INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Cryptocurrency Analysis based on User-Generated Social Media Content

Miguel de Guerra Narciso

Master in Computer Science and Business Management

Supervisor:

Doctor Luís Miguel Martins Nunes, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-supervisor:

Doctor Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2020

TECNOLOGIAS
E ARQUITETURA

Cryptocurrency Analysis based on User-Generated Social Media Content

Miguel de Guerra Narciso

Master in Computer Science and Business Management

Supervisor:

Doctor Luís Miguel Martins Nunes, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-supervisor:

Doctor Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October, 2020

Resumo

A Bitcoin é uma criptomoeda que começou a tornar-se interessante em 2017 através do entusiasmo que criou pelo mundo, afirmando ser a primeira criptomoeda totalmente digital e baseada na tecnologia de Blockchain. Existem pessoas que reagem e comentam diariamente sobre a Bitcoin em redes sociais e fóruns online. As redes sociais que mais impulsionam opiniões relacionadas com a Bitcoin são o Twitter e o Reddit. Podemos afirmar que a Bitcoin pode atribuir grande parte do seu crescimento através da rede social Reddit. A investigação efetuada no contexto desta tese é dedicada ao estudo da aplicação de uma análise de sentimento aos comentários dos utilizadores do Reddit contidos nas publicações de discussão diária da comunidade Bitcoin, com o objetivo de encontrar uma correlação entre os resultados obtidos e os indicadores conhecidos da Bitcoin. O estudo visa analisar o sentimento associado aos tópicos de discussão da comunidade Bitcoin, tendo duas principais questões de investigação: 1. Os comentários dos utilizadores das publicações de discussão diária têm uma correlação com os indicadores diários da Bitcoin? 2. Os comentários dos utilizadores das publicações de discussão diária têm uma correlação com os indicadores de Bitcoin se forem deslocados por alguns dias? A investigação permitiu analisar as correlações entre os diferentes indicadores.

Palavras chave

Reddit, Bitcoin, análise de sentimento, criptomoedas, correlação

Abstract

Bitcoin is a cryptocurrency that has started to become interesting in 2017 as a result of the massive hype the cryptocurrency has created around the world, claiming to be the first fully digital coin based on blockchain technology. There are people reacting and commenting everyday about Bitcoin on social networks and online forums. The known social networks that drive significant opinions related with Bitcoin are Twitter and Reddit. It can be said that Bitcoin may attribute much of its growth to the Reddit social forum. This research is dedicated to the study of the application of Sentiment Analysis to the Reddit users' comments on daily discussion publications from Bitcoin - Reddit community with the goal to find a correlation between these results and Bitcoin's known indicators. The study aims to analyze the sentiment of the Bitcoin discussion threads, having two main research questions: 1. Do users' comments of the daily discussion posts have a correlation with the daily indicators of Bitcoin? 2. Do users' comments of the daily discussion posts have a correlation when the indicators information of Bitcoin are shifted by a few days? The research has permitted to analyze the correlations between the different indicators.

Keywords

Reddit, Bitcoin, sentiment analysis, cryptocurrency, correlation analysis

Acknowledgements

Esta tese não teria sido possível sem a ajuda dos meus orientadores, os professores Luís Nunes e Fernando Batista. Foi através da partilha do seu conhecimento que foi possível o desenvolvimento desta dissertação. É gratificante o acompanhamento demonstrado em todas as fases da investigação, sendo determinante a sua participação para a conclusão deste estudo.

Quero agradecer aos meus familiares pelo apoio demonstrado, em especial à minha mãe, Laura, que proporcionou todo o apoio necessário no meu percurso académico, garantindo que tinha tudo para alcançar o meu sucesso. Outro agradecimento especial também para a minha namorada Léonie, por ter-me sempre ajudado e estado a meu lado neste desafio.

Aos meus amigos e colegas, agradeço por terem demonstrado disponibilidade para me ajudar a alcançar este desafio.

A realização deste trabalho foi parcialmente financiada por fundos nacionais através da FCT - Fundação para a Ciência e Tecnologia, I.P. no âmbito dos projetos UIDB/EEA/50008/2020 (Instituto de Telecomunicações) e UIDB/04466/2020 (ISTAR).

Por fim, um agradecimento a todos os que contribuiram para a concretização desta dissertação.

Lisboa, 31 de Outubro

Miguel Narciso

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid growth of the Internet and social networks introduced us to a new world of data generated online. Nowadays, people communicate more than ever before on the Internet. There are a considerable number of people expressing their opinions on brands and businesses through social networks which generate a huge amount of data every day. As digital transformation enters all areas of the business, this provides opportunities for how we can find available data in a more analytical way.

The adoption of Machine Learning comes as the right answer for the opportunity of data analysis. For businesses, processing all generated data related to the company is more important than ever, especially in online conversations such as news, blogs, social media, and social networks. This is something not physically possible for a human being to analyze due to the huge volume of data generated daily.

With the ability of machines to learn on their own by processing a large amount of data and identifying patterns to give interpretable information it makes a good point of start for an analytical environment. Indeed, through Natural Language Processing (NLP) domain it is possible for computers to understand, interpret and manipulate human language. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important.

The Sentiment Analysis (SA) appears as one approach of NLP to measure sentiment over texts. There are plenty of applications of SA in different domains such as political science, economics, and social sciences which are all affected by people's opinion. It is considered that financial market forecasting is one of the most attractive practical applications of SA. For this reason, this research is dedicated to the study of the application of Sentiment Analysis to the Reddit users comments on daily discussion publications from Bitcoin - Reddit community with the goal to find a correlation between these results and Bitcoin's known indicators.

First of all, the data available on the Bitcoin community from the social network Reddit is going to be analyzed . More specifically, it will be examined the comments of daily

discussion publication on the Bitcoin community. The comments will be examined with a Sentiment Analysis tool, objectively trying to interpret the sentiment of daily Bitcoin. Creating then a daily sentiment rate of the cryptocurrency. The selected algorithm to analyze the sentiment of the data will be the Valence Aware Dictionary and sEntiment Reasoner (VADER) which is a combined lexicon and rule-based sentiment analytic software.

Furthermore, after rating the sentiment of the cryptocurrency by day, the distribution of the SA output will be created to be compared with the market evaluation of the cryptocurrency. The main goal is going to find a correlation between the daily sentiment of the cryptocurrency and the Bitcoin market indicators.

1.1 Motivation and goals

Bitcoin is a cryptocurrency that has started to become interesting in 2017, as a result of the massive hype the cryptocurrency was creating around the world, claiming to be the first fully digital coin based on blockchain technology. Earlier on, the Bitcoin was known as an untraceable payment method that was used sometimes for illicit actions but in fact, it was the currency value that created the notoriety around it. This digital currency has the biggest share of the cryptocurrency market, it had started the year of 2017 under US$1,000 of value and had skyrocketed by more than 1,300% to over than $14,500 on December 29, 2017. The price of Bitcoin is notoriously volatile and susceptible to react strongly to geopolitical events and regulatory rulings concerning cryptocurrency. It was on December 17, 2017 Bitcoin achieved its record price high, which reached close to $20,000 and then experienced a series of crashes throughout 2018 that saw its value eventually drop below $4,000.

Initially, an investor can begin to keep track of changes of Bitcoin in a convenient way such as, analyzing charts of the cryptocurrency value evolution. Representing the technical analysis, used to identify trading opportunities by analyzing statistical trends gathered from trading activity, such as price movement and volume. When evaluating Bitcoin, there are benefits from harnessing fundamental, technical and sentiment analysis while learning more about the economics of Bitcoin, namely the various factors that affect supply and demand. There are plenty sources for technical analysis available on the Internet, such as, the Coindesk and CoinMarketCap which are referenced price-tracking websites that provide information of cryptocurrencies and digital assets through news and market data. Both websites are used on the investigation to obtain the Bitcoin historical market data.

Sentiment analysis is a significant approach to understand the perception of investors opinions. It can allow traders and investors to measure how crypto markets and their participants are feeling, whether confident or optimist by rising share prices (bullish) or pessimist associated with falling share prices (bearish). NLP tools enable to automatically understand and process natural human language such as, speech or text to interpret mean-

ing from it. NLP has many use cases on the data analysis field, one of its use cases is the Sentiment Analysis approach by identifying opinions and determining whether the author of an expression holds a positive, negative, or neutral opinion. Data generated from conversations, publications or tweets on the Internet are a representation of the source of information that can be used to be analyzed on NLP.

In order to obtain the most updated cryptocurrency news, it is most likely for people to turn to social media and online forums. There are known social networks that drive significant opinions related with Bitcoin such as Twitter and Reddit. Twitter allows the users to send and receive short posts of comments called tweets, having the feature of associating hashtags to define the related topic and creating a thread of users reactions. On the other hand Reddit is organized by communities where users can post, vote, and comment in the forums of their interest called "subreddits". When a post is created it generates a thread of conversation available for users to react and comment. Following the Bitcoin community on Reddit, there are daily discussion posts for users to discuss subjects related to the cryptocurrency day's events, comment technical analysis and ask quick questions to trade ideas.

Reddit is a huge source of shared information, it works as a place for users to interact within topic communities. Reddit aims to bring people sharing similar interests together. The popular website goes by the slogan "The front page to the Internet", it has earned this name by creating a platform that allows users to discuss and share content on the internet. The subreddits can be seen as engaging communities that are great sources for content based on a specific topic. On 31 December of 2017, the subreddit "/r/Bitcoin" reached 600,000 subscribers being ranked the 129 top subreddit community on Reddit. It can be said that Bitcoin may attribute much of its growth to the Reddit social forum due to the volume of subscribers posting and commenting about Bitcoin.

The known growth of Bitcoin subreddit community was one of the reasons Reddit was chosen as the source of information instead of Twitter. Additionally, Twitter is sometimes known for being used to spread misinformation and false rumors, often unintentionally but leading to misinterpretation of reactions and being a challenge for analysis. As Weninger, Zhu, and Han [1] research references the Reddit comments threads provides a user-generated and user-curated commentary on the topic at hand. Unlike Twitter discussions that are person-to-person and often times difficult to discern, comment threads in the social news paradigm are permanent (although editable), well-formed and hierarchical [1].

The goal of the investigation is to study the relationship between Bitcoin and the sentiment of comments on Reddit. The investigation starts by collecting the comments dataset from Bitcoin community and then classify the sentiment polarity by using a SA approach. The focus is to compare the daily Bitcoin indicators with the sentiment classification of users comments, in order to measure the correlation between them. Also, this research will attempt to verify if there are pre or post effects linked within the comments sentiment

and the Bitcoin value.

1.2 Research questions

Using Sentiment Analysis as an approach to classify the polarity of the comments and define them as positive, neutral or negative comments. After having the sentiment classification of the comments, the goal is to compute the correlation between the indicators. The correlation analysis is going to be represented on heat maps for data visualization across intensity color variations by cross-examining multivariate data.

Main research question: Does the result of the SA on the daily users comments of Reddit correlate with the daily Bitcoin market value?

To answer the question the first objective is to analyze the relation between the daily sentiment of Bitcoin with the evolution of the cryptocurrency market value.

In order to structure the analysis, the following questions must be answered:

1. Do users sentiment comments of the daily discussion posts have a correlation with the daily indicators of Bitcoin?
2. Do users sentiment comments of the daily discussion posts have a correlation when the indicators information of Bitcoin are shifted by a few days?

1.3 Methodology

The goal of the methodological approach is the classification of the social media Reddit comment sentiments from the daily discussion posts of Bitcoin community, creating alongside a daily, weekly and monthly indicators of sentiment classification, in order to compare the correlation with Bitcoin data.

The cryptocurrency market is a relatively new phenomenon and alternative data sources are limited. In the research, Reddit is considered to be one of the most comprehensive sources of investor sentiment relating to the cryptocurrency market.

The software supporting this research is programmed in Python, mostly using Jupyter notebook scripts for the analysis research. The planning for the research is described by the following processes:

1. Data collection;
2. Data cleansing;
3. Analyze comments data with Sentiment Analysis for sentiment polarity classification;

4. Correlation analysis between SA indicators and Bitcoin market information;
5. Compare the results of the Pearsons' correlation.

1.4 Document structure

The document is divided into five chapters, each of which focuses specifically on achieving the originally defined goals. It is followed by the methodological approach previously defined in section 1.3. In this introductory Chapter 1, it is presented the motivation and goals of the investigation defining as well the initial objectives and research questions. Then, Chapter 2, includes a review of the researched area with current information and previous studies surrounding the issue, besides introducing the history and background information on the thesis problem.

In Chapter 3 is given the state of the art of the analyzed research areas, as well as its development and recent applications. It allows a better understanding of the relevance of the problem researched in the first chapter, respecting the classification of Sentiment Analysis.

The Chapter 4 presents the development methods of data collection, as well as the data analysis by applying the Sentiment Analysis approach, following the third step of the methodology. It is demonstrated the experimentation of the Sentiment Analysis application and showed how can it solve the problem and the determined goal.

In Chapter 5 is described the research conclusions obtained through the investigation analysis. Also, it is demonstrated the limitations and proposals for future work.

Finally, the bibliographical references and the annexes are provided.

Chapter 2

Background

This chapter introduces the concepts related with the research topic on using Sentiment Analysis as an approach for correlation analysis with cryptocurrency. Previous studies surrounding the issue are presented, describing the methodologies of its applications and relevant background theory on the thesis problem.

2.1 Cryptocurrency

By definition, Rouse [2] says Cryptocurrency is a type of digital currency based on cryptography, or the process of converting plaintext into ciphertext, thus making readable text non decipherable. More recently, Gold and Mcbride [3] defines cryptocurrency as a strictly digital currency (and not merely the digital exchange of conventional currencies), typically overseen by a decentralized peer-to-peer community and are secured through cryptography.

Cryptocurrencies relies on cryptography techniques. To better understand the definition of cryptocurrencies it is necessary to comprehend what cryptography stands for.

Rouse [2] also defines cryptography as a method of protecting information and communications through the use of codes so that only those for whom the information is intended can read and process it.

As described in the Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction book [4], cryptography provides a mechanism for securely encoding the rules of a cryptocurrency system in the system itself. It can be used to prevent tampering and equivocation, as well as to encode the rules for creation of new units of the currency into a mathematical protocol [5].

Summarizing, cryptocurrency is a digital or virtual asset that functions as a manner of exchange currency in much the same way as more traditional forms of legal tender (like metal or paper currencies). It is a form of digital money that employs cryptography to

ensure security in the transfer of funds and creation of additional currency units. The foundation of any cryptocurrency lies in its security, anonymity, and accessibility to anyone with an Internet connection.

The cryptocurrencies differ from conventional currencies in different ways: they are not controlled or regulated by banks or governments; relies on a decentralized system (known as the “blockchain”); the exchange rates can be extremely volatile; the transactions are anonymous.

The most known cryptocurrency is Bitcoin. Established in 2008 by an individual with the name Satoshi Nakamoto, it was the first cryptocurrency created. As the creator Nakamoto [6] says Bitcoin is a purely peer-to-peer version of electronic cash that would allow online payments to be sent directly from one party to another without going through a financial institution.

Peer-to-peer consists of two or more computers that are connected and share resources without the use of a separate server [7]. This is one of the biggest differences when compared to traditional system banks. The Bitcoin system, unlike traditional banking and payment systems, is based on decentralized trust. Instead of a central trusted authority, in bitcoin, trust is achieved as an emergent property from the interactions of different participants in the bitcoin system [8].

As the research of [9] describes, in order for a consumer to begin interacting and conducting transactions utilizing Bitcoin, first it must download and setup a Bitcoin wallet. A Bitcoin wallet can show the total balance of all Bitcoins it controls and let a user pay a specified amount to a specific person, just like a physical wallet [6]. [9] complements that once the wallet is installed and configured, an address is generated which is similar to an e-mail or physical address, which provides to other users a numerical location to send Bitcoins to. In addition, the wallet contains a user private key which is located in the blockchain technology and allows users to transfer Bitcoins using the private key as an address.

