Universitat Oberta de Catalunya

uoc.edu

Predicting cryptocurrency prices with Twitter using sentiment analysis

Ricardo Sans Cipollitti

Master's Thesis

MSc in Computational Engineering and Mathematics

Artificial Intelligence

David Amorós Alcaraz (tutor)

Carles Ventura Royo (coordinator)

September 2022

Table of contents

1. Motivation

This paper examines to what extent influential Twitter users associated with cryptocurrencies can distort the market by writing —or tweeting— meaningful messages on this social network. To do so, we will build a machine learning (ML) model capable of reading, interpreting and scoring streams of tweets and then matching them with cryptocurrency prices on a minute-by-minute basis. Our goal is to test whether a sentiment score given by a computer to a tweet message is able to predict price movements, and to assess to what extent it works for high-valued digital currencies.

2. Methodology

Throughout our investigation we conducted quantitative analyses aimed at solving our thesis in an objective way. We built several machine learning algorithms and fed them with both tweet sentiment scores and crypto prices. Data extraction was possible thanks to CryptoArchive and the Twitter API v2 with elevated access for academic researchers.

Python was our main programming language for the creation of the source code. We firstly created a relational database using PostgreSQL to store all tweet extractions. Later, we wrote an API call code for extracting Twitter tweets, and manually downloaded the fifteen cryptocurrency prices from the CryptoArchive webpage. By using SQL-friendly python packages, we were able to pull SQL requests directly from our database for processing text and assigning sentiment scores. Finally, we developed several descriptive python notebooks for better understanding both the data and the result outputs. Graphs and tables cover most of these explanations for optimal comprehension.

3. Introduction

How did it become possible, as human beings, to invent and believe a system in which we can regularly make payments with a plastic card without any physical movement of money? How can we trust that there will be a transaction of capital between bank accounts? At which point in time did we start relying on a seemingly valueless round object for exchanging goods and services? Why do we trade with money instead of bartering?

Before the existence of fiat money people used to trade goods and services in exchange for other goods and services. As trading became much more complex and more products were added to the market, merchants tended to select one or two items, usually metals, as preferred commodities to trade because of their intrinsic properties, like robustness and durability (Davies, 2002). Besides, these new preferences needed some standardization, so governments who had already established a reputation acted as a reliable institution by minting metals into rounded tokens and assigning them a value which merchants then acknowledged and traded (Joseph A. Ritter, 1995).

Some hundreds of years later, instead of carrying around heavy-lifting metals, dealers started to trade with bills that guaranteed the property of those commodities. Again, since dealers could not trust each other, governments had to back up the underlying values of the bills. With this new methodology, governments improved both their economic system and the incentives to trade, although they were strictly tied to citizens' confidence. In the event that everybody decided to exchange their bill for the underlaying commodity, redemptions would not be fully honored. In that case, trade would fail and the currency would be completely useless because its issuer would not be able to fulfill the guarantees of the contract. According to Seghezza & Battista Pittaluga

(2021), the ultimate reason for the collapse of the Gold-Exchange Standard was the sterilization policies —acquiring vast amounts of gold— that countries like France and the USA adopted, which significantly shortened the supply of gold in the world's economy. Countries like England could not finance themselves and, more importantly, could not back up their currency in case of a confidence crisis.

Only after this breakdown the world moved towards fiat money: currency with no underlaying guarantee and solely based on citizens' credibility. On paper, disconnecting money from the commodity it used to be pegged to renders it useless, that is why nations had to reshape the idea of a secure and trustworthy economic system. Nowadays, the most important objective of fiat is that both sides of any transaction believe that the exchange has a positive value for them, and that its value is *universal* —does not depend on the bearer— and *immutable* —does not change from one transaction to the next—. For that, countries have had to build institutions and policies to protect the property rights of those who hold the currency, enforce the rule of law to show political confidence and apply moderate monetary policies to keep inflation rates at a minimum. Despite the application of all these tight rules, money still is, and will continue to be, directly affected by relations with foreign countries and their balance of power. Argentina's peso during the soaring inflation in 2001, Hungary's pengő and forint after World War II and, more recently, Russian's ruble with the invasion of Ukraine in 2022 are all examples of how a currency can dramatically lose its value due to the causes just noted before.

One of the most important bodies in charge of overseeing monetary policies and act according to the market environment is the central bank. Central banks thus have the ability to manipulate the liquidity of money in the financial system by issuing currency and setting interest rates on loans and bonds. When there is

a period of economic stress, the issuing of money into the national market is usually a satisfactory method to promote spending and reactivate the economy. On the other hand, during boom periods, central banks tend to absorb money to discourage borrowing and incentivize investment. Although these policies are meant to reduce the impact of economic cycles, they entail a great cost for citizens. Following the simple laws of supply and demand, the greater the quantity of money there is in the market, the lower its value must be. In other words, if all employees of an imaginary country suddenly got paid twice what they were paid before, vendors would have to raise their product prices to balance the market, otherwise they would lose wealth with respect to employees. However, note that if vendors where the ones to raise the prices of their goods then employees would have to get higher wages to keep having the same real wealth, an event that is unlikely in the real world.

After the financial crisis of 2008, and especially during the COVID-19 pandemic, governments and other institutions have been fighting for a steady recovery of the economic activity. Figure 3.1 shows a constant increase in U.S. dollar currency[1] and M1 U.S. dollar supply[2] since 1959, with a severe peak in 2020 as a result of Fed's extremely expansionary monetary policy. Although a higher difference in economic inequality is yet to be seen, inflation is already lifting up the consumer price index (CPI) at a higher rate than before, as it can be seen in figures 3.2 and 3.3 in the appendix.

[1] U.S. dollar currency includes all U.S. dollar cash flowing in the world market

[2] M1 money supply includes all U.S. dollar cash flowing in the world market and money inside bank accounts

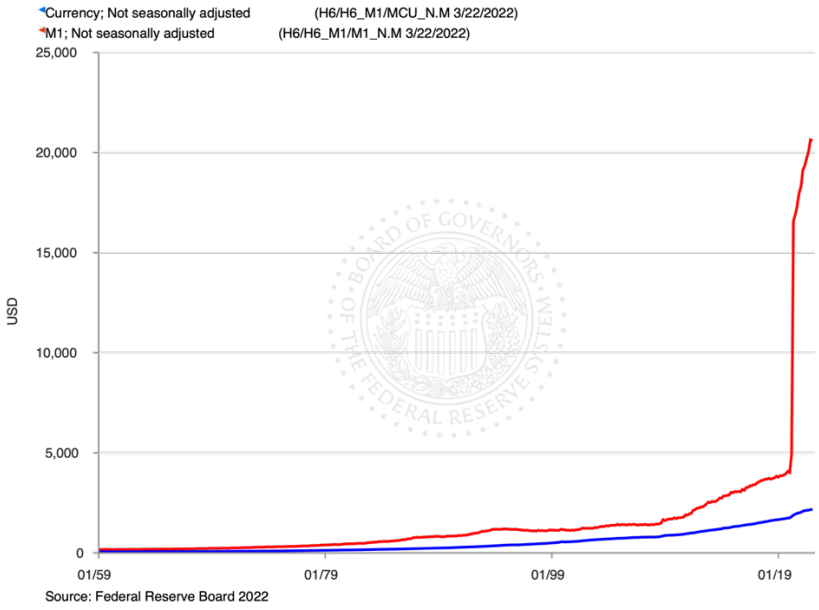


***Figure 3.1**, mass of dollar currency and M1 dollar's supply (in billions)*
***Source**: Federal Reserve Bank of St. Louis (FRED) webpage*

Because of all the previously-mentioned considerations, fiat money is currently under debate and faces serious flaws. Criticists have shifted their attention to cryptocurrencies, arguing that these solve some —to most— of the current money problems. In fact, the future prospects for this currency are so high that El Salvador and the Central African Republic have already approved bitcoin (BTC) as a legal tender. But, why is it a revolution in the first place?

Cryptocurrencies —in their vast majority— are free markets with no issuing authority at their back and are not tied to any political body, nation, or group of nations, which means that they are not subject to any state's movements and interests that may harm a share of the society. At the time this paper is being written, everyone with internet connection is able to open a crypto account and freely trade everywhere in the world at almost no cost, regardless of their race, economic and social condition. Senders and recipients of cryptocurrencies bear zero to almost no transaction costs and, surely enough, no exchange commissions. Indeed, transactions from dollars to dollars or euros to euros are

performed in a similar manner, but the idea that people are more and more frequently opting for exchanging value in cryptos is evidence of the fact that either transaction costs are lower or people trust cryptocurrencies more than fiat currencies.

Nevertheless, cryptocurrencies are far from perfect and are also subject of critiques which need to be addressed before moving on with our thesis. PhD Seele argues that the immoral use of cryptocurrencies threatens the virtuous potential that these may have in the present and near future, and warns that their secrecy and anonymity features are a double-edged sword which attracts both libertarians that believe economic transactions should not be recorded, and criminals which can now base their operations on a non-cash means of payment. The creation of cryptos has led to a proliferation of darknet marketplaces responsible for drug, cyber-arm, weapon, counterfeit currency, human trafficking and many other forms of illegal activities. According to Foley et al. (2019), monthly transactions by illegal bitcoin[3] users increased from $2B in 2014 to $11B in 2017, and the estimated dollar value of illegal user's bitcoin holdings increased from $1.5B in 2014 to almost $6B in 2017 —see figures 3.4 and 3.5 in the appendix—. These numbers represent a real threat to society and thus question the true viability of cryptocurrencies.

Not only illegal transactions are in the spotlight. Criticists claim cryptocurrencies are anything but stable and are subject to large losses of wealth. The truth is that they are prone to large price movements. Bitcoin has soared +900% in the last three years, +2,000% in the case of Ethereum and 2,100% in

[3] Bitcoin will be frequently mentioned throughout this paper as a proxy for describing cryptocurrencies' general behavior since it has the highest market capitalization and several studies only discuss about bitcoin.

the case of Binance coin[4], which, as of today, are the first, second and fourth crypto assets in terms of market capitalization. Cryptocurrencies are also known for having high volatility. In terms of the European Central Bank (ECB), they are "highly speculative investments whose valuations are based on extremely weak fundamentals[5]". In figure 3.6 we observe the daily return series of gold vs bitcoin and their histograms.

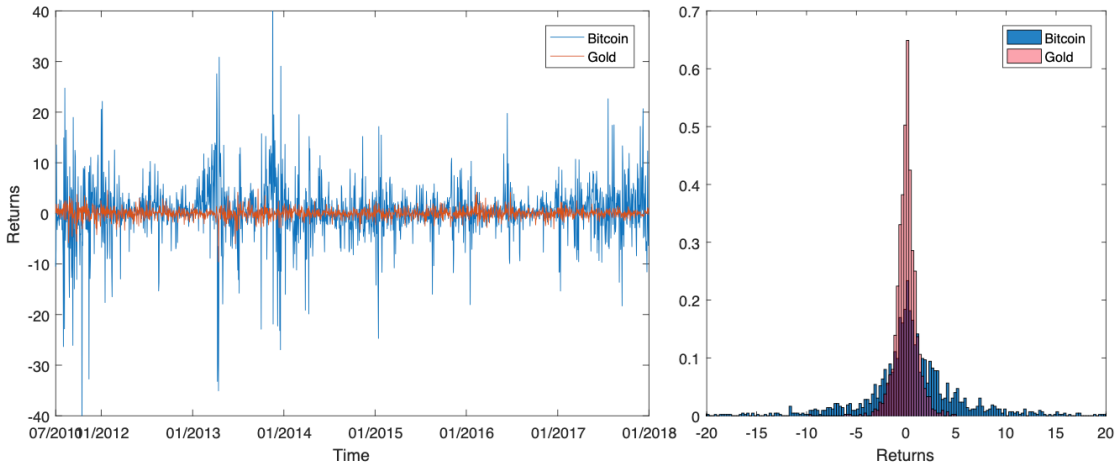


***Figure 3.6**, Plots of the daily return series of gold and bitcoin and resulting histogram*

***Source**: Bitcoin is not the New Gold A Comparison of Volatility, Correlation, and Portfolio Performance*

According to Klein et al. (2018), the standard deviation of the bitcoin was over 5 times higher to that of gold during the period from July 2, 2011 to December 31, 2017, meaning bitcoin was on average, and for a time period of almost 6 years, 5 times more volatile than gold. Compared to fiat currency pairs, cryptos also indicate high volatility. We can visibly note in figure 3.7 in the appendix the difference between the 30-day standard deviation of daily returns of bitcoin and that of the USD/EUR pair. The highest volatility of the USD/EUR pair in 10

[4] From April 2019 to April 2022, using data from Statista.com

[5] Interview to Luis de Guindos, Vice-President of the ECB, conducted by Frank Wiebe and Jan Mallien on 14 March 2020

years has been of 1.03%, whereas that of bitcoin has been 14.41%, 14 times greater.

Cryptocurrencies aim at solving some of the greatest challenges that fiat money has, by setting a maximum supply for reducing inflation, charging almost no transaction fees, allowing anyone to open an account to trade and —for some people—, breaking with the idea that money should always be governed by a central authority. On the contrary, since there is no limit to who can use them, their reputation has been undermined as they are a bargaining chip for illegal activities. They have also shown great volatility throughout their existence, posing a real threat for the savings of those who hold them for short periods of time, especially those newcomers with little to no experience in the investment sector.

Both cryptocurrencies and fiat money are traded in the market under similar conditions. Anyone can buy the BTC/EUR pair the same way as they would buy the USD/EUR pair. However, as this introduction has shown, both crypto and fiat currencies were built with different means and purposes, so naturally their market prices must depend on different factors. Later, we will analyze these main components that affect the price and volatility of cryptos. But before, we will briefly elaborate on the most technical characteristics: the internal algorithms that govern the realm of cryptos and their usages in the real world.

