 Data cleaning and pre-processing

Upon completing the first step of the analysis process - identifying the data requirements for this research and collecting this necessary data, the next step is to clean and pre-process the raw data to prepare it for the methods to be applied.

As mentioned in the Data Description section, the data collected is unstructured and contains many unnecessary variables and features, which is why it is crucial to clean and transform the content to obtain a more structured dataset, which can be used for classification.

Before diving into the data pre-processing procedure, the notion of text mining will be defined and described. It's important to gain an understanding on the concept of text mining, as it's the foundation of the whole cleaning process.

As mentioned by Dang (2014), a raw, unprocessed dataset which consists of mainly unstructured texts with a large amount of information cannot be easily used in its original form, without any further processing because the text will not be fully recognized by the computers. Text mining represents the process which has as primary objective uncovering valuable information from raw data. (Dang, 2014)

The text mining procedures allow data to be transformed in such way that it becomes accessible to different data mining and machine learning algorithms and tools. This means that text mining facilitates the detection and investigation of various relationships and patterns in the data, which would, in the lack of text mining, be too complicated or even impossible. (Zanini & Dhawan, 2015)

The field of text mining is considered fairly broad, with applications in many different areas, including both academic research, as well as business-oriented studies. A text mining process usually combines multiple tasks into one workflow. Based on this, four stages of text mining have been identified: information retrieval, natural language processing, information extraction and data mining. (Zanini & Dhawan, 2015)

As per the definition above, the data cleaning and pre-processing procedures are part of the broad file of text mining.

Preprocessing procedures are a common feature applied in natural language processing with the objective of preparing the text for the methods to be implemented. The preprocessing techniques consist of cleaning the texts, normalizing them and preserving only the information that is relevant for the classification process. (Duong & Nguyen, 2021)

Prior to describing how the data pre-processing was implemented in this paper, a discussion about the characteristics of the data is required to highlight the importance of this process. As mentioned in the previous section, the dataset consists of more than 2 million Tweets collected over a period of 4 years. Analyzing Tweets is considered a much more complex process, due to

some unique features of the Tweets, which differentiate them from other forms of data (such as reviews, interview, surveys etc).

Since Twitter is a social media platform, tweets have very different characteristics compared to conventional texts, producing some challenges for when trying to analyze them using machine learning methods. Tweets currently have a maximum limit of 280 characters, which is double the limit since November 2017. Social media content (including tweets) is informal with a more relaxed language style since individuals posting online do not always follow grammatical rules. Therefore, it is fairly common to encounter one of the following features in tweets: improper use of grammar, use of slang, abbreviations, use of sarcasm, humor, etc. Moreover, tweets also include a large amount of noisy information, such as punctuation marks, hashtags, URLs, stop words etc., information which proves to be irrelevant in the classification process. (Bao et al, 2014), (Duong & Nguyen, 2021) Given this, it can be understood why the cleaning and transformation of the dataset is relevant.

As the academic literature indicates, the preprocessing of the dataset is extremely valuable and can affect the performance of the model and its results. (Tajinder & Madhu, 2016). Consequently, a large amount of time has been dedicated to this stage in this study, in order to ensure the best results possible.

Browsing through the academic literature, the preprocessing phase contains different cleaning steps depending on the format of the data collected and the overall purpose of the research. Giachanou & Crestani (2016) recommend the following process for preprocessing a conventional text:

1. Tokenization
2. Stop word removal
3. Stemming and lemmatization

However, they distinguish between preprocessing conventional texts and tweets, so they advise on applying a different cleaning procedure for the tweets, which includes two extra intermediary stages:

1. Tokenization

2. Stop word removal
3. URL, hashtag, at symbol removal
4. Spelling correction
5. Stemming and lemmatization

As mentioned previously, the tweets have been collecting with a “scraper”, using a package from Python. Hence the tweets present some undesirable aspects, which need to be handled first. Based on the format of the tweets and the objective of this thesis, it has been determined that a different approach to preprocessing is more suitable, where the order of the steps has been changed and a few other steps have been added.

Moreover, since the paper combines two different natural language processing models, the cleaning and transformation process has to be adjusted for each of the models. Therefore, the data preprocessing will be divided into two parts: Preprocessing for the Sentiment Analysis and Preprocessing for the Topic Modelling.

The same dataset containing the tweets from 2017 to 2021 has been saved into two different variables in Python. Each variable contains the exact same information, but they will be processed differently, as per the required steps of the model being applied. Although the data preprocessing is similar for the two models, there are a few steps done distinctively to warrant a flawless performance of the models.

These are the preprocessing steps for Sentiment Analysis undertaken in this study:

1. Noise removal
    1.1. Emoji handling
    1.2. Remove URLs
    1.3. Target cleaning for emoji
    1.4. Replace contractions
    1.5. Remove punctuation marks and special characters
    1.6. Convert text to lowercase
    1.7. Remove stop words & negation handling
2. Normalization

2.1. Lemmatization

3. Tokenization

Data & Preprocessing for Sentiment Analysis

As a prime step in the cleaning process is installing and loading all the necessary packages in Python. The codes for loading the packages have been all grouped together into one line, which can be seen in Appendix 2. This includes all the packages that have been used throughout the study, incorporating the ones necessary for the sentiment analysis and topic modelling.

In the next step, the data, which has been collected and saved into a separate csv file for each year, has to be loaded and merged. Therefore, each year has been loaded into different Pandas data frames, which have been then merged into a single data frame using the concat function of Pandas. During the merging, the indexes of the data became irregular, so a reset was applied. The relevant code has been appended as Appendix 3.

All the data is now in one data frame and there is one step left before starting the preprocessing: eliminating the columns that are unnecessary. As mentioned in the previous sections, the data has 27 variables, most of which have no value for the analysis, therefore they have to be removed. Out of the 27 columns, only two have been deemed relevant: date column and content column, which contains the tweets, while the other 25 variables have been dropped. The date column has been transformed into datetime format. The corresponding codes can be seen in Appendix 4. The resulting data frame can be seen in Figure 7 below.

```
In [4]: df_btc.head(10)
```

Out[4]:

	Unnamed: 0	Date	content
0	0	2021-06-23	fooo isn't underwater\n\nfooo's entire net wor...
1	1	2021-06-23	Jejudoge CMC listing is Live! Currently Voted ...
2	2	2021-06-23	When people say “bitcoin is MySpace, something...
3	3	2021-06-23	@CryptoGodJohn #BTC Bullish & @CentricRise...
4	4	2021-06-23	HEX, BTC, ETH annual charts. B and E might be ...
5	5	2021-06-23	Fun facts to know and tell your friends: based...
6	6	2021-06-23	Inflation in both BTC and $ADA is ~2%, but in ...
7	7	2021-06-23	$METX .98 Bullish 8/21 MA cross Day 1 Close \n...
8	8	2021-06-23	@OTC_Bitcoin I first bought $jejudoge with Eth...
9	9	2021-06-23	RIP @officialmcafee 😢 Thank you for keeping us...

Figure 7 - First 10 rows tweets

For the analysis, two different methods have been used, corresponding to two Python packages: TextBlob and VADER. They will be described in more detail in the Data Analysis section, but they are mentioned here because there is one distinguishing feature which makes the data cleaning process slightly different: VADER is able to process emoji automatically, while TextBlob does not include such a feature. In consequence, when using TextBlob, the data preprocessing contains one extra step: handling emojis.

Noise Removal

Emoji handling

The preprocessing procedure starts with the emoji handling, step which is exclusively used for the TextBlob sentiment analysis.

Emojis are icons used in online messages, used to express an emotion or idea. In some previous studies, the emojis have been simply considered noisy information and have been deleted, however, this is erroneous because the emojis can contain essential information about the emotion of the tweet. (Duong & Nguyen, 2021) Removing emojis therefore can affect the results of the model and the overall sentiment. More recent studies have tried to analyze the connection between emojis and the sentiment polarity and they have found that some emojis are important for the sentiment polarity, but some others offer mixed signals. Moreover, they

also attempt to compare the sentiments with and without the inclusion of emojis, and are able to assess the significance of emojis in conveying sentiments. (Wang & Castanon, 2015)

Hence, in order to improve the accuracy of the classifier, the emojis must be translated into an input which is recognized by the model. The approach used in this study is to replace the emojis with their textual representation, which can be further subjected to the same treatment as the other words and therefore be recognized by the model.

To complete the emoji translation, the function demojize from the emoji package has been used. However, a layer of complexity has been added, as the function can be applied only to strings, not to data frames. The solution employed was to convert the data frame into a string, apply the demojize function and then convert the newly obtained content back into a data frame. The code has been presented into Appendix 5. Below is an example of how the new column with the translated emojis looks like:

	Date	content_ex \
0	2021-06-23	fooo isn't underwater\n\nfooo's entire net wor...
1	2021-06-23	Jejudoge CMC listing is Live! Currently Voted ...
2	2021-06-23	When people say “bitcoin is MySpace, something...
3	2021-06-23	@CryptoGodJohn #BTC Bullish & @CentricRise...
4	2021-06-23	HEX, BTC, ETH annual charts. B and E might be ...
...	...	...
5995	2021-06-22	"Proof-of-work is essentially one-CPU-one-vote...
5996	2021-06-22	🏆🏆🏆 WINNERS WINNERS WINNERS 🥳🥳🥳\n\n1- @ChAbuba...
5997	2021-06-22	BTC dump, who cares? Now it's a good time to ...
5998	2021-06-22	Happy to see all you who actually understand B...
5999	2021-06-22	If this is a spring, opportunity cost negates ...

	content_demoji
0	fooo isn't underwater\n\nfooo's entire net wor...
1	Jejudoge CMC listing is Live! Currently Voted ...
2	When people say “bitcoin is MySpace, something...
3	@CryptoGodJohn #BTC Bullish & @CentricRise...
4	HEX, BTC, ETH annual charts. B and E might be ...
...	...
5995	"Proof-of-work is essentially one-CPU-one-vote...
5996	:trophy::trophy::trophy: WINNERS WINNERS WINNE...
5997	BTC dump, who cares? Now it's a good time to ...
5998	Happy to see all you who actually understand B...
5999	If this is a spring, opportunity cost negates ...

Figure 8 - Processed emoji

Remove URL

The next step consists of removing URLs. These are links to the corresponding twitter posts or to some other materials or websites. They do not contain any information valuable for the sentiment analysis, so they have to be eliminated. This has been done by using the defining a function using the compile component of the re package. The effect of removing the URLs can be seen in Figure 9 below.CODE

```
In [77]: print(df_final.loc[28,'content_demoji'])

         print(df_final.loc[28,'content_url'])

         This daily close could be one for the ages with funding still negative. $36500 $BTC &amp;amp; $2250 $ETH not out of the question. F
         ire up, winter isn't coming. :fire::sparkler::chart_increasing: https://t.co/7webUh9voX
         This daily close could be one for the ages with funding still negative. $36500 $BTC &amp;amp; $2250 $ETH not out of the question. F
         ire up, winter isn't coming. :fire::sparkler::chart_increasing:
```

Figure 9 - Remove URL example

Replace contractions

Contractions, or contracted words, are shorter words obtained by combining two words and replacing letters with apostrophes. The most noticeable case of contraction is for negations, obtaining words like: don't, won't, haven't etc. Since the contraction combines two different words, the classifier model will not be able to identify them, so it is highly possible that the sentiment will not be determined correctly, especially for the case of negations.

To fix this shortcoming, an extensive dictionary has been created consisting of the majority of the contracted words and their corresponding expanded form. This dictionary (Appendix 6) has been used then to define a function which detects all the contractions and replaces them with their expanded form. An example of a transformation from contraction to full form is presented in Figure 10.

```
print(df_final.loc[0,'content_demoji_clean'])

print(df_final.loc[0,'content_contr'])
```

```
fooo isn't underwater

fooo's entire net worth is up 450x from March 2020 still

fooo hasn't been underwater in anything for years now

trading spot vs btc pair in 2020

literally hodling in 2021

without leverage this game becomes vomitously easy
fooo is not underwater

fooo's entire net worth is up 450x from march 2020 still

fooo has not been underwater in anything for years now

trading spot vs btc pair in 2020

literally hodling in 2021

without leverage this game becomes vomitously easy
```

Figure 10 - Expanding contractions

Remove punctuation marks and special characters

The next step in the noise removal phase is to remove all the punctuation marks, special characters such as hashtags, user references and numbers. None of these present any valuable traits as they do not affect the sentiment, so it is better to eliminate them to reduce noise and keep only the data that is meaningful for the sentiment models. (Duong & Nguyen, 2021)

Convert to lowercase

One common step in the data preprocessing is lowercase conversion. This technique consists in converting the text to lowercase form. The procedure ensures that the data is in the best form to be fed to the model, avoiding issues that could potentially arise if the sentiment classifier might be unable to recognize text in capital letters.

Remove stop words & negation handling

As a final step before moving to the lemmatization stage, the stop words must be removed. Stop words are the most common words in a language, usually carrying little to no valuable information. They appear with high frequency in the text and do not reflect any sentiments, so it is highly recommended to remove the stop words, in order to minimize the dimensionality, the computing time, and to obtain an overall enhancement of the model's performance. (Duong

& Nguyen, 2021) The removal of the stop words is being handled with NLTK from Python, which will be shortly introduced below.

NLTK, is the abbreviation used for Natural Language Toolkit, which is a collection of libraries and programs used, as the name suggests, for various natural language processing tasks. NLTK is widely used for projects and assignments within computational linguistics, due to its extensive range of available features: it can be used for anything from preprocessing to categorizing and analyzing the linguistic structure of a text. (https://www.nltk.org/)

In this study, NLTK has proven its versatility, as many of its libraries have been imported and used throughout the data preprocessing and data analysis stages. The main packages from NLTK are used for stop word removal, lemmatization, tokenization, corpus creation and sentiment analysis.

For stop word removal, the package stopwords has been imported from nltk.corpus. Besides this package, it was necessary to download a list of stop words, which will be used together with the package to help identify and delete the stop words from the data frame. Since the NLTK stop removal list also includes negation words, such as “no” and “not”, these had to be deleted from the list first, otherwise it could have affected the accuracy of the sentiment classifier. The code for this process can be seen in Appendix 7.

In the previous paragraph, the notion of negation has been shallowly touched upon, but given its high weight to the topic, it should be discussed more thoroughly. Negation handling is one of the most examined topics when it comes to sentiment analysis, because it is critical to the performance of the sentiment model. Negation words are essential in identifying the accurate sentiment polarity of a text. If not being handled properly, this can potentially affect the polarity of a text and even reverse it, which is certainly undesirable, as a negative text can be classified as positive and vice versa. As important as negation handling is, it is still a challenge for researchers, as no unique solution has been discovered so far. The academic studies suggest that the majority of the researchers are dealing with negation by reversing the polarity when encountering negative words. Of course, this solution is not perfect and has many critics to

how it is not always accurate and does not improve the model in all the cases. (Giachanou& Crestani, 2016)

Another approach introduced by Kiritchenko et al. (2014) was to create two lexicons, one consisting of words that generally are found in contexts with negations and one without. This approach seemed to improve the performance of the sentiment analysis, however, it is a very computational heavy task.

In this master thesis, a first attempt at dealing with negation was by using antonyms. A function has been defined that iterates over the content column of the dataset and replaces the words in the presence of negative forms such as “no” and “not” with their antonyms. However, this proved to be ineffective at times, as the antonyms used did not have the same contextual meaning as the words they were replacing, therefore having a higher chance to decrease the accuracy of the sentiment classifier.

The solution implemented was to keep the negative words by eliminating them from the stop words list, together with building bigrams. Just by keeping the negative words, the performance of the model still increases because the packages for sentiment analysis are able to recognize these words as negative and assign them a negative score. This is still not as good as having a solution that can interpret the negation contextually, but it is seen as an improvement from the basic procedure.

