The relationship between online factors and prices of cryptocurrencies.

January 2023

Msc. in Business Administration and Data Science

Master's Thesis

Author: **Krzysztof Koszewski**

Student number: **S120727**

Supervisor: **Somnath Mazumdar**

Number of characters: 179,056

Number of Pages: 78

Table of contents

Abstract

Cryptocurrencies are volatile assets that receive much attention in the media and academic literature. This paper focuses on the relationship between online factors from Twitter, Reddit, and Wikipedia and the five biggest cryptocurrencies by market capitalization: Bitcoin, Ether, Binance, Tether, and USD Coin. The analysis has three significant steps.

The first step consists of performing a 1,2- and 3-day price prediction analysis, with the models including only historical prices. The following six models were chosen: Time Series models: ARIMA and SARIMA; and Machine Learning models, RNN, LSTM, Bi-LSTM, and GRU. The results showed that the machine learning models had better accuracy than the time series models.

The second step encompasses the extraction of online data from Twitter, Reddit, and Wikipedia and the following analysis: performing the sentiment analysis with VADER, choosing the most important online variables through correlation analysis, and the Random Forest feature extraction. The results showed that Twitter variables were more correlated, and the Random Forest algorithm gave them more importance.

The third step of the analysis consists of extending the Machine Learning models from the first step by adding the online variables. The results showed that despite being less correlated and given less importance by the Random Forest feature extraction, Reddit variables had the best price prediction results. The positive, negative and neutral sentiment variables were equally successful at producing great predictions for all five cryptocurrencies. However, the non-sentiment variables from Twitter and Reddit, referred to as *engagement metrics,* delivered equal or better predictions than the VADER sentiments.

In addition, the Machine Learning models were run exclusively for the Covid-19 period. The results had worse accuracy for the models, which included only historical prices. However, the online variables offered more significant improvements; the results were also more consistent within each machine learning algorithm.

1. Introduction

The first cryptocurrency, Bitcoin, was introduced in 2008 in the wake of the financial crisis by an anonymous creator called 'Satoshi Nakamoto'. Bitcoin began trading at $0.0008 (investingnews.com); 14 years later, it trades at nearly $21,000 (finance.yahoo.com). There are thousands of cryptocurrencies, with the entire market reaching a staggering $1.025 trillion in July 2022 (investopedia.com). While Bitcoin is still the most popular cryptocurrency with the highest value, many others have been created, such as Ether, which offers users the ability to create Smart Contracts and Decentralized Apps (investopedia.com), or Tether, whose value is pegged to the US dollar. Cryptocurrencies work on blockchain technology, which ensures high transaction security and transparency under pseudo-anonymity (euromoney.com).

The supply of a cryptocurrency is determined by the amount of already existing tokens and the ones created through the validation of transactions, a process called 'mining' (simplilearn.com). Cryptocurrencies are highly volatile; even Bitcoin, an already-established cryptocurrency, is about five times more volatile than other currencies or gold (buybitcoinworldwide.com). The high volatility of an asset means that an investment is at elevated risk but could also be a high reward. For example, if one invested in Bitcoin in November 2020 and sold it a year later, they would effectively quadruple their money. However, since the peak of November 2021, Bitcoin has lost 70% of its value (finance.yahoo.com).

The high volatility of cryptocurrencies and a growing interest in them have led to many studies which attempt to predict the price of cryptocurrencies and establish which factors play important roles in their price fluctuations. Social media activity related to cryptocurrencies is a widely studied area. The content related to cryptocurrencies is increasing in popularity on social media, with not only individuals but financial institutions and CEOs of major companies discussing the topic of cryptocurrencies online. Social media data can be accessed through numerous channels, and having access to people's opinions on cryptocurrencies makes it possible to see how it affects the prices of this highly speculative asset.

1.1. Motivation

The motivation for this study is grounded in multiple sources. Firstly, the author's electives about blockchain technology and the growing popularity of cryptocurrencies in newspapers and on social media sparked the author's curiosity about cryptocurrencies.

Moreover, the events from the story of GameStop stock price increased by almost 8,000% through a collaborative effort of individual buyers who organized themselves through Reddit in January 2021, which resulted in significant losses for a financial institution that shorted GameStop stock (theprint.in). The impact of that collective action originating from social media made the author even keener to investigate social media's influence over the price of assets, especially digital assets - cryptocurrencies.