4. Understanding the technology

The word *cryptocurrency* comes from the Greek word kryptós —meaning "secret" or "hidden"— and the word currency, which we have already addressed at length in the introduction. As we shall see, cryptography plays a fundamental role in the process of transferring both money and information.

Every time we are buying or selling cryptocurrencies we follow a set of computer instructions, also known as a protocol, that governs the network of users. All cryptocurrencies, including those in our study, are secured by a specific type of instructions, commonly known as the blockchain. The term blockchain first appeared in a 2008 paper written by an anonymous group under the pseudonym Satoshi Nakamoto and it was the basis for Bitcoin[6]. The procedures that shape blockchain provide strong control of ownership, i.e., secure storage of money, while at the same time remove the issue of trusting a third party when we want to make any transaction. In the following lines we will briefly study some of the properties that make blockchain a unique protocol.

4.1. Read/write access to the ledger

Everyone in the blockchain network is allowed to record any transaction —e.g., Alice pays Bob 100 bitcoins— at any point in time. Because of its public nature, wrongdoers may unilaterally write unfair trades without the consent of the other users. That is why every user in the network relies on a public-private key, i.e., a cryptographic algorithm that verifies the approval of any user. While it is very easy for the network to verify the authenticity of a trade with the signature, it is virtually impossible to crack it. Furthermore, although anyone can write

[6] When talking about the blockchain protocol, we capitalize the words, as in Bitcoin. When talking about coins, we do not capitalize, as in bitcoin.

transactions, the system is made such that none of them are valid if the borrower owes more than what he/she has in the wallet. This method is used to prevent users from refusing to pay, similar to the functioning of a debit card.

4.2. Proof of work

Everyone in the network has an updated copy version of the ledger, reason why blockchain is also known as a distributed ledger technology (DLT). However, trusting incoming information from anyone in the network could lead to accounting errors, forgery and other failures in the system, that is why blockchain is based on the proof of work (PoW). When a group of transactions is created, *miners* —users with high computational power— spend time breaking a cryptographic hash[7] by brute force to find the number, i.e., the PoW, that validates the incorporation of those transactions into the distributed ledger. Since the breaking of the hash takes a considerable amount of time, as long as there are less than 50% of users trying to forge the network, it is virtually impossible in the long run for someone to fool the system by guessing the PoW correctly all the time.

4.3. Incentives for the miners

As mentioned before, miners are those users in the network with sufficient computational capacity to validate blocks of transactions and add them into the distributed ledger. In theory, everyone with a computer should be able to mine a correct PoW for a given block, but the best ones have more chances to do it first. For those fortunate ones, the protocol rewards them with a preestablished amount

[7] A hash is a text that has been converted, through a hash function, into a fixed-length array of ones and zeros. Although it is easy to transform a text into a hash, it is infeasible to do the opposite.

of (virtual) money. Since that money does not come from any of the users but the system itself, the supply of that cryptocurrency should always increase. However, most protocols are stablished such that every time a specified number of blocks are added to the ledger, the reward halves and thus the amount of work needed to mine the same number of crypto doubles. In the case of bitcoin, every time 210,000 blocks are added to the network, the reward is reduced by half, allowing for a maximum supply of 21 million bitcoins.

Let's digest all the previous characteristics in one example. Previously, Alice wanted to pay Bob 100 bitcoins. How does the system process and validate the transaction?

First of all, Alice sends the request through the network, signaling her willingness to transfer those funds to the rest of the users, together with her signature. The rest of the users validate the signature and approve the transaction, that is now registered in their ledgers.

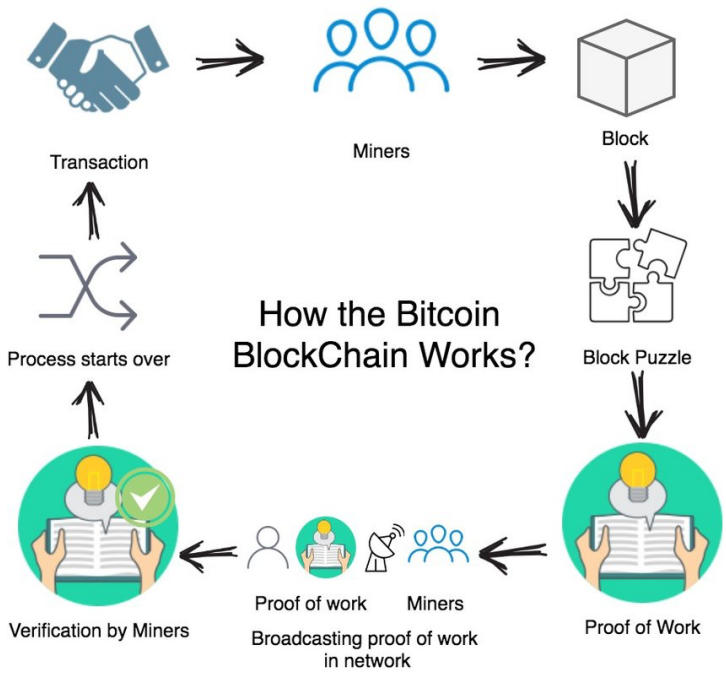


***Figure 4.1**, How the Bitcoin blockchain works?*

***Source**: Suman Ghimire on ResearchGate*

Finally, this transaction, along with others, is stacked in a "block" that, when full, is validated by guessing by brute force the solution of a hash. The further in time that transaction has been registered in a block and validated by others, the more certain we will be that it has not been forged. Surely, the issue that may arise with this protocol is that transactions in their initial stage should not be fully trusted, but the broader the network of users and the faster the blocks can be validated, the more confidence in any transaction. According to Nakamoto, the likelihood of an attacker attempting to forge a transaction decreases exponentially as more blocks are added into the chain.

5. State-of-the-art in cryptocurrencies

Due to the limited modifications and extensions that the Bitcoin can have, there is an approach leading to new currencies, the so-called altcoins. Most of these coins are very similar to Bitcoin, as they were created by forking the Bitcoin protocol and are based on its core functionality. However, there are also coins with a completely different design. Altcoins were created primarily to address Bitcoin's shortcomings. Some altcoins will only appeal to a smaller group, while others will appeal to a wider audience and can be considered real competition for Bitcoin. In the following section we will be discussing some of the most important and will explain some day-to-day usages.

5.1. Ethereum

It can be considered as the frontrunner of altcoins. This open-source project lets users create decentralized applications. One of the most widely used is that of smart contracts, a series of digital instructions that parties follow automatically. Indeed, smart contracts resemble “regular” contracts, but in this case, parties do not need to trust a third party for its execution, but rather the system. Besides, according to Wood, 2014, there is no risk of censorship, downtime, fraud, or third-party interference. Smart contracts run thanks to the Ethereum Virtual Machine (EVM); a piece of software built on top of the Ethereum nodes (Le, 2017). A successful application of Ethereum smart contracts is dApps, a platform for parties that lack trust in each other with a great variety of smart contracts. Some examples include financials, art, videogames and voting processes.

Another state-of-the-art application of Ethereum is that of tokens. According to di Angelo & Salzer, 2020, crypto tokens are digital assets built from the existence of other cryptocurrencies, and are usually the result of smart

contracts inside a dApps. In most of the cases, these tokens are used as a means of transaction between the two or more parties. Unlike cryptocurrencies, crypto tokens do not own their own blockchain protocol and depend entirely on the EVM. As we shall later see, these can be used as means of payment, although other uses include security and utility functions. With the introduction of ERC-20 tokens, parties can choose from several pre-established and standardized smart contracts, matter that accelerates the usage of such dApps.

5.2. Ripple

Ripple is a private blockchain protocol in which only some people can access and be part of the network. It was created as a transactions solution management software for transferring money through financial and non-financial institutions. One of the main challenges for banks these days is the transaction costs derived from moving funds from one country to another, especially if they do not have any branch or subsidiary in the destination country. Time, intermediaries and errors are the most relevant costs. Ripple removes this burden by allowing banks to become nodes in the private network so as to operate and validate transactions among themselves. Such activity improves communication, reduces costs and standardizes international transfers.

XRP is the crypto token derived from Ripple, and can be optionally used as an exchange method inside the network. XRP does not need to be mined because transactions are validated by the private nodes of the network. 100 billion (100.000.000.000) XRP coins were minted by the founders of Ripple, of which around 55 billion are currently in the network —the rest have been either kept by the founders or stored as future value inside Ripple—. Banks like Banco Santander (Spain) claim that implementing Ripple technologies could save them up to $20 billion every year (Moreno-Sanchez et al., 2016; Rella, 2020).

5.3. Litecoin

Litecoin was created in 2011 with the intention of being a cheap alternative to bitcoin. Its main purpose is to serve as coin for small quantities, hence 100.000.000 units of litecoin are equal to 1 bitcoin. However, litecoin is considered to be one of the cryptocurrencies with most market capitalization in the world. The main reason for its popularity is the reduced execution time and improved mining. Bitcoin encourages the use of expensive, powerful, dedicated hardware in its network, so not everyone can participate in mining. Litecoin facilitates the access and participation in this process. Instead of using Bitcoin's SHA-256 algorithm, Litecoin relies on his Scrypt-POW algorithm which is more memory intensive, resulting in shorter transaction times and more processed transactions (Le, 2017). To generate a block in Bitcoin, 10 minutes are needed. Instead, Litecoin needs 2.5 minutes.

5.4. Polkadot

Sometimes, protocol limitations stop blockchain networks from working smoothly. For example, if when using Bitcoin we register a great number of transactions, a bottleneck may occur, causing delays and higher transaction fees. Polkadot offers a solution by connecting different single-purpose blockchains and making them scalable by work together efficiently. In addition, they offer higher security and shorter development time than with a conventional protocol. The relay chain —see figure 5.1— can be considered as the "heart" of the network; where all blockchains —or parachains— connect. The idea behind the relay chain is that each of these unique blockchains can be optimized for a specific use case through the parallel communication of the parachains within the network. The relay chain is also capable of handling all transactions at the same time. Interoperability of

the parachains is vast and allows the interaction of all sorts of data: from tokens to stock data, account balances and football scores (Scott et al., 2021).

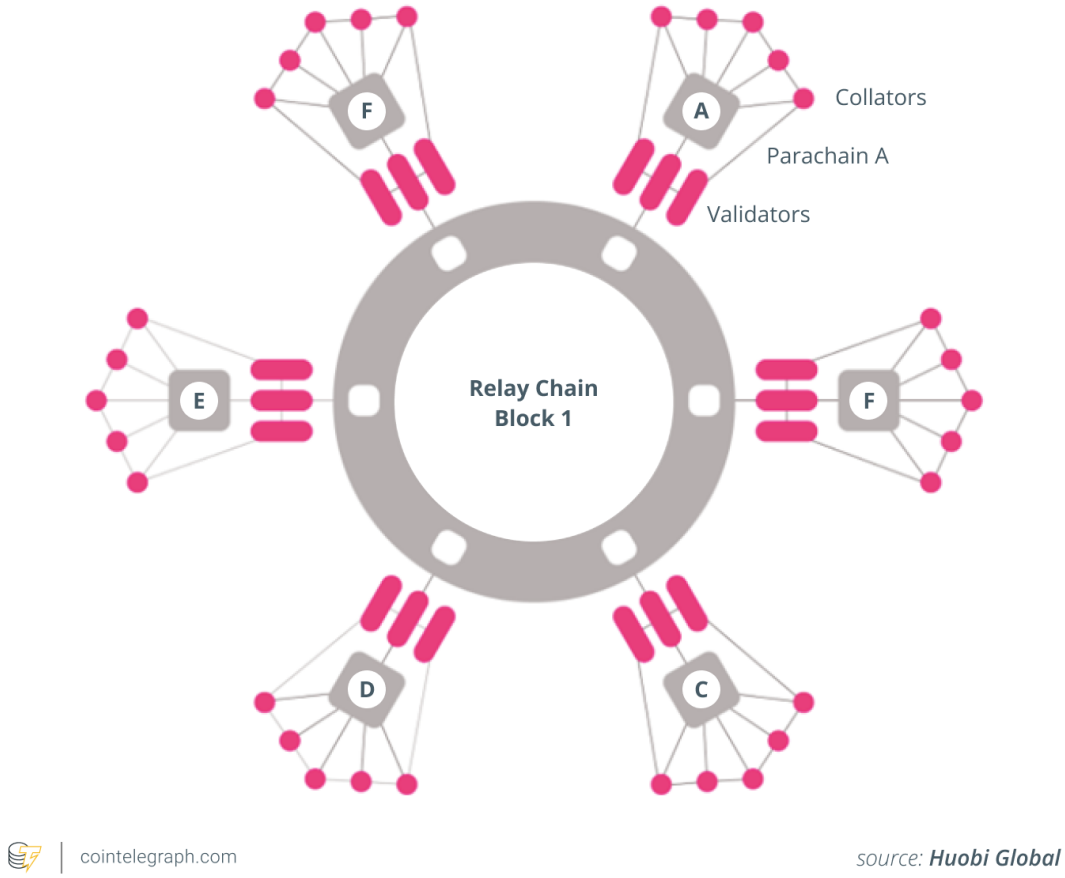


Figure 5.1

What we have just read are examples of usage that can be found today in four of the fifteen currencies that we will study later in our paper. Like these, many other cryptocurrencies have a clear objective and are continuously being improved to offer cutting-edge solutions. However, there are others that do not have any kind of objective but that of mockery or experiment. Some of them are Doge coin or Shiba Inu coin, which, surprisingly, have been among the coins with the highest market capitalization at the end of 2021. In the subsequent lines we will illustrate the characteristics inherent in cryptos that allow them to be traded and transacted in the same way as fiat currencies.