The results of applying the last three mentioned preprocessing procedures can be seen in Figures 11 and 12.

```
In [90]: print(df_final.loc[28,'content_ex'])

         print(df_final.loc[28,'content_clean_2'])

This daily close could be one for the ages with funding still negative. $36500 $BTC &amp;amp; $2250 $ETH not out of the question. F
ire up, winter isn't coming. 🔥🌻📈 https://t.co/7webUh9voX
daily close could one ages funding still negative btc amp eth not question fire winter not coming fire sparkler chart increasin
g
```

Figure 11 - Preprocessing example 1

```
In [91]: print(df_final.loc[100,'content_ex'])

         print(df_final.loc[100,'content_clean_2'])

         @FinancialTimes Anyone who can have there bitcoin taken from them has not done enough research in how to properly secure it.
         Leaving it on exchange is like leaving $s in a bank, it can be confiscated at any time.   Secure your keys and there is nothing
         that can be done to take them from you.
         financialtimes anyone bitcoin taken not done enough research properly secure leaving exchange like leaving bank confiscated tim
         e secure keys nothing done take
```

Figure 12 - Preprocessing example 2

Moreover, bigrams were created and added to the dataset. Bigrams represent sequences of two adjacent elements, such as words. The bigrams created consisted of two words that are frequently found together identified at least 50 times throughout the dataset. The frequent words were binded together using the underscore sign and these newly created bigrams replaced the separate words. Among these bigrams, there were many combinations of words which included a negation.

```
In [31]: filtered_bigram.head(10)
```

Out[31]:

	bigram	raw_freq
0	(btc, eth)	0.002958
1	(bitcoin, btc)	0.001581
2	(btc, bitcoin)	0.001226
3	(bitcoin, ethereum)	0.001175
4	(bitcoin, not)	0.000979
5	(eth, btc)	0.000887
6	(buy, bitcoin)	0.000842
7	(let, us)	0.000809
8	(bitcoin, crypto)	0.000804
9	(bitcoin, cash)	0.000753

Figure 13 - Example bigrams

Normalization

Text normalization can be defined as the process of transforming a word into its base form. During this process, the base is obtained by removing the inflectional form of the word. The aim of normalization processes is to remove the variations of a word to draw them nearer a single standard form. The main goal of normalization is to improve efficiency by decreasing the volume of information to be processed by the computer and model. (Bao et al, 2014)

Two of the most popular normalization techniques used in the preprocessing stage are stemming and lemmatization. Stemming is the process through which words are reduced to their stem form, with the goal of bringing related words to a common stem. (Bao et al, 2014) For example, the word “children”, through stemming, would be converted into “child”. There is one disadvantage associated with this normalization technique: the stem form is not necessarily a dictionary word, which can create complications during the sentiment analysis, as the words may not be recognized by the models

Lemmatization is similar to stemming, but instead of reducing the word to a stem form, it reduces it to the base word, a process that ensures the transformed word is found in the language. (Bao et al, 2014), As opposed to stemming, the lemmatized word, or lemma, has a dictionary form, so it does not affect the sentiment analysis performance. An example of lemmatization is the transformation of the word “bought” into its root form “buy”.

Given the discussion above, only the lemmatization process has been chosen for the normalization of the tweets, as the stemming can affect the accuracy of the sentiment classifiers.

To the apply lemmatization techniques to the dataset, the wordNetLemmatizer package from NLTK has been used. A function has been defined which incorporated multiple preprocessing techniques in order to facilitate the process and improve the time efficiency.

Following the lemmatization process there were still some noises identified in the dataset: single letters with no value for the sentiment analysis. Therefore, these were removed using a simple function.

Tokenization

The last step before starting the analysis of the collected data is tokenization. Tokenization is defined as splitting the text into smaller units, also known as tokens, generally represented by words. Tokenization is an essential step when preprocessing the date for sentiment analysis. (Wisdom, 2016)

In order to tokenize the tweets, Python's x.split function has been applied, separates the contents of a text using a comma as a separator.

Figure 14 shows the first 10 rows of the dataset after the tokenization has taken place.

```
In [158]: tokenized_tweet.head(10)

Out[158]: 0    [fooo, not, underwater, fooo, entire, net, wor...
          1    [jejudoge, cmc, listing, live, currently, vote...
          2    [people, say, bitcoin, myspace, something, gon...
          3    [cryptogodjohn, btc, bullish, amp, centricrise...
          4    [hex, btc, eth, annual, charts, might, sign, c...
          5    [fun, facts, know, tell, friends, based, sales...
          6    [inflation, btc, ada, latter, staking, risk, f...
          7    [metx, bullish, cross, day, close, low, bullis...
          8    [otc, bitcoin, first, bought, jejudoge, ethere...
          9    [rip, officialmcafee, thank, keeping, us, safe...
          Name: content_clean_3, dtype: object
```

Figure 14 - Tokenization example

Data & Preprocessing for Topic Modelling:

A second natural language processing model used in this paper is topic modelling. This technique will be more thoroughly described in the next section, but it is important to note the differences in data preprocessing. Although the cleaning and preprocessing procedure is similar to that for the sentiment analysis, described in the previous section, there are a few distinguishing steps fundamental for ensuring the best performance of the topic modelling.

Another dissimilarity compared to the sentiment analysis, is that the topic modelling will be conducted on a subset of data, consisting of one year of tweets, from June 2020 to June 2021. The explanation for choosing a one year subset instead of the whole dataset, is that topic modelling consists of a more in-depth analysis, which requires not only more time, but also more computational power.

Most of the steps are identical to the ones outlined in the sentiment analysis section. The same rationality applies in this section as well, therefore only the dissimilar steps will be discussed here.

Remove emojis

For the sentiment analysis, emojis are important as they can contain useful information about the emotions of the users. However, topic modelling is not concerned with identifying emotions, but only with determining the most popular topics in a certain text, which is done using word frequency. By converting emojis to a textual representation, this affects the topic identification, depending on the frequency of the emojis. Since the emojis were very frequent throughout the tweets, this influenced the topics found, as most of them were related to emojis. To prevent this situation which can lead to the misidentification of the topics, the emojis have been removed altogether from the dataset.

Removing words related to cryptocurrencies

The same reasoning described for emojis removal applies here as well: all the words related to cryptocurrencies and the words previously used as hashtags appear with a high frequency, affecting the results of the topic modelling. The cryptocurrency words such as “bitcoin”, “btc”, “crypto”, “Ethereum” etc. are present in a large number in the tweets and they do not reflect a topic, but rather the object of a topic. Accordingly, they were excluded from the text.

Remove words related to BITCOIN

In [41]:
```
# Remove the words: btc, BTC, Bitcoin, bitcoin etc.
remove_words = ['btc','bitcoin', 'xrp', 'eth', 'ethereum', 'cryptocurrencies', 'cryptocurrency',
                'crypto', 'bnb', 'binance', 'amp', 'doge','dogecoin', 'ada', 'cardano', 'xrp',
                'usdt', 'tether', 'usdc', 'usd coin', 'dot', 'polkadot', 'busd', 'uniswap',
                'uni', 'shib', 'shiba inu', 'nft', 'nfts', 'gt', 'ltc', 'litecoin', 'sol',
                'solana', 'coin', 'usd', 'blockchain', 'gon', 'na', 'wan']

pat = '|'.join([r'\b{}\b'.format(w) for w in remove_words])
df_final['content_clean_3'] = df_final['content_clean_2'].str.replace(pat, '')
```

Figure 15 - Remove Bitcoin words

Remove frequent words

Moreover, a list of the most frequent words has been compiled in order to evaluate the usefulness of these words. After a scrutinous evaluation, the first 20 words have been selected as having no value to the model and they have been eliminated to enhance performance of the topic modelling.

Get the most frequent words

```
In [67]: from collections import Counter
         cnt = Counter()
         for text in df_final["final_tweets"].values:
             for word in text.split():
                 cnt[word] += 1

         cnt.most_common(20)
```

```
Out[67]: [('like', 39330),
          ('buy', 37117),
          ('one', 37085),
          ('people', 35139),
          ('get', 34408),
          ('price', 32161),
          ('new', 30880),
          ('time', 30245),
          ('good', 29559),
          ('money', 29126),
          ('going', 26920),
          ('today', 25800),
          ('would', 25724),
          ('see', 25602),
          ('market', 23335),
          ('still', 23256),
          ('think', 22701),
          ('go', 22226),
          ('link', 22191),
          ('defi', 21765),
```

Figure 16 - Frequent words

```
In [69]: print(df_final.loc[100,'content'])

         print(df_final.loc[100,'content_freq'])
```

```
@FinancialTimes Anyone who can have there bitcoin taken from them has not done enough research in how to properly secure it.
Leaving it on exchange is like leaving $s in a bank, it can be confiscated at any time.   Secure your keys and there is nothing
that can be done to take them from you.
financialtimes anyone taken done enough research properly secure leaving exchange leaving bank confiscated secure keys nothing
done take
```

Figure 17 - Frequent words example

Remove negations

Negation words have been removed from the text together with the stop words, because they are the some of the most frequent words, but they do not have any influence on determining the topics.

Lemmatization

Although this step is identical to the one performed for sentiment analysis, the reasoning is slightly different. Lemmatization is an eminent step for topic modelling because of the nature of the task. The topic modelling algorithm is based on the distribution of the words, so identifying the correct frequency of a word is vital to the accuracy of the model. By

lemmatizing the words and bringing them to a standard variant which is consistent throughout the whole dataset, it is ensured that the topic modelling will correctly determine the frequency of words with different forms.

Split data monthly

After the whole data preprocessing has been completed, the last step before proceeding with the topic modelling is to split the dataset into smaller sets, using a monthly verge. By using the aggregate data, the frequency of some common words increases and therefore affects the topic identification. The one year dataset has been divided into 12 separate pandas data frames, one for each month.

Split data into 12 months

```
In [205]: june_20 = df_final[(df_final['Date'] >= '2020-06-01') & (df_final['Date'] < '2020-07-01')]
          june_20 = june_20[['Date','content_freq']]
          june_20.reset_index(drop=True, inplace=True)

          july_20 = df_final[(df_final['Date'] >= '2020-07-01') & (df_final['Date'] < '2020-08-01')]
          july_20 = july_20[['Date', 'content_freq']]
          july_20.reset_index(drop=True, inplace=True)

          aug_20 = df_final[(df_final['Date'] >= '2020-08-01') & (df_final['Date'] < '2020-09-01')]
          aug_20 = aug_20[['Date', 'content_freq']]
          aug_20.reset_index(drop=True, inplace=True)

          sep_20 = df_final[(df_final['Date'] >= '2020-09-01') & (df_final['Date'] < '2020-10-01')]
          sep_20 = sep_20[['Date', 'content_freq']]
          sep_20.reset_index(drop=True, inplace=True)

          oct_20 = df_final[(df_final['Date'] >= '2020-10-01') & (df_final['Date'] < '2020-11-01')]
          oct_20 = oct_20[['Date', 'content_freq']]
          oct_20.reset_index(drop=True, inplace=True)

          nov_20 = df_final[(df_final['Date'] >= '2020-11-01') & (df_final['Date'] < '2020-12-01')]
          nov_20 = nov_20[['Date', 'content_freq']]
          nov_20.reset_index(drop=True, inplace=True)
```

Figure 18 - Split data

4.7 Data Analysis

This chapter will demonstrate and justify the choice of data analysis methods proposed to answer the research question of this paper. At this stage, it is well-worth to recall that this study aims at investigating whether Bitcoin related tweets may contribute to the Bitcoin price volatility. The possible effect of Twitter posts will be measure by extracting sentiments from collected tweets, which will be further compared to the daily price of chosen cryptocurrency.

Given that, the following section will introduce the sentiment analysis as a sufficient method to address the research problem. Moreover, to improve the model and gain different perspective on the topic at hand, the researchers portrayed an aspect-based sentiment analysis, which also includes topic modeling.

Furthermore, to adequately comprehend the proposed methods, it is inevitable to introduce the field of computational linguistics and natural language processing at the beginning of this chapter. Finally, a discussion regarding validity and reliability of the research will be presented.

4.7.1 Computational linguistics

Computational linguistics is an interdisciplinary field that primarily focus on the computational modelling and mathematical properties of natural language as well as the design and analysis of natural language processing systems (McEnery, 1996). Computational linguistics requires expertise in various fields such as artificial intelligence, machine learning, cognitive computing, linguistics and neuroscience.

There are two components of computational linguistics, namely, theoretical and applied component. Theoretical computational linguistics concerns with issues in theoretical linguistics and cognitive science, whilst applied computational linguistics focuses on the practical side of modeling human language use (Och & Ney, 2003).

Majority of tasks in computational linguistics aim at improving the relationship between computers and basic language, which includes constructing artifacts that can be used to process and produce language. Consequently, it requires data scientists to analyze massive amounts of written and spoken language in both structured and unstructured formats (Motkov, 2004)

4.7.2 Natural language processing

Nowadays considerable amount of important information is contained in raw text in different human languages. Nevertheless, computers on their own are not able to understand and extract meaningful insights from these unstructured text data. Therefore, one has introduced Natural language processing as a perfect tool to make computers capable to understand the contents of various documents.

Natural language processing (NLP) is a sub-field of applied computational linguistics that focus on interactions between computers and human language, in particular how to program computers to process and analyze enormous amounts of natural language data. As a consequence, dedicated machines can extract relevant information and insights included in the documents as well as classify or sort the documents. The most common applications of NLP

incorporate speech recognition, natural language understanding, and natural language generation that could be seen in Amazon's Alexa or Apple's Siri. These virtual assistants recognize verbal commands and complete requested actions by humans, such as making a phone call or sending a text message.

As depicted by Liu (2015), with the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, and postings in social network sites), individuals and organizations are increasingly using the content in these media for decision making. Therefore, NLP found its another wide implementation in sentiment detection of social media data, which will be discuss in more details in the next subsection.

4.7.3 Sentiment Analysis

As defined by Liu (2015), sentiment analysis, also referred to as opinion mining, is "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes". The technology of sentiment analysis includes the use of NLP, Computational Linguistics, text analysis and text mining to determine the sentiment expressed in a given text, often scored within values for negative, neutral and positive attitude. (Pak & Paroubek 2010) Sentiment analysis has its applications in a wide variety of domains including analyzing news, movie or food reviews and social media content.

The main purpose of implementing sentiment analysis in this study is to determine whether the extracted Twitter posts are generally positive, negative or neutral in the users opinion of the chosen cryptocurrency. Hence, the model is extremely important since it allows to understand the sentiment of Twitter users towards Bitcoin. Consequently, sentiment analysis is a crucial part of the data analysis process conducted in this paper since the output of its module is used to identify any relations or patterns between Twitter interactions regarding Bitcoin and Bitcoin price variations.

Sentiment Analysis Levels

In general, sentiment analysis could be performed at three different levels: 1) Document level, 2) Sentence level, 3) Entity and Aspect Level. (Liu, 2015)

Firstly, application of the model at the document level is characterize by performing sentiment analysis on a one piece of text, which may vary from one line of a text to a one-hundred-page document. This level of analysis assumes that each document expresses opinions on a single entity. (Provost & Fawcett, 2013).

Secondly, task performed at sentence level requires to separate the documents into sentences and then determines whether each sentence expressed a positive, negative, or neutral opinion. Liu (2015) argue that presented level of analysis has common characteristics with subjectivity classification, which distinguishes objective sentences that express factual information from subjective sentences that express subjective views and opinions. Nevertheless, as pointed out by the author, the subjectivity is not equivalent to the sentiment since many objective sentences may contain opinions. (Liu, 2015)

Lastly, the goal of conducting sentiment analysis on the entity and aspect level is to identify and extract object features that have been commented on by the opinion holder. Therefore, instead of examining the language constructs, such as documents or sentences, aspect level directly focus on the opinion itself. Hence, it is based on the idea that an opinion consists of a positive or negative sentiment and a target of that opinion. (Liu, 2015)

For the purpose of this study, the researchers found it relevant to conduct sentiment analysis on both document as well as entity and aspect levels. The application of the entity and aspect level will be discussed in more details in the section regarding aspect-based sentiment analysis. Although, this section will focus on implementation of the document level model. The document is defined as a text from one day, which was achieved by grouping by all tweets from that particular day, whilst a single entity is represented by Bitcoin.

Sentiment Analysis Algorithms

It is essential to acknowledge that sentiment analysis can be performed in various ways since there are many methods and algorithms to implement sentiment-analysis systems. The most general approaches refers to the rule-based method as well as automatic method.

Rule-based system focus on performing sentiment analysis based on set of manually crafted rules according to which the words are classified as either positive or negative along with their corresponding intensity measure. Rule-based approach uses a lexicon, which is a component of NLP system that contains information about individual words or word strings. (Guthrie et al., 1996) This may incorporate a dictionary of positive and negative words with a sentiment value assigned to each of the words.

Automatic systems utilizes machine learning and statistical techniques to learn from data to determine the sentiment orientation. This approach is known for its precision and accuracy since it involves unsupervised machine learning algorithms to explore the data as well as supervised machine learning for classification algorithms. (Hassan et al. 2014)

Despite that automatic approach may perform the best results, its application requires larger datasets than in the rule-based system as well as the existence of labelled training set. (Pang et. al., 2002) Even though dataset gathered for this analysis contains over two millions tweets, none of them is primarily labeled as positive, negative or neutral, which means that additional training dataset will be needed to train the machine learning algorithm. Moreover, if the research would like to construct the training and testing datasets on their own, they need to ensure equal distribution of positive, negative, and neutral tweets in the training set. It will be hard to achieve as there is a tendency for more positive tweets than negative or neutral, which may lead to unbalanced data. In this way machine learning algorithm may classify significant majority of the tweets as positive based on the biased learning. Therefore, considering time constrain as well as the lack of resources, which includes distributed-servers needed for training sophisticated machine learning models, it was decided to undertake rule-based approach to perform sentiment analysis for this paper. Nevertheless, with the use of currently available advanced tools, the accuracy of rule-based technique is comparable to automatic

approach but at lower cost level. Moreover, rule-based approach can be more easily understood and modified by humans, which is considered a significant advantage for this study.