A thorough literature review revealed that the most relevant factors to consider are: Twitter's and Reddit's posts sentiments, the number of posts on both platforms, and the Wikipedia trend. Furthermore, cryptocurrencies are currently going through a prolonged and volatile price decline, a so-called 'bear market.' The bear market of cryptocurrencies could be the result of cryptocurrencies encountering their first major global crisis since their inception. Hence, in light of the recent Covid-19 pandemic, it is even more interesting to investigate how social media data can help investors navigate these challenging times.

1.2. Research question

Therefore, the author decided to investigate the influence of social media activity on the price of the five biggest cryptocurrencies by market capitalization: Bitcoin, Ether, Binance, Tether and USD Coin (coinmarketcap.com). This has led the author to ask the following research question:

"How do Twitter, Reddit, and Wikipedia trends influence prices of Bitcoin, Ether, Binance, Tether, and USD Coin?"

1.2.1. Sub-Research questions

To facilitate answering the research question, the following sub-research questions were formulated:

1) *"How do machine learning models for price prediction compare to the time series models?"*

2) *"How important is each chosen social media and its sentiment in each cryptocurrency?"*

3) *"How effective are online factors in cryptocurrency price prediction?"*

4) *"How do current results compare to the Covid-19 period's results?"*

1.3. Delimitations

The aim of this study is to examine the relationship between the online variables relating to only the five most valuable cryptocurrencies. Therefore, this study will focus exclusively on the five biggest cryptocurrencies by market capitalization; this is a limitation as currently, over 20,000 cryptocurrencies exist (explodingtopics.com).

The second limitation is the Reddit data, which was only acquired for Bitcoin, Ether, and Binance, as the Tether and USD Coin subreddits had few members and posts. Moreover, only the most popular subreddit per cryptocurrency was included, while many subreddits, for instance, Bitcoin, have multiple subreddits with more than 1 million members.

Another limitation is the choice of social media platforms; while Twitter and Reddit are among the most popular social media forums, there exist other means through which people communicate their thoughts about cryptocurrencies, for instance, Facebook groups and Telegram channels. The reason to include only Twitter and Reddit data is their accessibility, as they can be acquired through a simple Python script, and most of the data is textual.

The usage of purely textual data is a limitation since much of the sentiment about cryptocurrencies is expressed in graphic screenshots of one's portfolio, memes, or videos on TikTok or YouTube. Furthermore, this paper only analyses the textual data in English, excluding the posts made in other languages, which poses a linguistic limitation. Although English is one of the most popular languages online, it is far from being the only one.

2. Background

This section will explain the concepts, analytical tools, and data used in the analysis; the subsections are presented graphically in Figure 1. The section consists of two major topics; models and data. First, the model topic will be discussed, including the time series and machine learning models, and sentiment analysis tools used for the data analysis process. Secondly, the data used in the study is presented, which includes the chosen five cryptocurrencies; Bitcoin, Ether, Binance, and the two stablecoins; Tether and USD Coin, as well as the websites and social media used; Twitter, Reddit, and Wikipedia.