There are four methods to acquire Bitcoins [6]: as payment for goods or services, purchase of coins through a Bitcoin exchange, exchanging them with another user or earn the coins through competitive mining. The easiest method to obtain Bitcoins is to purchase them from an online vendor like Coinbase. The cryptocurrency has a market value to identify the actual price of the currency. The Bitcoin is the cryptocurrency with the biggest percentage of market share. Currently, there are a good number of web sites that provides the current market rate of cryptocurrency, Coin Market Cap is one example of cryptocurrency market cap with rankings, charts and more.

2.1.1 How it works

The cryptocurrencies are based in the blockchain technology. [10] defines blockchain as essentially a distributed database of records, or public ledger of all transactions or digital

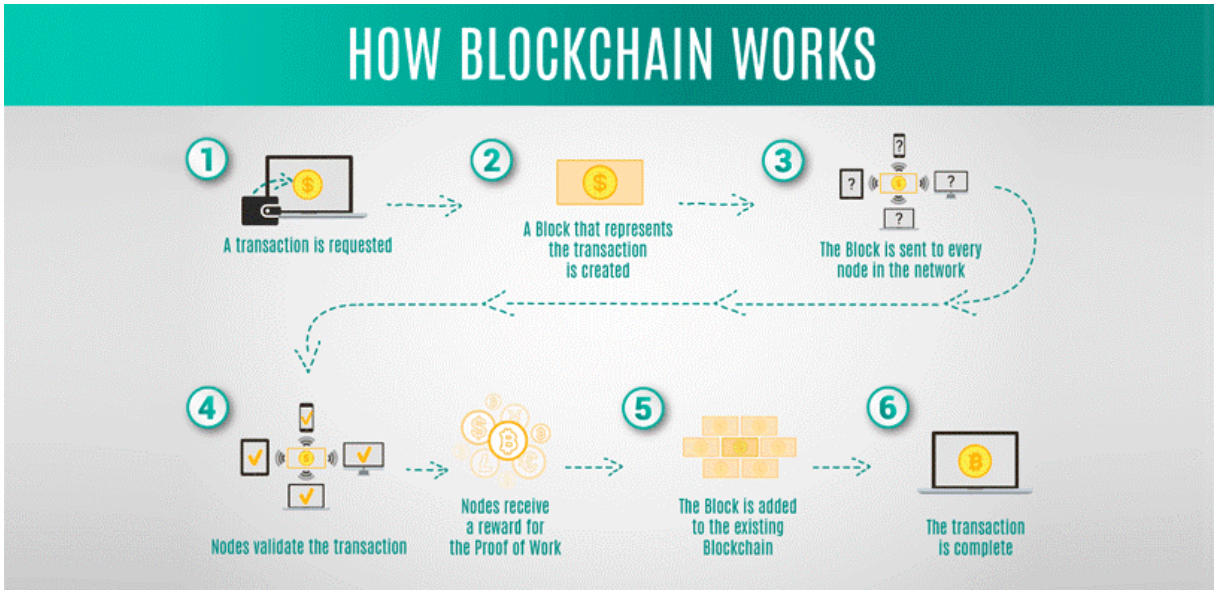

Figure 2.1: How blockchain works (Source: Zignuts Technolab)

events that have been executed and shared among participating parties. Each transaction in the public ledger is verified by the consensus of most of the participants in the system. The blockchain contains a certain and verifiable record of every single transaction ever made. [11] indicates blockchain technology ensures the elimination of the double-spend problem, with the help of public-key cryptography, whereby each agent is assigned a private key kept secret like a password and a public key shared with all other agents. A transaction is initiated when the future owner of the coins sends his public key to the original owner.

As the Figure 2.1 shows, the technology works as a workflow of block transactions by placing them in groups called blocks and then linking these blocks through what is called blockchain [10]. Nakamoto [6] defines the technology of Bitcoin as it provides documentation for each of the confirmed transactions as well as providing Bitcoin wallets a way to calculate their spendable balance and a way for new Bitcoin transactions to be verified.

Antonopoulos [8] describes users of bitcoin own keys which allow them to prove ownership of transactions in the bitcoin network, unlocking the value to spend it and transfer it to a new recipient. Those keys are often stored in a digital wallet on each users computer. Possession of the key that unlocks a transaction is the only prerequisite to spending bitcoins, putting the control entirely in the hands of each user.

If a user is successful in obtaining Bitcoins through purchase, trade or by mining, the users Bitcoins remain in their worker account until they are transferred to their individual Bitcoin wallet says Doran [9] . When a user wishes to conduct a transaction, three pieces of information are required:

1. An input - this is the record of which Bitcoin address was used to send Bitcoins to the user [12];

2. An amount - this is the amount of Bitcoins that the user is sending to another user [12];

3. An output - this is the address of the recipient of the Bitcoins to be sent [12].

Citing Coindesk [12] if a person wants to send the Bitcoins to an intended user and complete a transaction, the person needs to have a Bitcoin address, which is automatically generated when the Bitcoin wallet software is installed and a private key generated. A private key serves as a cryptography signature that validates a users right to send Bitcoins from a specific wallet. Nakamoto [6] described if a user is utilizing a software wallet, the private key is stored on the users computer, whereas if the user makes use of a web-based wallet the private key is stored on a separate server.

Doran [9] added saying that with the addresses of the sender and the recipient, the amount and the private key, the user can then conduct a Bitcoin transaction. Coindesk [12] concludes that the users private key signs a message with the input, amount, and output of Bitcoins before it is sent from their Bitcoin wallet out to the wider Bitcoin network where the transaction is placed on the transaction block where it is eventually verified by Bitcoin miners.

Crosby, Nachiappan, Pattanayak, et al. [10] indicates that the digital currency Bitcoin itself is highly controversial but the underlying blockchain technology has worked flawlessly and found wide range of applications in both financial and non-financial world.

2.1.2 Applications

The blockchain, the technology behind the cryptocurrency system, is considered to be both alluring and critical for ensuring enhanced security and (in some implementations, non-traceable) privacy for diverse applications in many other domains. Intensive research is currently being conducted in both academia and industry applying the blockchain technology in multifarious applications [13].

Crosby, Nachiappan, Pattanayak, et al. [10] notes the system of using the blockchain algorithm to achieve distributed consensus on a particular digital asset. As the system may share miners with a parent network such as Bitcoin's, which is called merged mining. These alternative blockchains have been suggested to implement applications such as DNS, SSL certification authority, file storage and voting.

Crosby, Nachiappan, Pattanayak, et al. [10] says the companies such as IBM, Samsung, Overstock, Amazon, UBS, Citi, Ebay, and Verizon Wireless, to name a few, are all exploring alternative and novel uses of the blockchain for their own applications.

Tapscott and Tapscott [14] indicated blockchain to be the "World Wide Ledger", enabling many new applications beyond verifying transactions such as in: smart deeds, decentralized and/or autonomous organizations/ government services, etc.

For example, Miraz and Ali [13] stated NASDAQ is using its 'Linq blockchain' to record its private securities transactions. The Depository Trust & Clearing Corporation (DTCC, USA) is working with Axoni in implementing financial settlement services such as post-trade matters and swaps. Regulators are also interested in blockchain's ability to offer secure, private, traceable real-time monitoring of transactions.

Also, Crosby, Nachiappan, Pattanayak, et al. [10] on his work a systematic literature review of blockchain-based applications, identify Augur as a decentralized prediction market that allow users to buy and sell shares in anticipation of an event having the objective to create a prediction on market based on any topic.

[13] concludes the application of the blockchain concept and technology has grown beyond its use for Bitcoin generation and transactions. The properties of its security, privacy, traceability, inherent data provenance and time-stamping have seen its adoption beyond its initial application areas.

2.2 Sentiment analysis

Before defining the Sentiment Analysis concept, it is necessary to identify that it is a Text Classification process. This process is also known as text tagging or text categorization, it categorizes text into organized groups. Text classifiers can automatically analyze text and then assign a set of preset tags or categories based on its content.

Through SA, there are more techniques of Text Classifier, which are Intent Analysis, Contextual Semantic Search (CSS) and Emotion Analysis. Shortly resumed, Intent Analysis is the analysis of acknowledging the intentions from a text, CSS tries to interpret a contextual relation to a concept (e.g. price as an input) between messages that closely match with the given concept and Emotion Analysis is the ability to accurately detect the emotion involved from any textual data [15].

The first interpretation as we read the name Sentiment Analysis is perhaps the analysis of the sentiment of an expression or analyzing an attitude towards an opinion.

According to [16] SA is about finding the underlining opinions, sentiment, and subjectivity in texts, which all are important factors in influencing behavior.

In the computational area, [17] research refers that it is the use of natural language, text analysis and computational linguistics which identify or extract subjective information from the attitude of a speaker/writer or from a set of customer opinion in order to classify the polarity.

SA is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral [18].

It is then considered that SA is the automatic process of detecting positive, neutral and negative sentiments within text using machine learning algorithms. It is a known text classification tool that analyses an incoming message and tells the underlying sentiment.

Besides identifying the opinion, usually these systems extract attributes of the expression e.g.:

- Polarity: if the speaker expresses a positive, negative or neutral opinion;
- Subject: the thing that is being talked about;
- Opinion holder: the person, or entity that expresses the opinion. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

SA is an application of Natural Language Processing (NLP), NLP is defined as a technology used to aid computers to understand the human's natural language. The objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable in order to interpret it. Most NLP techniques rely on machine learning to derive meaning from human languages.

The Sentiment Analysis it has been used for long as a research method, quoting [19] although linguistics and NLP have a long history, little research had been done about people's opinions and sentiments before the year 2000. Since then, the field has become a very active research area and there are plenty of reasons for that.

[19] stated that first, it has a wide arrange of applications in different types of domains such as political science, economics, and social sciences as they are all affected by people's opinions. Second, it offers many challenging research problems, with new ways of analyzing opinions which had never been studied before. Third, with all the data available online, there is a huge volume of opinionated data in the social media on the Web.

This way it can be seen that SA offers us a lot of capabilities in which it can be analyzed quantitatively any possible opinion. But with these capabilities it also comes limitations, furthermore it will be described some.

SA can be applied at different levels of scope based on the level of granularity of the information researched. It can often be conducted in three levels known as Document Level, Sentence Level and Entity and Aspect Level.

Preethi, Uma, and Kumar [18] affirmed that Document Level sentiment analysis classifies the entire document into either positive or negative which at this level, the analysis is evaluating only one referenced subject, there is no comparisons of subjects.

On the same page for Liu [19] the task at the Document Level is to classify whether a whole opinion document expresses a positive or negative sentiment. Liu used the example of a "given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product". This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g. a single product).

For Preethi, Uma, and Kumar [18] sentence level classification classifies the sentence into positive, negative or neutral category. Additionally, SA is closely related to subjectivity classification, which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions [19].

Preethi, Uma, and Kumar [18] defines Entity and Aspect level as it gives the summary about which feature of a product did the user like or dislike. Aspect level was earlier called feature level (feature-based opinion mining and summarization) by Minqing Hu and Bing Liu [20]. Liu [19] says aspect level performs finer-grained analysis which means it is more precise about the level of polarity of the opinion. Instead of looking at the language construct (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion).

2.2.1 How it works

There are different Sentiment classification methods and algorithms approaches for using SA. Kaur and Gupta [17] notes that the most popular approaches used are:

- Subjective lexicon – is a list of words where each word is assigned a score that indicates nature of word in terms of positive, negative or objective. It can also be called as semantic approach.
- Using N-Gram modeling - for given training data, it creates a N-Gram model (unigram, bigram, tri-gram or combination of these) for classification.
- Machine learning – perform the supervised or semi- supervised learning by extracting the features from the text and learn the model. This approach can be considered as an automatic system that relies on its techniques to learn from data.

The semantic approaches are characterized by the use of dictionaries of words (lexicons) with semantic orientation of polarity or opinion says Villena-Román [21]. Systems typically

pre-process the text and divide it into words, with proper removal of stop words and a linguistic normalization with stemming or lemmatization, and then check the presence or absence of each term of the lexicon, using the sum of the polarity values of the terms for assigning the global polarity value of the text.

Hatzivassiloglou and McKeown [22] were the first to develop empirical method of building sentiment lexicon for adjectives. The key point was based on the nature of conjunctive joining the adjectives.

Alongside with that, it was created a sentiment lexicon for English words by Bradley and Lang [23] that developed the Affective Norms for English Words (ANEW) to provide a set of normative emotional ratings for a large number of words in the English language. The goal was to develop a set of verbal materials that have been rated in terms of pleasure, arousal, and dominance to complement the existing International Affective Picture System.

For the classification of positive and negative opinion, Pang, Lee, and Vaithyanathan [24] proposed the idea of Thumbs Up and Thumbs Down. For better problem formalization, there was the necessity of an automated system, which could be employed for electronic documents.

Esuli and Sebastiani [25] created the semantic approach SentiWordNet that relies on the quantitative analysis of the glosses associated to , and on the use of the resulting vectorial term representations for semi-supervised synset classification. Synset classifiers (be them individual classifiers or classifier committees) are ternary classifiers, i.e., they attempt to predict whether a synset is Positive, Negative, or Objective.

The machine learning classification approach usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines (SVM), or Neural Networks.

The Naïve Bayes classification is a family of probabilistic algorithms that uses Bayes's Theorem to predict the category of a text. An SVM is a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. On the other hand, logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). These examples are mapped so that the examples of the different categories (sentiments) belong to distinct regions of that space. Then, new texts are mapped onto that same space and predicted to belong to a category based on which region they fall into.

Preethi, Uma, and Kumar [18] introduced a prediction model based on temporal sentiment analysis to identify the causal relation between the events and uses it to predict the event sentiment and duration between the events. The work used SVM for sentiment classification and the causal relation is found using support and confidence.

Pang, Lee, and Vaithyanathan [24] applied supervised machine learning methods (Naïve Bayes, maximum entropy classification and support vector machine) for sentiment classification and evaluated its effectiveness in movie domain and concluded that SVM method

showed better performance compared to other methods while Naïve Bayes showed the worst performance.

Mullen and Collier [26] proposed an SVM based sentiment classification method that assigned values to selected phrases and words then used a technique for bringing them together to create a model for classification of texts. For these classification models there are processes that must be implemented in order to create the SA algorithm: Kaur and Gupta [17] describes five different processes the Process of Sentiment Analysis for Text (Lexicon Generation), Subjectivity Detection, Sentiment polarity Detection, Sentiment Structuration, Sentiment Summarization-Visualization-Tracking.

Certainly, in general there is the training process, which the model can learn to associate a particular input (i.e. a text) to the corresponding output (a tag) based on a defined interpretation criteria. Then with a feature extractor technique it transfers the text input into a feature vector. Pairs of feature vectors and tags (e.g. positive, negative, or neutral) are fed into the machine learning algorithm to generate a model that can be considered as the lexicon basis.

Springer Science [27] Feature Extraction aims to reduce the number of features in a data set by creating new features from the existing ones (and then discarding the original features). From building derived values (features) it intends to be informative and non-redundant, facilitating the subsequent learning and generalization step.

At the end, the first step in a machine learning text classifier is to transform the text into a numerical representation, usually a vector. Where each component of the vector represents the polarity of a word or expression in a predefined dictionary (e.g. a lexicon of polarized words). This process is known as feature extraction or text vectorization and the classical approach would be bag-of-words or bag-of-n-grams with their polarity.

More recently, new feature extraction techniques have been applied based on word embedding (also known as word vectors). This kind of representations makes it possible for words with similar meaning to have a similar representation, which can improve the performance of classifiers.

Mohammad, Dunne, and Dorr [28] proposed a technique to increase the scope of sentiment lexicon. It includes the identification of individual words as well as multi-word expressions with the support of a thesaurus and a list of affixes.

Also, there is the rule-based approach a system that perform SA based on a set of manually crafted rules. This can be defined as a set of rules in a scripting language that identify subjectivity, polarity, or the subject of an opinion. The rules may use a variety of inputs, such as: Classic NLP techniques: like stemming, tokenization, part of speech tagging and parsing; and other resources, such as lists of words and expressions (i.e. lexicons).

Yang and Shih [29] created a rule-based sentiment analysis technique that employs the class association rule mining algorithm to automatically discover interesting effective rules

capable of extracting product features or opinion sentences for a specific product feature interested.