As mentioned before, rule-based systems are built on lexicons that may vary from simple one to more complex, which includes negation rules, distance calculations, added-variance, and other sufficient rules. (Liu, 2015) Since real world data, such as social media content, is very intricate and nuanced as it includes slang, emojis, abbreviations, misspelled words, punctuation or shorthand, there is a high need for sophisticated lexicons to handle this complex data. Fortunately, there are several sufficient Python libraries that enable to conduct rule-based sentiment analysis process, which not only assess the polarity of the word but they also consider the intensity of the words, meaning that they assess how much positive or negative the given word is. Additionally, those advanced lexicons also determine the subjectivity or objectivity of the word. The most popular and reliable tools for rule-based sentiment analysis are TextBlob and VADER.

TextBlob

TextBlob is an extremely powerful NLP library that offers its own sentiment analysis functionality, which is based on Natural Language ToolKit. NLTK, as previously described during the pre-processing stage, is a Python library that interfaces to 50 corpora and lexical resources, which enable to precisely conduct categorization, classification and many other tasks. (nltk.org)

TextBlob returns texts' sentimental features, such as polarity and subjectivity by using predefined word scores. (Loria, 2018) The polarity score is a measurement that states how positive or negative the given content is. Polarity value ranges between -1 to 1, where 1 means positive statement and -1 means a negative statement. On the other hand, subjectivity refers to the personal opinion or judgment as well as factual information contained in the text. (Loria, 2018) It lies in the range between 0 and 1. The higher the subjectivity value for particular instance, the less objective it becomes. Additionally, TextBlob contains the intensity parameter, which assess whether the given word modifies the next word, for example 'extremely perfect'. Moreover, it is portrayed that the TextBlob can detect negation words and

handle them by multiplying the polarity by -0,5, which results in reversing the polarity of a given instance. It is also worth to notice that when a negation combines with modifiers, the inverse intensity of the modifier enters for subjectivity and polarity. Lastly, the library also offer access to semantic labels, which are remarkably beneficial for performing precise analysis. (Loria, 2018)

Abovementioned TextBlob's characteristic made it a perfect tool to support complex analysis and operations on textual data. Therefore, it will be implemented for sentiment analysis in this paper. Nevertheless, in order to gain broader overview on the topic as well as a point of comparison for TextBlob, it was also decided to conduct sentiment analysis with VADER, which is introduced below.

VADER

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a well-known lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. (Hutto & Gilbert, 2014) It relies on a dictionary that contains words and other numerous lexical features that are common for sentiment expression in microblogs. Each word in the dictionary is assigned a score measured on a scale from -4 to +4, where -4 stands for the most 'Negative' sentiment and +4 for the most 'Positive' sentiment. Intuitively the midpoint of 0 represents 'Neutral' sentiment. Subsequently,

VADER maps the word under the consideration with the corresponding one from the dictionary, which is already labeled as positive or negative according to its semantic orientation. As a result, VADER returns the probability of a given input sentence to be positive, negative, and neutral. (Hutto & Gilbert, 2014) Moreover, it also provides a fourth score, which is a compound sentiment score that ranges from -1 to 1. The compound score is computed by normalizing the three aforementioned probability scores.

VADER, similarly to TextBlob, is not only sensitive to polarity but also to intensity. Furthermore, proposed tool is perfectly adjusted to detect and handle negations as well as conjunctions. Lastly, VADER's resource-efficient approach helps to decode and quantify the emotions widely contained in social media content. (Hutto & Gilbert, 2014)

Aspect-Based Sentiment Analysis

In contrast to basic sentiment analysis, which extracts the general sentiment expressed in a piece of text, aspect-based sentiment analysis aims to obtain both the entity described in the text, in this case, Bitcoin related event or situation, as well as the sentiment expressed towards such entity. Aspect-based sentiment analysis can be divided into two main tasks, which are feature extraction and sentiment classification. (Liu et al., 2012) Each of these stages will be thoroughly discussed in separate paragraphs.

Feature extraction

During the first phase of feature extraction, aspect-based sentiment analysis relies on the feature engineering, which is an essential process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data. (Chandrakala & Sindhu, 2012) It enables to identify aspects of the entity and is also known as information extraction task. For example, in the tweet stating 'Bitcoin Miami conference was awesome', the aspect is represented by 'Miami conference', while 'awesome' is the opinion word towards this aspect.

In order to conduct feature extraction process, it is inevitable to detect the dominant aspects of entity in a specific domain. There are four primary methods for feature determination, which are: 1) extraction by frequent nouns and noun phrases, 2) extraction by opinions and aspects relation, 3) extraction by supervised methods, and 4) extraction by topic modeling. (Li & Li, 2015)

As depicted by Liu (2012), in recent years, statistical topic models have emerged as a principled method for discovering topics from a large collection of text documents. Therefore, the later section will exclusively focus on introducing topic modeling as a chosen method for feature extraction for this paper.

Topic modeling

Topic modeling "is an unsupervised learning method that assumes each document consists of a mixture of topics and each topic is a probability distribution over words". (Liu, 2015) The main

purpose of topic modeling is to uncover latent variables that govern the semantics of a document, these latent variables representing abstract topics. Consequently, the output of topic modeling is a set of word clusters, where each cluster forms a topic that is a probability distribution over words in the document collection. (Liu, 2015)

There are two main statistical topic models, namely, probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). (Blei & Jordan, 2003) As opposed to pLSA, LDA does not require time consuming data training to estimate model parameters and thus is more scalable. Moreover, LDA is the most popular technique for topic modeling and is effectively implemented on variety of documents, such as collections of news articles, policy documents, social media posts or tweets. (Liu, 2015) Considering the above, LDA model has been used in this study to find relevant Bitcoin related topics discussed by Twitter users.

Sentiment Classification

After identifying the meaningful topics, one should proceed to the second phase of aspect-based sentiment analysis, which is a sentiment classification. The process of sentiment classification broadly refers to binary or multi-class categorization that includes two important tasks, namely sentiment polarity assignment and sentiment intensity assignment. (Chandrakala & Sindhu, 2012) The former task determine whether the expressed opinion on different aspects are positive, negative or neutral, whilst the later task includes the analysis whether the positive or negative sentiments are mild or strong. As already discussed in the previous subsection, various tools and techniques could be undertaken to conduct meaningful Sentiment Classification. However, based on the aforementioned arguments from previous section, it was decided to apply VADER as the best sentiment tool.

4.8 Quality of research

Assessing research quality is an essential task while conducting scientific papers, especially the ones with quantitative methods. (Veal, 1997) There are two main aspects that need to be taken into consideration to evaluate research quality, namely validity and reliability. Both of them not only refers to the quality of the collected data but also to the quality of conducted study. In general, validity is defined as “the extent to which the information presented in the research truly reflects the phenomena which researcher claims it reflects”. (Veal, 1997)

There are two main types of validity: internal validity and external validity. Internal validity aims to measure how accurate the phenomena under consideration is represented by the chosen measures and collected data. (Veal, 1997) Therefore, to ensure high level of accuracy, the researchers extract data exclusively from trusted and academic, peer reviewed sources. Moreover, the data collection process was manually performed by the researchers and then double-checked to avoid any errors.

External validity focus on generalisability or representativeness, which means to what extent findings can be applied beyond the specific case. Considering that this paper is a single case study, as it investigates the price volatility of one cryptocurrency, applied methods do not seek to produce findings which are generally representative. (Veal, 1997) Nevertheless, sufficient relationship could be established between the case study and applied theories. Hence, this paper may not yield in generalisations about the population, but it may portray valid insights in relation to the theory.

Similarly to validity, one can also distinguish between internal and external reliability. Internal reliability indicates whether there is a consistency throughout all stages of the paper, whilst external reliability refers to the reproducibility of the findings. (Saunders et al., 2007). According to Veal (1997), a reproducible study means that undertaken data collection techniques and analysis methods will yield the same results if they were repeated by the same researchers at different occasion or by someone else.

In this paper, the internal reliability was accomplished by involving two researchers in data gathering and analysis. On the other hand, external reliability was ensured by setting special parameter, called ‘random state’, which guarantees that all of the random numbers, used in the analytical models, are generated in the same order.

5. Analysis & Results

The following section introduces the implementations of the models as well as the corresponding results. Figure 19 displays the visual illustration of the process undertaken during analysis.

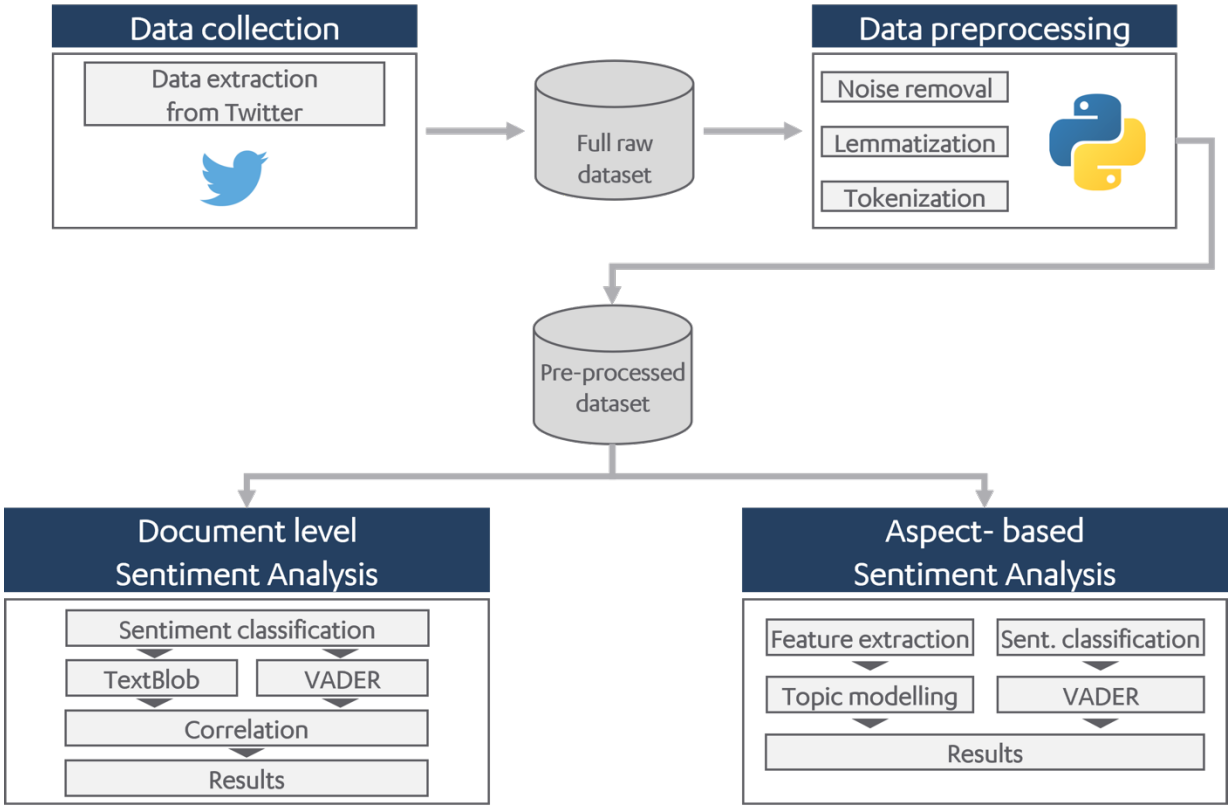

Figure 19 - Data Analysis diagram

5.1 Document Level Sentiment Analysis

As discussed in the methodology section, the first data analysis method, undertaken to address the research question, refers to the sentiment analysis conducted at the document level. It is well-worth to recall that for the purpose of this study a document is defined as a content of a single tweet. As a consequence of sentiment analysis application, each tweet will be labeled as positive, negative or neutral according to its semantic orientation.

Since Twitter data is known for its lack of structure and its high levels of noise, prior to sentiment analysis implementation, collected tweets were extensively preprocessed, as described in the former chapter. The preprocessing was conducted with the use of Natural Language Toolkit (NLTK), which is known as one of the most powerful Python libraries that supports text preprocessing for sentiment analysis application. (nltk.org)

Finally, all unnecessary columns have been removed, which results in obtaining the final dataset with 2 columns and 2,195,405 rows. Remaining columns, which are 'Date' and 'content_tokenize' refer to the date of the Twitter post and actual tweet text after tokenization. Figure 20 illustrate first 10 rows of the final dataset.

```
In [164]: #Remove unnecessary columns
          df_3 = df_final[['Date', 'content_tokenize']]
          df_3.head(10)
```

Out[164]:

	Date	content_tokenize
0	2021-06-23	fooo not underwater fooo entire net worth marc...
1	2021-06-23	jejudoge cmc listing live currently voted comm...
2	2021-06-23	people say bitcoin myspace something gon na co...
3	2021-06-23	cryptogodjohn btc bullish amp centricrise bull...
4	2021-06-23	hex btc eth annual charts might sign come sinc...
5	2021-06-23	fun facts know tell friends based sales millio...
6	2021-06-23	inflation btc ada latter staking risk free way...
7	2021-06-23	metx bullish cross day close low bullish rever...
8	2021-06-23	otc bitcoin first bought jejudoge ethereum jun...
9	2021-06-23	rip officialmcafee crying face thank keeping u...

Figure 20

Among various sentiment analysis algorithms, the researchers have selected TextBlob and VADER as the most relevant tools to identify the emotional tone behind a body of text. Both libraries are built upon Natural Language ToolKit (NLTK) to compute Natural Language Processing (NLP) tasks. It is important to notice that since VADER effectively handles emoji detection, emojis were not substituted with the corresponding textual meaning, as it was a case for TextBlob. Hence, emojis remain untouched during the preprocessing for VADER.

The subsequent two paragraphs will present the implementation of TextBlob and VADER as well as their corresponding results.