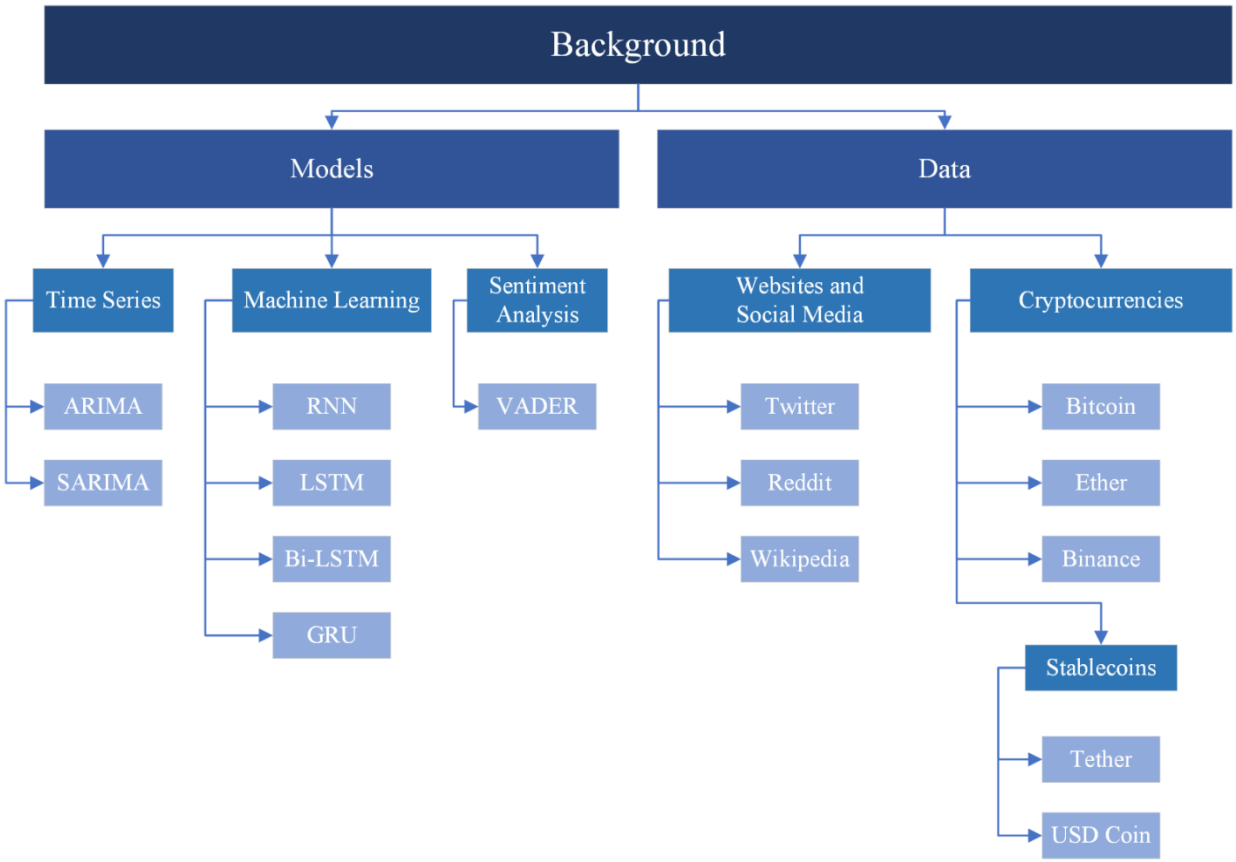


Figure 1 (The structure of the Background section)

2.1. Times series models

Auto-Regressive Integrated Moving Average (ARIMA) is a stochastic time series model that can be used to train and forecast future values based on historical values of a given time series (Medium.com).

An ARIMA model, in accordance with its name, has three components: autoregressive, differencing and moving average parts. The order of the autoregressive part, denoted by p, forecasts the target variable using a linear combination of the target variable's previous values. In a regression-like model, the moving average part uses historical forecast errors, and its order is denoted by the letter q. Moreover, the autoregressive and moving average components of ARIMA are invertible. Differencing is the process of calculating the differences between a series of consecutive observations. The main goal of this task is to make the data stationary, which means the time series properties are independent of the observation time. Consecutively, data that is trending or seasonal is not stationary. Time series' stationarity is particularly important as an

AIMA model will not accept data that is not stationary (Hyndman, R. J., & Athanasopoulos, G., 2018, chapter 8.).

Subsequently, in an ARIMA (p,d,q) model, the first component, p, represents the autoregressive part, the second component, d, represents the degree of first differencing, and lastly, the third component, q, represents the order of the moving average part (Hyndman, R. J., & Athanasopoulos, G., 2018, chapter 8).

An ARIMA model that includes additional seasonal terms is called SARIMA (Seasonal Autoregressive Integrated Moving Average). By including the seasonal component, the order changes to SARIMA (p,d,q)(P,D,Q)[m], where m represents the number of observations per year, and the P, D, and Q are analogous to their non-seasonal counterparts but include backshifts to the seasonal period (Hyndman, R. J., & Athanasopoulos, G., 2018, chapter 8)

Hyndman, R. J., & Athanasopoulos, G. (2018) give a detailed guide over how to apply the ARIMA models in R. However, as all other components of this paper are made with Python, the author also decided to use Python for the ARIMA models. The order of an ARIMA or SARIMA model can be determined by looking at the time series decomposition, the ACF, and PACF. However, Hyndman, R. J., & Athanasopoulos, G. (2018) mention that R has a built-in auto.arima() function, which determines the order automatically by employing the maximum likelihood estimation, is used to determine the best model. The auto.arima() function has its Python counterpart in the Python package pmdarima in the model_selection function (alkaline-ml.com/pmdarima/), which was used in this project.