A very basic example of a rule-based implementation would be the following:

1. Define two lists of polarized words (e.g. negative words such as bad, worst, ugly, etc. and positive words such as good, best, beautiful, etc.);
2. Given a text:
   (a) Count the number of positive words that appear in the text;
   (b) Count the number of negative words that appear in the text.
3. If the number of positive word appearances is greater than the number of negative word appearances return a positive sentiment, conversely, return a negative sentiment. Otherwise, return neutral.

Denecke [30] introduced the use of SentiWordNet in terms of prior polarity scores. The author proposed two methods: rule-based and machine learning based. The Accuracy of rule-based was 74% which was less than 82% accuracy of machine learning based. Finally, was concluded that there were needed more sophisticated techniques of NLP for better accuracy.

Those sophisticated techniques may appear as the hybrid systems approaches, these usually combines both rules based and machine learning approaches then by combining both it can improve the accuracy and precision of the technique.

A hybrid approach was developed by Hutto and Gilbert [31] the VADER, a combined lexicon and rule-based sentiment analytic software. VADER uses a combination of qualitative and quantitative methods, that were first constructed and empirically validated as a gold standard list of lexical features (along with their associated sentiment intensity measures) which are specifically attuned to sentiment in micro blog-like contexts. VADER is capable of both detecting the polarity (positive, neutral, negative) and the sentiment intensity in text [31].

2.2.2 Applications

It is estimated that 80% of the world's data is unstructured and not organized in a predefined manner [32]. Most of this comes from text data, like emails, chats, social media, surveys, articles and documents. These texts are usually difficult, time-consuming and expensive to analyze, understand, and to sort through. With SA techniques it is possible to interpret these opinion data in an efficient way that was never scalable possible.

Getting information from unstructured data nowadays is a great way to explore reactions from a business or a brand to support decision making. With online sentiment tools,

businesses are able to monitor consumer sentiments and understand their customer's needs by automatically assigning sentiment categories to their data.

Coyne, Smit, and Güner [33] investigated if sentimental analysis was feasible for the classification of product reviews from Amazon.com. The performance of three different algorithms were compared to determine the most accurate classification with the best results for reviews on furniture products.

On Minqing Hu and Bing Liu [20] research developed a study where they analyze customer reviews of a particular product, involving three sub-tasks: (1) identifying features of the product that customers have expressed their opinions on (called product features); (2) for each feature, identifying review sentences that give positive or negative opinions; and (3) producing a summary using the discovered information.

Social networks are a good source of data for SA, Twitter is one example of a know used data source. Twitter is a social network that has available a huge volume of unstructured data. This data involves every topic being talked around the world with constant people's reactions. The opinions expressed by online users are considered user-generated content about some topic.

Tumasjan, Sprenger, Sandner, et al. [34] used Twitter data for SA using Linguistic Inquiry and Word Count text analysis software, that conducted a content analysis of over 100,000 messages containing a reference to either a political party or a politician. The results demonstrated close correspondence to the parties and politicians' political positions indicating that the content of Twitter messages plausibly reflects the offline political landscape.

Also, O'Connor, Balasubramanyan, Routledge, et al. [35] analyzed several surveys on consumer confidence and political opinion over the 2008 to 2009 period to find how they were correlated to sentiment word frequencies in contemporaneous Twitter messages. The results highlighted the potential of text streams as a substitute and supplement for traditional polling.

HUANG [36] demonstrated on his research the superiority of SVM in forecasting weekly movement directions of the NIKKEI 225 index.

There are other social networks that are gaining dimension online. Which means a growth of user-generated content available and more data to analyze. One that is getting great relevance is the network Reddit. Reddit is a network of communities based on people's interests. A user can share content by posting statements, links, images, and videos related to topics and then it can be commented by other users from the community. Allowing people to communicate with others in a form of threaded conversations.

Lubitz [37] measured whether the sentiment in financial news articles posted on Reddit can predict future stock market returns. Evaluated the predictive power of Reddit by comparing the sentiment of all submitted articles posted on a trading day to the corresponding

closing value of the S&P 500 stock index. The result suggested that the predictive power of Reddit is slightly better than using standard newspaper analysis.

Mudinas, Zhang, and Levene [38] investigated the potential of using sentiment attitudes (positive vs negative) also sentiment emotions (joy, sadness, etc.) extracted from financial news or tweets to help predict stock price movements. The experiment results indicated that the interaction between sentiment and price is complex and dynamic: while some stocks in some time periods exhibited strong cross-correlation, it was absent in other cases.

On Salac [39] research he mentions that there is a clear research gap on utilizing Reddit as forecasting source through sentiment analysis. No adequate research could be found on the particular combination of comparing Reddit posts for sentiment analysis and prediction in regard to cryptocurrencies. It shows there are not much research's developments on using Reddit as data source for Sentiment Analysis to predict cryptocurrencies.

There are limitations which need to be taking into concern when doing Sentiment Analysis, factors that can cause challenges to the accuracy of the sentiment classifiers.

For Hussein [40] there are several challenges faced the sentiment analysis and evaluation process. These challenges become obstacles in analyzing the accurate meaning of sentiments and detecting the suitable sentiment polarity.

As such for example, the sarcasm challenge Liu [19] stated that sarcastic sentences with or without sentiment words are hard to deal with e.g., "What a great car! It stopped working in two days.". In sarcastic text, people express their negative sentiments using positive words. This leads to Sentiment Analysis to misunderstand the meaning of the opinion. It occurs most often in user-generated content such as Facebook comments, tweets, etc.

But as human language, sarcasm can be hard for a human to understand as well as for a machine. Common topics, interests, and historical information must be shared between two people to make sarcasm understood.

The polarity of an expression can be a challenge as well, considering the Pang and Lee [16] example "Jane Austen's books madden me so that I can't conceal my frenzy from the reader". Just as the topic of this text segment, the presence of words like "madden" and "frenzy" suggests negative sentiment, but the statement is indicating the opposite, a positive sentiment about the Jane Austen author.

Dhuria [41] identified the contextual information as a challenge: the actual sense of the text varies from domain to domain; this property is referred as contextual property. So, based on the context, the behavior of the word changes.

To address the context issue, the machines cannot learn about contexts if they are not mentioned explicitly. Creating inputs to a model that recognize context, tone, and previous

indications of sentiment can help increase accuracy and get a better overall sense of what the author is trying to express.

Defining statements with implied expressions can be a challenge for SA classification. Liu [19] says sentences without sentiment words can also imply opinions. Many of these sentences are objective sentences that are used to express some factual information. The sentence “This washer uses a lot of water” implies a negative sentiment about the washer since it uses a lot of resource (water). But it has no associate sentiment words which can mislead the SA to neutral classification.

It also shows, that defining neutral polarity is another challenge to tackle in order to perform accurate Sentiment Analysis. As in all classification problems, defining the categories - and, in this case, the neutral tag - can be one important part of the problem.

Chapter 3

Related Work

The literature review will mainly focus on two domains: Forecasting Cryptocurrency and Reddit as source of information. As a result, it will allow to identify the state of the art that highlights the current state of knowledge on the subjects addressed in the dissertation. The accomplishment of a literature review protocol is a critical point for the investigation development in order to systematically achieve the intended objectives. The scoping review protocol was the methodology used on the literature review.

3.1 Forecasting cryptocurrency

Hyndman and Athanasopoulos [42] define forecasting as a common statistical task in business, where it helps to inform decisions about the scheduling of production, transportation and personnel, moreover, it provides a guide to long-term strategic planning. It is about predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts.

Alicia Tuovila [43] describes forecasting as a technique that uses historical data as input to make informed estimates that are predictive in determining the direction of future trends. Businesses use forecasting to determine how to allocate their budgets or plan for anticipated expenses for an upcoming period of time. This is typically based on the projected demand for the goods and services offered. Furthermore, Alicia Tuovila [43] complements that investors use forecasting to determine if events affecting a company, such as sales expectations, will increase or decrease the price of shares in that company. Forecasting also provides an important benchmark for firms, which need a long-term perspective of operations.

Bitcoin is an online currency that is used worldwide to make online payments [44]. It has consequently become an investment vehicle in itself and is traded in a way similar to other open currencies. The ability to predict the price fluctuation of Bitcoin would therefore facilitate future investment and payment decisions. Kim, Lee, Park, et al. [44] state

that some recent researches had focused on the characteristics of Bitcoin online forums. People who share common interests tend to post comments concerning certain topics on online forums. Bitcoin is mostly traded on the web with many users making buying/selling decisions based on information acquired on the Internet. Therefore, it is possible to observe how users respond to daily Bitcoin price fluctuations, and to identify or predict future fluctuations in the Bitcoin price and trade volume.

Forecasting cryptocurrency is reported in a number of recent studies. Tetlock [45] was the first to measure the interactions between (mass) media reporting and financial market movements [46].

Kim, Lee, Park, et al. [44] stated that research on cryptocurrencies, particularly on Bitcoin, has been extensively conducted from diverse perspectives, e.g. the analysis of user sentiment as manifested by social media including Twitter. Matta, Lunesu, and Marchesi [47] studied whether social media activity or information extracted by web search media could be helpful and used by investment professionals in Bitcoins. And presented an analysis of a corpus of tweets about Bitcoin, considering a total amount of 1.9 million tweets. Then applied automated Sentiment Analysis on these tweets in order to evaluate whether public sentiment could be used to predict Bitcoin's market.

Stenqvist and Lönnö [48] studied if sentiment analysis on Twitter data, relating to Bitcoin, can serve as a predictive basis to indicate if Bitcoin price will rise or fall. Developing a model for forecasting Bitcoin price evaluation based on Twitter posts over a time frame of 31 days, gathering 22 million tweets and subsequently analyzing them using the VADER lexicon method. Stating that a naive prediction model was presented, based on the intensity of sentiment fluctuations from one time interval to the next. The model showed that the most accurate aggregated time to make predictions over was 1 hour, indicating a Bitcoin price change 4 hours into the future [48]. Thus, the primary conclusion is that even though the presented prediction model yielded a 83% accuracy, the number of predictions were so few that venturing into prediction model conclusions would be unfounded.

Burnie and Yilmaz [49] created a new Data-Driven Phasic Word Identification methodology to determine which words matter as the bitcoin pricing dynamic changes from one phase to another which found out that Reddit submissions are both correlated with Google and have a comparable relationship with a variety of bitcoin metrics, using Spearman's rho. The article established that Reddit submissions are similar to Google searches in capturing Internet activity but are an improvement in providing greater textual content for more in-depth analyses [49]. The authors claim that Reddit submissions' strong correlation with Google searches suggests that either users from Reddit and Google searchers react similarly to external events or the same people are interacting with one website and then the other. Reddit submissions had a comparable relationship with a variety of bitcoin metrics [49].

More recently, Salac [39] in his study evaluated the possibility of using data from social

media sites to run sentiment-analysis-based predictions on the Bitcoin price developments. In contrast to preexisting literature, it also aimed to compare the feasibility of Reddit data in comparison to the current benchmark source of sentiment data, namely Twitter. It evaluated using the VADER sentiment analysis and compared in different time intervals to historical Bitcoin price developments over the course of three months.

3.2 Reddit as a source of information

From forums to mainstream social media channels like Facebook, Twitter, Pinterest, Instagram, and others, cryptocurrency is a trending subject. The relationship between social media and cryptocurrency continues to evolve in new and thrilling ways.

Reddit is one of the most known platform of news aggregation and discussion forums in the world, it was created in 2005 by Steve Huffman and Alexis Ohanian, it has a monthly average of 430 million active users, which makes it the number 6 most-visited site in the United States and the number 21 in the world, indicates the information on the about page. It is continuously growing as an online platform of shared content between users. It is divided into more than a million communities known as subreddits, each one covers a different topic. The name of a subreddit begins with /r/, which is part of the URLs that Reddit uses. For example, /r/NBA is a subreddit where people talk about the National Basketball Association, while /r/Bitcoin is a subreddit for people to discuss about Bitcoin.

The community of Bitcoin was created in 9th of September of 2010 and it held 1.4 million members, this community has a daily post where users can discuss, share knowledge, answer and ask any question about Bitcoin. From crypto-enthusiasts to crypto-analysts, they can all be sharing insights on Reddit.

It can be recognized the amount of information it results from Reddit activity, the data generated from users comments on the social network permits new ways of data analysis. The Gómez, Kaltenbrunner, and López [50] study was one of the firsts researches using social networks and discussion threads as a source of information for analysis. On their study, it was analyzed the emerging from user comments activity on the website Slashdot. Using Kolmogorov-Smirnov statistical tests, they showed that the degree distributions are better explained by log-normal instead of power-law distributions [50]. Also, the investigation exhibited some special features that deviate from traditional social networks: neutral mixing by degree, almost identical in and out degree distributions, only moderated reciprocity. They conjectured that most of the reactions in Slashdot arise when high diversity in opinions occur and users are therefore more inclined to be linked to people who express different points of view [50].

On Weninger, Zhu, and Han [1] research, it was investigated the dynamics of Reddit discussions threads asking two main questions: firstly, to what extent do discussion threads resemble a topical hierarchy? And secondly, can discussion threads be used to enhance

web search? It was then shown interesting results for these questions on a very large snapshot of several sub-communities on Reddit. Discussing the implications of the results and suggest ways by which social news Web site's can be used to perform other tasks [1]. Glenski and Weninger [51] studied user interactions such as likes, votes, clicks, and views are assumed to be a proxy of a content's quality, popularity, or news-worthiness. In their paper the research question was: how predictable are the interactions of a user on social media? To answer the question they recorded the clicking, browsing, and voting behavior of 186 Reddit users over a year [51]. Presenting interesting descriptive statistics about their combined 339.270 interactions and finding that relatively simple models are able to predict users individual browse or vote interactions with reasonable accuracy [51].

Thukral, Meisheri, Kataria, et al. [52] created methods to systematically analyze individual and group behavioral patterns observed in community driven discussion platforms like Reddit where users exchange information and views on various topics of current interest. They conducted the study by analyzing the statistical behavior of posts and modeling user interactions around them. They have chosen Reddit as an example, since it has grown exponentially from a small community to one of the biggest social network platforms in the recent times [52].

3.3 Summary of related work

In this research, it was recognized that the first authors to develop an empirical method of building sentiment lexicon in order to use Sentiment Analysis were Hatzivassiloglou and McKeown [22]. As it was seen there are plenty of other researches defining the concept of SA, most of the papers were published after 2004. The studies of Pang and Lee [16], Liu [19] and Preethi, Uma, and Kumar [18] are considered the fundamentals ones when it concerns the Sentiment Analysis approach.

The research of Hutto and Gilbert [31] comes with the creation of the algorithm VADER. This algorithm was the one considered to be used on this study, it is going to be the chosen method for the use of Sentiment Analysis. One of the main reason for this decision was that it demonstrates remarkable results when analyzing sentiment of social media content.

The first digital currency created was Bitcoin which established the concept of blockchain technology for the first time, stating that "the network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work." [6]. The identity of the person who has created Bitcoin is still unknown, but his/her creation was a huge development of blockchain technology. Since the technology behind Bitcoin is open-source and its design is public access, it was through Bitcoin that other applications of blockchain came to practice. Moreover, there are cryptocurrencies which were created based on Bitcoin, such as the example of

Dash cryptocurrency developed by Evan Duffield in 2013 when he discovered key weaknesses in the Bitcoin technology.

Nowadays, there are a considerable number of researches on using Sentiment Analysis to forecast cryptocurrency. The regarding Table 3.1 describes the features of the relevant related work in comparison with the actual research:

The research of Salač is considered the work that best relate with the current study, as it uses the algorithm VADER to analyze the sentiment of Twitter and Reddit content related to Bitcoin. Afterwards, Salac [39] evaluates the link between social media sentiments and forecasted value of cryptocurrencies. It is an interesting research reference although it differs from the current investigation considering the defined main objective, as it is to identify if the daily sentiment of Bitcoin on Reddit can relate with the market value of Bitcoin. Additionally, during this study the data which are going to be analyzed are only the comments from the daily discussion publications of Reddit Bitcoin community demonstrating different niche and specific focus when compared with the Salač research.