5.1.1TextBlob Analysis & Results

As a starting point, it is essential to import TextBlob liabrary with the following code:

```
In [4]: from textblob import TextBlob
```

Figure 21

The algorithm used by TextBlob provides the sentiment scores in the form of polarity and subjectivity that were discussed in the previous section. However, for the purpose of this study, subjectivity score is not considered relevant to address the research question since the researchers are interested in the emotions of the given tweets rather than assessing whether the tweet is subjective or objective. Consequently, it was decided to solely focus on the polarity score that could be achieved by means of the presented function:

```
In [116]: #create a function to get the polarity
          def getPolarity(text):
              return TextBlob(text).sentiment.polarity
```

Figure 22

Once the function is generated, it can now be applied to the preprocessed text contained in the column ‘content_tokenize’. With the use of NLTK library, TextBlob is able to assign an individual polarity score to each of the word within the single tweet. Following that, by means of averaging technique, it calculates the pooled polarity score for the whole Tweet content. In this way each row is assigned a score that lies in the range from -1 to 1. Computed score will be stored in the newly created column, called ‘Polarity’.

```
In [166]: # Create new 'Polarity' column and apply the getPolarity function
          df_3['Polarity'] = df_3['content_tokenize'].apply(getPolarity)
```

Figure 23

```
In [175]: df_3.head(10)
```

Out[175]:

	Date	content_tokenize	Polarity
0	2021-06-23	fooo not underwater fooo entire net worth marc...	0.066667
1	2021-06-23	jejudoge cmc listing live currently voted comm...	0.068182
2	2021-06-23	people say bitcoin myspace something gon na co...	0.285714
3	2021-06-23	cryptogodjohn btc bullish amp centricrise bull...	-0.050000
4	2021-06-23	hex btc eth annual charts might sign come sinc...	0.222222
5	2021-06-23	fun facts know tell friends based sales millio...	0.300000
6	2021-06-23	inflation btc ada latter staking risk free way...	0.150000
7	2021-06-23	metx bullish cross day close low bullish rever...	0.000000
8	2021-06-23	otc bitcoin first bought jejudoge ethereum jun...	0.159091
9	2021-06-23	rip officialmcafee crying face thank keeping u...	0.100000

Figure 24

In order to assess the sentiment orientation of each tweet, *getAnalysis* function has been created and applied on the 'Polarity' column. As indicated from below, a tweet is considered as positive when the polarity score is higher than 0, negative when it is lower than 0 and finally neutral when it is equal 0.

```
In [179]: #Create a function to compute negative, neutral and positive analysis
def getAnalysis(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'

df_3['Analysis'] = df_3['Polarity'].apply(getAnalysis)
```

Figure 25

```
In [180]: df_3.head(5)
```

Out[180]:

	Date	content_tokenize	Polarity	Analysis
0	2021-06-23	fooo not underwater fooo entire net worth marc...	0.066667	Positive
1	2021-06-23	jejudoge cmc listing live currently voted comm...	0.068182	Positive
2	2021-06-23	people say bitcoin myspace something gon na co...	0.285714	Positive
3	2021-06-23	cryptogodjohn btc bullish amp centricrise bull...	-0.050000	Negative
4	2021-06-23	hex btc eth annual charts might sign come sinc...	0.222222	Positive

Figure 26

After the successful classification of tweets' sentiment, the whole dataset will be grouped by date and analysis. As a result, the newly created dataset will provide the information regarding how many negative, neutral and positive tweets are there per each day.

```
In [181]: df_group = df_3.groupby(['Date','Analysis'], as_index = False).count()
```

Figure 27

```
In [189]: df_group.head(5)
```

Out[189]:

	Date	Analysis	Count
0	2017-06-24	Negative	44
1	2017-06-24	Neutral	167
2	2017-06-24	Positive	126
3	2017-06-25	Negative	44
4	2017-06-25	Neutral	189

Figure 28

At this point, there are three polarities with their corresponding value counts for each day. Therefore, to gain an overall daily sentiment, the researchers check which polarity count value is the highest during the given day. Consequently, as depicted by the graph below, there are 1325 days where the dominant polarity is positive and 136 when the dominant polarity is neutral. Surprisingly, none of the day has been assigned a negative sentiment.

```
In [197]: items_count_2 = df_group_max['Analysis'].value_counts()
          print(items_count_2)

          Positive    1325
          Neutral      136
          Name: Analysis, dtype: int64
```

Figure 29

```
In [199]: values = [1325, 136, 0]
          names = ['Positive', 'Neutral', 'Negative']

          plt.bar(names, values)
          plt.show()
```

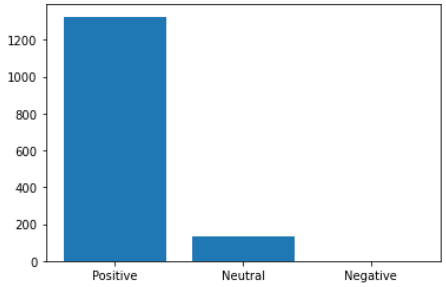

Figure 30

The above results demonstrate an overview of the prevailed daily sentiment of the Bitcoin related Tweets during the 4 years period. Even though the vast majority of the tweets occur to be positive, the researchers still endeavor to obtain the insight whether there is a correlation between polarity score and Bitcoin price. Therefore a new data frame has been created, which includes tweets grouped by date and the polarity score mean for each day.

```
In [201]: df_4 = df_3.groupby(['Date'], as_index = False).mean()
          df_4.head(5)
```

Out[201]:

	Date	Polarity
0	2017-06-24	0.073669
1	2017-06-25	0.087462
2	2017-06-26	0.060505
3	2017-06-27	0.086527
4	2017-06-28	0.104599

Figure 31

Following that, the sentiment orientation has been assigned to each day based on the average polarity score. Next, the Bitcoin price has been added to the data frame and the last step refers to the computation of two charts, one illustrating the Polarity over time and the second one demonstrating Bitcoin Price over time. The aim of the following graphs is to determine whether the change in those two variables follow similar pattern.

```
In [214]: btc.plot(x='Date', y='Polarity')
Out[214]: <AxesSubplot:xlabel='Date'>
```

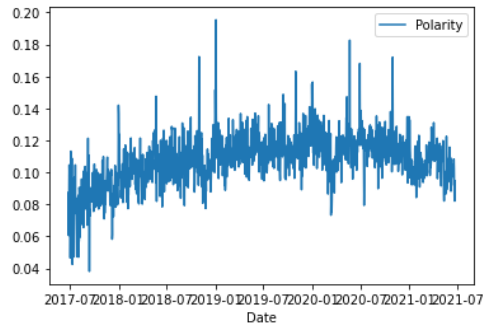

Figure 32

```
In [216]: btc.plot(x='Date', y='Price')
Out[216]: <AxesSubplot:xlabel='Date'>
```

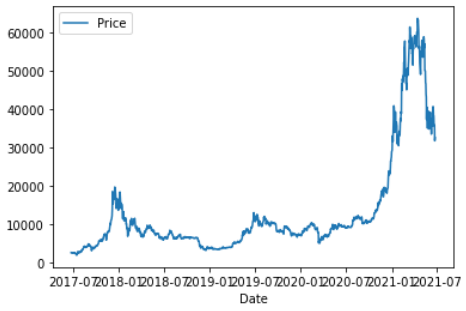

Figure 33

Looking at the graph 1, it could be noticed that there is an overall increasing trend in Polarity with several significant peaks. However, while compering to the graph 2, it cannot be stated that the price change follow the same pattern as the polarity change over the given period of time.

Furthermore, result from the graphs was supplemented by a calculation of the correlation score since as depicted by Kumar (2015), correlation analysis is dedicated to discover the relationship between two aspects of a situation.

```
In [228]: # correlation
          data = btc[['Polarity', 'Price']]
          correlation = data.corr(method='pearson')
          print(correlation)

                    Polarity     Price
          Polarity  1.000000  0.015837
          Price     0.015837  1.000000
```

Figure 34

Presented table shows that the correlation between price and polarity is 0,0158. Since the achieved value is above 0, the correlation is positive. Nevertheless, it is considered eminently low.

5.1.2 VADER Analysis & Results

This section will describe the process of conducting the sentiment analysis using the VADER package in Python and it will present the corresponding results.

The VADER sentiment analysis is conducted similarly to the one in TextBlob, with small differences in the code used, which is dependent on the Python package. The main dissimilarity in the analysis is how the packages work in identifying the sentiment, as described in the the Methodology section.

The process starts by loading the necessary packages and the already preprocessed data frame. Then, the Sentiment Intensity Analyzer from VADER is being loaded into a variable called “analyzer”. Using this analyzer, four new columns will be added into the already existing data frame: one column for the negative score, one for positive, one for neutral and one for the compound score. The first three columns, as the name suggests, represent the scores attributed to each sentiment category, in terms of proportion. The compound column is a metric based on the three scores.

```
In [42]: analyzer = SentimentIntensityAnalyzer()

In [ ]: df_final['neg'] = df_final['tweets_w_ngrams'].apply(lambda x:analyzer.polarity_scores(x)['neg'])
        df_final['neu'] = df_final['tweets_w_ngrams'].apply(lambda x:analyzer.polarity_scores(x)['neu'])
        df_final['pos'] = df_final['tweets_w_ngrams'].apply(lambda x:analyzer.polarity_scores(x)['pos'])
        df_final['compound'] = df_final['tweets_w_ngrams'].apply(lambda x:analyzer.polarity_scores(x)['compound'])
```

Figure 35

```
In [5]: df_final.head(10)
```

Out[5]:

	content	tweets_w_ngrams	compound	neg	neu	pos
0	fooo isn't underwater\n\nfooo's entire net wor...	fooo not underwater fooo entire net worth marc...	0.5859	0.000	0.827	0.173
1	Jejudoge CMC listing is Live! Currently Voted ...	jejudoge cmc listing live currently voted comm...	0.5994	0.000	0.860	0.140
2	When people say “bitcoin is MySpace, something...	people say_bitcoin myspace something gon_na co...	0.0000	0.000	1.000	0.000
3	@CryptoGodJohn #BTC Bullish & @CentricRise...	cryptogodjohn btc bullish amp centricrise bull...	0.3400	0.000	0.876	0.124
4	HEX, BTC, ETH annual charts. B and E might be ...	hex btc_eth annual charts might sign come sinc...	0.7506	0.057	0.652	0.291
5	Fun facts to know and tell your friends: based...	fun facts know tell friends based sales millio...	0.7506	0.000	0.758	0.242
6	Inflation in both BTC and $ADA is ~2%, but in ...	inflation btc ada latter staking risk free way...	0.6249	0.162	0.588	0.249
7	$METX .98 Bullish 8/21 MA cross Day 1 Close \n...	metx bullish cross day close low bullish rever...	0.4215	0.062	0.794	0.144
8	@OTC_Bitcoin I first bought $jejudoge with Eth...	otc bitcoin_first bought jejudoge ethereum jun...	0.6369	0.000	0.794	0.206
9	RIP @officialmcafee 😢 Thank you for keeping us...	rip officialmcafee thank keeping us safe cyber...	0.7783	0.000	0.536	0.464

Figure 36

As it can be seen in Figure 36 after applying the Sentiment Intensity Analyzer on the data frame, each tweet has been attributed a negative, positive, neutral and a compound score.

A similar step included also for TextBlob analysis is defining a function which will determine the sentiment label based on the compound score. The function is identical to the one presented previously and the results of its application can be seen in Figure 37 below.

```
In [5]: df_final.head(10)
```

Out[5]:

	content	tweets_w_ngrams	compound	neg	neu	pos	Sentiment
0	fooo isn't underwater\n\nfooo's entire net wor...	fooo not underwater fooo entire net worth marc...	0.5859	0.000	0.827	0.173	Positive
1	Jejudoge CMC listing is Live! Currently Voted ...	jejudoge cmc listing live currently voted comm...	0.5994	0.000	0.860	0.140	Positive
2	When people say “bitcoin is MySpace, something...	people say_bitcoin myspace something gon_na co...	0.0000	0.000	1.000	0.000	Neutral
3	@CryptoGodJohn #BTC Bullish & @CentricRise...	cryptogodjohn btc bullish amp centricrise bull...	0.3400	0.000	0.876	0.124	Positive
4	HEX, BTC, ETH annual charts. B and E might be ...	hex btc_eth annual charts might sign come sinc...	0.7506	0.057	0.652	0.291	Positive
5	Fun facts to know and tell your friends: based...	fun facts know tell friends based sales millio...	0.7506	0.000	0.758	0.242	Positive
6	Inflation in both BTC and $ADA is ~2%, but in ...	inflation btc ada latter staking risk free way...	0.6249	0.162	0.588	0.249	Positive
7	$METX .98 Bullish 8/21 MA cross Day 1 Close \n...	metx bullish cross day close low bullish rever...	0.4215	0.062	0.794	0.144	Positive
8	@OTC_Bitcoin I first bought $jejudoge with Eth...	otc bitcoin_first bought jejudoge ethereum jun...	0.6369	0.000	0.794	0.206	Positive
9	RIP @officialmcafee 😢 Thank you for keeping us...	rip officialmcafee thank keeping us safe cyber...	0.7783	0.000	0.536	0.464	Positive

Figure 37

In order to be able to draw some insights and results from the model, it was necessary to aggregate the tweets. Therefore, these have been grouped by using the Date and Sentiment column and count as an aggregation function. The results shown in Figure 38 display the number of positive, negative and neutral tweets in each day.

```
In [ ]: df_4 = df_final.groupby(['Date', 'Sentiment'], as_index = False).count()

In [17]: df_4.head(10)
```

Out[17]:

	Date	Sentiment	Count
0	2017-06-24	Negative	46
1	2017-06-24	Neutral	134
2	2017-06-24	Positive	157
3	2017-06-25	Negative	58
4	2017-06-25	Neutral	167
5	2017-06-25	Positive	150
6	2017-06-26	Negative	104
7	2017-06-26	Neutral	217
8	2017-06-26	Positive	197
9	2017-06-27	Negative	104

Figure 38

Moreover, only the sentiment with the highest count has been selected for each day, as this would become representative for the general sentiment displayed in that day.

```
In [132]: df_5.head(8)
```

Out[132]:

	Date	Sentiment	Count
0	2017-06-24	Positive	157
1	2017-06-25	Neutral	167
2	2017-06-26	Neutral	217
3	2017-06-27	Neutral	204
4	2017-06-28	Neutral	205
5	2017-06-29	Neutral	218
6	2017-06-30	Neutral	207
7	2017-07-01	Neutral	178

Figure 39

The graph below shows the distribution of sentiments among the 1481 days. As it can be clearly seen, the results are similar to the ones obtained using TextBlob, with only a slight difference between them.

```
In [81]: items_count_2 = df_5['Sentiment'].value_counts()
         print(items_count_2)

         Positive    1357
         Neutral      103
         Negative       1
         Name: Sentiment, dtype: int64

In [83]: values = [1357, 103, 1]
         names = ['Positive', 'Neutral', 'Negative']

         plt.bar(names, values)
         plt.show()
```

Figure 40

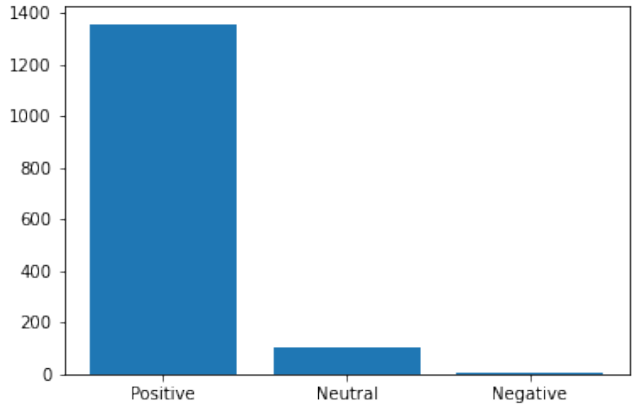

Figure 41

Additionally, a correlation score can be calculated to gain better insight into the relationship between the Bitcoin price volatility and the dominant sentiment on Twitter. First, the tweets have been grouped by again, this time using mean as an aggregation function. The tweets have been grouped by day and the compound score is now an average of all the scores in that day.

```
In [ ]: df_3 = df_final.groupby(['Date'], as_index = False).mean()

In [140]: df_3.head(8)
```

Out[140]:

	Date	compound	neg	neu	pos	Sentiment
0	2017-06-24	0.151061	0.052970	0.796344	0.150688	Positive
1	2017-06-25	0.106899	0.051709	0.816043	0.132253	Positive
2	2017-06-26	0.061696	0.082112	0.792736	0.125158	Positive
3	2017-06-27	0.068359	0.071179	0.808489	0.120326	Positive
4	2017-06-28	0.083228	0.064500	0.806732	0.128768	Positive
5	2017-06-29	0.109159	0.052272	0.821198	0.126532	Positive
6	2017-06-30	0.077838	0.072006	0.800446	0.127544	Positive
7	2017-07-01	0.109515	0.062652	0.807340	0.130019	Positive

Figure 42

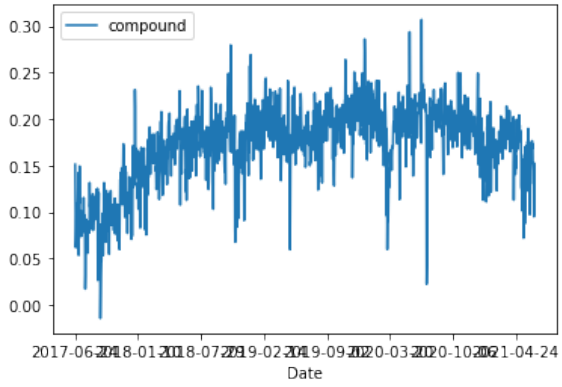

Figure 43

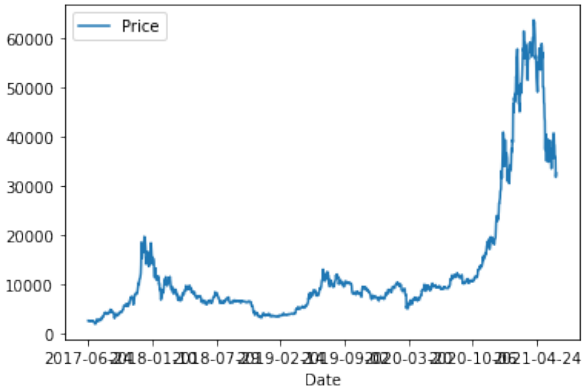

Figure 44

The two graphs above show the change over time in the compound score and Bitcoin price. Already by looking at the graphs it can be observed that there does not seem to be any similar

pattern between the two variables that could indicate a significant relationship. To verify this, the correlation score has been calculated below:

```
In [111]: # Correlation
          data = btc[['compound', 'Price']]
          correlation = data.corr(method='pearson')
          correlation
```

Out[111]:

	compound	Price
compound	1.000000	0.024253
Price	0.024253	1.000000

Figure 45

The correlation between the Bitcoin price and the compound is 0.024, which is very close to 0. The low correlation score implies that there is no relationship between the two variables, the price and the compound. This is in line with the results of the previous sentiment analysis conducted using TextBlob, where a similar low but negative score has been identified.