6. Monetary characteristics of blockchain

The value of cryptocurrencies can be assimilated to that of precious metals. Both coincide in that they are scarce resources, as we have seen above, and are precious within their respective fields. Copper, for example, is needed in various applications in the industrial sector. The underlying application of some cryptocurrencies within their technology makes them also valuable. In addition, their cryptographic system may enable them, in the not-too-distant future, to perform key functions, such as being a medium for retail transactions, adding more value.

When discussing about Ethereum in the previous section, we noted the meaning of crypto tokens. Unlike cryptocurrencies, they do not own a blockchain, but instead exist inside Ethereum or other blockchains that allow for multiple tokens. As they are not the means of payment of the protocol itself, their valuation depends on different reasons than those of cryptocurrencies. For example, a token for a baseball ticket depends on its supply and demand, as well as the expectations on that baseball match. Tokens are used as a payment method when parties exchange a product or service for a specific quantity of tokens. Nevertheless, it should be noted that payments or exchanges of these assets are considered to be "outside the regulatory perimeter" in some countries, such as the United Kingdom and Switzerland (FCA, 2019; FINMA, 2018). Utility tokens, on the other hand, are developed so as to provide access to do certain actions in a specific ecosystem. For instance, if we were to gamble on an online casino game powered by blockchain technology, most probably they would offer utility tokens as chips. Lastly, security tokens aim at validating the ownership of financial securities. The most popular example is the share token, which evidences ownership of a portion of a company, just as a stock contract would. Their regulation is tight and many

countries, like Germany and Hong Kong, already consider security tokens as financial instruments under their jurisdiction (The Tokenizer, 2021).

The three token categories mentioned above: payment, utility and security, all fall inside a macro category called "fungible". This means that they can be replicated and are divisible, e.g., someone may have multiple $50 casino chip-tokens, which at the same time can be exchanged for many other chips of lesser value. Unlike fungible tokens, "non-fungible" tokens, better known as NFTs, are unique and non-divisible, so their valuation depends mainly on their exclusivity. The NFT market has seen a sharp increase in 2021. Daily trading volume in 2020 was $183.000, a far cry from the $38 million traded daily in 2021. They represent ownership of a digital asset and because they work along with smart contracts, the artist of the asset gets a royalty every time the NFT changes hands. Some of these artworks can be truly valued as masterpieces of art. *Everydays: The First 5000 Days*, an NFT from the artist Beeple, was sold in March 2021 for 38.525 ETH, the equivalent of $69.3 million. The reasons why people value these pieces at such extreme prices is worthy of another paper, but it is clear that some of them can even buy a house —or a palace—.

Albeit we have evidenced the fact that cryptocurrencies, as well as derivatives like crypto tokens and NFTs can be fully considered as a transaction method and even as a means of payment, doubts arise as to how financially secure they are. High volatility and speculative prejudices have become one of their main challenges. In the next lines we will deepen into the drivers of such volatility and analyze how their prices are set by the market.

7. Price and volatility of cryptocurrencies

Throughout the years, cryptocurrencies have experienced several bubbles and bursts[8]. Bitcoin, for example, was valued $1 in 2011, then surged to almost $20.000 in 2017, plummet to the $3.200 levels and later skyrocketed to values over $60.000 in 2021. Certainly, the possibility of investing in *tenbaggers*[9] sounds more than appealing to anyone, but precisely because of that, there is a high risk associated with it. As in a vicious circle, the more explosive the return are, the more people will invest in the asset and hence the more its value will grow, to the point where a burst is inevitable; whether it is due to a financial crisis, a rise in the interest rates, or simply because of a general loss of confidence in the asset.

When studying the factors that may influence the price and volatility of cryptocurrencies, we find different opinions within academia. Kristoufek (2015) argues that there are three main types of drivers affecting bitcoin: economic, transaction and technical.

Economic drivers refer to those whose impact is related to the economy. For example, he correlates the variation of bitcoin prices with a proxy for its demand to find out that the coin appreciates when it is used as a means of trade —i.e., to buy things— instead of as a purely speculative method; and that when its price increases, the speculative transactions increase in the short term too. This follows a common sense because when bitcoin is used for buying a car, for example, its value increases as demand increases and, as a side effect, it shortly increases the volume for speculative purposes. Commodity prices also play an important role in the price of bitcoin. Theory tells us that when the commodity

[8] Considered as massive fluctuations in investors' sentiment and in market capitalization.

[9] A tenbagger is an investment that generates an explosive return, reaching 10 times its initial value. The word was coined by the US investor Peter Lynch.

prices of a currency devalue relative to other currencies, the latter increase in value. He reasons that bitcoin does follow this standard monetary theory. Another metric for analyzing variation in prices is the money supply of bitcoin. In this case, and probably because the mass of money supply is known in advance —as we have already seen— it does not generate a direct impact, which most certainly implies that money supply is already fused in present prices.

When he studies transaction drivers, he mentions the trade volume of bitcoin, which turns from positive to negative depending on the period, therefore giving no concrete answer. He also insists on the fact that the more trading there is —understood as purchases with the cryptocurrency— the more positive correlation with its price.

The technical drivers he references include those which he considers to have an impact on prices because of the underlaying mechanism of Bitcoin. He analyzes two sources for price variation, which are the increase in difficulty and the hash rate, the last one understood as the number of computational operations that a network of miners is capable of performing. For both of them there is a positive correlation, as expected, since the more difficult it becomes to mine a bitcoin and the higher the hash rate, the less supply there is.

Another source of price variation he mentions which is of predominant importance in this paper is the interest of people on cryptocurrencies. Easy access, coupled with low initial investment and media hype create a perfect combination capable of attracting both amateurs and speculative investors. To quantify the interest in bitcoin, he uses Google and Wikipedia search engines to quantify how many people were looking for such cryptocurrency. With the help of wavelet graphs, he concludes that for most of the bubbles there is a strong and positive correlation between prices and interest when they rise, and a weaker and negative

correlation when prices fall, thus showing an asymmetric effect in interest during the bubble formation and its bursting.

On the other hand, Hayes (2014), doubts Kristoufek's idea that the Google engine results are a trigger for bitcoin's increase in price. He argues that a rapid rise in price attracts media attention and word of mouth, leading to more people becoming interested in the cryptocurrency, who eventually turn to the internet for more information. He claims that those who actively mine or trade bitcoin do not need to search for the word "bitcoin" as a search term on Google on a regular basis, but rather people who are hearing about it for the first time or who want to learn more about it. According to him, there are three drivers that are able to explain 84% of the variability in bitcoin prices: computational power, the number of coins minted per minute and the algorithm being used. The computational power variable is defined as the amount of energy and work that a computer needs to perform in order to get one unit of a crypto coin; the number of coins minted per minute is the quantity of money supply that the blockchain is able to reward; and the algorithm being used consists on whether the cryptocurrency is using a scrypt or SHA-256 cypher algorithm for securing the network.

Certainly, both authors have valid reasons when arguing price movements through popularity, and it is complex to determine if there is simultaneity bias between Google trend's variable and the increase —or decrease— in the cryptocurrency price. Other authors, however, have moved forward with the idea of internet trends reasoning that tweets from the social network Twitter are indeed drivers of change and infer causality. On the one hand, Dibakar Raj et al. (2018) made a model for analyzing whether tweets had positive or negative impact, and a recurrent neural network (RNN) to predict the price trend for bitcoin given the tweet input. Their main conclusion was that about a 77% of

bitcoin price movements could be predicted with those short messages. On the other hand, Stenqvist & Lönnö (2017) stated the opposite by mentioning that, for the period between May 11 to June 11, 2017, they was no significant correlation between the 2.271.815 tweets they mined and the price of bitcoin, although for small subsets of data, especially those with more prominent fluctuations, there was a >50% prediction accuracy, indicating partial correlation for those periods. We will now begin our journey towards the creation of a model that will provide yet another conclusion to all this academic work, in addition to responding to the thesis of our project.

8. Overview and implementation of our model

Below, we will proceed to explain the hands-on section of this project. As mentioned earlier, our purpose is to design a model and find out whether there is any predicting logic between tweets' sentiment score and crypto prices. Since it would be infeasible to include all known cryptocurrency prices in our models, we decided to limit our study to the fifteen cryptocurrencies with the largest market capitalization, excluding all those that are pegged to the U.S. dollar —see figure 8.1—.

Cryptocurrency	Symbol	Market cap. Ranking
Bitcoin	BTC	1
Ethereum	ETH	2
Binance coin	BNB	3
Solana	SOL	4
Cardano	ADA	5
XRP	XRP	6
Luna	LUNA	7
Polkadot	DOT	8
Avalanche	AVAX	9
Dogecoin	DOGE	10
Shiba Inu	SHIB	11
Polygon	MATIC	12
Litecoin	LTC	13
Cosmos	ATOM	14
Chainlink	LINK	15

***Figure 8.1**, the 15 currencies of interest for our study*

8.1. Gathering data

There is a very well-known sentence in the field of data science that states:

Garbage in, garbage out

This principle applies to data that is faulty, incomplete, or that needs some preprocessing before feeding it into an algorithm. Aware of this, we searched different public and free webpages where we could draw accurate and reliable

data, until we found CryptoArchive, whose main purpose is to provide cryptocurrency prices for free and with a simple interface. All its data is retrieved on a daily basis from Binance, a crypto exchange platform which, as of 2020, is considered to be the one with most trading volume in the world. Since we could not find on that website any official pairings with the US dollar, we had to rely on the 15 cryptocurrencies with the largest market capitalization paired with USDT. What this means is that instead of analyzing the BTC/USD pair, we did so with the BTC/USDT pair, and since the USDT is a crypto pegged to the U.S. dollar, we will assume —throughout the rest of the paper— that 1 USDT is equal to 1 USD.

The information is given in the usual tabular form and presents several columns: timestamp, open, high, low, close, volume, taker buy quote asset volume, taker buy base asset volume, quote asset volume and number of trades. The open timestamp represents the unix time[10] of the information given. The open, high, low, close and volume columns display the standard OHLC prices minute by minute with their volume of shares traded. Finally, we dropped from our table the following variables: number of trades, taker buy quote asset volume, taker buy base asset volume and the quote asset volume; as they were not of concern for our study.

Regarding Twitter data, we extracted tweets from relevant personalities related to digital currencies (they can be found in figure 8.2 in the appendix). Some of them have been found by general knowledge —e.g., Elon Musk and Vitalik Buterin— due to their strong relationship with this field, and the rest

[10] It is the number of seconds that have elapsed since the Unix epoch, excluding leap seconds. The Unix epoch is 00:00:00 UTC on 1 January 1970. (Wikipedia)

have been located from different internet newspapers and blogs (more in figure 8.3 in the appendix). By limiting our tweets to specific users, we have solved two different problems that otherwise would have risen.

Firstly, if we had focused our search on the tickers —or symbols— of each currency rather than the Twitter users, we would have biased our database, since many of those tickers have other meanings too —i.e., LINK and link—. It is also the case in other non-English languages —i.e., SOL and *sol*; LUNA and *luna* (both sun and moon in Spanish)—, fact that could have added non-related tweets to our model. In figure 8.4 we find the total number of tweets in the social network when we look for the following keywords:

Coin	Number of tweets	Key words	Market Cap. Ranking
Litecoin	5.627.930.018	LTC	13
Cardano	2.314.523.060	ADA	5
Chainlink	1.462.561.510	LINK	15
Solana	380.214.595	SOL	4
Bitcoin	122.985.654	BTC	1
Polkadot	113.700.619	DOT	8
Luna	95.203.823	LUNA	7
Ethereum	94.144.002	ETH	2
Dogecoin	50.291.279	DOGE	10
XRP	48.408.295	XRP	6
Binance coin	47.085.036	BNB	3
Shiba Inu	29.810.431	SHIB	11
Polygon	24.978.156	MATIC	12
Cosmos	22.279.154	ATOM	14
Avalanche	5.465.938	AVAX	9

Figure 8.4

As we can see, the number of tweets is not only vastly distorted with the market capitalization at the end of 2021, but also the number of tweets increases massively.

Secondly, and as seen in figure 8.4, there was a clear storage problem. Not filtering by users would have meant storing more than 10.000.000.000 tweets. Considering that a tweet can occupy up to 280 bytes we would have needed a hard disk with 2.722 GiB of space. Assuming a better scenario where all the tweets

took up half their maximum space, we would have needed 1.361 GiB, an impossible assignment for the general usage laptop we were employing, a 500GiB M1 MacBook Air.

Lastly, and after these considerations, we gathered 1.143.993 tweets, of which we later erased all those prior to the date of creation of the cryptocurrencies. For example, Solana was created on January 3, 2009, so all those tweets that were written before that date and contained the word "solana" or "SOL" were deleted. The same is done for all the rest.

Now that we have seen the collection of the data for both time series of prices and tweets, let's move ahead with the wrangling and preprocessing of data, very much needed to feed the model correctly.

8.2. Preprocessing data

For the transformation of data, as well as for all the models, we have used python and SQL queries. Python is a well-known programming language that, along with some libraries like Pandas and Regex, makes programming rather straightforward. After creating the main table with 1.141.993 tweets, we proceeded to take each one of them, categorize them according to the type of crypto they were talking about, and analyzing a sentiment score for each one of them.

By using Regex and a loop we were able to search from each tweet the cryptocurrency(ies) they were talking about and, if there was more than one of them, we duplicated the tweet into different rows, one for each crypto. After that, we proceeded to clean the text as much as possible. The following criteria was used:

1. Transforming the text to lower caps.
2. Substituting the hashtag "#crypto[11]" to just the word "crypto".
3. Removing any mention to another user, e.g., deleting "@elonmusk".
4. Erasing hashtags other than "#crypto".
5. Deleting all links to external websites.
6. Dropping the new line character "\\n", present in the tweet extraction.
7. Erasing any special character, mainly "|-+*'#,;\", and the ampersand "&".
8. Deleting any number

All these changes were taken because a computer is not able to read them and output a sentiment score as a result. Finally, and as a consequence of reading text from an informal social network with a limit of 280 characters, we chose to translate several internet slang acronyms into their real meaning, e.g., *idk* into *I don't know* (see more in figure 8.5 in the appendix).