Research	Features	Comparison
Kim, Lee, Park, et al. [44]	**Data source:** Bitcoin-related online forum **SA approach:** To create a user opinion analysis based on Bitcoin online forum data **Objective:** Extract keywords from Bitcoin-related user comments posted on the online forum with the aim of analytically predicting the price and extent of transaction fluctuation of the currency	Both data source and SA approach are different, however the identical objective is to analyze the keywords from Bitcoin related user comments in order to predict the fluctuation of the currency
Stenqvist and Lönnö [48]	**Data source:** Twitter social network **SA approach:** VADER Sentiment Analysis **Objective:** To analyze if there is a correlation between Twitter sentiment and Bitcoin price fluctuation	The SA approach used and the objective are the same, but the analysis is based on a different data source
Karalevicius [46]	**Data source:** Database of Bitcoin relative news articles as well as blog posts **SA approach:** A lexicon-based sentiment analysis on the document level **Objective:** To measure the interaction between media sentiment and the Bitcoin price	The research has the identical objective as it aims to identify if the value of Bitcoin can be determined by sentiment
Burnie and Yilmaz [49]	**Data source:** Reddit and Google search data **SA approach:** Developed a new Data-Driven Phasic Word Identification methodology **Objective:** To determine what was being discussed on social media during the phase of falling prices compared with the phases before (rising prices) and after (stable prices) in order to delineate significant words and the context in which they are being used	The data source is partly the same and the objective has some similarities in the way it determines words that are most significant in Bitcoin context
Salac [39]	**Data source:** Reddit and Twitter social networks **SA approach:** VADER Sentiment Analysis **Objective:** To use data from social media sites to run sentiment-analysis-based predictions on the Bitcoin price developments, also it aims to compare the feasibility of Reddit data in comparison to other source of sentiment data, Twitter	Both data source and SA approach are going to be the same, but the objective is slightly different as it is not a daily analysis of Bitcoin price

Table 3.1: Overview of related work features

Chapter 4

Experiments and Results

This chapter outlines the different methods used to analyze the sentiment comments from the daily discussion publications on the Bitcoin Reddit community. First of all, it begins by describing the process of collecting data from Reddit. It was performed by using a Application Programming Interfaces (API) in order to extract the contents. Secondly, the data were cleansed depending on the existent inconsistencies of the users comments.

Additionally, the data were analyzed through applying a sentiment analysis approach as the goal is to obtain a polarity classification based on the expressed users comments. Lastly, it was demonstrated the correlation between the sentiment analysis results and the Bitcoin information using a heat map to compare the correlation of the different variables.

4.1 Data collection

Following the scope of the investigation, the object of the study is the users comments from the social network Reddit, more precisely the comments on the daily discussion posts from the subreddit Bitcoin community, so called “r/Bitcoin”. The process of data collection was made through the Jupyter Notebook using Python programming language to create scripts for the intended objectives.

The period analysis of the investigation begins on 02/01/2018 until 29/02/2020, representing the time frame of the data collection and analysis. The daily discussion posts are identified on the subreddit r/Bitcoin having the title of “Daily Discussion, [Month] [Day], [Year]” meaning Month, Day and Year are the correspondent actual date of the post. The initial process for collecting the respective data starts by using the Python Reddit API Wrapper (PRAW)[1] and the pushshift.io Reddit API[2]. While pushshfit provides full functionality for the search of Reddit data, it was created a simple function to have a filter in order to search

[1] PRAW, is a python package that allows simple access to Reddit’s API content

[2] Pushshift.io was designed and created by the /r/datasets mod team to help provide enhanced functionality and search capabilities for searching Reddit comments and submissions

for a contained term on the title field of a post that will return a list of PRAW submission objects (posts) during a particular period from a defined subreddit.

To identify and extract the daily discussion posts from the subreddit Bitcoin into the period analysis, the function was used to collect the identification numbers of the posts. Resulting on a total of 665 publications to further collect associated comments. It was then created an array with the posts identification number.

The Submission function model of PRAW initializes with an id of a Reddit publication serving as an input and returns the objects comment contained in the publication. The array of identification numbers were iterated using this function to collect all the associated comments with the identification posts. An example of users comments on a Daily discussion thread from Reddit is shown in the following Figure 4.1.

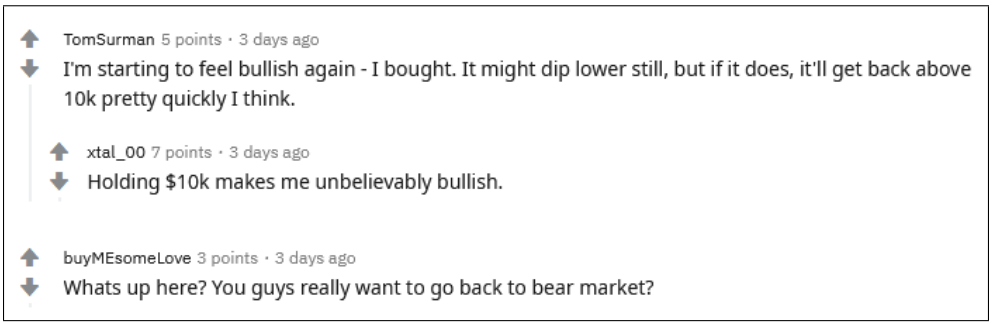

Figure 4.1: Reddit users comments thread example

After applying the Submission function it was retrieved the contained comments on the daily discussion posts. The function allowed to collect data as it returns the attributes that characterize the object comment. These comments adhere a pre-defined data model, the proprieties of the data constitutes a structured data set that conforms to a tabular format in relationship between the different rows and columns. On Table 4.1 it is represented the data model for each comment:

The collected data comments were inserted into a data frame in which the rows represents the comments to be analyzed. The date of the post identifies the correspondent day of the users comments. Table 4.1 presents the object of study which is the field Body. The field contains the expressed user comment, which means the data text that is going to be applied in the Sentiment Analysis.

The historical Bitcoin data value information was collected from CoinMarketCap website which is a price-tracking platform for the cryptocurrencies market. While the information of the number of Bitcoin transactions were provided by Bitcoinity, which gives exchange data and has aggregated information of the top cryptocurrency exchanges of Bitcoin. Shortly introducing about the sources of information, CoinMarketCap was founded by Brandon Chez in May 2013 and has quickly grown to become the most trusted source by Bitcoin enthusiasts, institutions, and media. It is mainly used to compare thousands of

Field	Description
Id	The unique identification number of the comment
Author	Designates the user of the comment
Body	Contains the expressed comment of the author
Score	Represents the number of up votes or down votes of the comment
Timestamp	Indicates the exact time the comment was published, in timestamp format
Parent_id	Corresponds to the id of a comment reply where, through it, it is possible to identify the "mother" comment and its successive responses

Table 4.1: users comments data model structure

cryptocurrencies which are commonly cited by CNBC, Bloomberg, and other major news outlets. Bitcoinity was created by Kacper Cieśla, he aggregates cryptocurrency market information by fetching the data directly from the exchanges through their APIs providing graphics and raw data to be extracted and analyzed.

Furthermore, the Bitcoin market data information was collected using the same period of analysis, then it was extracted from the websites CoinMarketCap and Bitcoinity data of the number of transactions and the value by day of Bitcoin in US dollars currency. Finally, it enabled to create a data frame containing the daily Bitcoin market information having the represented structure of the Table 4.2.

Field	Description
Date	The identification of the day of the Bitcoin information
Number of transactions	The total number of transactions
Open	The Bitcoin open price
High	The Bitcoin highest price reached
Low	The Bitcoin lowest price
Close	The Bitcoin closed price

Table 4.2: Bitcoin market information data structure

As a result two data sets were created, one data set with Bitcoin market information and another one with the generated comments by users on the daily discussion posts from Bitcoin community of Reddit.

4.2 Data cleansing

The data cleansing is a crucial process to ensure that the data is correct, consistent and usable for analysis. One of the methods to guarantee cleansed data is to remove or modify data that are incorrect, incomplete or irrelevant. Irrelevant data correspond to noticed observations that do not fit into the specific problem that is going to be analyzed.

To obtain the comments data sets with users reactions, some of them had to be cleansed because they were considered as irrelevant for the analysis. There are users comments containing unwelcome content or offensive language that the Reddit comments policy remove and replace the content with DELETED or REMOVED quotes.

When this kind of situation was observed, the comments which had been removed and deleted previously from the Reddit comments policy were cleansed from the data set, as so the comments which contain the text “[DELETED]” and “[REMOVED]” were deleted. It was removed 10,675 irrelevant comments.

It was verified that the Bitcoin market information collected data did not contained any missing or incoherent value meaning that it was not needed to clean the data after the extraction.

4.3 Data analysis

The data collection and cleansing processes are now completed which means the data is ready to be analyzed. As referred previously, the period date between January 02, 2018 and February 29, 2020 represents the period of analysis. With regard to the section 4.1, the content generated by users comments on the daily discussion posts of Bitcoin community and the historical market information of Bitcoin are the two data sets for data analysis. The first data set is going to be evaluated by the VADER Sentiment Analysis tool, resulting on a output score classification for each comment analyzed. After having the SA output, it will be verified the correlation between the first data set and the second one, being the date as the key reference between each other.

After a brief analysis of the data, it was observed the following number of comments by year:

Year	Number of comments	Avg comments per day
2018	119,748	333
2019	63,439	174
2020 (Jan-Feb)	8,115	135

Table 4.3: Number of comments by year

It can be seen that on the analysis period there were a total of 191,302 comments to be analyzed, as the Table 4.3 indicates that the year of 2018 was the year with most comments

on the daily discussion posts of Bitcoin community, holding a total of 119,748 comments which contrasts with the year after 2019, that had half of the number of comments. It is also necessary to remind that is was on the year of 2018 that the price of Bitcoin reached the highest evaluated price ever, specifically it was on the 6th of January the highest day of Bitcoin with a closed price of $17,527 each. On the other hand was on the 15th December of 2018 that Bitcoin registered the lowest market price with a value of $3,236.

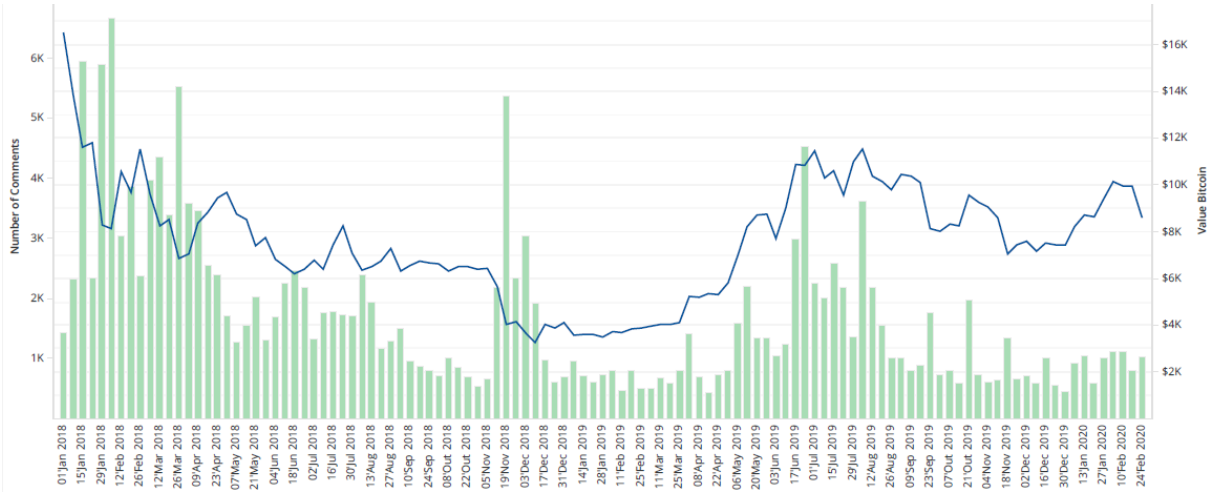

Figure 4.2: Number of comments and Bitcoin value evolution

As it can be observed on Figure 4.2 the top four months with the most comments were between January 2018 and April 2018 counting more than 12,000 of users comments on each month.

The week with most comments was the 6th week of the year 2018 with 6,661 comments having a correspondent closed price value on the week of $8,129 a Bitcoin. The following weeks of 01'Jan 2018 and 08'Jan 2018 have seen the highest value of Bitcoin, respectively closed the week on $16,477 and $13,772 for a Bitcoin.

4.3.1 Using sentiment analysis

The literature review permitted to understand that VADER is a python library built especially for sentiment analysis of social media texts. VADER-Sentiment-Analysis was the tool selected for the sentiment analysis as it demonstrated optimized results when using social media type text. The tool was applied to evaluate each generated comment by the users on the daily discussion publications allowing to create a score classification with the sentiment polarity of the expressed sentence.

The comments were examined by the tool giving a score to each comment, as so the score represents a result computed by the lexicon and rule-based sentiment analysis tool. It functions by summing the valence scores of each word in the lexicon, adjusted according

to the rules and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive) as the creators Hutto and Gilbert [31] described. Referring also that, accurately the compound score should be called “normalized, weighted composite score”. As demonstrated on the original VADER code [31], the compound score is calculated by a sum of the sentiment arguments passed on the input text. The sum of the sentiment arguments represents a float for the sentiment strength on the input text and the sentiment lexicon defined, where positive values are positive valence and negative value are negative valence. The result is then normalized to be between -1 and 1, which is used an hyper-parameter alpha that approximates the maximum expected value.

In order to classify sentences as either positive, neutral, or negative it was needed to set standardized thresholds for the compound score. The thresholds defined were:

- Positive sentiment: compound score >= 0.05;
- Neutral sentiment: (compound score > -0.05) and (compound score < 0.05);
- Negative sentiment: compound score <= -0.05.

Indeed, each class of the thresholds categorized the users comments into positive, negative and neutral polarity based on the compound score of each expressed sentence.

When looking into the users comments data set, the Table 4.1 demonstrates the data model for each comment. It is the field body containing the text that is going to be the object of Sentiment Analysis classification. The data contained text in the field body which was iterated working as an input for the VADER Sentiment Analysis tool. After the iteration, it gives a Sentiment Analysis result for each comment as demonstrated on the Sentiment Analysis compound result column on the Table 4.4.

The following Table 4.4 describes an example of the application of the Sentiment Analysis tool, on the compound result the “pos, neu, and neg” scores according to Hutto and Gilbert [31] are ratios for the proportions of text that fall in each category as so, these should all add up to be one or close to it with float operation. The compound score represents the final sentiment analysis classification, it is based on the thresholds previously stated that the polarity classification is categorized. These examples demonstrate how the Sentiment Analysis tool performs, the first comment is classified as neutral linked to not express an strong positive or negative emotion. On the other hand, the third comment is rated as negative possibly justified by the user expressing strong negative words like burned, unpopular and bad timing.

Date	Comment	SA compound result	Comment classification
04/03/2018	The more you use the more likely you'll get some right	'neg': 0.0 'neu': 1.0 'pos': 0.0 'compound': 0.0	Neutral
04/03/2018	So basically typical bull market manipulation?	'neg': 0.306 'neu': 0.694 'pos': 0.0 'compound': -0.296	Negative
04/03/2018	I don't want to get burned again, ATM I am very unpopular at work due to some bad timing investing advice. It was actually good investing advice however the timing was off.Now that we are looking at a bull market for the next few years I am wondering if I should start gathering investors for the next run.	'neg': 0.081 'neu': 0.87 'pos': 0.049 'compound': -0.2076	Negative
04/03/2018	Everyone in the world wins, as the people of the world have a currency that increases in value over time, rather than decreases and is electronic (unlike gold) and decentralized.	'neg': 0.0 'neu': 0.816 'pos': 0.184 'compound': 0.7269	Positive
04/03/2018	This isnt true and it gets repeated far too much. If people value it more tomorrow than they did today, the only people who "lose" are the ones who get in late and get fewer bitcoins for their $. They did not lose actual dollars	'neg': 0.048 'neu': 0.857 'pos': 0.095 'compound': 0.3239	Positive
04/03/2018	So going by your analogy, every dollar spent on bitcoin is a loss?	'neg': 0.173 'neu': 0.827 'pos': 0.0 'compound': -0.3182	Negative
04/03/2018	Could still be a bull trap. Careful!	'neg': 0.316 'neu': 0.488 'pos': 0.195 'compound': -0.2481	Negative

Table 4.4: Comment classification

The compound score for the comment is used to determinate whether the expressed comment is positive, neutral or negative. The Sentiment Analysis score of the comments enabled to characterized them by the polarity classification based on the thresholds defined. It was finally obtained the classification for each comment. To summarize the comment classification, the number of comments were summed as the correspondent classification,

aggregating the number of positive, negative and neutral comments by week, month and day.