2.2. Neural Network models

The RNN, LSTM, Bi-LSTM, and GRU models, used in this paper are all Artificial Neural Networks. Their underlying architecture is similar and will be discussed in this section. Each of the four neural networks will have its short subsection, providing an in-depth explanation of the algorithm.

Artificial neural networks are a class of nonlinear statistical models which resemble biological neural networks to replicate their learning ability. The fundamental element of a neural network is a perceptron, which is the sum dot products of two equal length vectors; the input vector, which comes from the transformed input data and the weights vector to the result of that operation bias is added, the results is afterward treated as an input for the activation function. The second essential element of an artificial neural network is a layer consisting of many perceptrons, whose sum is stored in the outputs. Consequently, all densely connected layers are called dense layers (Hastie et al., 2009).

An Artificial Neural Network can consist of at least three layers; the input, hidden layer, and output layer; it is important to note that there may be more than one hidden layer. The input layer consists of the input data, whereas the hidden layer consists of the bias added to the sum of the dot produced between input data and weights. Lastly, the output layer consists of the output of the activation function, with the input being the output of the last hidden layer (Hastie et al., 2009).

Artificial Neural Networks use the training set to calculate how accurate its predictions are; to determine the performance of these training set predictions, the loss is calculated. Subsequently, a function based on the variance in losses named the loss function can display the performance of the training set prediction.

The weights vector is adjusted, as it minimizes the loss function and consequently improves the model's performance. To minimize the loss function, the gradient of the loss function must be calculated as it enables the model to find the loss function's minimum. This process is performed recursively, and weights get adjusted in each following layer. However, the performance does not necessarily improve with each added layer; the issue is with the vanishing gradient. The vanishing gradient is brought on by backpropagation-

induced diminishing gradient values; the smaller the gradient, the slower the learning rate, and eventually, the network stops learning (Hastie et al., 2009).

2.2.1. RNN

A regular feedforward neural network has no memory of past inputs and only considers the current input to make a prediction. A Recurrent Neural Network (RNN) differs from a regular Artificial Neural Network by running the information cycles through a loop. Hence, the RNN prediction takes as input not only the current input but also past inputs. Consequently, an RNN needs to apply weights not only to current inputs but also to past inputs. The main issue of an RNN is the vanishing gradient, which occurs when the gradient's values are too small, and the model stops learning (builtin.com).

2.2.2. LSTM

A Long Short-Term Memory neural network (LSTM) is considered an extension of an RNN. An LSTM can read, write, and delete information from its memory enabling the network to remember inputs longer. LSTM's memory is a gated cell, which assigns weights to information based on its importance and decides to store or delete that information based on its importance. The model also learns and updates the weights through time (builtin.com).

A long short-memory cell has three gates: the input gate, responsible for letting the new input in; forget gate, responsible for the deletion of the information; and the output gate, which lets the information impact the output of the current timestep (builtin.com).

Lastly, LSTM solves the problem of vanishing gradient by keeping the gradient steep enough, which keeps the training relatively short and the accuracy high (builtin.com).

2.2.3. Bi-LSTM

Within a Bi-directional LSTM (Bi-LSTM), the information flows in both directions, considering both past and future values, which is especially helpful when dealing with an NLP problem (baeldung.com), however, times series predictions can also greatly benefit from it. A Bi-LSTM model can be described as two LSTM models with the information flow from left-to-right and right-to-left; this improves the long-term dependencies and consequently should improve the time series' prediction accuracy (Siami-Namini et al., 2019).

2.2.4. GRU

A Gated Recurrent Unit (GRU) Neural Network is another extension of RNN, newer than LSTM. The key difference between GRU and LSTM is that GRU does not use the cell state but instead the hidden state to transfer information. The hidden state has two gates; a reset gate, which decides what information should be kept and what information should be added, and an update gate, which decides how much of the past inputs to forget (towardsdatascience.com). The reset gate first activates; it stores pertinent data from the previous time step in new memory content. The input vector and hidden state are then multiplied by their respective weights. After that, it multiplies the multiple of the previously hidden state and the reset gate element by element. The following sequence is generated by applying the non-linear activation function after adding up the abovementioned steps (analyticsindiamag.com).