Once the wording was cleaned up, it was time to tokenize and stem our words. Tokenization means that we split a whole text into smaller units. For example, a tweet with the form "Hello world!" would have become ["Hello", "world", "!"]. This method is useful for performing several changes to the words, like stemming, which consists of removing the part that is not considered its root or stem. For instance, *information* is transformed into *inform*, and *computers* into *comput*. An alternative to stemming is lemmatization, whereby we transform a word into its lemma, or origin word. Following the previous examples, *information* would transform into *information* and *computers* into *computer*. Which method we choose is indifferent since the sentiment algorithm is able to read and output

[11] "crypto" can be any of the 15 digital currencies analyzed

a score regardless of the method we use. However, it is important that we do use one of them to remove data redundancy when using it in machine learning models.

There is a natural language processing (NLP) python package —NLTK— that helps us in the procedure of stemming by offering 4 different types of algorithms: Porter stemmer, Snowball stemmer, Lancaster stemmer and Regexp stemmer. We decided to use the Snowball stemmer because it is the most widely used algorithm within academia and enjoys prestige on a wide variety of internet blogs.

Lastly, once all of the above steps were completed, it was time to execute the sentiment analysis algorithm. Certainly, creating one by our own would have been the best option due to the fact that our field of study and the words we are looking for are quite specific, but doing so would have taken much longer to implement, so we turned to a built-in sentiment model called Sentiment Intensity Analyzer (SIA) from the same NTLK package. We decided on this particular one because it is one of the best at reading from social media, as it can read emojis, slangs and emoticons like ":)". The NTLK algorithm outputs sentiment scores to 4 different sentiments: negative, neutral, positive —these three add to 100%— and compound, the latter being an aggregated score. As a general rule of thumb on the data science community, when the compound score is greater than 0.05 then the sentiment is considered to be positive; negative for those lower than -0.05 and neutral otherwise.

In the following lines we will present a descriptive analysis on the behavior of crypto prices, and later, we will do the same with Twitter tweets.

8.3. Crypto price analysis

Before analyzing the core results of this project, it is important to understand the price conduct of cryptocurrencies and assess the extent to which they are similar to a closed-book statistical definition. A common hypothesis in the realm of quantitative finance is that of the random walk, according to which market prices move in a random pattern and therefore cannot be predicted. If, when performing this analysis, we see that they follow this same pattern, we will most likely have no chance at predicting price movements with our models. Oddly enough, the academia has not yet found any irrefutable study showing pure randomness in stocks.

Financial markets usually replicate random walks with normally distributed step sizes, which means prices increase according to a Gaussian random variable. If we know P_t, then P_{t+1} must be equal to $P_t + N(\mu,\ \sigma^2)$, where $N(\mu,\ \sigma^2)$ is a Gaussian —or normally distributed random variable— with mean μ and variance σ^2. In this type of studies, it is better to analyze the returns of prices instead of the prices themself, as they have better characteristics, mainly constant mean and variance. Algebraically speaking, if we know P_t , then we could know the return on P_{t+1} with this formula: $R_{t+1} = \frac{P_{t+1}}{P_t}$. However, we still need to further tune the returns if we want them to behave like a normal random variable. Prices range from \$0 to, theoretically, \$∞. Consequently, returns must range from 0% to +∞%. On the contrary, minimum and maximum values for a normal distribution are -∞ and +∞, respectively. By using the formula $r_t = log(R_t)$ we solve this issue, as now we are able to bring all those returns close to 0% towards -∞, while keeping the upper limit at +∞, as seen in figure 8.6.

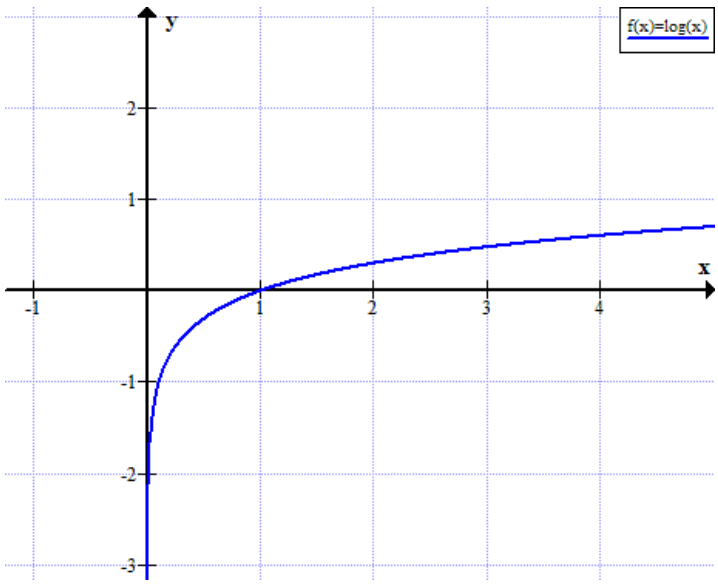


Figure 8.6

The histogram of bitcoin log returns can be found in figure 8.7.a with an orange line representing the kernel density function for a normal distribution with same mean and same variance as the bitcoin sample. Significant disparity can be found between the two distributions. Bitcoin histogram exhibits a substantial amount of mass in the center and thin tails, advising that extreme returns are not as likely as a normal distribution would suggest. On the other hand, figure 8.7.b displays the histogram of a random walk sample following a normal distribution. As we can see in this case, the realization does fit the kernel density function as they both behave similarly.

Another way to check if log returns are normally distributed is by performing a kurtosis test that checks how differently shaped are the tails of a distribution as compared to the tails of the normal distribution. If there is a conclusive result that the kurtosis is different from a normal distribution then the null hypothesis must be rejected. In figure 8.8 we see the test for the bitcoin sample of figure 8.7.a and for the normally distributed sample of figure 8.7.b. As results show, the p-value for the bitcoin sample is well beyond the <0,05 threshold,

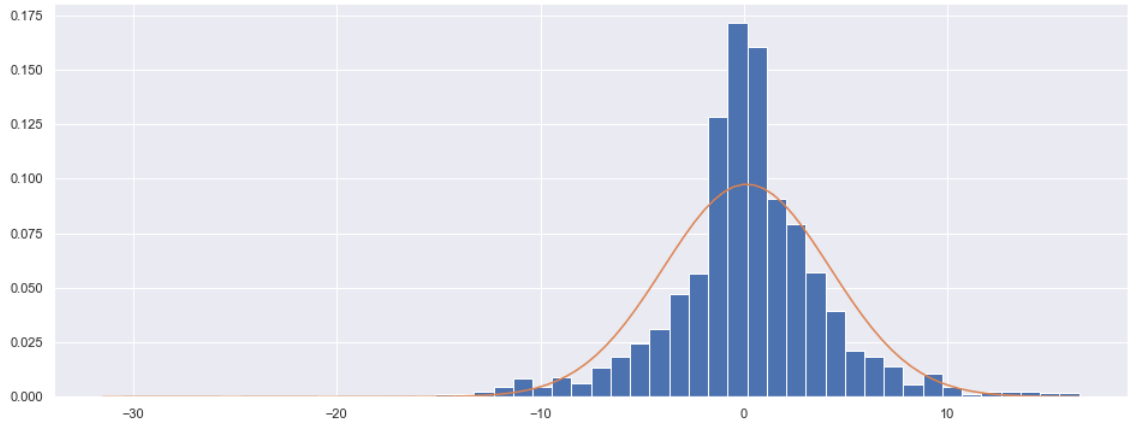


Figure 8.7.a

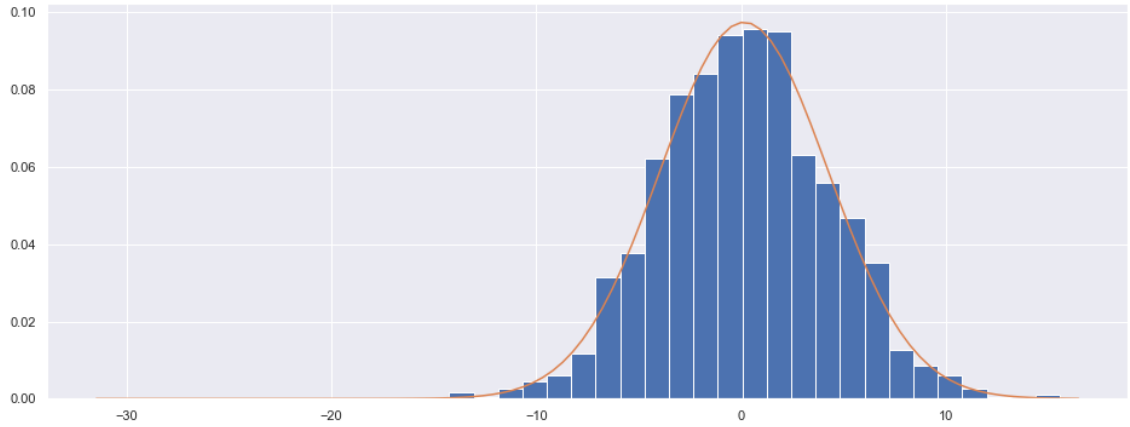


Figure 8.7.b

exposing no signs of normality; whereas the normally distributed sample —represented with an "x"— does fall within the null hypothesis conditions.

	Test statistic	p-value
x:	1.02	0.3100
BTC:	13.80	0.0000

Figure 8.8

After examining the kurtosis test and p-value for the rest of the cryptos we can confirm that none of them has a distribution similar to that of a Gaussian random variable. Accordingly, there is some chance that we can end up building a model with predictive power. The full study of this quantitative analysis can be found in the attached folders of this paper. Next, we will do a similar study to the one done in this section but with Twitter data.

8.4. Tweet analysis

We will now judge Twitter data by plotting various graphs and relating them to the project. The first descriptive graph we proceed to draw is a bar chart with the number of tweets of each cryptocurrency —see figure 8.9—. Whenever any tweet contained its full name (e.g., bitcoin), or its symbol (e.g., BTC), it was added to the count. Note that some tweets may be counted more than once, as they may talk about more than one cryptocurrency at a time.

Although not fully aligned with the market capitalization ranking—that of 2021—, graph in figure 8.9 truly represents a proxy of the market purchasing habits, since bitcoin and ethereum are on the top positions and the following ones, generally, do have more market capitalization than the bottom ones. Another takeaway to be noted is the popularity of bitcoin in the social network. Ethereum, the second-most popular cryptocurrency in the world, is only talked about 20% as much as bitcoin, with the rest having substantially lower percentages.

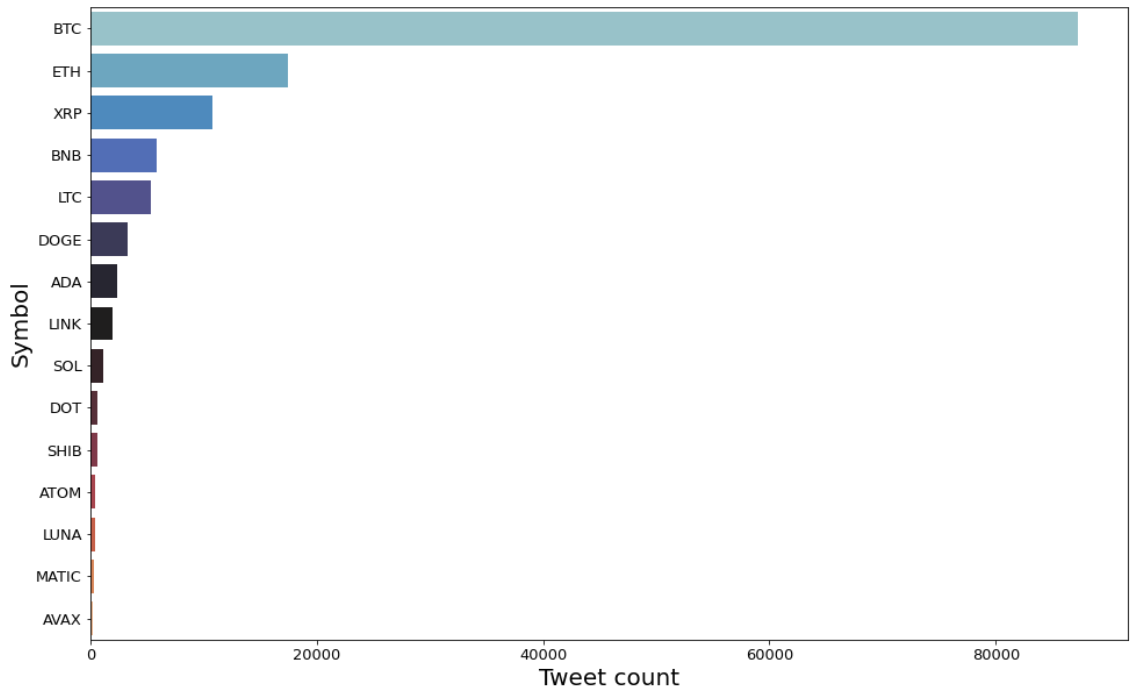


Figure 8.9

This may be one explanation why Google Scholar website has plenty of academic papers talking about bitcoin rather than any other cryptocurrency.

Another interesting analysis is that of the association between cryptocurrencies. When someone talks about crypto "X", do they also talk about crypto "Y" in that same tweet? For that, we took all those texts that talked about more than one currency and proceeded to plot the percent stacked bar graph we see in figure 8.10.