Considering that the results obtained from the document level sentiment do not yield significant correlation, the researchers agree to undertake different approach that focus on sentiment analysis on the entity and aspect level, which is further referred to as an aspect-based sentiment analysis.

5.2 Aspect-Based Sentiment Analysis

While reviewing social media content, it is noticeable that Twitter society not only use Bitcoin hashtag while referring to the Bitcoin related situation, but also while expressing general thoughts regarding other cryptocurrencies. As a result, in the day associated to the relevant Bitcoin situation, the aggregated daily polarity score may be not accurate, since it includes bias from Twitter users posting common thoughts under #bitcoin, which was used as the key phrase when extracting Twitter data. Therefore, to tackle this constrain, it was decided to apply aspect-based sentiment analysis.

In contrast to document level sentiment analysis, which extracts the general sentiment expressed in each tweet, aspect-based sentiment analysis focus on determining both the entity described in the text as well as the sentiment articulated towards that entity. For the purpose of

this paper, an entity is defined as any Bitcoin related event, situation or attribute. Hence, the model help to group tweets by the most popular topics and assign sentiment orientation to each of them.

Prior to the implementation of the aspect-based sentiment analysis, it is essential to recall that the analysis is performed on the subset of collected data, consisting of one year of Twitter posts, from June 2020 to June 2021. The reason for decreasing the number of tweets from the initial dataset is driven by the fact that the topic modelling, which is part of the aspect-based sentiment analysis, requires not only more time but also more computational power. Moreover, the process of topic modeling is based on the in-depth analysis. Therefore, in order to increase credibility of the results, it was decided to conduct 12 separate aspect-based sentiment analysis for each month. The following section will exclusively portray the results from June 2021, as they were considered the most interesting and adequate to present.

As discussed in the methodology section, text aspects were extracted by the means of topic modeling techniques, whilst sentiment classification was performed with the use of VADER.

5.2.1 Topic modelling

Bigrams

After the cleaning and preprocessing of the raw data extracted from Twitter, the researches have decided to implement n-gram model to increase the accuracy of the topic modeling. More specifically, the process focus on determining 2-word phrases, known as bigrams that have been frequently occurring together. Once such complementary words have been identified, they were further concatenated, so that topic modeling algorithm can perceive such a bigram as a one word. Example of the most relevant bigrams are depicted in figure below.

	bigram	raw_freq
0	(el, salvador)	0.001746
1	(elon, musk)	0.001062
2	(legal, tender)	0.000850
3	(michael, saylor)	0.000782
4	(let, us)	0.000743
5	(long, term)	0.000570
6	(market, cap)	0.000464
7	(looks, like)	0.000441
8	(bull, market)	0.000381
9	(bear, market)	0.000359

Figure 46

As indicated in Figure 46, the first four bigrams refers to 1) 'El Salvador', which is a country in Central America that has recently adopted Bitcoin as a legal tender, 2) 'Elon Musk', who is the CEO of SpaceX and Tesla, and is the active Twitter user especially within cryptocurrency field, 3)'legal tender', which is a legally recognized payment instrument, that could also refer to the recent announcement of El Salvador president, 4) 'Michael Saylor', who is the co-founder of MicroStrategy, that is constantly increasing its Bitcoin holdings.

Based on the above, construction of the bigrams seems sufficient for the purpose of this paper. Thus, with the use of *genism* package, 2-word phrases have been implemented into the preprocessed tweets, which led to the final dataset demonstrated in Figure 47.

Out[295]:

	Date	content
0	2021-06-23	fooo underwater fooo entire net worth march fo...
1	2021-06-23	jejudoge cmc listing live currently voted comm...
2	2021-06-23	say myspace something come facebook theyre rig...
3	2021-06-23	cryptogodjohn centricrise cnr cns long_term ...
4	2021-06-23	annual charts might sign whats come since be...
5	2021-06-23	fun facts tell friends based sales million ip...
6	2021-06-23	inflation latter staking risk free way offset ...
7	2021-06-23	metx cross day close reversal tail bottom p...
8	2021-06-23	otc first bought jejudoge june_th exact today ...
9	2021-06-23	rip officialmcafee thank keeping us safe cyber...

Figure 47

The next stage will incorporate the identification of dominant topics that Twitter users talk about with respect to Bitcoin. As argued in the methodology section, topic modeling was performed by means of LatentDirichletAllocation. The LDA model could be implemented through *genism* library and *sklearn* library. In order to increase accuracy, the researchers have decided to undertake both approaches.

LDA with Gensim

After successful installation of genism library as well as all necessary packages, the researchers need to specify several crucial parameters for the LDA model.

There are two main inputs, namely dictionary and corpus. Dictionary contains the mapping of all words from preprocessed Tweets to their unique word ID and is used to determine the size of the vocabulary. On the other hand, corpus consist of abovementioned word ID and its frequency in each document. Hence, by the use of dedicated codes (Figure 48), preprocessed Twitter text has been represented as a bag of its words, known as Bag of words (BoW).

```
In [277]: # dictionary
          dictionary1 = corpora.Dictionary(final_tweets)

In [278]: # corpus
          doc_term_matrix = [dictionary1.doc2bow(doc) for doc in final_tweets]
```

Figure 48

In addition to corpus and dictionary, the researchers has specified the subsequent parameters:

30 passes through the corpus during the training, 200 iterations through the corpus and 10 000 documents to be used in each training chunk. Moreover, to assure the reproducibility of results, a random state of 0 has been established. Finally, the number of topics was determined by computing the coherence score for each topic within the range of [2,11]. It means that the LDA model was performed 9 times by the use of the following code:

```
In [89]:
coherence = []
for k in range(2,11):
    print('Round: '+str(k))
    Lda = gensim.models.ldamodel.LdaModel
    ldamodel = Lda(doc_term_matrix, num_topics=k, id2word = dictionary1, passes=30,\
                   iterations=200, chunksize = 10000, eval_every = None, random_state=0)

    cm = gensim.models.coherencemodel.CoherenceModel(model=ldamodel, texts=final_tweets,\
                                                     dictionary=dictionary1, coherence='c_v')
    coherence.append((k,cm.get_coherence()))
```

```
Round: 2
Round: 3
Round: 4
Round: 5
Round: 6
Round: 7
Round: 8
Round: 9
Round: 10
```

Figure 49

The results outlined in Figure 50, indicate that the coherence score keep increasing with the number of topics and it reaches a peak at 8 topics with the score value of 0,44. Therefore, it is reasonable to pick the LDA model that yields the highest coherence score.

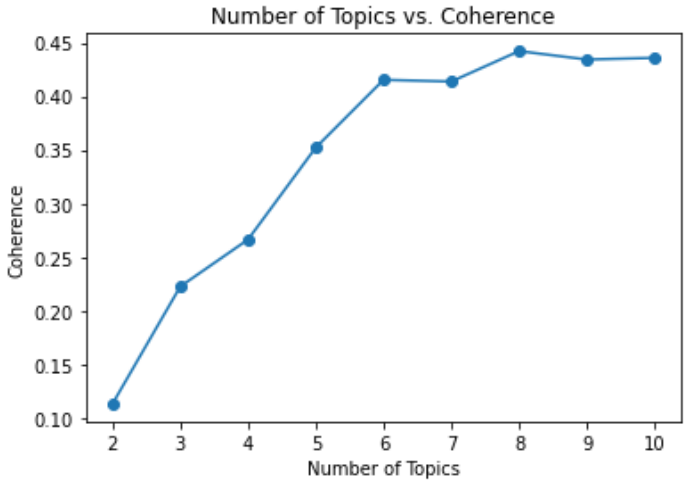

Figure 50

Consequently, the final topic modeling by means of LDA from genism package was performed to gain 8 dominan. Each topic is a combination of keywords and each keyword contributes a certain weightage to the topic. Figure 51 demonstrate 8 topics with their 7 corresponding keywords.

```
In [281]: # To show initial topics
          ldamodel.show_topics(8, num_words=7, formatted=False)

Out[281]: [(0,
            [('nano', 0.011192198),
             ('sats', 0.009568444),
             ('twitter', 0.00931558),
             ('bch', 0.008128903),
             ('jack', 0.0073969862),
             ('jeju', 0.0069150575),
             ('triangle', 0.0066903825)]),
           (1,
            [('live', 0.009556936),
             ('follow', 0.0066367793),
             ('join', 0.006418532),
             ('team', 0.0060746176),
             ('win', 0.005873548),
             ('miami', 0.0058111437),
             ('today', 0.005486306)]),
           (2,
            [('el_salvador', 0.028531365),
             ('legal_tender', 0.010241227),
             ('country', 0.00808361),
             ('currency', 0.0075034066),
             ('world', 0.006460301),
             ('president', 0.0064366483),
             ('million', 0.0064118877)]),
           (3,
            [('support', 0.007293147),
             ('today', 0.0066914074),
             ('daily', 0.006393181),
             ('may', 0.006293728),
             ('long', 0.00616451),
             ('looking', 0.0061340677),
             ('profit', 0.005896302)]),
           (4,
            [('mining', 0.014717756),
             ('energy', 0.008354036),
             ('china', 0.0074881),
             ('use', 0.0074820607),
             ('network', 0.0067448695),
             ('world', 0.0059789736),
             ('power', 0.005558105)]),
           (5,
            [('think', 0.009284709),
             ('conference', 0.0064643756),
             ('go', 0.006039259),
             ('want', 0.005638712),
             ('never', 0.005557733),
             ('really', 0.0054936013),
             ('much', 0.005309935)]),
           (6,
            [('sec', 0.008374533),
             ('thebitcoinconf', 0.0082947835),
             ('trading', 0.00763005),
             ('theta', 0.0071139056),
             ('cake', 0.0065978062),
             ('iota', 0.006371789),
             ('kucoin', 0.0056100218)]),
           (7,
            [('elon', 0.015385555),
             ('hodl', 0.014999668),
             ('elonmusk', 0.013169219),
             ('safemoon', 0.012286811),
             ('moon', 0.007871539),
             ('altcoin', 0.007229437),
             ('altcoins', 0.0060583185)])]
```

Figure 51

It could be noticed that not all key words seem relevant. Nevertheless, there are still some meaningful results that may contribute to extracting few prominent topics.

In order to gain some visual interpretation, all topics have been displayed in the interactive chart that depicts marginal topic distribution. The larger the circle on the graph the higher the value of the marginal distribution.

The most appealing bubble in Figure 52 includes words such as ‘conference’ and ‘miami’, which clearly refers to the biggest Bitcoin conference held in Miami in June 2021.

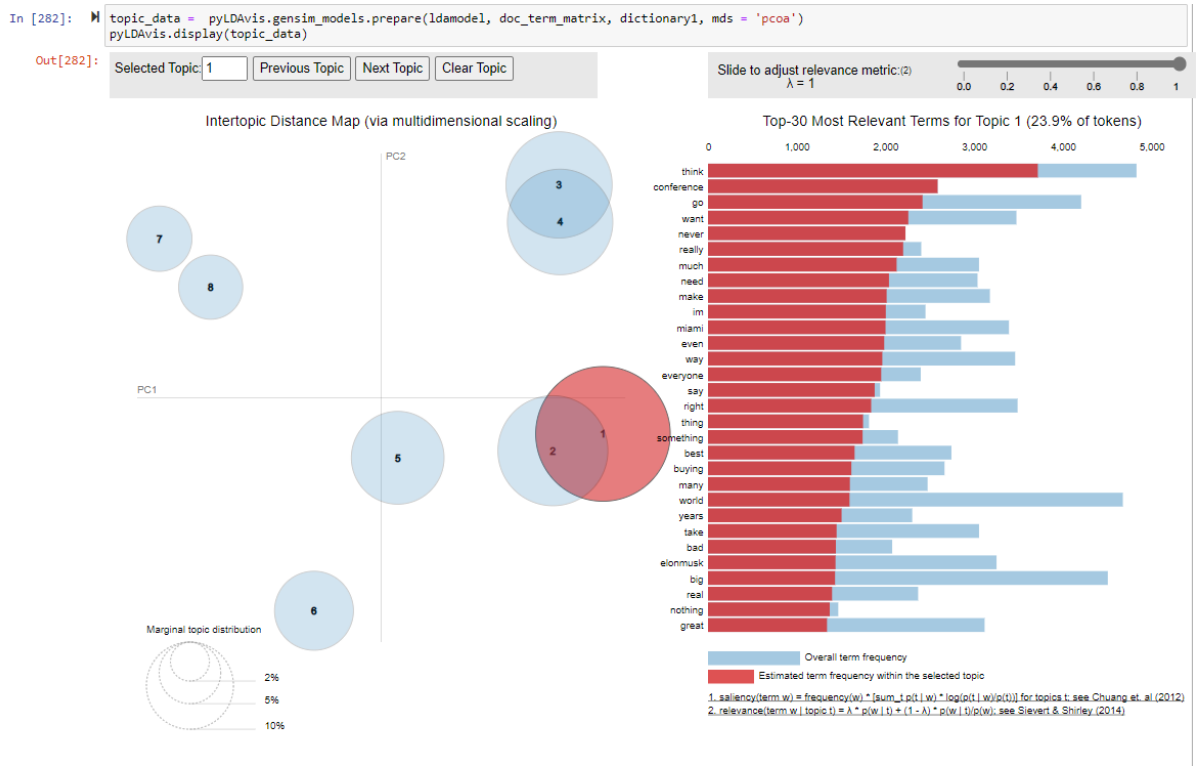

Figure 52

Another topic that is important to mention involve words such as ‘mining’, ‘energy’, ‘china’, ‘power’, ‘miners’ and ‘government’. Hence, there is a strong association with the recent announcement of Chinese government to ban Bitcoin mining.

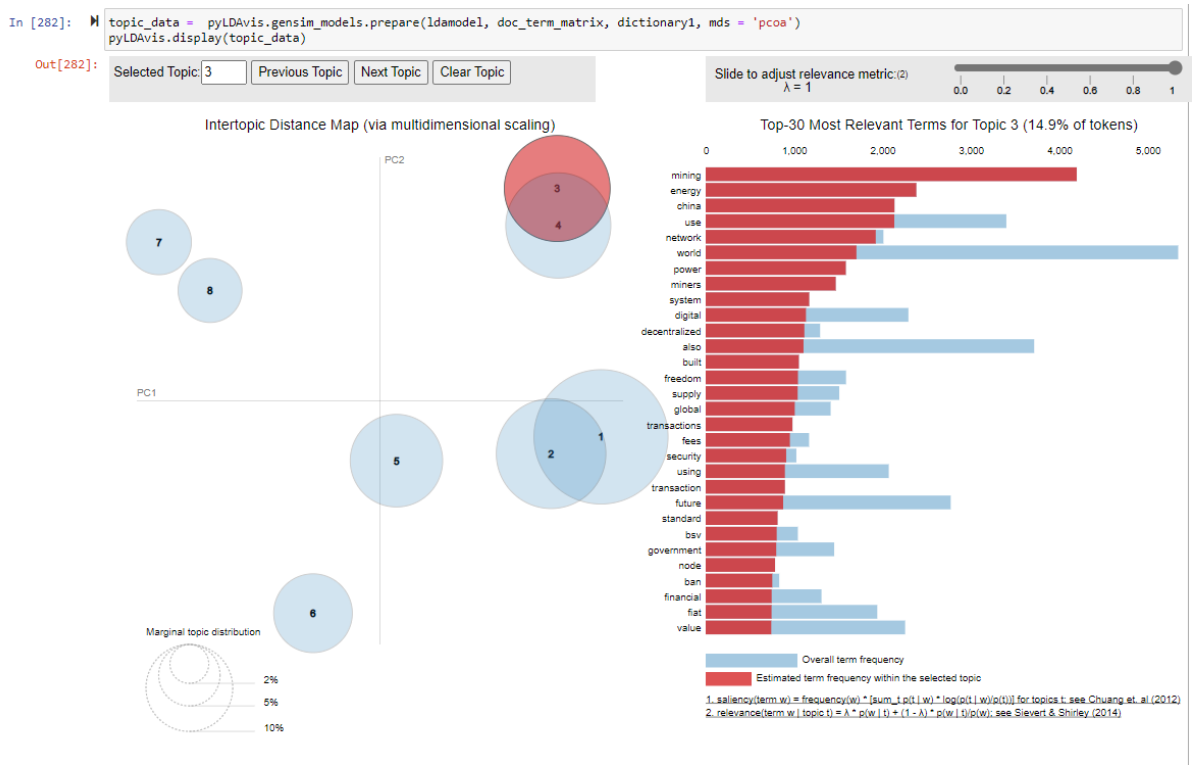

Figure 53

Third prominent topic refers to ‘el_salvador’, ‘legal_tender’, ‘country’, ‘currency’, ‘president’, ‘nayibbukele’. All of these words point out to the declaration of Nayib Bukele, president of El Salvador, that the Bitcoin will be considered as a legal tender.