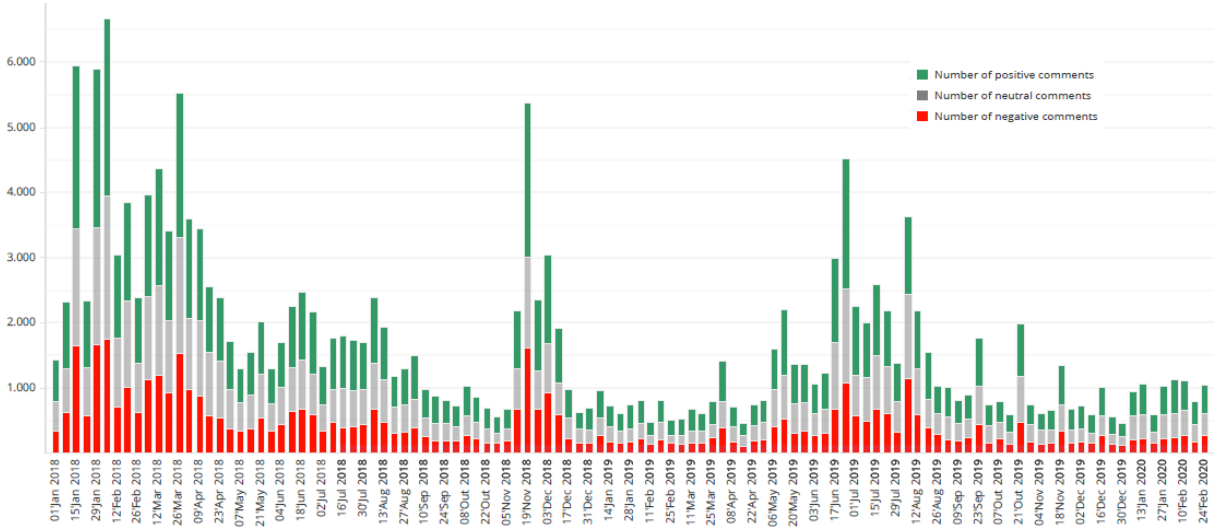

Figure 4.3: Number of users comments by polarity classification

Figure 4.3 highlights the evolution by week of the number of users comments based on the polarity classification from the Sentiment Analysis tool. The Sentiment Analysis demonstrated a considerable number of comments classified as neutral, these comments indicated an impact over the weeks. Most of the weeks had one third of comments classified as neutral. The period between Jan'2018 and Mar'2018 is the period with the most negative and positive comments, in which the 05'Feb week of 2018 faced the most positive and negative comments. It was verified that the most weight of positive comments was on the week of 11'Mar 2019, having half of the comments classified as positive. On the other hand, the week with most weight of negative comments was the 05'Aug 2019 week, indicating that one third of the comments were rated as negative.

VADER Sentiment Analysis score is computed by a lexicon and rule-based sentiment analysis approach. This approach have a determined preassigned score to words and vocabularies, classifying the words as positive or negative. These preassigned scores are based on a pre-trained model that assigns the words classification as human reviewers. For this reason, it can demonstrate some challenges concerning to the sentiment classification, especially for user expressions on social media. The misspellings and grammatical mistakes may cause a limitation for the sentence analysis. On subsection 2.2.2, from the previous researches it was noticed that also the sarcasm and irony can be misinterpreted by a Sentiment Analysis tool, being one of the main challenges for the analysis.

Type of indicator	Indicators
General count	Number of comments
Sentiment Analysis	Number of positive comments
	Number of negative comments
	Number of neutral comments
	Average of the sentiment
	Standard Deviation of the sentiment
	Ratio positive comments on negative
	Percentage of positive comments
	Percentage of negative comments
	Percentage of neutral comments
Sentiment stats	Min sentiment
	25th percentile
	50th percentile
	75th percentile
	Max sentiment
Bitcoin info	Bitcoin value
	Bitcoin value variation
	Number of Bitcoin transactions

Table 4.5: Indicators of the correlation analysis

4.3.2 Correlation analysis

The sentiment analysis classification provided indicators of the users comments. These indicators are going to be focus of a correlation analysis with the historical data of Bitcoin. In order to quantitatively describe the indicators, it was computed descriptive statistics of the users comments. The indicators were summarized by aggregating them into day and week. To perform the summary few calculations were made, in which the number of users comments were summed, the calculation of the average and standard deviation of the sentiment classification was made and it was obtained the minimum, maximum and quartiles of the sentiment classification of the users comments. After, it was calculated the ratio of positive comments on negative by the calculation of the number of positive comments divided by the sum of the number of positive and negative comments.

The summary of the previous calculations resulted on the dimensions day and week. The objective of the analysis is to find correlations between the summarized users comments indicators and the historical Bitcoin data which are represented on Table 4.2. Both data sets were linked by the field date, it allowed to unify the indicators into one data set as the Table 4.5 demonstrates. There are now one data set with three different views contained from the indicators, divided by day, week and month.

In addition, to have different points of view and to provide insights into the possible post or pre correlation effect, different data sets were created where the Bitcoin indicators information was shifted by days in comparison with the users comments indicators. It means that the data was shifted plus and less days among the reference day of the com-

ments. Bitcoin indicators were then shifted for +1, +2, +3 and -1, -2, -3 days between the occurrence of the comments creating seven different tables with the represented data. The views are presented as shown on the Table 4.6, where each data set constitutes each view.

Event	Shifted information
T-n	The event date of Bitcoin information is shifted less n days then the users comments
T-1	The event date of Bitcoin information is shifted less one day then the users comments
T	Bitcoin information have the same event date of users comments
T+1	The event date of Bitcoin information is shifted plus one day then the users comments
T+n	The event date of Bitcoin information is shifted plus n day then the users comments

Table 4.6: Event analysis views

The correlation analysis was verified by using a Heat Map analysis. It allowed to examine the correlation between each variable. The correlation was calculated using Pearsons' standard correlation coefficient. The analysis is represented by a table with the relations between each variable, where each color identifies the intensity of the relation. It places the indicators variables in rows and columns coloring the cells within the table based on its correlation strengths. In which, if the value is close to 1 or -1, the correlation between variables is stronger and if closer to zero, the smaller the correlation will be. If the relation is close to 1, then when one of the variables either increases or decreases the value, the other variable will follow the same trend, on the other hand, if the value is close to -1, it means that when one of the variables increases or decreases the frequency, the other variable will follow the opposite trend. The correlation of the variables by day resulted into the following heat map as represented on the Figure 4.4.

One of the objectives of the study is to analyze if the Bitcoin value, the variation of Bitcoin value and the number of Bitcoin transactions correlate with the indicators of the comments sentiment classification. When analyzing the results of the correlation, as the Figure 4.4 demonstrates, the variables with most intensity of correlation were the Bitcoin transactions with the number of comments on the daily publication which have a positive correlation of 0.80 between them which represent that when the number of transactions increases the number of comments follows the same trend.

The number of positive comments also verified a positive correlation of 0.81 with the indicator of Bitcoin transactions. While looking at the indicator of Bitcoin value, the correlation with the number of the different polarity such as number of positive and negative comments confirmed a low correlation for both variables with a respectively 0.19 and 0.22 of correlation. The average sentiment by day have a negative low correlation of -0.2 with the number of transactions and even lower when comparing with the Bitcoin value.

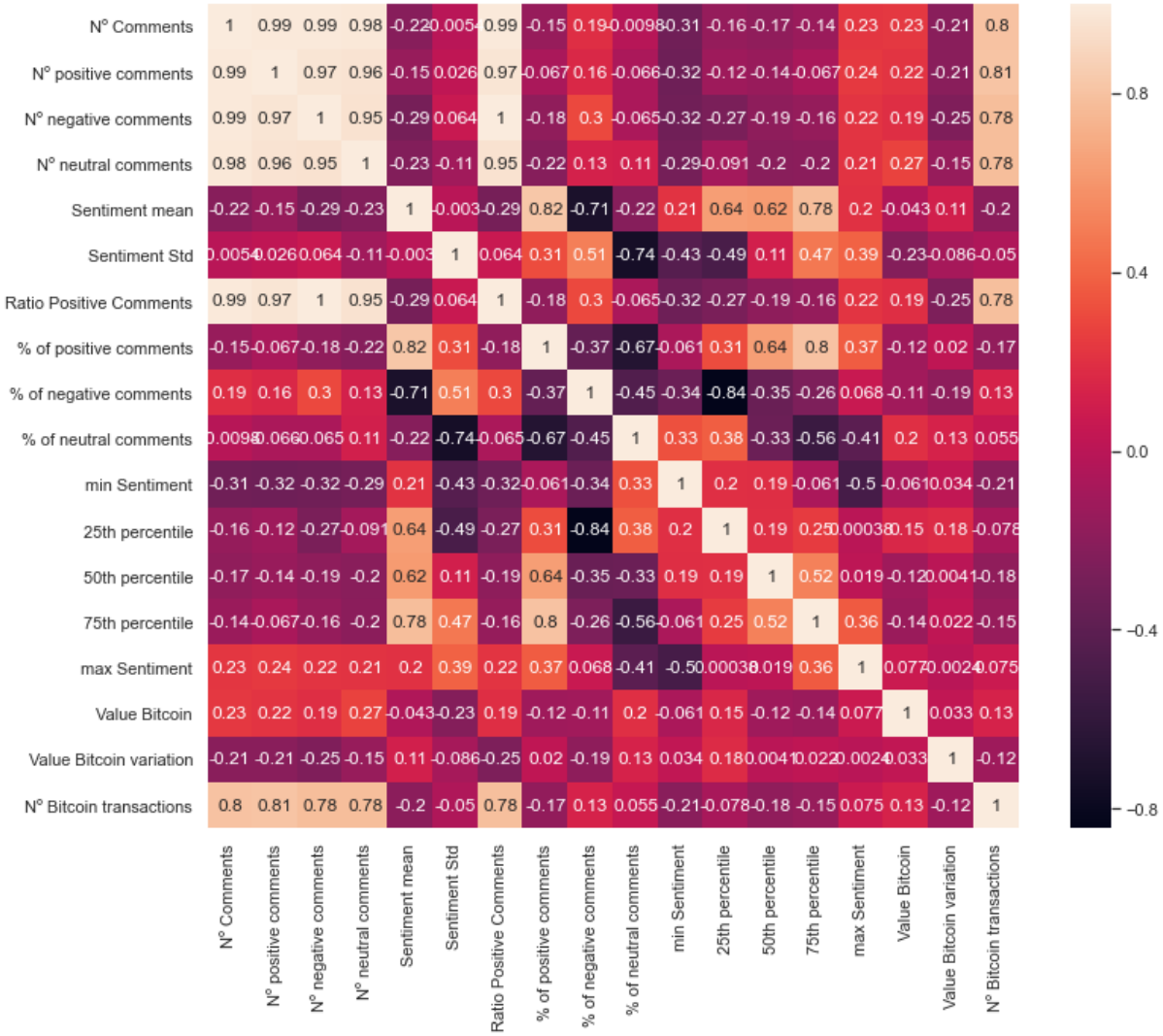

Figure 4.4: Heat map correlation by day

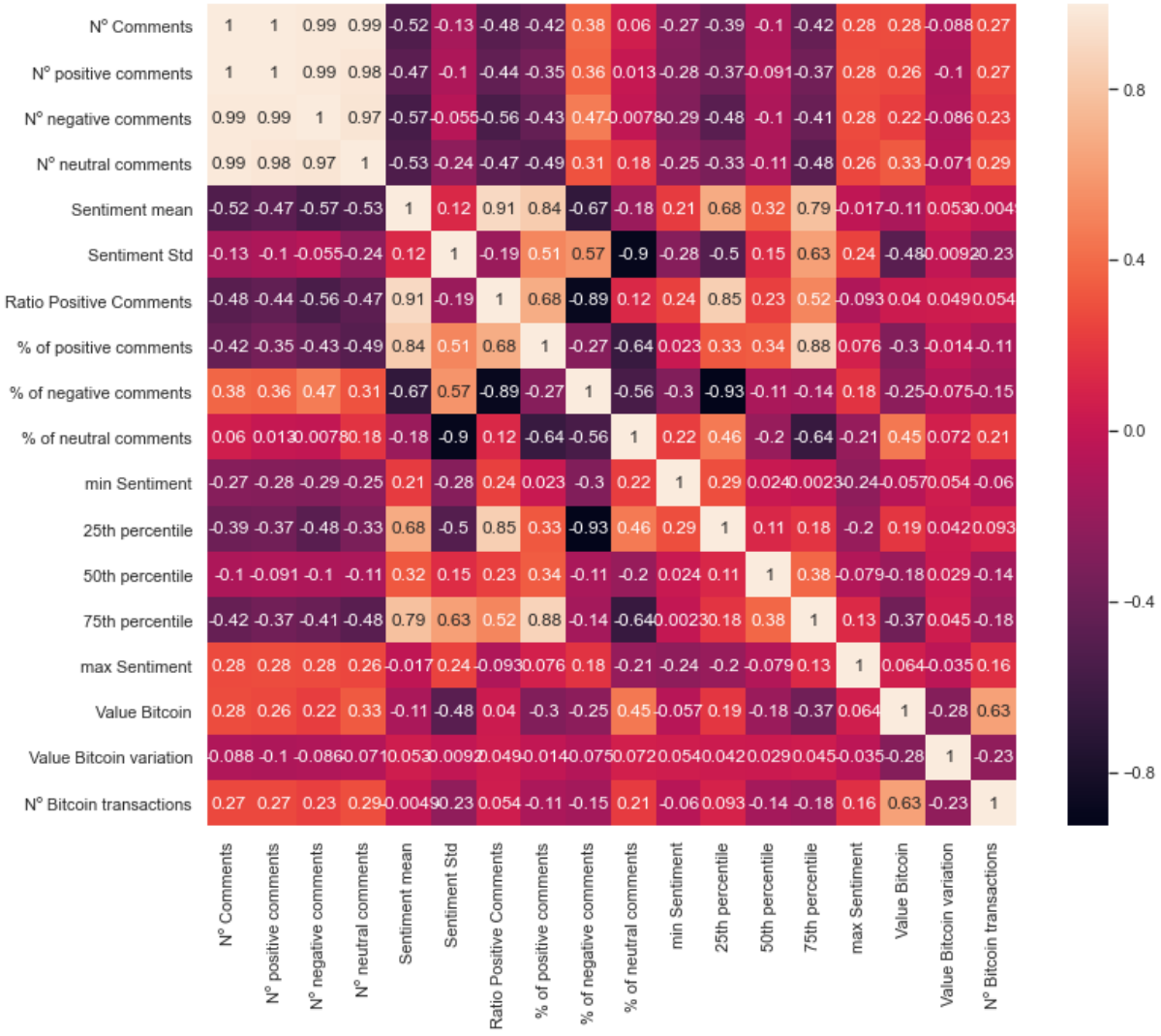

Figure 4.5: Heat map correlation by week

When observing the indicators by week as the Figure 4.5 shows, the sentiment mean verified a negative strength correlation of -0.52 with the number of comments. The indicator of the Bitcoin transactions verified a positive correlation between the Bitcoin value with 0.63 of strength and a low correlation with the numbers of comments having a result of 0.27 correlation weight. The variable Bitcoin value had a low correlation with the sentiment standard deviation of the week having -0.48 of intensity. Also, the percentage of neutral comments demonstrated a moderate correlation of 0.38 with the Bitcoin value.