On the y axis we find the crypto of interest; on the legend the color of the other cryptocurrencies that are also being talked about; and, on the x axis, their percentage. For example, if we are interested in knowing which other coins are being talked about when someone writes about XRP, we just need to take a look at its horizontal bar. When mentioning XRP, most people also talk about bitcoin —roughly 45% of them—. A key takeaway is the intertwine between shiba inu with dogecoin, ethereum with bitcoin, and bitcoin with ethereum.

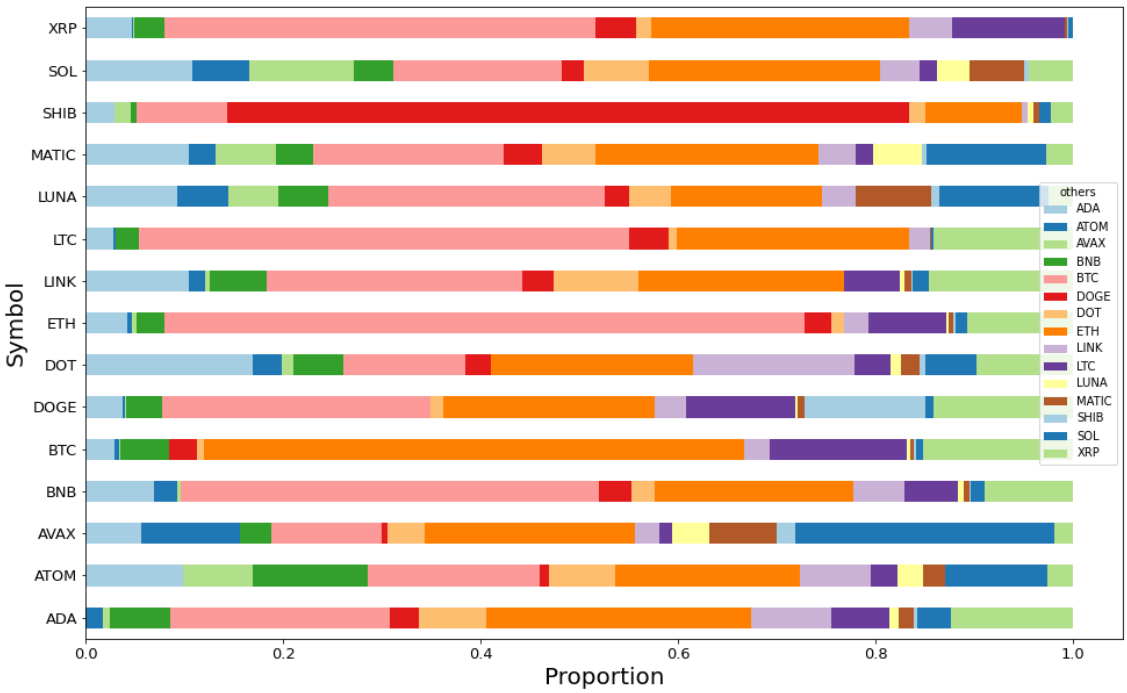


Figure 8.10

The first phenomenon most probably occurs because shiba is much younger than doge —2020 and 2013, respectively—, and both arose as a joke payment system. The second and third are likely to be caused by the importance they represent in the world of cryptos and because of their high positive correlation.

Moving on to sentiment scoring, it is time to look at how positive or negative Twitter users understand each crypto. In figure 8.11 we notice the sentiment score distribution per crypto. An unusual 25% percentile at 0% seems to be the norm for all coins and, upon further investigation, we realized that all scores between [-100%, 0%) represented far less than 25% than the sample. Likewise, positive scores represent a small part of the sample, therefore most tweets have exactly the same score: 0%. In practice this is bad news, as our model might be fed with a great number of unscored tweets, generating more noise and poorer predictive power. Figure 8.12 further identifies this dilemma by plotting the bitcoin sentiment score distribution. Too many scores with 0% means that we had to delete them from our database and not take them into account when building our model.

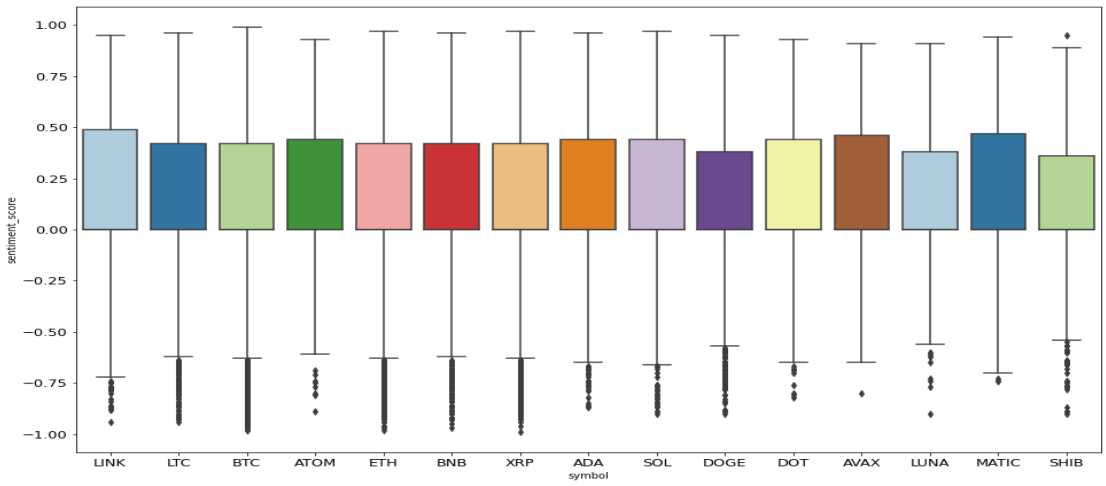


Figure 8.11

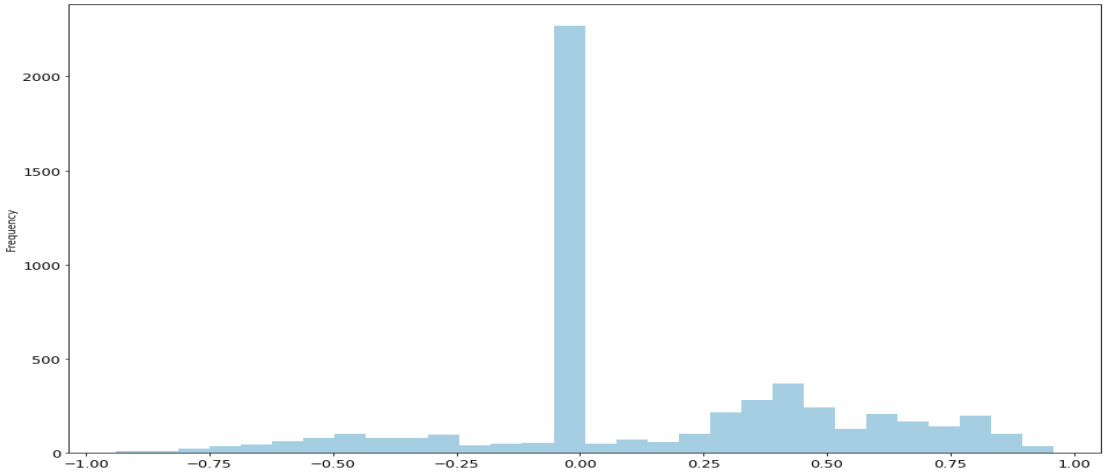


Figure 8.12

Finally, we took a look at those tweets that had the greatest number of retweets. To do this, we first ranked the retweet variable from highest to lowest and then grabbed the top 5. To our surprise, these were all about the same topic: giveaways. An example can be found below:

> I'm giving away a DOGE Tesla & $50,000 in $DOGE. Which do you want?
>
> To enter:
> 🔶Retweet
> 🔶Follow @elonmusk, @binance, @cz_binance
> 🔶Have a KYC'd Binance account
> 🔶Tweet the prize you want & why with hashtags #DOGEorTesla & #Binance
>
> Our favorite answer from each 'team' will win. https://t.co/A2VLB6fh4a

This tweet from *cz_binance*, Binance CEO, is the tweet with most retweets in our database, with a total of 94.113 accounts sharing the post. Interested in this topic, we decided to inquire into this type of tweets, and so we investigated all those that had at least 15.000 retweets. To our surprise, 15 of the 20 texts that met these conditions were giveaways. Besides, 18 of them presented highly

positive scores, 1 neutral, and 1 negative. Since these types of promotions can still alter the market and they mostly present positive scores —which is the correct sentiment for promotions— we did not drop them from our database.

8.5. Generating the model

Finally, after cleaning, processing and understanding our data, we produced a model with the intention to solve the proposal of our paper:

Is it possible to predict the price movement of cryptocurrencies using sentiment score in tweets?

For that we did compute our models twice, as we not only inputted the sentiment score as the only feature, but also the number of retweets —see figure 8.13—. Because of that, we will be referring them as scenario A and scenario B, respectively.

	Scenario A	Scenario B
Features	Sentiment score	Sentiment score and number of retweets
Target	Price movement	Price movement

Figure 8.13

At first, we wanted to make a model that directly linked the timing of tweets with that of prices, assuming that a tweet would have a direct impact on the market. That is, if Elon Musk posts at 8:56 am, it should have an impact on price movements at 8:56 am. However, we then realized that there was a possibility that it could take a few minutes for the market to correct itself, so we added extensions to scenario A where we considered a delay of 1, 2 and 5 minutes between the tweet being posted and the price movement. So, for example, we

added the possibility that the same Elon Musk tweet at 8:56 am could have an impact on the market at 8:57 am, 8:58 am and 9:01 am.

For the sake of simplicity, the model tried to predict price movements —up or down— rather than the magnitude of the price difference. That means that the target variable consists of 1s in case log returns go up and 0s otherwise. In addition, we did not take into account time periods, i.e., we did not segment recessionary periods from non-recessionary periods, nor did we try to define periods with higher predictive accuracy than others.

Choosing to predict price movements over price differences means that we needed to use classification models instead of regression models. Thus, the following six classification algorithms were proposed:

- Logistic regression (LR)
- Support Vector Classification (SVC)
- K-Nearest Neighbor (KNN)
- Naive Bayes (NB)
- Decision Tree Classifier (DTC)
- Random Forest Classifier (RFC)

All these ML algorithms can be found in the Scikit-Learn library and were used applying a train-test split of the data. Only a percentage of the information —the train— was fitted into the model to find the optimal parameters; and the rest of the data —the test— was used to check that those parameters do indeed fit the unseen data correctly. We used 80% of the data for the train sample and 20% for the test. Finally, we plotted receiver operating characteristic (ROC) curves for each studied crypto in the 0-minutes delay extension. The greater the area under the curve (AUC), the better the model is at predicting price movements.

Figure 8.14 and 8.14 bis in the appendix show ROC curves for all ML algorithms and cryptocurrencies using scenario A. For all those whose test set is greater than 100 samples, their AUCs never exceed 52%, except for cardano, whose random forest AUC tops 56%. Conversely, coins with less than 100 samples in the test set perform noisily with AUCs that reach maximums of 71% in the case of Polygon (MATIC) and minimums of 37% in the case of avalanche (AVAX). However, when assessing the extension of scenario A for the 1-, 2- and 5-minutes delays —figures 8.16 to 8.18 in the appendix—, we observe better results. In figure 8.16, the best performer with less than 100 test samples is luna with a 59% AUC, and out of those who have more than 100 samples, litecoin is the best with almost 56% AUC. It is also worth mentioning that litecoin's AUC surges to a whopping 86% and 75% AUC in the 2- and 5-minutes delay, respectively. Other notorious performances are those of Avalanche in the 2-minute delay with 76%— and Shiba inu in the 2-minute delay —with 67%—.

Figures 8.15 and 8.15 bis in the appendix contain the same graphs as figures 8.14 and 8.14 bis, although in this case for scenario B. Clarifications on these are almost the same to those on scenario A. Again, cardano is top performer out of those with more than 100 test samples, as it achieves 56% AUC with its random forest. In addition, cryptocurrencies whose test samples are less than 100 samples again indicate high variability with maximum AUCs of 67% and minimum AUCs of 35%.

9. Conclusions and outlook

Money can be seen as a flawed mechanism where participants blindly believe that owning a mass of tokens will generate wealth and peace of mind. As we cannot produce everything we need by ourselves —be it fruits, a car or a home—, we need to generate such wealth through these currencies and thus we assign them a value according to market consensus. The more desired a product is, the more coins we need to redeem to purchase that item. However, we have seen the inverse is also possible. Currencies are also able to change value depending on their supply and demand, and that is why, just as there is a real estate market for people who want to buy or sell a house, there is also a market for people who want to buy and sell currencies. So far, entities known as central banks have been the only authority to manage most of the money market by controlling their supply. The rest has been driven by investor sentiment, the psychological idea of robustness and confidence behind the coin, and the performance of the nation linked to it. If a government defaults on its sovereign debt, if a country's inflation fails to stabilize sufficiently, or if there are any other factors that cast doubt on the effectiveness of a currency, then these lose value and become less attractive than others. The advent of cryptocurrencies has distorted this status quo by challenging the rules of the central banks themselves. Through cryptographic algorithms and distributed networks, cryptocurrencies have been born as an alternative to fiat money. Although far from being a conventional payment system, they are already quoted in the markets and suffer price fluctuations like any other financial derivative.

Given these conditions, we wondered if it was possible to predict the price of these crypto assets with Twitter social network. It is known that tweets can be very powerful, and that is why we chose certain crypto celebrities and analyzed

their tweets to then find the ones that were more focused on our study. After, we searched for cryptocurrency prices through a website that extracted data from Binance, and merged both databases —the tweets and the prices— in order to continue our investigation. Finally, and after processing an NLP algorithm to each of the social network texts, we built a predictive model capable of responding our project thesis.