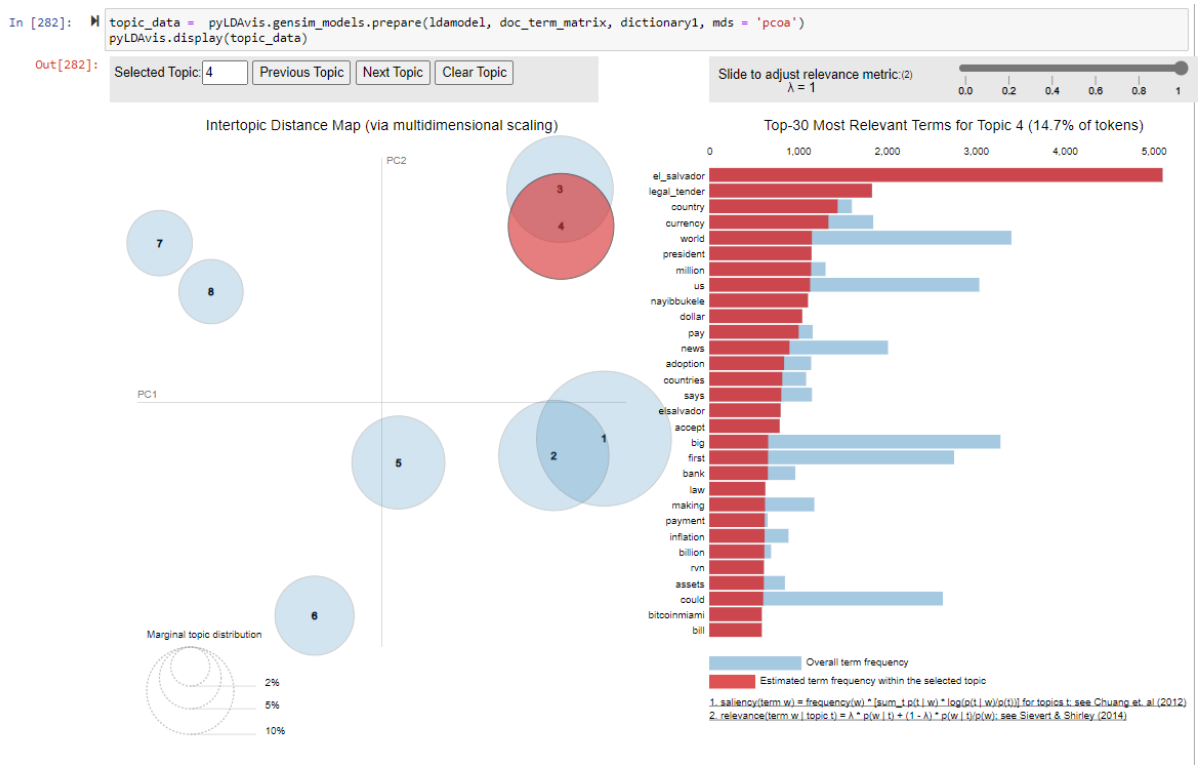

Figure 54

Lastly, it is noteworthy that phrases such as ‘elon’ and ‘elonmusk’ illustrates the influential power of Elon Musk’s Twitter posts regarding Bitcoin.

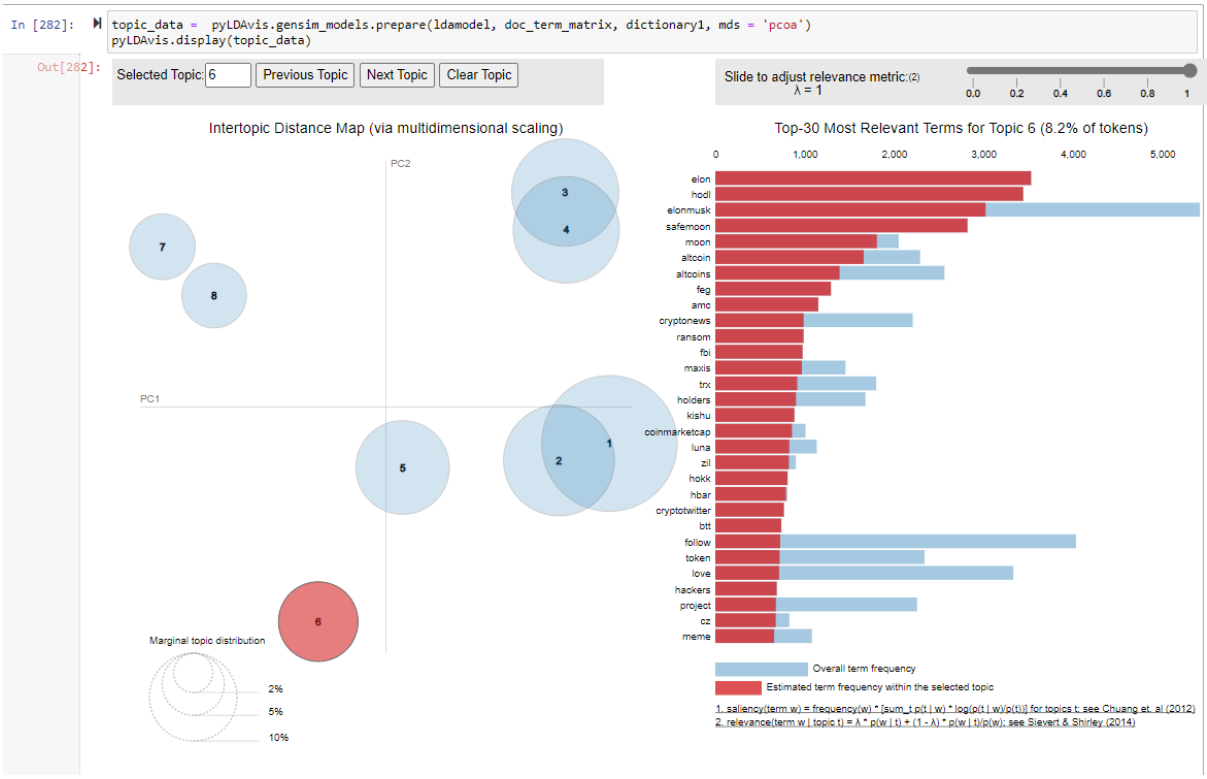

Figure 55

Taking into account the considerable insights derived from proposed result, it could be concluded that there were 4 dominant topics in June 2021:

1) Bitcoin conference in Miami

2) Ban of Bitcoin mining in China

3) Bitcoin as a legal tender in El Salvador

4) Influencing power of Elon Musk

LDA with Sklearn

The second approach to the LDA model will be conducted by means of Sklearn library. The main distinction from the genism example is that the current model incorporates *TF-IDF* as a more sophisticated method for bag of words creation. Previously explored *doc2bow* technique

only describes a document in a standalone fashion without considering the context of the corpus. Nevertheless, TF-IDF takes into account the relative frequency of words in the document against their frequency in other documents. Hence, the process of document vectorization with the Tf-IDF score was performed by the use of *TfidfVectorizer* transformer.

Figure 56 covers all TF-IDF steps.

```
In [241]: from sklearn.feature_extraction.text import TfidfVectorizer
          tf_idf_vectorizer = TfidfVectorizer(tokenizer=lambda doc: doc, lowercase=False)
          tf_idf_arr = tf_idf_vectorizer.fit_transform(clean_corpus)
          vocab_tf_idf = tf_idf_vectorizer.get_feature_names()
```

Figure 56

Next stage refers to the specification of the LDA parameters. For this model, the researchers have defined only 3 features: number of topics being 8, maximum iterations of 30 and 'random_state' equals 0. Other input values remained as default. Once the LDA model has been requested, it could be fit into vectorized version of the text. Consequently, the following 8 dominant topics have been determined:

	Dominant_topic	topic_name
0	Topic3	[miami, conference, youtube, trx, cryptonews, ...
1	Topic4	[million, network, use, cash, wallet, digital,...
2	Topic1	[el_salvador, world, currency, fiat, us, count...
3	Topic2	[mining, support, green, daily, china, miners,...
4	Topic8	[im, never, say, want, really, even, right, ma...
5	Topic7	[elonmusk, tesla, elon_musk, elon, live, tsla,...
6	Topic5	[trx, follow, trading, win, giveaway, airdrop,...
7	Topic6	[think, year, long, since, months, alts, big, ...

Figure 57

Examining the figure above, it is important to perceive that majority of key words, identified with genism library, are also present in the current approach. The most remarkable phrases includes: ‘el_salvador’, ‘mining’, ‘china’, ‘miners’, ‘miami’, ‘conference’, ‘elonmusk’, ‘elon_musk’, ‘tesla’ and ‘elon’.

It could be concluded that both libraries were successful in their capabilities, even though they yield slightly different results. Nevertheless, the main topics remain the same.

Considering that the Sklearn library provides more sophisticated and advanced tools, it was decided to pursue with the output generated by this library. Hence, the achieved topics served as the main aspects for the further process of aspect-based sentiment analysis. Consequently, the following steps endeavor to determine one prominent topic as well as a sentiment score for each tweet.

Having said that, each document has been assigned one dominant topic based on the highest relevance score computed for each topic, as outlined in Figure 58.

	Date	content_freq	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Dominant_topic
0	2021-06-23	fooo underwater fooo entire net worth march fo...	0.03	0.03	0.03	0.03	0.74	0.10	0.03	0.03	5
1	2021-06-23	jejudoge cmc listing live currently voted comm...	0.02	0.02	0.18	0.02	0.39	0.10	0.24	0.02	5
2	2021-06-23	say myspace something come facebook theyre rig...	0.74	0.04	0.04	0.04	0.04	0.04	0.04	0.04	1
3	2021-06-23	cryptogodjohn centricrise cnr cns long_term ...	0.35	0.12	0.38	0.03	0.03	0.03	0.03	0.03	3
4	2021-06-23	annual charts might sign whats come since be...	0.03	0.03	0.03	0.58	0.26	0.03	0.03	0.03	4

Figure 58

5.2.2 Sentiment Classification

The process of detecting sentiment orientation of each Twitter text has been computed with VADER algorithm. More detailed description of the VADER technique was portrayed in the previous result section regarding document level sentiment analysis. However, it is essential to recall that VADER analyzes the documents with the *SentimentIntensityAnalyzer()* function. Subsequently, based on the compound sentiment score, calculated for each tweet, VADER assign positive, negative or neutral label to every document.

The next step incorporates the merger of the topic related dataframe with the one including sentiment scores. The first 10 rows are displayed in the Figure 59.

	Date	content_freq	sentiment score	sentiment	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Dominant_topic
0	2021-06-23	fooo underwater fooo entire net worth march fo...	0.5859	Positive	0.03	0.03	0.03	0.03	0.74	0.10	0.03	0.03	5
1	2021-06-23	jejudoge cmc listing live currently voted comm...	0.5994	Positive	0.02	0.02	0.18	0.02	0.39	0.10	0.24	0.02	5
2	2021-06-23	say myspace something come facebook theyre rig...	0.0000	Neutral	0.74	0.04	0.04	0.04	0.04	0.04	0.04	0.04	1
3	2021-06-23	cryptogodjohn centricrise cnr cns long_term ...	0.3400	Positive	0.35	0.12	0.38	0.03	0.03	0.03	0.03	0.03	3
4	2021-06-23	annual charts might sign whats come since be...	0.7506	Positive	0.03	0.03	0.03	0.58	0.26	0.03	0.03	0.03	4
5	2021-06-23	fun facts tell friends based sales million ip...	0.7506	Positive	0.62	0.03	0.21	0.03	0.03	0.03	0.03	0.03	1
6	2021-06-23	inflation latter staking risk free way offset ...	0.6249	Positive	0.02	0.02	0.02	0.56	0.24	0.02	0.02	0.08	4
7	2021-06-23	metx cross day close reversal tail bottom p...	0.4215	Positive	0.02	0.02	0.02	0.25	0.62	0.02	0.02	0.02	5
8	2021-06-23	otc first bought jejudoge june_th exact today ...	0.6369	Positive	0.03	0.03	0.18	0.03	0.50	0.03	0.18	0.03	5
9	2021-06-23	rip officialmcafee thank keeping us safe cyber...	0.7783	Positive	0.03	0.03	0.03	0.03	0.03	0.03	0.37	0.45	8

Figure 59

Taking into account that not all of the 8 dominant topics are implicating meaningful results, the researchers have chosen 4 with the most relevant keywords. As a consequence, the final analysis includes the following topics:

Dominant_topic	topic_name
Topic1	[el_salvador, world, currency, fiat, us, count...
Topic2	[mining, support, green, daily, china, miners,...
Topic3	[miami, conference, youtube, trx, cryptonews, ...
Topic7	[elonmusk, tesla, elon_musk, elon, live, tsla,...

Figure 60

Finally, by the use of *groupby()* function, all documents have been grouped by the dominant topic and sentiment orientation. Subsequently, the *count()* function has been used to calculate how many tweets there are for particular topic and for each sentiment.

	Dominant_topic	Sentiment	Count
0	Topic1	Negative	24886
1	Topic1	Neutral	16291
2	Topic1	Positive	40814
3	Topic2	Negative	15250
4	Topic2	Neutral	6812
5	Topic2	Positive	10084
6	Topic2	Negative	1885
7	Topic3	Neutral	6117
8	Topic3	Positive	11610
9	Topic7	Negative	2582
10	Topic7	Neutral	6089
11	Topic7	Positive	6892

Figure 61

For simplicity of the results interpretation, the researchers referred to the topics as follows:

Topic 1: ‘El Salvador’

Topic 2: ‘China Mining’

Topic 3: ‘Miami Conference’

Topic 7: ‘Elon Musk’

It could be observed from Figure 61 that the most popular aspect of Bitcoin discussed during June 2021 relates to El Salvador. This topic has been an area of interest for 81 991 Twitter users. More than a half of tweets, which is 40 814, associate positive emotions towards the topic. Nevertheless, there is also quite significant number of negative posts that sum up to 24 886.

Second the most frequent topic, among Twitter users, concerns the Bitcoin mining in China. As opposed to the first topic, the highest value count is now represented by negative expressions included in 22 250 tweets. The positive sentiment is determined in 10 084 posts, which stands for the half of the negative values.

Third topic, with the total number of tweets being 19 612, refers to the Bitcoin conference in Miami. The event mostly contributed to the positive mood among Twitter society.

Lastly, it is noteworthy that Topic 7 does not refer to any Bitcoin event, but rather to the well-known person, Elon Musk.

In order to gain visual insights of the sentiment discribution among chosen topics, Figure 61 has been plotted in form of a bar graph, shown in Figure 62.

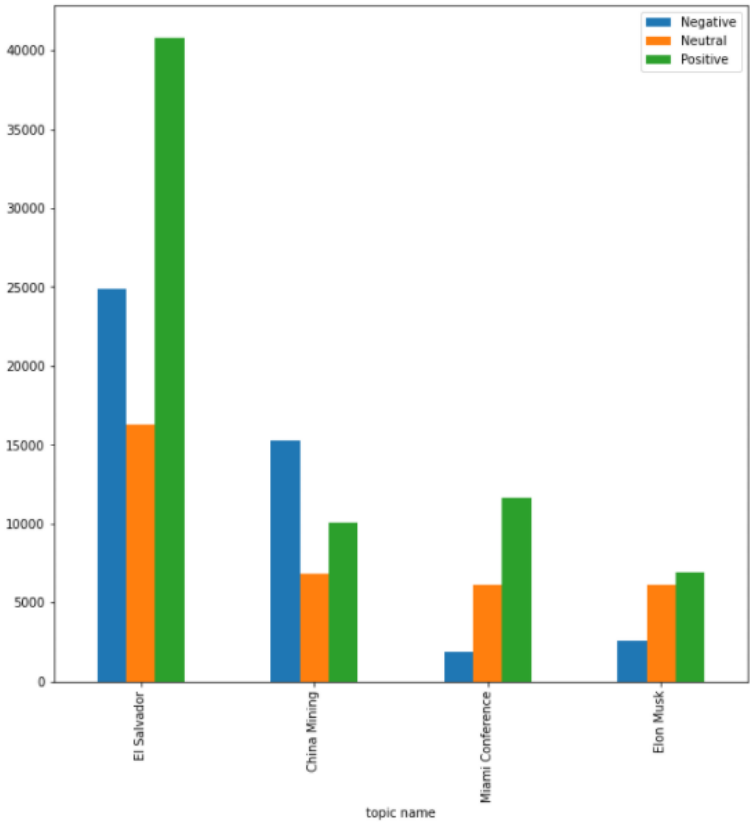

Figure 62

6. Discussion

The main purpose of this study was to investigate whether the social media mood extracted from Bitcoin related tweets help to understand the Bitcoin price volatility. In order to reach research objectives, two different data analysis models have been implemented. The results of the document level sentiment analysis show that more than 90% of the days yield positive sentiment orientation, while only one day out of the 1481 days have been identified as negative. Comparing the daily polarity score and the daily prices, it can be observed that the sentiment of the tweets tended to stay positive, regardless of price changes. Moreover, correlation score computed by two algorithms, TextBlob and VADER, was significantly low, with the value close to 0. Consequently, it could be stated that there is no correlation between daily Twitter sentiment and Bitcoin price movements.