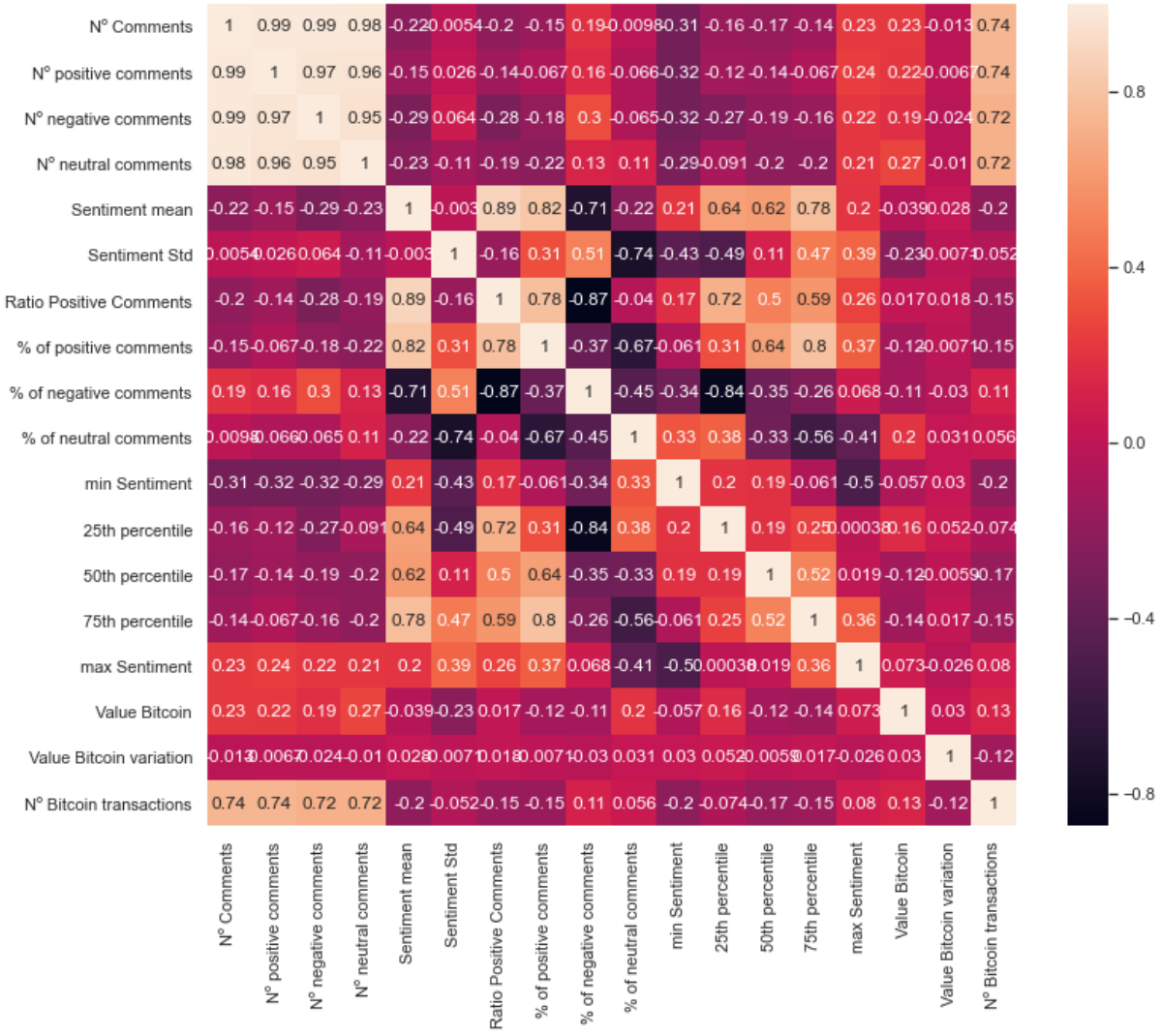

Figure 4.6: Heat map correlation Bitcoin indicators shifted less one day

The view of the Bitcoin shifted information less one day then users comments indicators (e.g. users comments on 4 Jan and Bitcoin value verified on 3 Jan) shows that the number of Bitcoin transactions had a positive correlation of 0.74 with the number of comments of the next day as well as the number of different polarity comments. The number of positive and negative comments verified a low correlation between the variable Bitcoin value having a correlation of 0.22 and 0.19 respectively. Also, with low negative correlation were the variables of the sentiment standard deviation and the Bitcoin value with an intensity of -0.23.

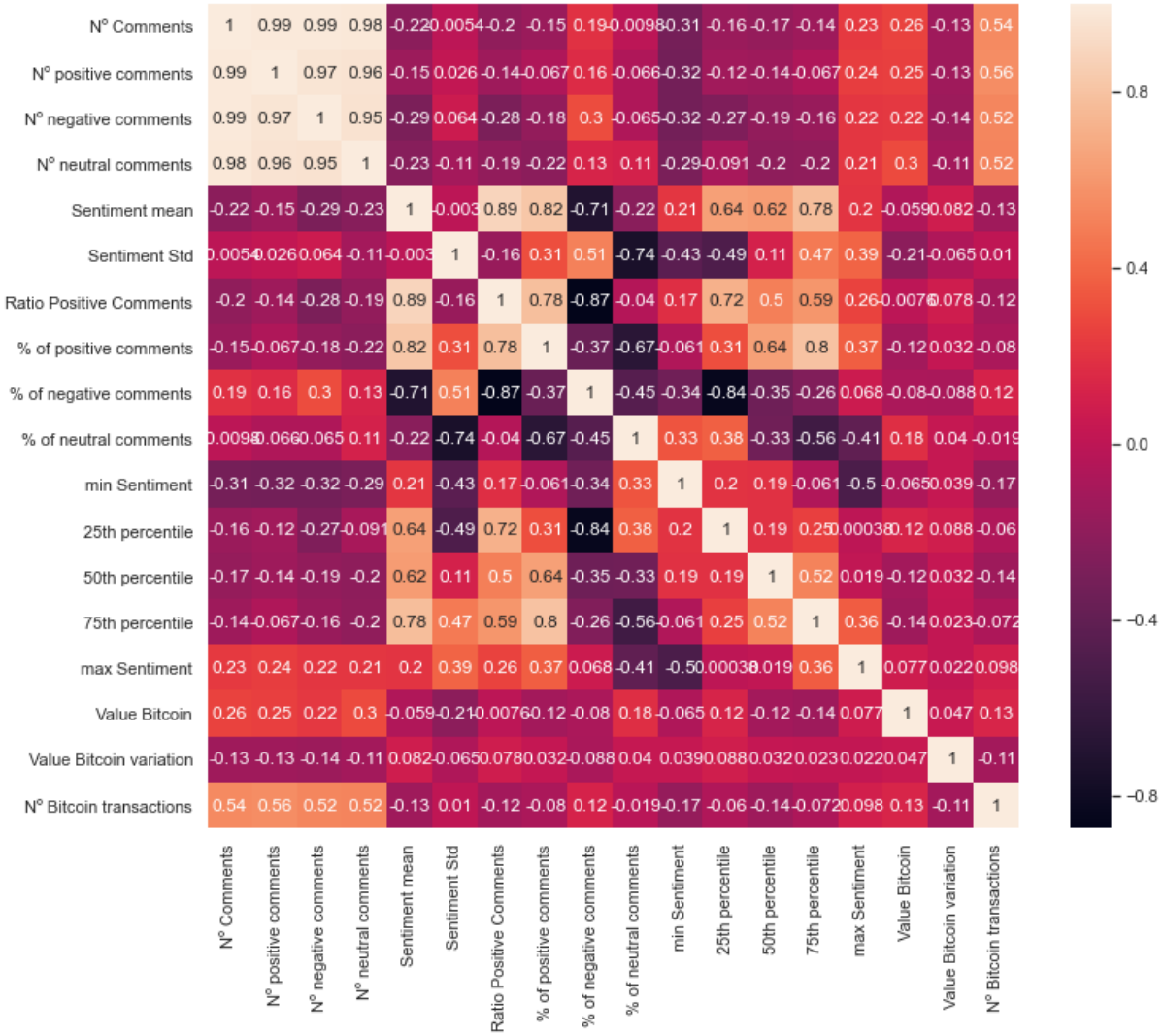

Figure 4.7: Heat map correlation Bitcoin indicators shifted plus one day

Looking now for the view of the Bitcoin shifted information plus one day then users comments indicators (e.g. users comments on 4 Jan and Bitcoin value verified on 5 Jan) indicated that the number of Bitcoin transactions had a positive correlation of 0.54 with the number of comments. The Bitcoin value had a low correlation between the number of comments with all the different polarity having a range of 0.2 and 0.3 intensity between them.

As viewed previously, the indicators with most intensity of correlation were the number of Bitcoin transactions and the number of comments being the most intriguing indicators. To have a better perspective of the correlation effect, an analysis was created focused on the correlation between both indicators, compiling the correlation result between the different views, representing the Figure 4.8.

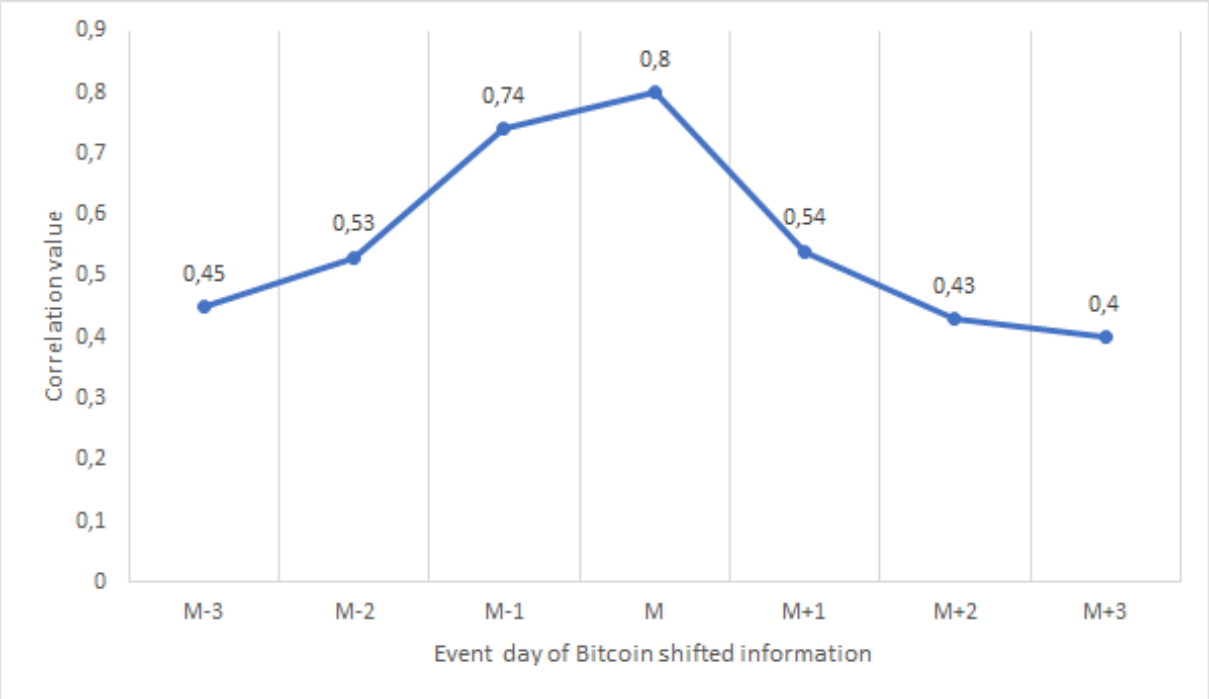

Figure 4.8: Correlation between number of Bitcoin transactions and number of comments

The Figure 4.8 presents the correlation strength between the indicators of the number of Bitcoin transactions and the number of users comments. In which, the axis X represents the different views of the information of Bitcoin transactions being day shifted than the comments actual date, meaning that M is the point where both indicators have the same reference date. It can be seen that the transition of the correlation has more impact when the information of the number of Bitcoin transactions are shifted less days then the number of the comments possible indicating that the number of transactions can have a correlation effect of the increase or decrease of the number of users comments on the daily discussion posts .

Chapter 5

Conclusions and Future Work

This thesis aimed to find correlations between the users comments on the daily discussion posts of the Bitcoin Reddit community and the daily indicators of Bitcoin cryptocurrency. We have used sentiment analysis to extract high level sentiments from the daily discussion posts that provided additional insights about the daily posts.

Chapter 4 presents the results of this study. The investigation started by collecting data from the daily discussion publications among Bitcoin Reddit community, each associated comment of the publications was extracted. It has provided a data set of users comments linked with the daily reactions related on Bitcoin topic. The generated users comments on the Internet may sometimes contain misspell or ironic words which subsequently create difficulties for interpretation, especially in terms of Sentiment Analysis as referred previously in the subsection 4.3.1. This factor was one of the main challenges encountered when Sentiment Analysis was used for sentiment classification of the users comments. Additionally, the Sentiment Analysis tool presented different interpretations when analyzing some words as it may not be interpreted the way it should have had. The words *bullish* and *bearish* which can be connected with respectively a positive and negative sentiment about Bitcoin are examples of words that the Sentiment Analysis tool does not interpret the different associated sentiment.

Attending the research questions of the investigation, the first question is defined as: Do users comments of the daily discussion posts have a correlation with the daily indicators of Bitcoin? It can be seen from the investigation that the daily users comments indicators evidenced the most correlation strength between both number of comments and number of Bitcoin transactions. It highlighted that while using Pearsons' correlation coefficient, the number of comments on the daily posts of Bitcoin Reddit community verified a positive correlation compared with the daily number of transactions of Bitcoin (0.8 of correlation value). On the other hand, it was found out that the daily Bitcoin value did not demonstrated a correlation strength with the daily users comments indicators. While the daily positive, negative and neutral comments or the daily sentiment mean and standard deviation re-

mained indifferent to the correlation of daily Bitcoin value and number of transactions.

When it comes to the second research questions which is: Do users comments of the daily discussion posts have a correlation when the indicators information of Bitcoin are shifted by a few days? It can be said that despite the various attempts, the method used and investigations which were developed along the study in order to answer this research question, the indicators number of Bitcoin transactions and the number of users comments remained the only indicators that holds a correlation. It was not found out any additionally correlation between the Bitcoin shifted information indicators and the users comments as when the Bitcoin information indicators were shifted plus or less days towards the users comments indicators did not demonstrated correlations.

To sustain the demonstrated results on the investigation it would be complementary to examine the same intended data set for analysis but while using a different Sentiment Analysis tool for sentiment classification, to then obtain the correlation values and compare them with the current investigation. The goal would be to analyze if the indicators establish the same correlation strength values.

For further research, it might be interesting to use additional sources of information as a context complement of the users comments, with the objective of linking the comments along with Bitcoin related events as it would increase the input information which could contribute for the sentiment analysis. Perhaps, collecting available news online related with Bitcoin and including it to the analysis could possibly create a better interpretation of a user comment.

Bibliography

[1] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. “An exploration of discussion threads in social news sites: A case study of the Reddit community”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013* (2013), pp. 579–583. DOI: 10.1145/2492517.2492646.

[2] Margaret Rouse. “Cryptography”. In: *Techtarget* (2014). URL: https://searchsecurity.techtarget.com/definition/cryptography.

[3] Zack Gold and Megan Mcbride. “Cryptocurrency : A Primer for Policy-Makers”. In: August (2019).

[4] W. Bolt. “Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction”. In: *The Journal of Economic Literature* 55.2 (2017), pp. 467–469. ISSN: 0022-0515.

[5] Edward Felten Arvind Narayanan, Joseph Bonneau and Steven Goldfeder Andrew Miller. “Bitcoin and Cryptocurrency Technologies”. In: *Journal of Information Technology* 21.4 (2016), pp. 284–298. ISSN: 02683962. DOI: 10.1057/palgrave.jit.2000080.

[6] Satoshi Nakamoto. “Bitcoin: A User-to-User Electronic Cash System”. In: (2008), p. 9.

[7] James Cope. “QuickStudy: Peer-to-Peer Network”. In: *Computerworld* (2002). URL: https://www.computerworld.com/article/2588287/networking-peer-to-peer-network.html.

[8] Andreas M. Antonopoulos. “Mastering Bitcoin: Unlocking Digital Cryptocurrencies”. In: *O’Reilly Media, Inc.* 50.4 (2016), pp. 675–704. ISSN: 10116702. DOI: 10.1002/ejoc.201200111. arXiv: arXiv:1011.1669v3. URL: https://www.bitcoinbook.info/.