Upon execution and extraction of the results, we can proceed to conclude this work. Judging by the information given on figures 8.14 and 8.14 bis, there seems to be no predictive power whatsoever for all those cryptos with more than 100 test samples. What is more, one of them, the binance coin in figure 8.14, only shows AUCs below 50%. In another scenario one could argue that turning around the feature variable —in this case, $(\text{sentiment score})^{-1}$— would improve the predictive outcome. Notwithstanding, we cannot use this method as then we would be questioning the original definition of the sentiment score —we would be changing positive for negative scores, and vice versa—. Instead, we accept the idea that our model has no predictive power for these types of assets. Turns out to be different for the other subset of cryptos —those with less than 100 test samples. They present several performance metrics with opposite results. Polygon, for example, is a perfect case that matches high predictability with a low number of observations. On the one hand, we believe that —similar to the law of large numbers— there is not enough data to assess with certainty what is happening, as we see that the bigger the data, the smaller the AUCs are. On the other hand, we are aware that there is also the possibility that these currencies, being less known and having less market capitalization, may be easier to manipulate through social networks and therefore the scenario actually shows effective predictive power.

Figures 8.16, 8.17 and 8.18[12] provide more interesting results, indicating AUCs of up to 86% and of as low as 28%. Again, this high volatility generates confusion to the point that we do not know whether we are generating good or bad models. In our opinion, as much as we have such high metrics, if they are accompanied by others much lower than 50%, it may be a sign of randomness in the model's prediction. In the case of cryptocurrencies with higher samples, their AUCs seem to hold up with respect to those on figure 8.14, validating the previous idea of randomness.

Figures 8.15 and 8.15 bis suggest analogous conclusions to those mentioned above. They show virtually no difference compared to the performance of scenario A and suggest the same concerns as before: AUCs below 50% for some cryptos and AUCs closer to 100% for smaller samples.

All in all, we assert that, with the publicly available data we have found and with the study conducted in this paper, Twitter does not appear to have predictive power over cryptocurrency prices, at least not for those with most mentions. As a next step, we believe that either the number of influencers should be expanded to accommodate an analysis with more tweets, or a predictive model capable of differentiating tweets related to cryptocurrencies should be trained —as discussed with the LINK and link analogy—. In addition, we believe that a more exhaustive study should also be carried out for all those cryptocurrencies with few observations, since they are most likely to be more influenced by social networks.

[12] Figures 8.16, 8.17 and 8.18 have cryptos with test samples of less than 100 observations filled in red on the side.

10. Appendix

***Figure 3.2**, annual inflation rate in the US*

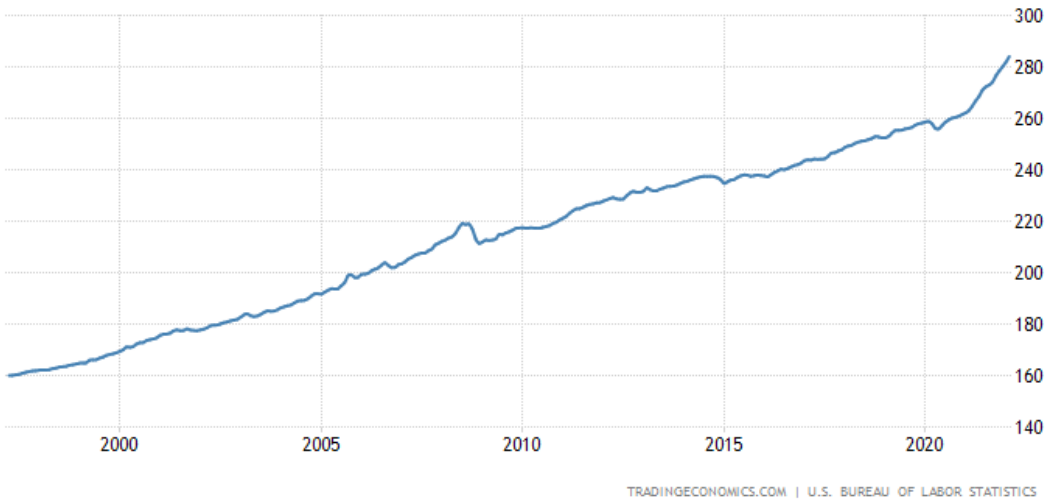


***Figure 3.3**, Consumer Price Index (CPI) YoY in the US*

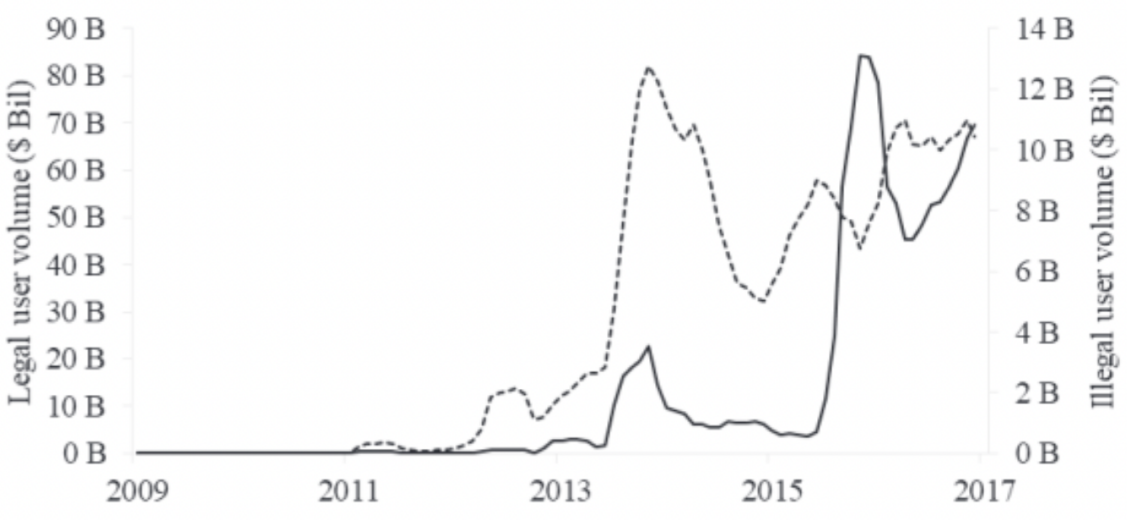


***Figure 3.4**, Estimated dollar volume of illegal and legal bitcoin users' transactions per month*

***Source**: Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies?*

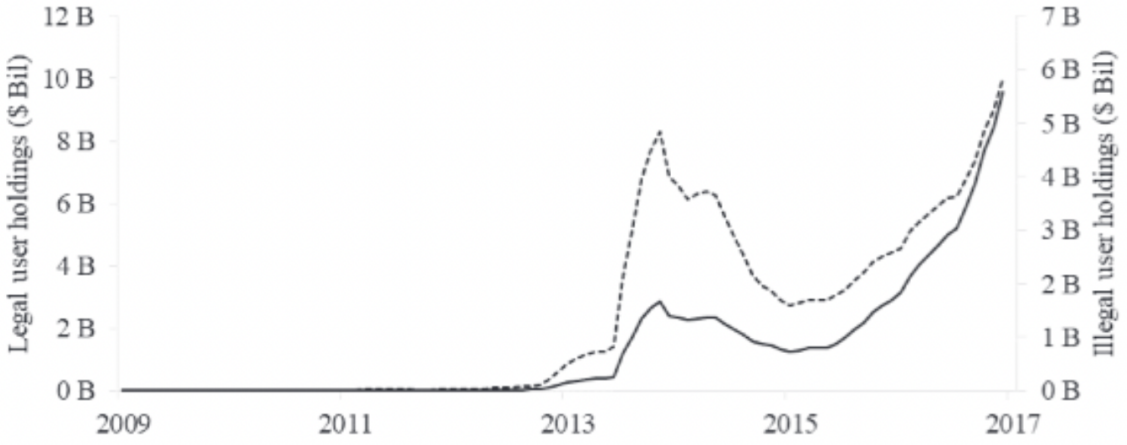


***Figure 3.5**, Estimated dollar value of illegal and legal users' bitcoin holdings*

***Source**: Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies?*

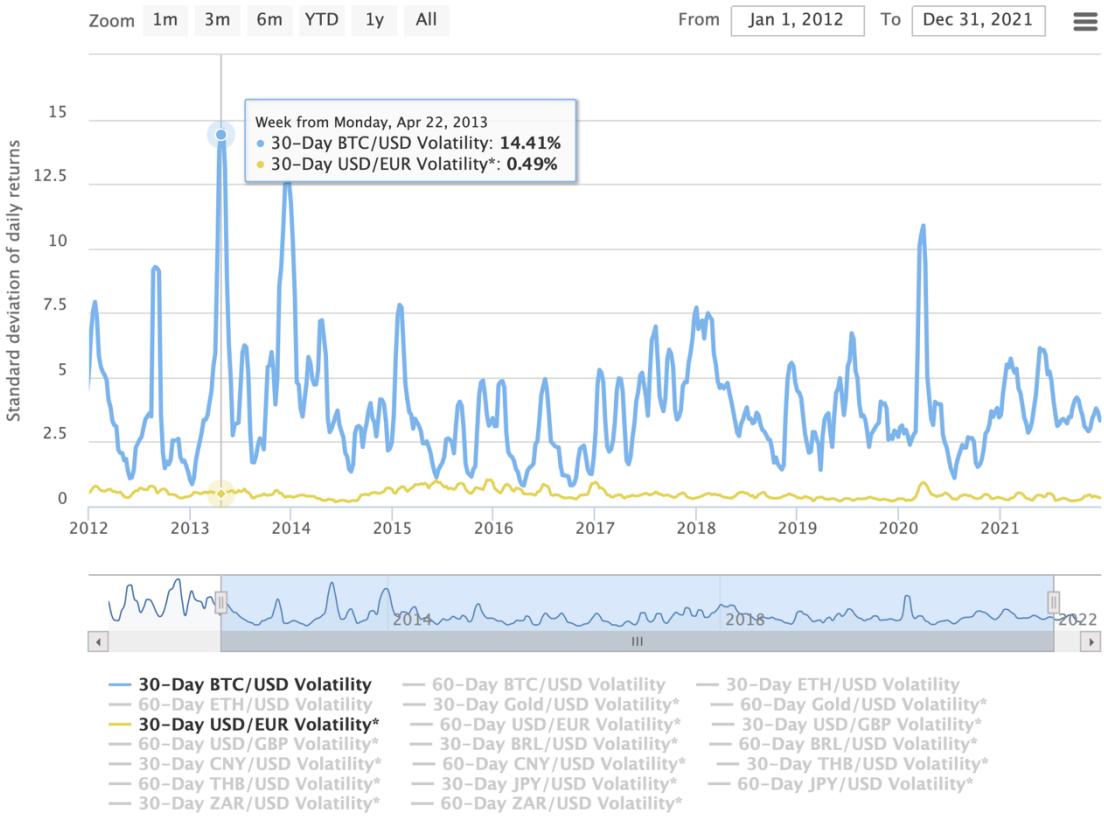


***Figure 3.7**, Bitcoin and USD/EUR 30-day volatility*

***Source**: Buy Bitcoin Worldwide,*

https://www.buybitcoinworldwide.com/es/indice-de-volatilidad/

influencer	username	influencer	username
Vitalik Buterin	@VitalikButerin	Coin Bureau	@coinbureau
Roger Ver	@rogerkver	BoxMining	@boxmining
Andreas M. Antonopoulos	@aantop	Lark Davis	@TheCryptoLark
Tim Draper	@TimDraper	BlockchainLeaks	@LeaksBlockchain
Charlie Lee	@SatoshiLite	CryptoLove	@TheCryptoLove
Anthony Pompliano	@APompliano	Aimstone	@Aimstone5
Erik Voorhees	@ErikVoorhees	Hashoshi	@hashoshi4
Tone Vays	@ToneVays	Philakone	@PhilakoneCrypto
John McAfee	@officialmcafee	Cryptonauts	@CryptonautsShow
Ivan on Tech	@IvanOnTech	Jason Pizzino	@jasonpizzino
CryptoBrekkie	@BVBTC	Andreas Antonopoulos	@aantonop
Dan Held	@danheld	Roger Ver	@rogerkver
Layah Heilpern	@LayahHeilpern	Nick Szabo	@NickSzabo4
Kenn Bosak	@KennethBosak	CryptoCred	@CryptoCred
Ben Horowitz	@bhorowitz	Erik Voorhees	@ErikVoorhees
Elon Musk	@elonmusk	PlanB	@100trillionUSD
Ty Smith	@TyDanielSmith	Brian Armstrong	@brian_armstrong
CryptoWendyO	@CryptoWendyO	Loomdart	@loomdart
Euclid and Oaks	@EuclidAndOaks	Naval	@naval
David Gokhshtein	@davidgokhshtein	Credible Crypto	@CredibleCrypto
Hailey Lennon	@HaileyLennonBTC	Josh Olszewicz	@CarpeNoctom
Justin Sun	@justinsuntron	Marty Bent	@MartyBent
Ivan on Tech	@IvanOnTech	Tim Draper	@TimDraper
LayahHeilpern	@LayahHeilpern	Documenting Bitcoin	@DocumentingBTC
Coinbound	@coinboundio	Adam Back	@adam3us
Sheldon Evans	@SheldonEvans	Messari	@MessariCrypto
CryptoBusy	@CryptoBusy	Nick Szabo	@NickSzabo4
JRNY Crypto	@JRNYcrypto	Cred	@CryptoCred
BitBoy Crypto	@Bitboy_Crypto	Changpeng Zhao	@cz_binance
Whale Panda	@WhalePanda	Gavin Andresen	@gavinandresen
Camila Russo	@CamiRusso	Balaji Srinivasan	@balajis
Nicholas Merten	@Nicholas_Merten	The Wolf Of All Streets	@scottmelker