One of the limitations that could affect the performance of the models is related to the behavior of the Twitter users who are not always using the hashtags as expected. For example, Twitter society not only use Bitcoin hashtag while referring to the Bitcoin related situation, but also while expressing general thoughts about cryptocurrencies. This implies that the accuracy of the models might be reduced due to the bias from Twitter users posting common thoughts under #bitcoin, which was used as the key word when extracting Twitter data.

Since the aggregated daily sentiments occur to be not an accurate indicator of Bitcoin price volatility, it was decided to implement more sophisticated method, which is aspect-based sentiment analysis. This model requires more computational power and resources, thus it was decided to limit the dataset to 1 year, from June 2020 to June 2021. To yield more accurate results, the subset has been divided into 12 different data frames, one for each month.

The aspect-based sentiment analysis consists of two essential stages, namely feature extraction and sentiment classification. Although, the model has been applied to 12 months, the most interesting results were identified in June 2021.

The first relevant topic contains key words that point towards the Miami Bitcoin conference, which was held in Miami, USA on the 4th and 5th of June 2021. This was the biggest Bitcoin event in history, with more than 12,000 attendees and specialized speakers from the Bitcoin community. The most remarkable element of the conference was the announcement made by Nayib Bukele, the president of El Salvador. He communicated a partnership with digital wallet company Strike, to develop the country's financial infrastructure based on the Bitcoin technology. As a result, El Salvador takes the lead by being the first country in the world to adopt Bitcoin as legal tender. It is well worth to notice that even though the announcement took place during the Miami conference, it was identified as a separate topic during feature extraction phase, which demonstrates the magnitude of the news.

Both topics display positive prevailing sentiments, which is aligned with the nature of the events and their implications. A positive sentiment for the Miami conference was expected, since it is an event for Bitcoin enthusiasts, where they can share innovative insights within the field. The same reasoning could be applied to the announcement made by the El Salvador president, as it represents a promising step into the future for Bitcoin.

However, while looking at the Bitcoin price (Figure 63) from 4th to 8th June 2021, in contrary to the expectations, there has been a significant price decrease. The explanation of this unanticipated drop can be found by exploring another dominant topic, which refers to Elon Musk. It is not surprising that he was identified within several key words given his social status as well as his active engagement in Bitcoin content on social media.

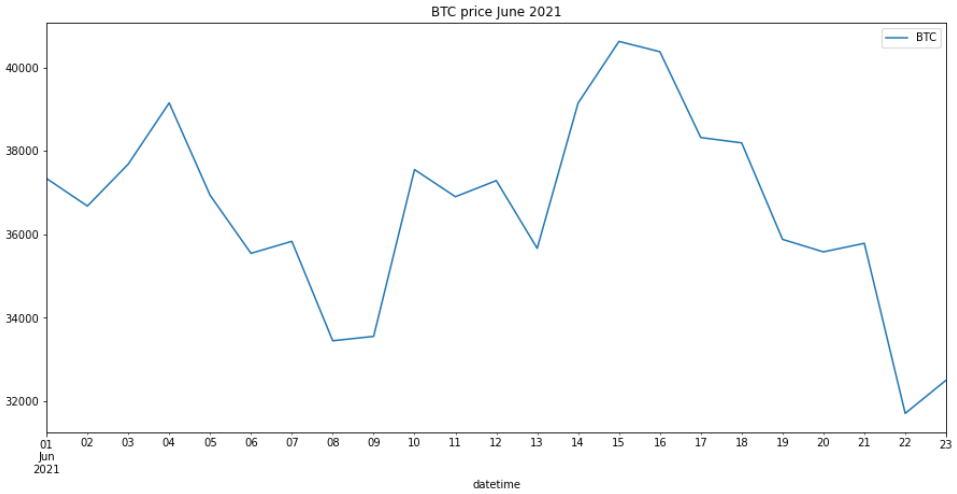

Figure 63

Reviewing Elon Musk's most popular posts from June 2021, an interesting tweet (Figure 64) has been found, which was posted on the first day of Miami Conference. This tweet embodies a meme suggesting a 'break up' with Bitcoin. It is out of the scope of this study to analyze the exact meaning behind this tweet, nevertheless, there is a strong implication that this post express negative sentiment.

Figure 64 - Post extracted from Elon Musk's Twitter https://twitter.com/elonmusk/status/1400620080090730501

Considering the above, it could be hypothesized that Elon Musk has such a strong influence, that with one negative post he was able to eliminate the positive effects associated with Miami conference, which may have led to price increase.

The last dominant topic points towards Bitcoin mining in China. On the 20th June 2021, the Bitcoin mins in South-West China were closed due to the regulatory scrutiny. This was a consequence of the Bitcoin ban announcement by the Chinese government. It is estimated that more than 90% of mining capacity will be shut down due to the ban. The event was broadly discussed among Twitter users, who have expressed their negative sentiment as exemplified in Figure 65.

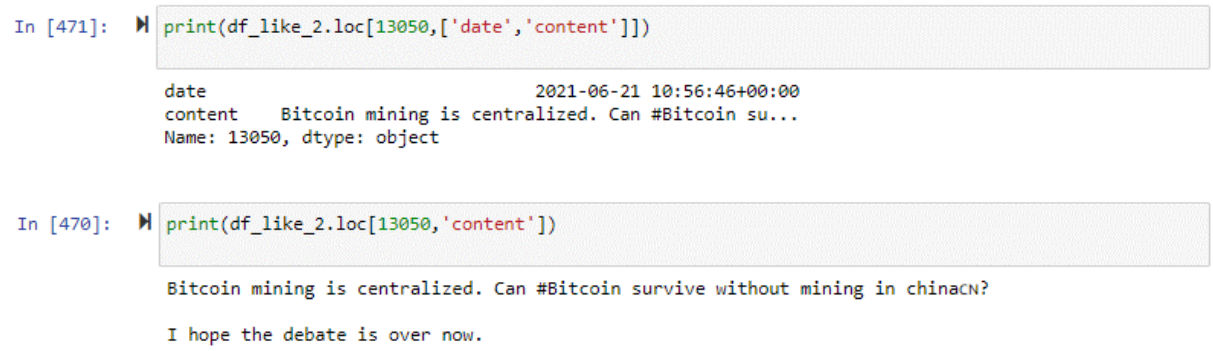

Figure 65

While looking at the graph from Figure 63, it could be discovered that there is a significant drop in Bitcoin price on June 21st, which is aligned with the results from the sentiment classification yielding negative emotion.

All things considered, aspect-based sentiment analysis, in comparison to document level sentiment analysis, is more suitable for determining the link between the price and social media content. However, the aspect-based model is still not a perfect tool to fully explain the cryptocurrency price movements due to the several limitations that can affect the accuracy of the model.

One of the first limitations identified is the lack of proper techniques to analyze multimedia content included in the tweets, especially pictures such as memes. Many users post pictures

instead or along text to convey opinions and emotions, therefore they are relevant for the sentiment analysis and by omitting them, the accuracy of the model can be affected.

Another limitation focuses on the language style used by individuals on social media platforms, such as slang and sarcasm. The use of slang is very common among Twitter community, which cannot be handled properly by the natural language processing techniques implemented in this paper. Sarcasm detection is another challenging task when working with sentiment classification, that needs to be addressed correctly, since it has the power to reverse the polarity of the sentiment.

Overall, there is not enough evidence to conclude that the price movements follow the sentiment orientation of the given topics.

Lastly, it is an already well-known fact that news have a powerful influence on the price of securities, but with the increasing popularity of social media, news are also breaking on these platforms. The conclusions made throughout this paper do not imply the lack of importance of the news announcements on the price of Bitcoin, since it has been acknowledged that they are highly influential factors. News have the ability to affect the sentiments, as well as to express emotions on their own. From the concepts proposed by behavioral finance as well as the empirical findings, it can be recalled that emotions have an effect on the decision making process of an investor. Therefore, it can be argued that the sentiment analysis conducted on the tweets is still relevant, as it can provide insight on the sentiments stemming from the news. Moreover, social media can also have an intensifying effect on the sentiments by spreading the information at a faster pace.

7. Conclusion

This master thesis examined the relationship between the sentiments extracted from Twitter posts and the Bitcoin price volatility. Drawn from behavioral finance, concepts such as cognitive biases, herd behavior and emotions have been explained and linked to the decision making process of investors. Moreover, the value of sentiment analysis as a significant tool has been presented, based on the evidence that emotions affect the decision making process. In order to answer the research question, two main approaches have been combined: applying a document-level sentiment analysis and identifying the correlation between the polarity score and the price movements and an aspect-based sentiment analysis, which proposes to determine dominant topics discussed on Twitter and their prevailing sentiments.

After conducting an extensive literature review, the data collection process was initiated. The data was collected from Twitter for 2017-2021, using a package from Python. The raw dataset consisted of more than 2.1 million tweets related to Bitcoin and 27 variables. Preprocessing the dataset was an essential step to ensure a proper performance of the classifiers and it consisted of three main stages: noise removal, normalization and tokenization.

Once the data has been cleaned and preprocessed, it was ready for the analysis stage of this study. The analysis started with the document-level sentiment analysis, which was conducted using two different Python packages. Both models exhibited similar results: preponderant positive sentiments for the aggregated daily data and no correlation determined between polarity score and price changes.

Given the shortcomings of the first tool to provide some insights on the link between social media sentiment and Bitcoin prices, a more complex technique has been implemented, the aspect-based sentiment analysis. This model has been applied only on a subset of the data, consisting of 1 year of tweets from June 2020 to June 2021. First, the dominant topics are being extracted from the dataset using topic modelling, then, the sentiment classifier labels the data as either positive, negative or neutral. The results of this analysis are rather inconsistent, since for the China mining topic, the movement in price seems to follow the negative sentiment

of the event, but for the other topics, the price development does not appear to align with the prevailing emotion.

Several limitations are being described which can affect the accuracy of the models: lack of proper techniques to interpret multimedia content, use of slang and use of sarcasm.

To sum everything up and to answer the research question formulated at the beginning of this paper: there was not enough evidence to confirm a significant relationship between the sentiment derived from Twitter and the price volatility of the cryptocurrency, however, sentiment analysis on social media can still be valuable, as discussed previously, perhaps using different approaches and techniques.

7.1 Further studies

On the basis of the above, several suggestions for further studies can be proposed. First, future research can focus on implementing techniques that are able to handle multimedia content, for example using image processing in the data transformation phase. This can help improve the accuracy of the classifier, by including more relevant content in the analysis. Another suggestion for further research is to use a machine learning sentiment model. This involves building from scratch a model that is being trained to classify data using the a sample of the dataset. This is a more complex approach that requires a balanced dataset, more computational power and time.

8. Bibliography

Baker, H. Kent and Ricciardi, Victor, How Biases Affect Investor Behaviour (2014). The European Financial Review, February-March 2014, pp. 7-10, Available at SSRN: https://ssrn.com/abstract=2457425

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The Role of Pre-processing in Twitter Sentiment Analysis. Intelligent Computing Methodologies, 615–624. https://doi.org/10.1007/978-3-319-09339-0_62

Becker, H.S. (1998) Tricks of the trade: how to think about your research while you're doing it. University of Chicago Press.

Berentsen, A., & Schar, F. (2018). A Short Introduction to the World of Cryptocurrencies. Review - Federal Reserve Bank of St. Louis, 100(1), 1–19. https://doi.org/10.20955/r.2018.1-16

Blaikie, N. (2010) "Designing Social Research" Polity Press

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Bouri, E., Gupta, R., & Roubaud. D. (2019). Herding behaviour in cryptocurrencies, Finance Research Letters, Volume 29, Pages 216-221

Bryman, A. (2012) "Social Research Methods" 4th edition, Oxford University Press

Buyya, R., Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (2016). Big data : principles and paradigms (1st edition). Morgan Kaufmann.

Chandrakala, S., & Sindhu, C. (2012). Opinion Mining and sentiment classification a survey. ICTACT journal on soft computing, 3(1), 420-425.

Crotty, M. (1998) The Foundations of Social Research. London: Sage.

Dang, S., (2014). Text Mining : Techniques and its Application. International Journal of Enginerring & Technology Innnovation. 1. 22-25.

Daniel, K., Hirshleifer, D. Subrahmanyam, A. (1998) Investor psychology and security market under- and overreactions, Journal of Finance 53, 1839-1885.

Darity, W.A., Jr., (2008), International Encyclopedia of the Social Sciences (2nd ed., Vol. 3, pp. 459-460). Macmillan Reference USA.

Denzin, N. K., & Lincoln, Y. S. (2017). The Sage handbook of qualitative research (5. edition.). Sage.

Duong Huu-Thanh, & Nguyen-Thi Tram-Anh. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. Computational Social Networks, 8(1), 1–16.

Fenton-O'Creevy, M., Soane, E., Nicholson, N. and Willman, P. (2011), Thinking, feeling and deciding: The influence of emotions on the decision making and performance of traders. J. Organiz. Behav., 32: 1044-1061.

Franco, P. (2015). Understanding bitcoin : cryptography, engineering and economics. (1st edition). John Wiley & Sons.

Gema Bello-Orgaza, Jason J. Jung, David Camachoa (2016) "Social big data: Recent achievement and new challenges" Information Fusion 45 – 59

Gerring, J. (2008). Case study research : principles and practices (Repr.). Cambridge University Press.

Giachanou, A., & Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys, 49(2), 1–41. https://doi.org/10.1145/2938640

Guthrie, Louise; Pustejovsky, James; Wilks, Yorick; Slator, Brian M. 1996 The role of lexicons in natural language processing. *The Free Library* (January, 1), https://www.thefreelibrary.com/The role of lexicons in natural language processing-a017923993

Härdle, W. K., Harvey, C. R., & Reule, R. C. G. (2020). Understanding Cryptocurrencies. Journal of Financial Econometrics, 18(2), 181–208. https://doi.org/10.1093/jjfinec/nbz033

Hofmann, T. (1999). Probabilistic latent semantic indexing. In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50–57). New York, NY, USA: ACM.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

Kallinterakis, V & Gregoriou, G.N (2017). Herd behaviour: A survey. Aestimatio: the IEB International Journal of Finance.

Kelly, Makena (August 19, 2019). "Facebook and Twitter uncover Chinese trolls spreading doubts about Hong Kong protests". The Verge.

Kietzmann, Jan H.; Kristopher Hermkens (2011). "Social media? Get serious! Understanding the functional building blocks of social media". Business Horizons (Submitted manuscript). 54 (3): 241–251. doi:10.1016/j.bushor.2011.01.005.

Kim, Y. B., Lee, J., Park, N., Choo, J., Kim, J.-H., & Kim, C. H., (2017). "When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation," PLoS ONE, vol. 12, no. 5, May 2017

Kiritchenko, S., Zhu, X., & Mohammad, M. S., (2014). Sentiment analysis of short informal texts. J. Artif. Intell. Res. 50 (2014), 723–762

Kumar, A. (2009). Dynamic style preferences of individual investors and stock returns, Journal of Financial and Quantitative Analysis, 44, pp. 607-640

Lee, N. & Lings, I. (2008) "Doing Business Research: A Guide to Theory and Practice" SAGE Publications

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. Journal of Personality and Social Psychology, 81, 146–159.

Li, S., Zhou, L., & Li, Y. (2015). Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. Information Processing & Management, 51(1), 58-67.

Lim, K. W. & Buntine, W. (2014). Twitter Opinion Topic Model. 1319-1328. 10.1145/2661829.2662005.

Lim, L. C. (2012). The relationship between psychological biases and the decision making of investor in Malaysian share market. In Unpublished Paper International Conference on Management, Economics and Finance 2012 Proceeding.