[9] Michael Doran. “A Forensic Look at Bitcoin Cryptocurrency”. In: *SANS Institute InfoSec Reading Room* (2015), p. 35. URL: https://www.sans.org/reading-room/whitepapers/forensics/forensic-bitcoin-cryptocurrency-36437.

[10] Michael Crosby, Nachiappan, Pradhan Pattanayak, et al. “Blockchain Technology - BEYOND BITCOIN”. In: *Berkley Engineering* (2016), p. 35. DOI: 10.1515/9783110488951. URL: http://www.degruyter.com/view/books/9783110488951/9783110488951/9783110488951.xml.

[11] F Xavier Olleros, Majlinda Zhegu, F Xavier Olleros, et al. *Research Handbook on Digital Transformations*. Edward Elgar Publishing, Incorporated, 2016. ISBN: 1784717754.

[12] Coindesk. "How do Bitcoin Transactions Work?" In: (2015), pp. 1–5. URL: https://www.coindesk.com/information/how-do-bitcoin-transactions-work/.

[13] Mahdi H. Miraz and Maaruf Ali. "Applications of Blockchain Technology beyond Cryptocurrency". In: *Annals of Emerging Technologies in Computing* 2.1 (2018), pp. 1–6. ISSN: 2516-0281. DOI: 10.33166/aetic.2018.01.001. arXiv: 1801.03528.

[14] Don Tapscott and Alex Tapscott. *Blockchain Revolution: How the Technology Behind Bitcoin Is Changing Money, Business, and the World*. Portfolio, 2016. ISBN: 1101980133.

[15] Newgenapps. "What is Intent analysis and how can it be used?" In: (2018), p. 1. URL: https://www.newgenapps.com/blog/what-is-intent-analysis-how-can-it-be-used.

[16] Bo Pang and Lillian Lee. "Opinion mining and sentiment analysis". In: *Foundations and Trends in Information Retrieval* 2.1-2 (2008), pp. 1–135. ISSN: 15540669. DOI: 10.1561/1500000001.

[17] Amandeep Kaur and Vishal Gupta. "A survey on sentiment analysis and opinion mining techniques". In: *Journal of Emerging Technologies in Web Intelligence* 5.4 (2013), pp. 367–371. ISSN: 17980461. DOI: 10.4304/jetwi.5.4.367-371.

[18] P. G. Preethi, V. Uma, and Ajit Kumar. "Temporal sentiment analysis and causal rules extraction from tweets for event prediction". In: *Procedia Computer Science* 48.C (2015), pp. 84–89. ISSN: 18770509. DOI: 10.1016/j.procs.2015.04.154.

[19] Bing Liu. "Sentiment analysis and opinion mining". In: *Synthesis Lectures on Human Language Technologies* 5.1 (2012), pp. 1–184. ISSN: 19474040. DOI: 10.2200/S00416ED1V01Y201204HLT016.

[20] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews". PhD thesis. 2004.

[21] Julio Villena-Román. "An Introduction to Sentiment Analysis (Opinion Mining)". In: *Meaningcloud* (2015).

[22] Vasileios Hatzivassiloglou and Kathleen R. McKeown. "Predicting the semantic orientation of adjectives". In: (1997), pp. 174–181. DOI: 10.3115/979617.979640.

[23] M M Bradley and P J Lang. "Affective norms for English words (ANEW)". In: (1999).

[24] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: (2002). arXiv: 0205070 [cs]. URL: http://arxiv.org/abs/cs/0205070.

[25] Andrea Esuli and Fabrizio Sebastiani. "SENTIWORDNET: A high-coverage lexical resource for opinion mining". In: *Evaluation* (2007), pp. 1–26. URL: http://ontotext.fbk.eu/Publications/sentiWN-TR.pdf.

[26] Tony Mullen and Nigel Collier. “Sentiment Analysis using Support Vector Machines with Diverse Information Sources”. In: *Proceedings of EMNLP 2004* (2004), pp. 412–418. URL: http://research.nii.ac.jp/%7B~%7Dcollier/papers/emnlp2004.pdf.

[27] Springer Science. “Feature Extraction Techniques”. In: *Computer-Based Design and Manufacturing* (2007), pp. 101–124. DOI: 10.1007/978-0-387-23324-6_5.

[28] Saif Mohammad, Cody Dunne, and Bonnie Dorr. “Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus”. In: *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009* (2009), pp. 599–608. DOI: 10.3115/1699571.1699591.

[29] Chin Sheng Yang and Hsiao Ping Shih. “A rule-based approach for effective sentiment analysis”. In: *Proceedings - Pacific Asia Conference on Information Systems, PACIS 2012* (2012).

[30] Kerstin Denecke. “Are SentiWordNet scores suited for multi-domain sentiment classification?” In: *4th International Conference on Digital Information Management, ICDIM 2009* (2009), pp. 32–37. DOI: 10.1109/ICDIM.2009.5356764.

[31] C. J. Hutto and Eric Gilbert. “VADER: A parsimonious rule-based model for sentiment analysis of social media text”. In: *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (2014), pp. 216–225.

[32] Christie Schneider. “The biggest data challenges that you might not even know you have”. In: *IBM Blog AI for the Enterprise* (2016). URL: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.

[33] Emilie Coyne, Jim Smit, and Levent Güner. *Sentiment analysis for Amazon.com reviews*. 2019. DOI: 10.13140/RG.2.2.13939.37920.

[34] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, et al. “Predicting elections with Twitter: What 140 characters reveal about political sentiment”. In: *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media* (2010), pp. 178–185.

[35] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, et al. “From tweets to polls: Linking text sentiment to public opinion time series”. In: *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media* (2010), pp. 122–129.

[36] W HUANG. “Forecasting stock market movement direction with support vector machine”. In: *Computers & Operations Research* (2004). ISSN: 03050548. DOI: 10.1016/s0305-0548(04)00068-1.

[37] Michael Lubitz. “Who drives the market? Sentiment analysis of financial news posted on Reddit and the Financial Times”. In: (2017), pp. 1–39. URL: http://ad-publications.informatik.uni-freiburg.de/theses/Bachelor%7B%5C_%7DMichael%7B%5C_%7DLubitz%7B%5C_%7D2018.pdf.

[38] Andrius Mudinas, Dell Zhang, and Mark Levene. “Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward”. In: (2019). arXiv: 1903.05440. URL: http://arxiv.org/abs/1903.05440.

[39] Adam Salac. “Forecasting of the cryptocurrency market through social media sentiment analysis”. In: (2019). URL: http://essay.utwente.nl/78607/.

[40] Doaa Mohey El Din Mohamed Hussein. “A survey on sentiment analysis challenges”. In: *Journal of King Saud University - Engineering Sciences* 30.4 (2018), pp. 330–338. ISSN: 10183639. DOI: 10.1016/j.jksues.2016.04.002.

[41] Shabina Dhuria. “Sentiment Analysis: An approach in Natural Language Processing for Data Extraction”. In: *International Journal of New Innovations in Engineering and Technology* 2.4 (2015), pp. 27–31. ISSN: 2319-6319.

[42] Rob J Hyndman and George Athanasopoulos. “Forecasting: Principles and Practice”. In: *Principles of Optimal Design* (2018), pp. 421–455. DOI: 10.1017/9781316451038.010.

[43] Alicia Tuovila. *Forecasting*. 2019. URL: https://www.investopedia.com/terms/f/forecasting.asp.

[44] Young Bin Kim, Jurim Lee, Nuri Park, et al. “When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation”. In: *PLoS ONE* 12.5 (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0177630.

[45] Tetlock. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market”. In: 62.3 (2007), pp. 1139–1168.

[46] Vytautas Karalevicius. “Using sentiment analysis to predict interday Bitcoin price movements”. In: *Journal of Risk Finance* 19.1 (2018), pp. 56–75. ISSN: 09657967. DOI: 10.1108/JRF-06-2017-0092.

[47] Martina Matta, Ilaria Lunesu, and Michele Marchesi. “Bitcoin spread prediction using social and web search media”. In: *CEUR Workshop Proceedings* 1388 (2015). ISSN: 16130073.

[48] Evita Stenqvist and Jacob Lönnö. “Predicting Bitcoin price fluctuation with Twitter sentiment analysis”. In: *Diva* (2017), p. 37. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-209191.

[49] Andrew Burnie and Emine Yilmaz. “Social media and bitcoin metrics: Which words matter”. In: *Royal Society Open Science* 6.10 (2019). ISSN: 20545703. DOI: 10.1098/rsos.191068.

[50] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. “Statistical analysis of the social network and discussion threads in Slashdot”. In: *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08* (2008), pp. 645–654. DOI: 10.1145/1367497.1367585.

[51] Maria Glenski and Tim Weninger. “Predicting user-Interactions on reddit”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017* (2017), pp. 609–612. DOI: 10.1145/3110025.3120993. arXiv: 1707.00195.

[52] Sachin Thukral, Hardik Meisheri, Tushar Kataria, et al. “Analyzing behavioral trends in community drivendiscussion platforms like Reddit”. In: *abs/1809.07087* (2018).

Appendix A

Heat map correlation analysis

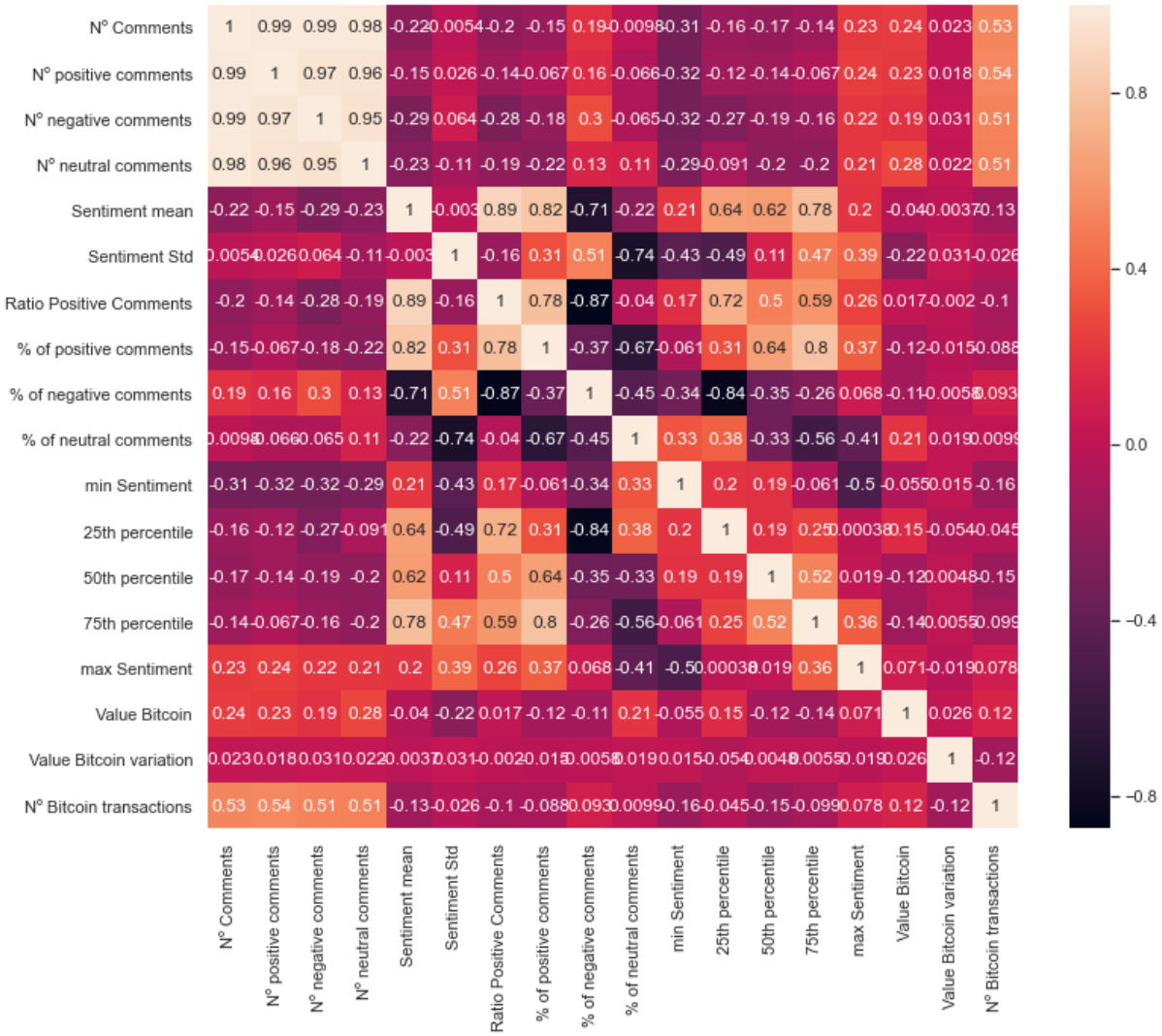

Figure A.1: Heat map correlation Bitcoin indicators shifted less two days then users comments indicators

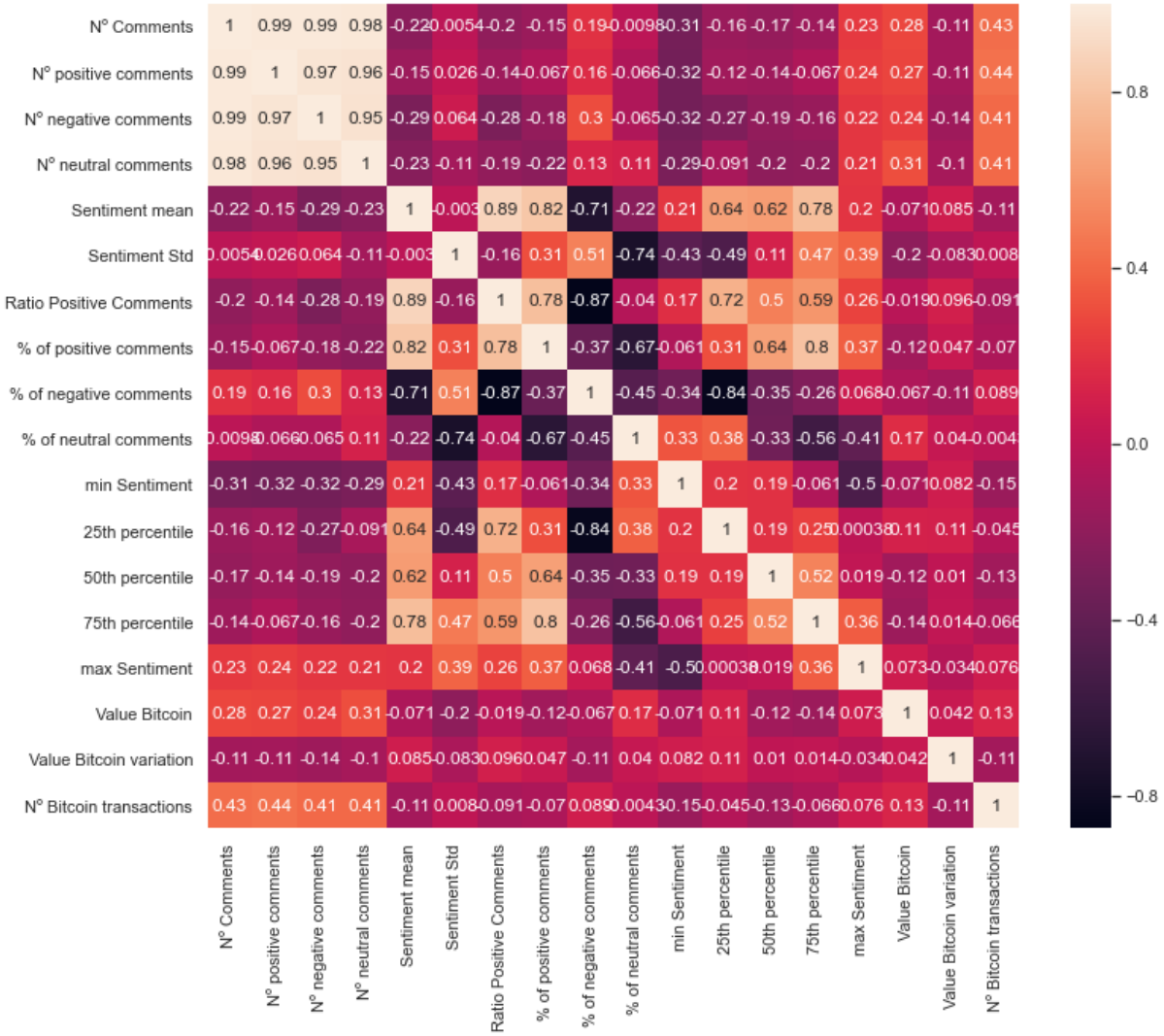

Figure A.2: Heat map correlation Bitcoin indicators shifted plus two days then users comments indicators