***Figure 8.2**, Personalities used for twitter sentiment analysis*

***Source**: own source*

Sources

https://influencermarketinghub.com/top-crypto-influencers/

https://coinbound.io/best-crypto-influencers-on-twitter/

https://crowdcreate.us/top-crypto-influencers-on-twitter/

https://blockwiz.com/crypto-marketing/top-crypto-twitter-influencers/

https://www.glyph.social/blog/top-50-crypto-influencers-and-thought-leaders-to-follow-on-twitter

***Figure 8.3**, Sources for crypto-personalities*

Acronym	Sentence
idk	I don't know
smh	Shaking my head
ikr	I know, right?
immd	It made my day
snh	sarcasm noted here
ama	as me anything
icymi	in case you missed it
dr	double rainbow
mfw	my face when
rofl	rolling on the floor laughing
stfu	shut the f**k up
nvm	never mind
tbh	to be honest
btw	by the way
aka	also known as
asap	as soon as possible
np	no problem

***Figure 8.5**, Slang acronyms and their real meaning*

***Source**: own source*

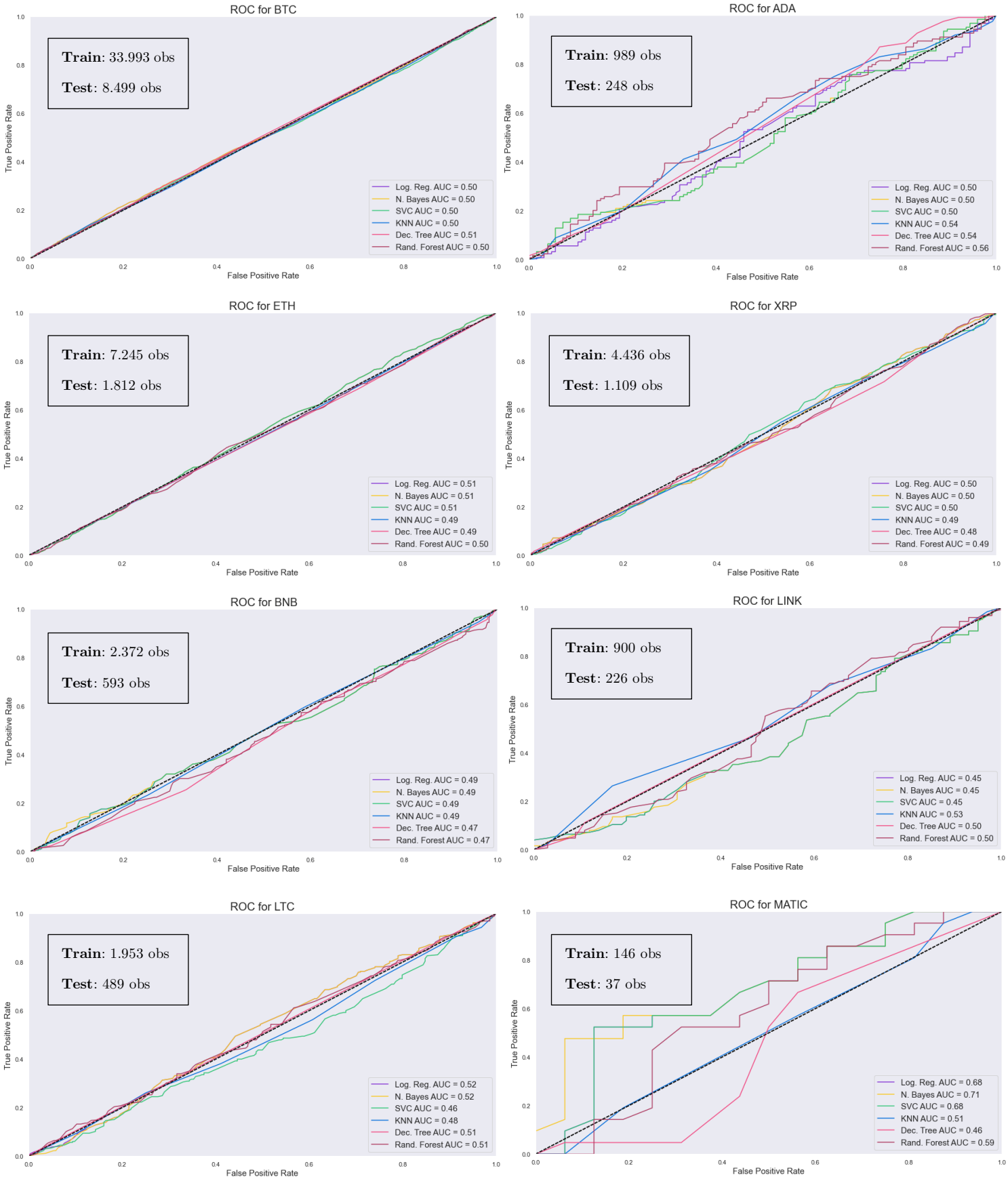


***Figure 8.14**, ROC curves for the 15 cryptocurrencies*

(Feature: sentiment score)

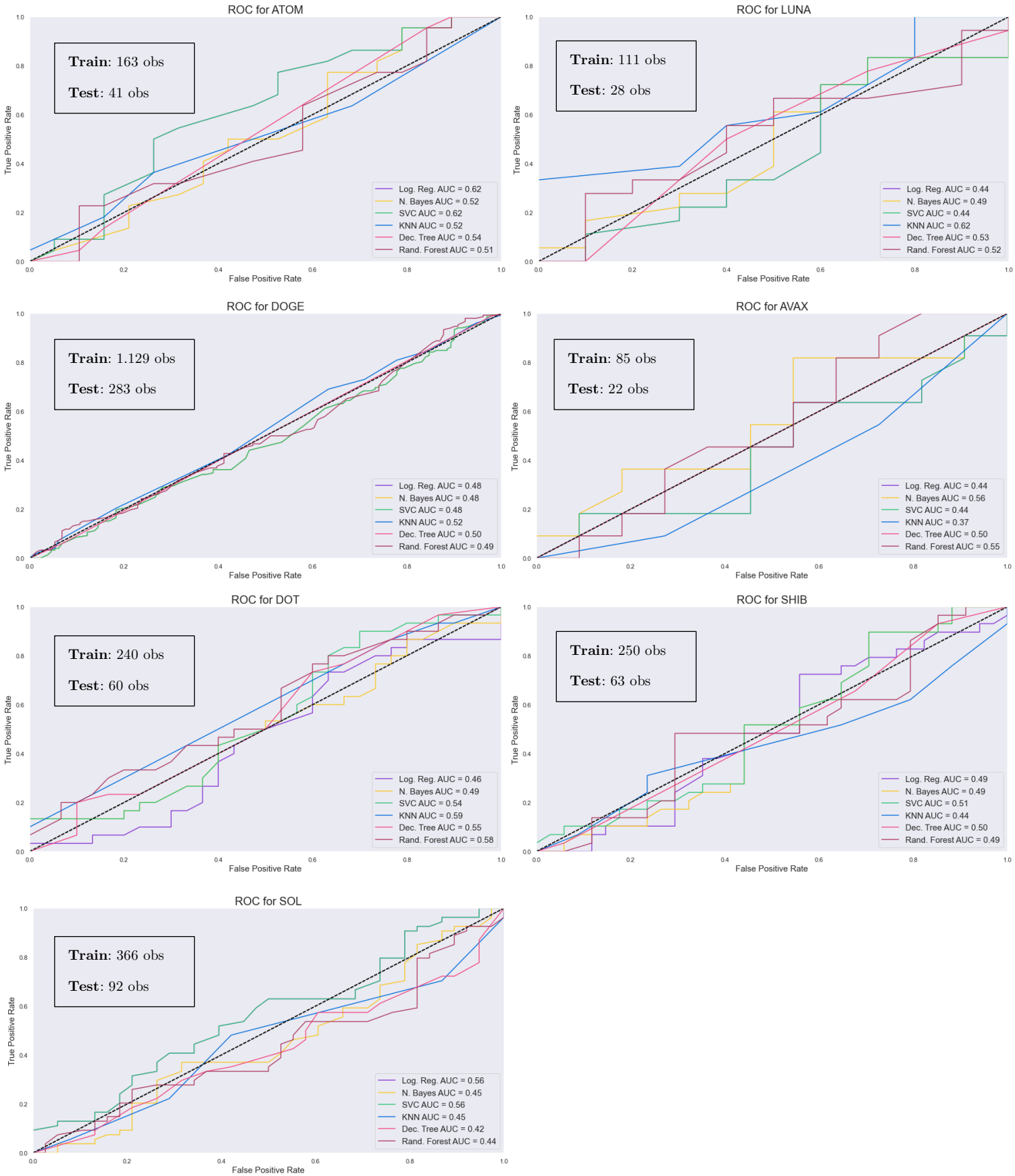


Figure 8.14 bis, *ROC curves for the 15 cryptocurrencies*

(Feature: sentiment score)

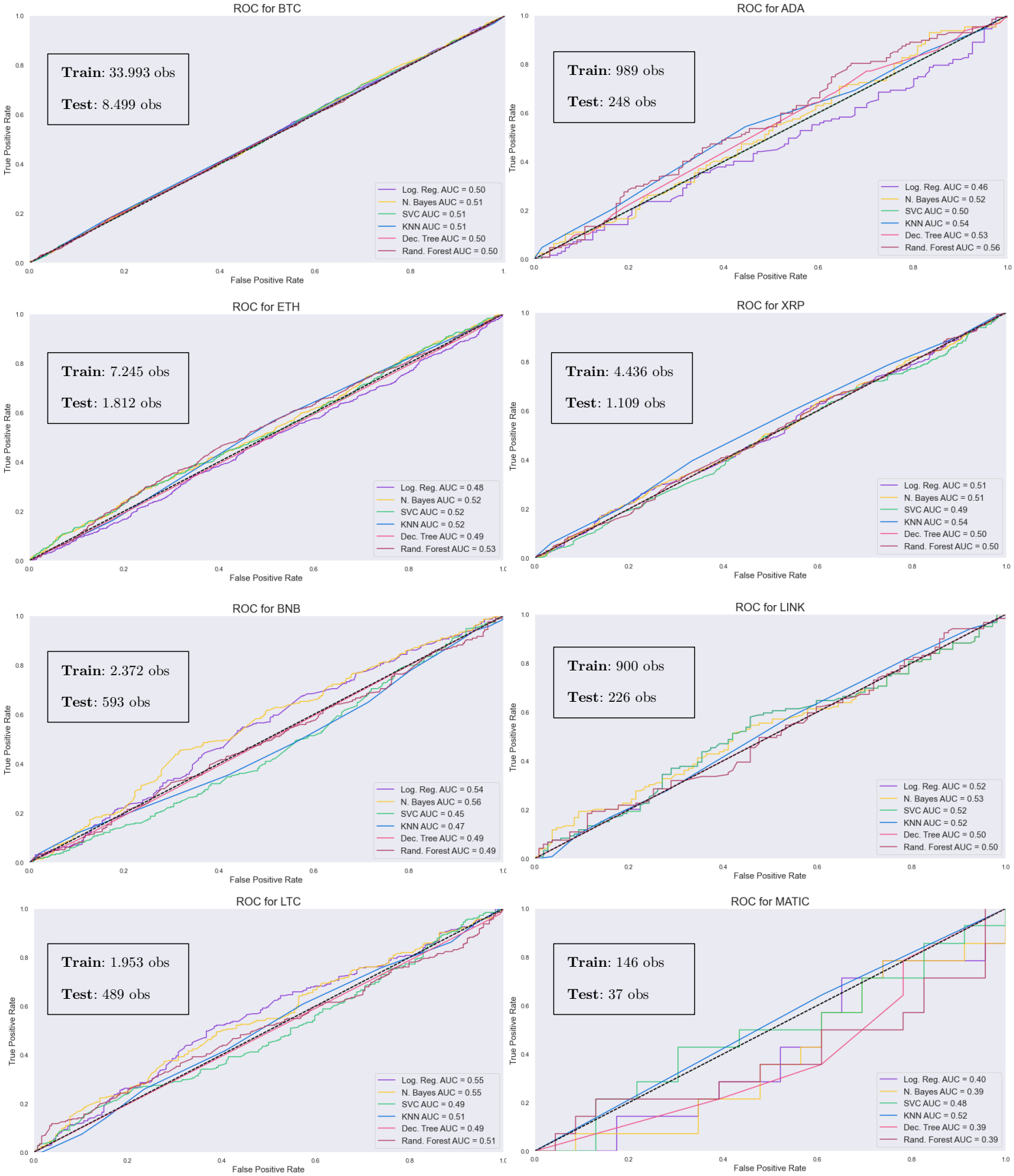
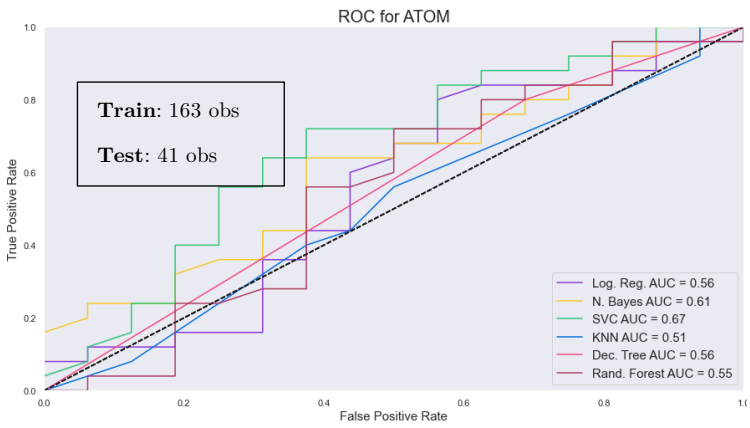
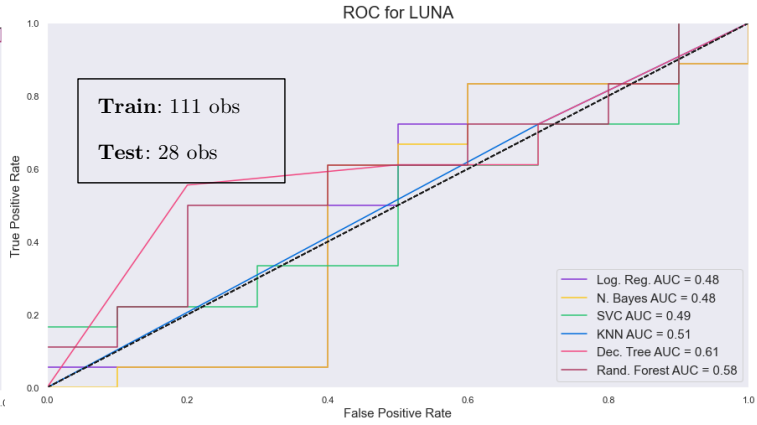
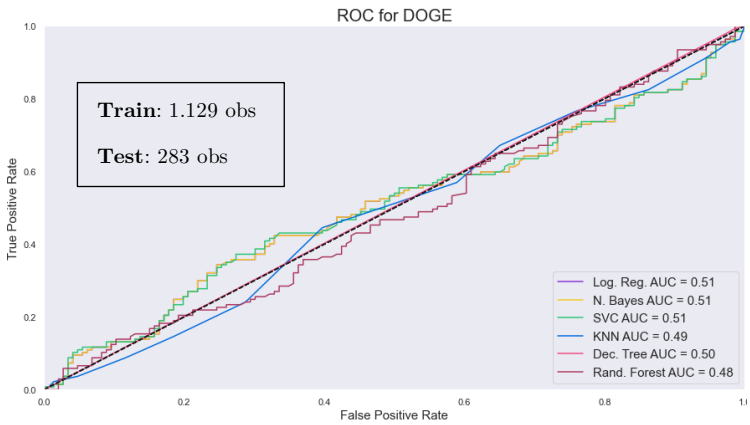