Liu, B. (2015). Sentiment analysis : mining opinions, sentiments, and emotions . Cambridge University Press.

Liu, K. L., Li, W. J., & Guo, M. (2012, July). Emoticon smoothed language models for twitter sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 26, No. 1).

Loria, S. (2018). textblob Documentation. Release 0.15, 2.

Malhotra, Naresh K, Nunan, Dan, & Birks, David F. (2017). Marketing Research (5.th ed.). Pearson Education UK.

Matthews, B., & Ross, L. (2010). Research Methods. (1. ed.). Pearson Education UK.

McEnery, Thomas (1996). Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University Press. p. 114.

Mirtaheri, M., Abu-El-Haija, S., Morstatter, F., Steeg, G. V., & Galstyan, A. (2021). Identifying and Analyzing Cryptocurrency Manipulations in Social Media. IEEE Transactions on Computational Social Systems, 8(3), 607–617. https://doi.org/10.1109/TCSS.2021.3059286

Nunan, D., Birks, D. F., & Malhotra, N. K. (2020). Marketing Research. Pearson Education, Limited.

Obar, Jonathan A.; Wildman, Steve (2015). "Social media definition and the governance challenge: An introduction to the special issue". Telecommunications Policy. 39 (9): 745–750. doi:10.1016/j.telpol.2015.07.014. SSRN 2647377.

Och, F. J.; Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models". Computational Linguistics. 29 (1): 19–51.

Oumayma B. (2020) Social Media Made Me Buy It: The Impact of Social Media on Consumer Purchasing Behavior and on the Purchase Decision-Making Process. In: Ben Ahmed M., Boudhir A., Santos D., El Aroussi M., Karas İ. (eds) Innovations in Smart Cities Applications Edition 3. SCA 2019.

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

Patton, MQ (2002). Qualitative research and evaluation methods (3rd ed.). Sage.

Philips, R. C., & Gorse, D., (2017). "Predicting cryptocurrency price bubbles using social media data and epidemic modelling," 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-7, doi: 10.1109/SSCI.2017.8280809.

Pompian, M. M., & Pompian, M. (2012). Behavioral Finance and Wealth Management: How to Build Investment Strategies That Account for Investor Biases. John Wiley & Sons, Incorporated.

Provost, F., & Fawcett, T. (2013). Data Science for Business (1st ed., pp. 26-34, 91-93,187-208, 233- 248, 249-275, 9). Beijing: O'Reilly.

Qureshi, S. A., & Hunjra, A. I. (2012). Factors affecting investment decision making of equity fund managers.

Rahman, S., Hemel, J. N., Junayed Ahmed Anta, S., Muhee, H. A., & Uddin, J. (2018). Sentiment Analysis Using R: An Approach to Correlate Cryptocurrency Price Fluctuations with Change in User Sentiment Using Machine Learning. 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 492–497. https://doi.org/10.1109/ICIEV.2018.8641075

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu (2014) “Social Media Mining” Cambridge University Press

Riffe, D., Lacy, S., Fico, F. G. (2005). Analyzing media messages: Using quantitative content analysis in research. New York, NY: Routledge.

Robson, C. (2002). Real World Research: A Resource for Social Scientists and Practitioner-Researchers (2nd ed.). Oxford: Blackwell Publishers Ltd.

Sadi, R., Asl, H.G., Rostami, M., Gholipour, A., & Gholipour, F. (2011). Behavioral Finance: The Explanation of Investors' Personality and Perceptual Biases Effects on Financial Decisions. International journal of economics and finance, 3, 234.

Saunders, M., Lewis, P., & Thornhill, A. (2007). Research methods for business students: 4. ed . Prentice Hall / Financial Times.

Segendorf, B. (2014). What is bitcoin. Sveri gesRiksbankEconomicReview, 2014, 2-71.

Simon, A. (1997) Models of Bounded Rationality, Vol. 3, Empirically Grounded Economic Reason, Cambridge: The MIT Press.

Smuts, N. (2019). What Drives Cryptocurrency Prices? An Investigation of Google Trends and Telegram Sentiment. SIGMETRICS Perform. Eval. Rev. 46, 3 131–134. DOI: https://doi.org/10.1145/3308897.3308955

Tajinder S., Madhu K., (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, Volume 89, 549-554, https://doi.org/10.1016/j.procs.2016.06.095

Tseng, K.C.. (2006). Behavioral Finance, Bounded Rationality, Neuro-Finance, and Traditional Finance. Investment Management and Financial Innovations. 3.

Veal, Anthony James. (1997). Research Methods for Leisure and Tourism. Pearson Education M.U.A.

Wang H., Castanon J.A., (2015). Sentiment expression via emoticons on social media. IEEE International Conference on Big Data (Big Data), Santa Clara. 2015; pp. 2404-2408, https

Wisdom, V., (2016). An introduction to Twitter Data Analysis in Python. 10.13140/RG.2.2.12803.30243.://doi.org/10.1109/BigDa ta.2015.73640 34.

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Systems; 37:e12493. https://doi-org.esc-web.lib.cbs.dk:8443/10.1111/exsy.12493

Wulfenia Journal, 19(10), 280-291.

Yin, R. (2009). Case Study Research: Design and Methods (Vol. Fourth). London: Sage.

Zanini, Nadir & Dhawan, Vikas. (2015). Text Mining: An introduction to theory and some applications. Research Matters. 38-44.

Websites:

Natural Language Toolkit (n.d). *Natural Language Toolkit — NLTK 3.5 Documentation*, Retrieved from www.nltk.org at 10 August 2021

Our company, (n.d). Retrieved from https://about.twitter.com/en/who-we-are/our-company at 5 August 2021

Monthly Active Twitter Users, (n.d). Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ at 15 August 2021

Documentation., (n.d). Retrieved from https://www.coingecko.com/en/api/documentation at 10 June 20

9. Appendices

1. Appendix 1 – Code for Bitcoin price retrieval

Install and load packages

In [46]:
```
!pip install pycoingecko
import pandas as pd
from pycoingecko import CoinGeckoAPI
```

Create API call

In [47]:
```
cg = CoinGeckoAPI()
timePeriod = 1480

gecko_list = ["bitcoin"]

data = {}
for coin in gecko_list:
    try:
        nested_lists = cg.get_coin_market_chart_by_id(
            id=coin, vs_currency="usd", days=timePeriod
        )["prices"]
        data[coin] = {}
        data[coin]["timestamps"], data[coin]["values"] = zip(*nested_lists)

    except Exception as e:
        print(e)
        print("coin: " + coin)
```

In [48]:
```
crypto = [
    pd.DataFrame(data[coin]["values"], index=data[coin]["timestamps"], columns=[coin])
    for coin in gecko_list
    if coin in data
]
```

In [49]:
```
crypto_price = pd.concat(crypto, axis=1).sort_index()
```

Convert timestamp to standard date time

In [55]:
```
crypto_price["datetime"] = pd.to_datetime(crypto_price.index, unit="ms")
crypto_price["date"] = crypto_price["datetime"].dt.date
crypto_price["hour"] = crypto_price["datetime"].dt.hour
```

Format the data frame

In [57]:
```
crypto_price = crypto_price.melt(
    id_vars=["datetime", "date", "hour"])
```

```
<ipython-input-57-254d47341808>:1: FutureWarning: This dataframe has a column name that matches the 'value_name' column name of the resulting Dataframe. In the future this will raise an error, please set the 'value_name' parameter of DataFrame.melt to a unique name.
  crypto_price = crypto_price.melt(
```

In [58]:
```
btc_price = crypto_price[['date', 'variable', 'value']]
```

In [58]:
```
btc_price = crypto_price[['date', 'variable', 'value']]
```

In [59]:
```
# Remove NA rows
btc_price.dropna(subset = ["value"], inplace=True)
```

```
<ipython-input-59-2f37d26e480f>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  btc_price.dropna(subset = ["value"], inplace=True)
```

In [60]:
```
# Reset Index
btc_price = btc_price.reset_index()
```

In [131]:
```
# Save as csv file
btc_price.to_csv("btc_price.csv")
```

2. Appendix 2 – Code snip for Installing and loading Python packages

1. Install and load packages

In [1]:
```
import pandas as pd
import snscrape.modules.twitter as sntwitter
import itertools
from datetime import datetime,date
import demoji
import emoji
demoji.download_codes()
import string
import re
import nltk
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('vader_lexicon')
nltk.download('twitter_samples')
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
from io import StringIO
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from textblob import TextBlob
import sys
import os
from nltk.corpus import wordnet
nltk.download('averaged_perceptron_tagger')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.corpus import twitter_samples
from nltk import ngrams
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import warnings
```

3. Appendix 3 – Code snip for Loading and merging data

2. Load and merge data

In [8]:
```
df_17 = pd.read_csv('BTC_2017_NEW')
df_18 = pd.read_csv('BTC_2018_NEW')
df_19 = pd.read_csv('BTC_2019_v2')
df_20_07 = pd.read_csv('BTC_2020_07')
df_20 = pd.read_csv('BTC_2020_v2')
```

In [9]:
```
df_list = [df_20, df_20_07, df_19, df_18, df_17]
```

In [11]:
```
df_btc = pd.concat(df_list)
```

In [12]:
```
df_btc.reset_index(drop=True, inplace=True)
```

4. Appendix 4 – Code snip for removing unnecessary columns

3. Remove unnecessary columns

In [5]:
```
# Convert date column to datetime format
df_btc = pd.to_datetime(df_btc['date']).dt.date

# Keep only necessary columns
df_final = df_btc[['date', 'content']]

# Rename column
df_final.rename(columns={'date': 'Date'}, inplace=True)
```

5. Appendix 5 – Code snip for emoji handling

4. Emoji handling

In [70]:
```
str_emoji = df_final['content'].to_csv()
```

In [71]:
```
str_demoji = emoji.demojize(str_emoji)
```

In [73]:
```
df_demoji = pd.read_csv(StringIO(str_demoji))

df_final['content_demoji'] = df_final['content']
```

6. Appendix 6 – Code snip for expanding contractions

Replace contractions with the full form

In [8]:
```
cList = {
  "ain't": "am not",
  "aren't": "are not",
  "can't": "cannot",
  "can't've": "cannot have",
  "'cause": "because",
  "could've": "could have",
  "couldn't": "could not",
  "couldn't've": "could not have",
  "didn't": "did not",
  "doesn't": "does not",
  "don't": "do not",
  "hadn't": "had not",
  "hadn't've": "had not have",
  "hasn't": "has not",
  "haven't": "have not",
  "he'd": "he would",
  "he'd've": "he would have",
  "he'll": "he will",
  "he'll've": "he will have",
  "he's": "he is",
  "how'd": "how did",
  "how'd'y": "how do you",
  "how'll": "how will",
  "how's": "how is",
  "i'd": "i would",
  "i'd've": "i would have",
  "i'll": "i will",
  "i'll've": "i will have",
  "i'm": "i am",
  "i've": "i have",
  "isn't": "is not",
  "it'd": "it had",
  "it'd've": "it would have",
  "it'll": "it will",
  "it'll've": "it will have",
  "it's": "it is",
  "let's": "let us",
  "ma'am": "madam",
  "mayn't": "may not",
  "might've": "might have",
  "mightn't": "might not",
  "mightn't've": "might not have",
  "must've": "must have",
  "mustn't": "must not",
  "mustn't've": "must not have",
  "needn't": "need not",
  "needn't've": "need not have",
  "o'clock": "of the clock",
  "oughtn't": "ought not",
  "oughtn't've": "ought not have",
  "shan't": "shall not",
  "sha'n't": "shall not",
  "shan't've": "shall not have",
  "she'd": "she would",
  "she'd've": "she would have",
  "she'll": "she will",
  "she'll've": "she will have",
  "she's": "she is",
  "should've": "should have",
  "shouldn't": "should not",
  "shouldn't've": "should not have",
```

```
  "so’ve": "so have",
  "so’s": "so is",
  "that’d": "that would",
  "that’d’ve": "that would have",
  "that’s": "that is",
  "there’d": "there had",
  "there’d’ve": "there would have",
  "there’s": "there is",
  "they’d": "they would",
  "they’d’ve": "they would have",
  "they’ll": "they will",
  "they’ll’ve": "they will have",
  "they’re": "they are",
  "they’ve": "they have",
  "to’ve": "to have",
  "wasn’t": "was not",
  "we’d": "we had",
  "we’d’ve": "we would have",
  "we’ll": "we will",
  "we’ll’ve": "we will have",
  "we’re": "we are",
  "we’ve": "we have",
  "weren’t": "were not",
  "what’ll": "what will",
  "what’ll've": "what will have",
  "what’re": "what are",
  "what’s": "what is",
  "what’ve": "what have",
  "when’s": "when is",
  "when’ve": "when have",
  "where’d": "where did",
  "where’s": "where is",
  "where’ve": "where have",
```

```
"who’ll": "who will",
"who’ll've": "who will have",
"who’s": "who is",
"who’ve": "who have",
"why’s": "why is",
"why’ve": "why have",
"will’ve": "will have",
"won’t": "will not",
"won’t've": "will not have",
"would’ve": "would have",
"wouldn’t": "would not",
"wouldn’t've": "would not have",
"y’all": "you all",
"y’alls": "you alls",
"y’all’d": "you all would",
"y’all’d’ve": "you all would have",
"y’all’re": "you all are",
"y’all’ve": "you all have",
"you’d": "you had",
"you’d’ve": "you would have",
"you’ll": "you you will",
"you’ll’ve": "you you will have",
"you’re": "you are",
"you’ve": "you have",
  "ain't": "am not",
"aren't": "are not",
"can't": "cannot",
"can't've": "cannot have",
"'cause": "because",
"could've": "could have",
"couldn't": "could not",
"couldn't've": "could not have",
"didn't": "did not",
```

```
"doesn't": "does not",
"don't": "do not",
"hadn't": "had not",
"hadn't've": "had not have",
"hasn't": "has not",
"haven't": "have not",
"he'd": "he would",
"he'd've": "he would have",
"he'll": "he will",
"he'll've": "he will have",
"he's": "he is",
"how'd": "how did",
"how'd'y": "how do you",
"how'll": "how will",
"how's": "how is",
"i'd": "i would",
"i'd've": "i would have",
"i'll": "i will",
"i'll've": "i will have",
"i'm": "i am",
"i've": "i have",
"isn't": "is not",
"it'd": "it had",
"it'd've": "it would have",
"it'll": "it will",
"it'll've": "it will have",
"it's": "it is",
"let's": "let us",
"ma'am": "madam",
"mayn't": "may not",
"might've": "might have",
"mightn't": "might not",
"mightn't've": "might not have",
```

```
"must've": "must have",
"mustn't": "must not",
"mustn't've": "must not have",
"needn't": "need not",
"needn't've": "need not have",
"o'clock": "of the clock",
"oughtn't": "ought not",
"oughtn't've": "ought not have",
"shan't": "shall not",
"sha'n't": "shall not",
"shan't've": "shall not have",
"she'd": "she would",
"she'd've": "she would have",
"she'll": "she will",
"she'll've": "she will have",
"she's": "she is",
"should've": "should have",
"shouldn't": "should not",
"shouldn't've": "should not have",
"so've": "so have",
"so's": "so is",
"that'd": "that would",
"that'd've": "that would have",
"that's": "that is",
"there'd": "there had",
"there'd've": "there would have",
"there's": "there is",
"they'd": "they would",
"they'd've": "they would have",
"they'll": "they will",
"they'll've": "they will have",
"they're": "they are",
"they've": "they have",
```

7. Code snip for stop words removal

```
In [18]: stop_words = set(stopwords.words('english'))

In [19]: #Removing 'no' and 'not' to handle word negations
         stop_words.remove('not')
         stop_words.remove('no')

In [20]: print('no' in stop_words)

         False
```

In [21]:
```
def preprocess_tweet_text(tweet):

    #convert to lowercase
    tweet = tweet.lower()

    #remove any urls
    #tweet = re.sub(r"https\S+|www\S+|https\S+","", tweet, flags=re.MULTILINE)

    #remove punctuations
    tweet = tweet.translate(str.maketrans(" ", " ", string.punctuation))

    #remove user @ refreneces and '#' from tweet
    tweet = re.sub(r'\@\w+|\#', "", tweet)

    # remove special characters, numbers and punctuations
    #tweet = re.sub(r"[^a-zA-Z]"," ",tweet)

    #remove stopwords
    tweet_tokens = word_tokenize(tweet)
    filtered_words = [word for word in tweet_tokens if word not in stop_words]

    #Lemmatizing
    lemmatizer = WordNetLemmatizer()
    lemma_words = [lemmatizer.lemmatize(w, pos='a') for w in filtered_words]

    return " ".join(lemma_words)
```