2.3. VADER sentiment analysis

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based sentiment analysis tool made especially for the sentiment analysis of social media (GitHub.com). Each word has its sentiment score, which ranges from –4 (most negative) to +4 (most positive), and 0 represents the neutral

sentiment (medium.com). A text analyzed with VADER yields a vector of positive, neutral, negative, and compound sentiments. The positive, neutral, and negative sentiments add up to 1, while the compound sentiment ranges from –1 (the most negative) to +1 (the most positive).

In order to categorize the tweets based on sentiments, the compound sentiment will be used as it is made based on the normalization of the sum of dictionary scores of each VADER-dictionary-listed word in the sentence (medium.com). This paper will adopt the same classification as Pano, T., & Kashef, R. (2020) so that a compound score of above 0.05 will be classified as positive, below –0.05 as negative, and the rest as neutral. The acquired sentiments will be later used in the results

2.4. Cryptocurrencies

Cryptocurrencies are digital assets intended to be used as forms of exchange, mimicking traditional currencies. They rely on blockchain technology, a distributed ledger system that records all transactions across the network; each instance of the cryptocurrency has a unique identifier with all its transactions, since its creation, stored on the ledger, ensuring the transparency and security of the transactions (Tredinnick, 2019). The transactions are accepted in blocks, which are then added to the chain of transactions, hence the name of the technology. Security is ensured by adding a complicated one-way algorithmic process, a cryptographic hash function, which needs to be solved to accept the block of transactions. The hash function can only be solved by trying different numbers in the equation. After the hash function is solved, the network participants 'vote' on accepting the block of transactions; if the transactions are non-fraudulent and the block is accepted, the network participant that solved the hash function gets awarded a specified, newly created amount of cryptocurrency. This process, called 'mining,' is the only way new cryptocurrency instances can be created (Tredinnick, 2019).

The number of transactions per block, the amount of cryptocurrency offered as rewards, and the specific cryptographic functions vary between different cryptocurrencies. Security and social acceptance constitute a cryptocurrency's value (Van Alstyne, 2014). Hence it is crucial that the system works well and detects fraudulent behavior to instigate trust.

According to Forbes, the five most valuable cryptocurrencies by market capitalization are Bitcoin ($323.1 billion), Ether ($148.0 billion), Tether ($66.2 billion), USD Coin ($44.5 billion), and Binance ($39.8 billion) (forbes.com).

2.4.1. Bitcoin

Bitcoin was created in the wake of the 2008 financial crisis; it followed the ideas of the whitepaper of an unknown creator under the pseudonym 'Satoshi Nakamoto'. Bitcoin initially became popular on illegal websites like "Silk Road," which contributed to the media and the general public's initial bad perception of it. Despite that and the fact that it had been illegal in many countries, Bitcoin entered mainstream media again in 2017 when its value skyrocketed to $20,000. This event made investors and the public reevaluate Bitcoin as not only an illegal payment system on the dark web but also as a real investment opportunity with significant potential. The evolution of Bitcoin's price over time can be seen in Figure 2.

Bitcoin was the first cryptocurrency; by convention, there is a distinction between Bitcoin and all the other cryptocurrencies, with the latter being named altcoins. Bitcoin has no tangible form and is fully decentralized, which makes it independent from all countries, companies, and financial institutions. It can be used as a medium of exchange; however, due to its significant valuation, it is often divided into smaller parts, as the smallest part of one Bitcoin is one 'Satoshi', which constitutes one-millionth of a Bitcoin (Berentsen & Schar, 2018).

Bitcoin's underlying technology is Blockchain; it uses a Proof-of-Work (PoW) consensus mechanism and SHA-265 hashing algorithm to ensure the safety of transactions (Narayanan et al., 2016). The main idea of the PoW mechanism is to achieve consensus in a decentralized manner, which prevents malicious actors from overtaking the network. The network participants that try to solve the hash function need to provide evidence that they have used considerate computational power while solving the function (Narayanan et al., 2016). Bitcoin requires each block to be generated every ten minutes; as the computing power increases, so does the hash function's difficulty, so that each block is generated once every ten minutes. The network participant who solves the hash function receives a reward. However, the reward per block is halved every 210,000 blocks, which is about every four years and is currently at 6.25 BTC (Narayanan et al., 2016). Thus, the PoW consensus mechanism requires increasingly more computing power to solve increasingly complex hash functions, which led to heavy criticism of the PoW consensus mechanism, urging the cryptocurrency community to look for alternative consensus mechanisms (Schinckus, C., 2021). The argument about Bitcoin mining resulting in a large consumption of fossil fuels was the reason on May 13th, 2021, Elon Musk tweeted about Tesla no longer looking to accept payments in Bitcoin, which, along with many other tweets about Bitcoin, has had considerable consequences for the price of Bitcoin throughout the year (vox.com).