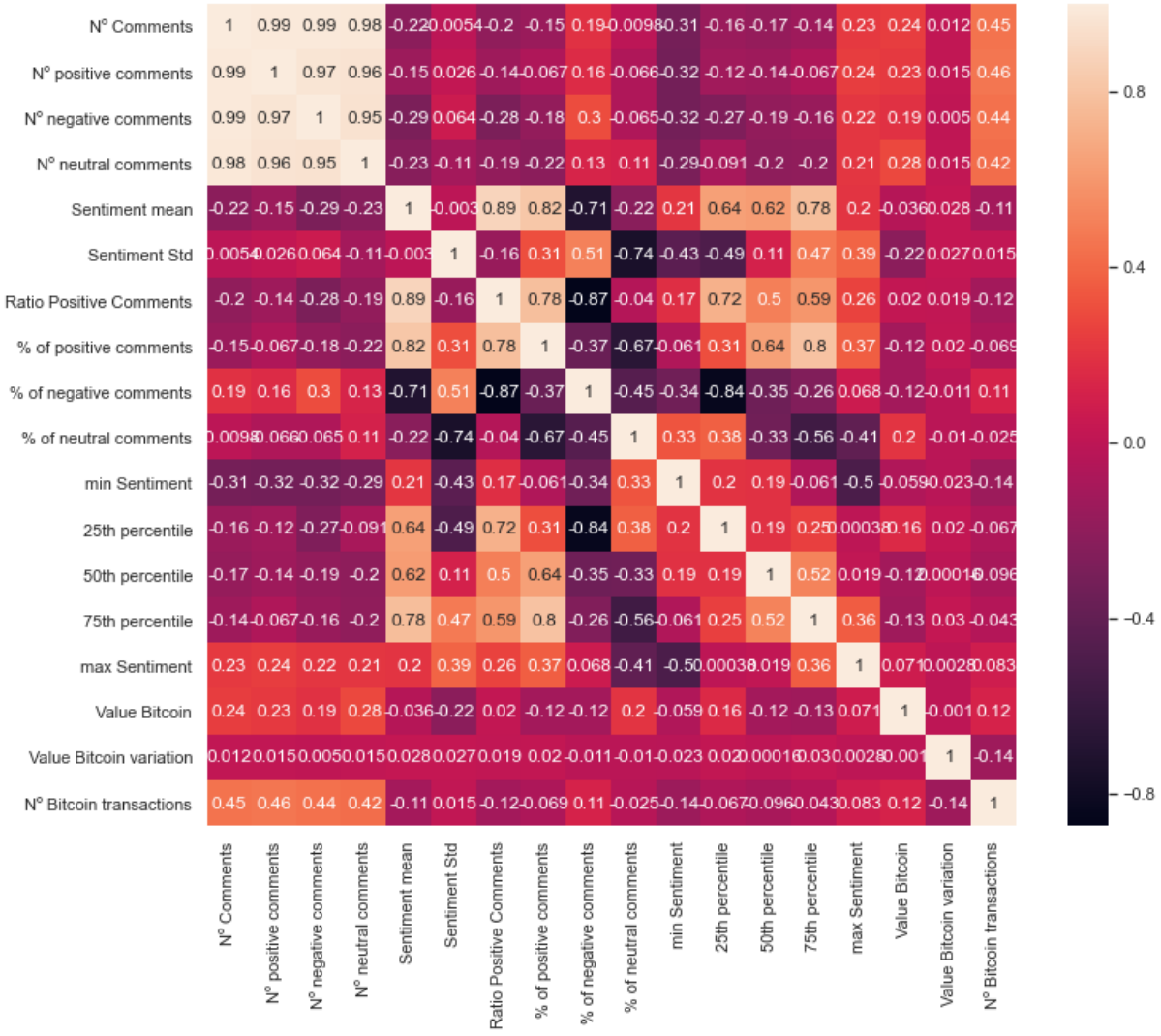

Figure A.3: Heat map correlation Bitcoin indicators shifted less three days then users comments indicators

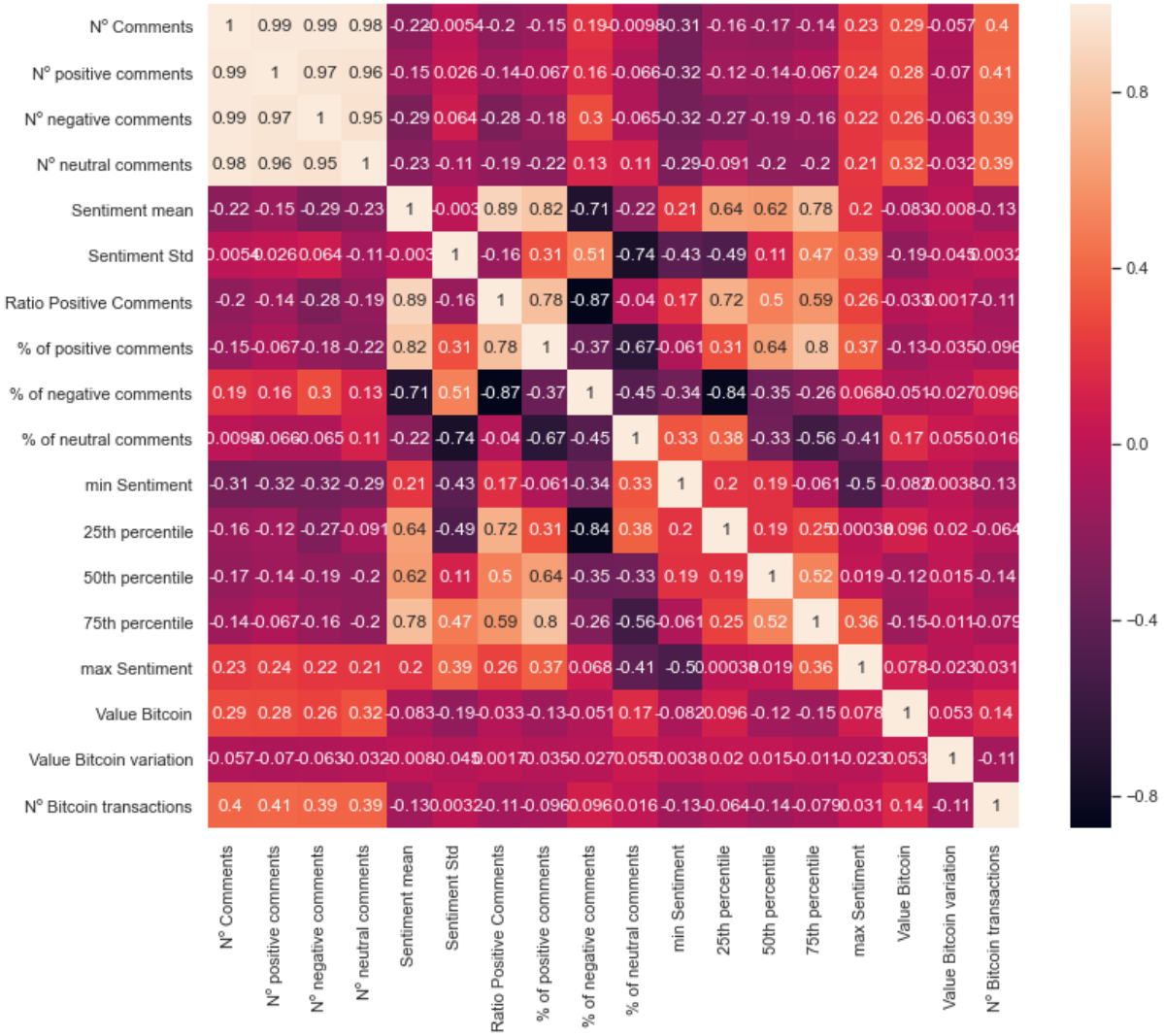

Figure A.4: Heat map correlation Bitcoin indicators shifted plus three days then users comments indicators

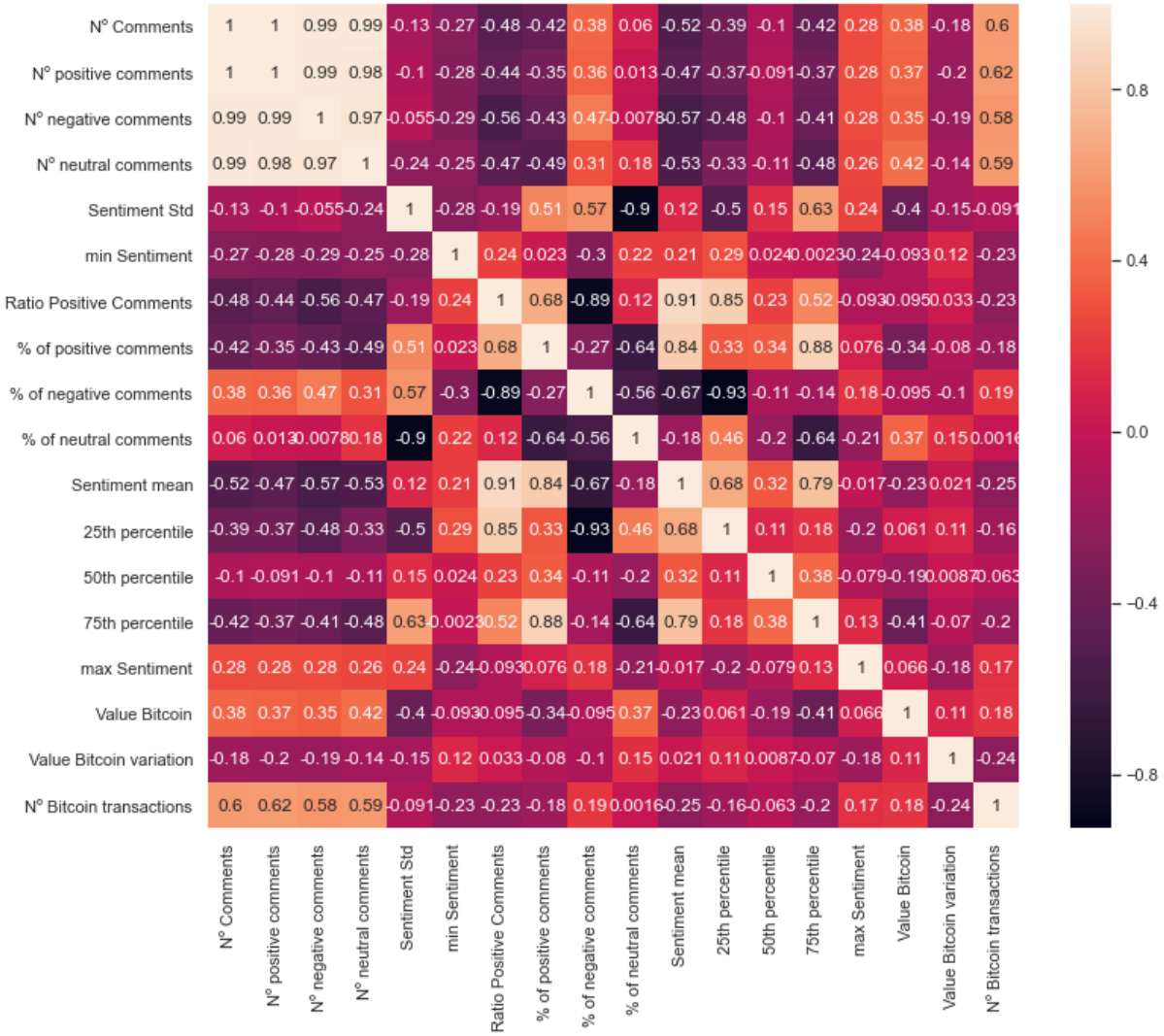

Figure A.5: Heat map correlation Bitcoin indicators shifted plus one week then users comments indicators

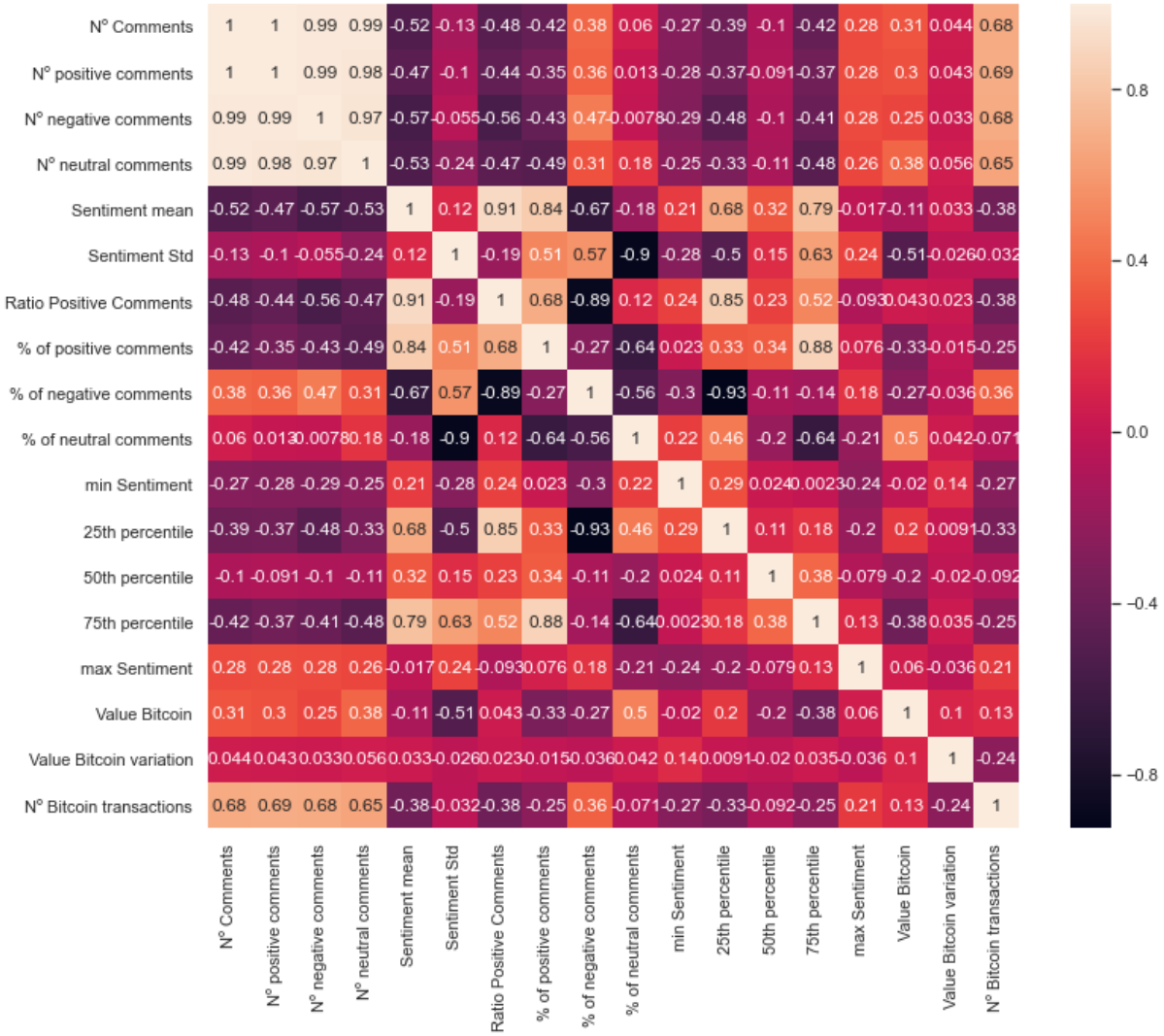

Figure A.6: Heat map correlation Bitcoin indicators shifted less one week then users comments indicators