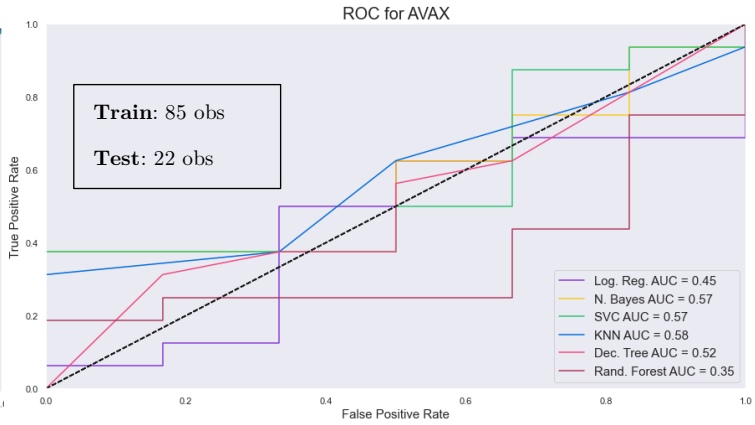
***Figure 8.15**, ROC curves for the 15 cryptocurrencies*

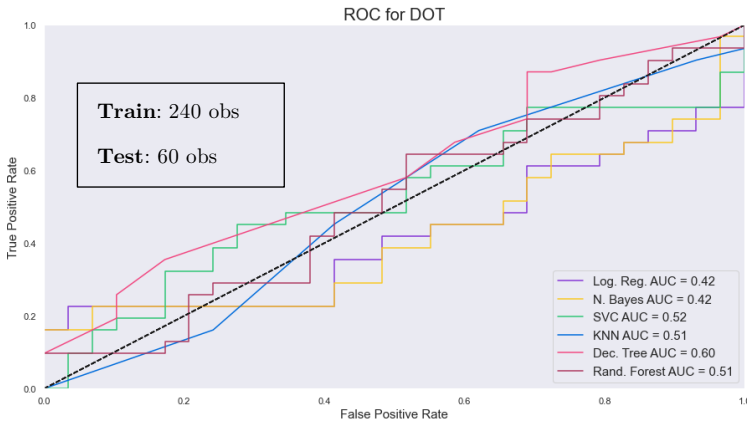
(Features: retweets and sentiment score)

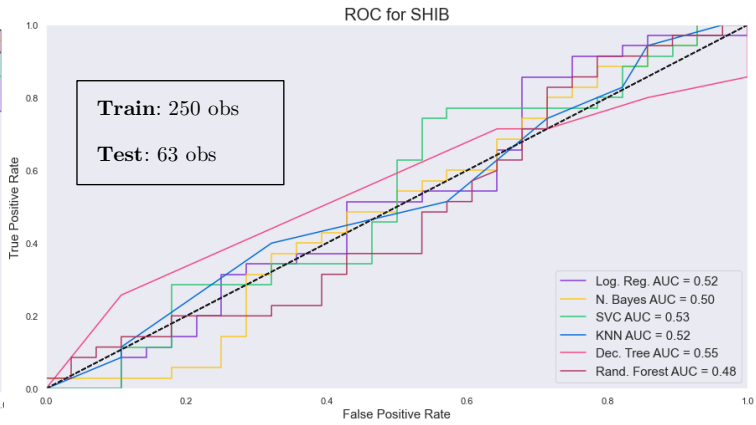












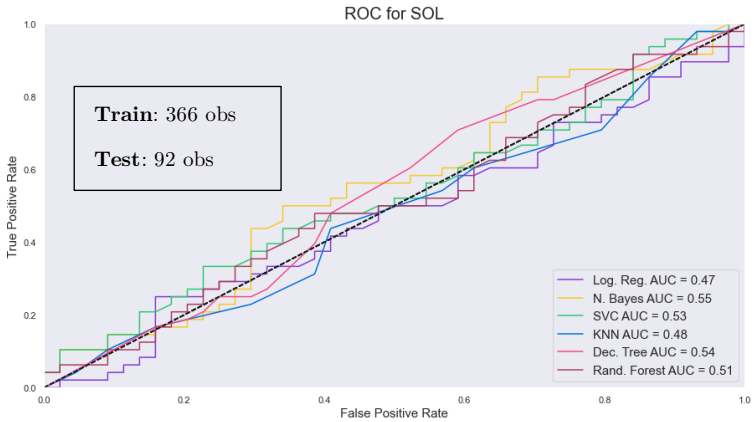


Figure 8.15 bis, *ROC curves for the 15 cryptocurrencies*

(Features: retweets and sentiment score)

Delay: 1 minute						
Crypto	Logistic regression	Naive Bayes	SVC	KNN	Decision tree	Random Forest
BTC	49,8%	49,8%	50,2%	50,0%	50,4%	50,4%
ETH	46,6%	53,4%	46,6%	50,1%	51,3%	51,1%
BNB	52,4%	52,4%	52,4%	52,8%	52,4%	51,5%
LTC	51,4%	50,7%	51,4%	55,9%	50,1%	53,0%
ADA	52,5%	50,8%	50,0%	53,2%	49,3%	51,4%
XRP	51,2%	48,4%	50,0%	50,2%	49,1%	49,9%
LINK	48,8%	48,5%	50,5%	47,7%	49,5%	52,7%
MATIC	46,8%	46,8%	46,8%	41,5%	45,8%	37,6%
ATOM	56,5%	56,5%	42,0%	42,9%	56,2%	48,3%
DOGE	54,7%	54,7%	50,0%	51,4%	53,2%	50,9%
DOT	55,5%	52,6%	55,5%	54,2%	47,4%	37,6%
SOL	56,5%	50,0%	41,6%	50,6%	48,3%	50,3%
LUNA	57,1%	57,1%	42,9%	59,4%	53,6%	54,8%
AVAX	50,0%	35,8%	41,7%	39,6%	31,7%	38,3%
SHIB	56,5%	56,5%	43,9%	46,6%	42,3%	48,1%

Figure 8.16

Delay: 2 minutes						
Crypto	Logistic regression	Naive Bayes	SVC	KNN	Decision tree	Random Forest
BTC	49,7%	49,7%	50,0%	51,0%	49,3%	49,7%
ETH	49,4%	49,4%	50,6%	51,5%	49,9%	50,2%
BNB	49,8%	49,8%	50,2%	52,0%	50,3%	51,8%
LTC	52,9%	47,1%	52,9%	47,6%	50,1%	45,7%
ADA	46,6%	54,1%	53,4%	58,5%	50,3%	58,4%
XRP	51,0%	51,0%	47,2%	50,9%	51,2%	51,7%
LINK	52,5%	47,5%	47,5%	53,7%	50,6%	50,6%
MATIC	50,4%	49,6%	49,6%	42,5%	50,0%	54,0%
ATOM	60,8%	40,4%	29,0%	61,2%	50,0%	43,7%
DOGE	59,1%	59,1%	40,9%	46,1%	44,4%	44,2%
DOT	39,9%	55,3%	39,9%	45,2%	51,7%	59,3%
SOL	46,0%	54,0%	46,0%	50,7%	53,6%	54,5%
LUNA	41,9%	58,1%	41,9%	61,9%	61,7%	86,4%
AVAX	45,4%	54,6%	54,6%	76,2%	49,2%	48,8%
SHIB	32,4%	41,0%	67,6%	46,3%	56,9%	56,3%

Figure 8.17

Delay: 5 minutes						
Crypto	**Logistic regression**	**Naive Bayes**	**SVC**	**KNN**	**Decision tree**	**Random Forest**
BTC	49,3%	49,5%	49,3%	49,9%	50,0%	49,6%
ETH	49,7%	49,7%	49,7%	47,0%	50,3%	49,7%
BNB	53,5%	46,5%	46,5%	50,9%	50,4%	50,2%
LTC	48,6%	50,7%	51,4%	49,2%	50,5%	49,0%
ADA	46,9%	50,0%	46,9%	50,7%	50,1%	48,7%
XRP	48,9%	49,2%	48,9%	50,6%	50,9%	50,4%
LINK	49,5%	49,5%	50,7%	52,6%	51,8%	52,5%
MATIC	52,2%	46,3%	47,8%	55,0%	50,3%	47,4%
ATOM	49,3%	56,5%	49,3%	55,7%	50,0%	53,1%
DOGE	51,0%	51,0%	51,0%	53,2%	52,7%	54,9%
DOT	46,6%	46,6%	46,6%	44,8%	49,4%	47,6%
SOL	50,0%	50,0%	45,0%	44,8%	48,1%	45,4%
LUNA	28,3%	28,3%	71,7%	30,0%	41,4%	35,8%
AVAX	47,4%	52,6%	52,6%	40,6%	47,0%	35,5%
SHIB	53,8%	46,2%	46,0%	47,1%	53,4%	49,6%

Figure 8.18

11. Bibliography

Davies, G. (2002). History of Money. *University of Wales Press.*

di Angelo, M., & Salzer, G. (2020). *Tokens, Types, and Standards: Identification and Utilization in Ethereum.* https://doi.org/10.1109/DAPPS49028.2020.00-11

Dibakar Raj, P., Prasanga, N., Anuj, P., Anup Kumar, P., & Bishnu Kumar, L. (2018). *Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis.*

FCA. (2019). *CP19/3: Guidance on Cryptoassets.* www.fca.org.uk/cp19-03-response-form

FINMA. (2018). *FINMA Annual Report 2018.*

Foley, S., Karlsen, J. R., & Putnins, T. J. (2019). Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies? In *Review of Financial Studies* (Vol. 32, Issue 5, pp. 1798–1853). Oxford University Press.

Hayes, A. (2014). *What Factors Give Cryptocurrencies Their Value: An Empirical Analysis.* https://willyreport.wordpress.com/

Joseph A. Ritter. (1995). The Transition from Barter to Fiat Money. *The American Economic Review*, *85*(1), 134–149.

Klein, T. ;, Thu, H., Pham, ;, & Walther, T. (2018). *Bitcoin is not the New Gold A Comparison of Volatility, Correlation, and Portfolio Performance.* www.econstor.eu

Kristoufek, L. (2015). What are the main drivers of the bitcoin price? Evidence from wavelet coherence analysis. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0123923

Le, Y. (2017). *The State of the Art in Cryptocurrencies.*

Moreno-Sanchez, P., Zafar, M. B., & Kate, A. (2016). Listening to Whispers of Ripple: Linking Wallets and Deanonymizing Transactions in the Ripple

Network. *Proceedings on Privacy Enhancing Technologies*, *2016*(4), 436–453. https://doi.org/10.1515/popets-2016-0049

Nakamoto, S. (n.d.). *Bitcoin: A Peer-to-Peer Electronic Cash System*. www.bitcoin.org

Rella, L. (2020). Steps towards an ecology of money infrastructures: materiality and cultures of Ripple. *Journal of Cultural Economy*, *13*(2), 236–249. https://doi.org/10.1080/17530350.2020.1711532

Scott, I. J., Neto, M., & Pinheiro, F. L. (2021). *Bringing trust and transparency to the opaque world of waste management with blockchain: a Polkadot parathread application*. https://ssrn.com/abstract=3825072

Seele, P. (2018). Let Us Not Forget: Crypto Means Secret. Cryptocurrencies as Enabler of Unethical and Illegal Business and the Question of Regulation. *Humanistic Management Journal*, *3*(1), 133–139.

Seghezza, E., & Battista Pittaluga, G. (2021). *Building Trust in the International Monetary System*.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis. In *DEGREE PROJECT TECHNOLOGY*.

The Tokenizer. (2021). *The Security Token RegRadar Report-a comparative regulatory analysis of nine countries*.

Wood, G. (2014). *Ethereum: A Secure Decentralised Generalised Transaction Ledger*.