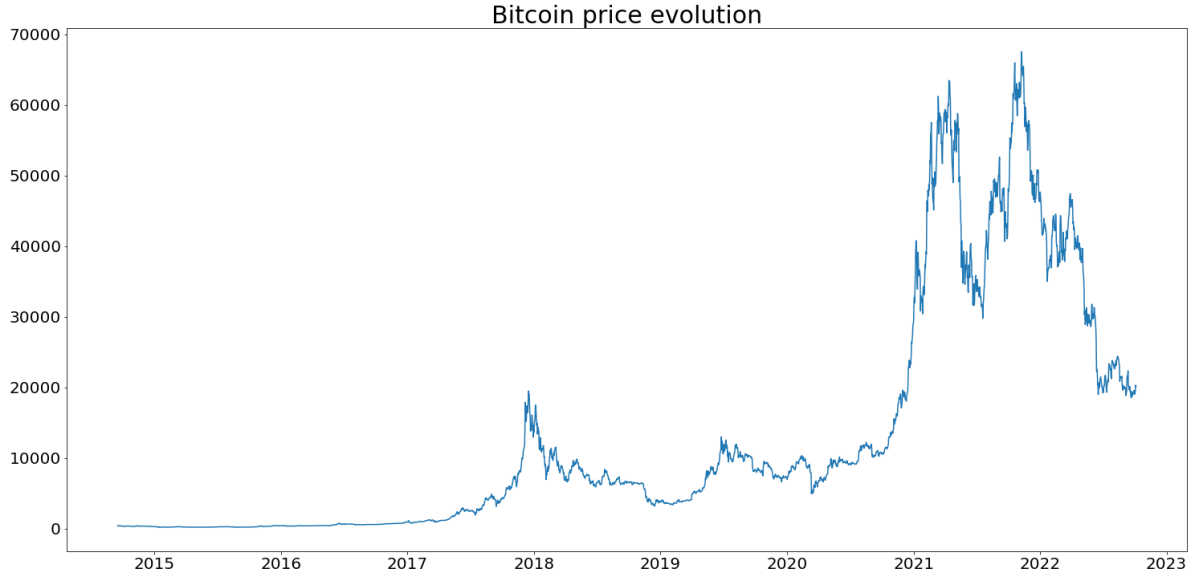


Figure 2 (The evolution of Bitcoin prices)

2.4.2. Ether

Ether is a cryptocurrency native to the Ethereum blockchain, launched in 2015 by Vitalik Buterin. One of the main motivations for creating Ethereum was the constraints of Bitcoin's blockchain as being intended to be only a peer-to-peer payment network. Ethereum is programmable, which means it can be used to build decentralized applications and code smart contracts (ethereum.org).

Ether and Ethereum are often used interchangeably, but for this paper's purpose, a distinction must be made; Ethereum is an open-source, decentralized blockchain network that is Turing-complete and has a built-in programming language: Solidity (Dannen C., 2017). The Ethereum network, like Bitcoin's, is powered by the network participants, who need monetary incentives to stay on the network and lend their computing power. In order to do that, the Ethereum network has its cryptocurrency, Ether (ETH). Ether is used to reward the miners and make transactions across the Ethereum network. Consequently, Ether's price is highly dependent on the Ethereum network; hence a few key parts of the network need to be explained: smart contracts, NFTs and DApps.

Smart Contracts are executable codes written with business logic, in the case of Ethereum, written in the Solidity programming language, which transfers value from one account (or contract) to another (Dannen C., 2017). Smart Contracts are immutable by design; hence, once deployed, they cannot be altered, and no change can be made to the business logic embedded in them. However, Ethereum does offer certain ways in which the executable code can be altered (ethereum.org). Smart contracts are stored on the Ethereum blockchain and are, therefore, public. Since a smart contract needs to be executed by all Ethereum nodes, it causes an increase in computational power required to run it; hence, an additional fee is required to run it, called 'gas'. Developers of a smart contract can set a 'gas limit', which is the maximum amount one is willing to pay to execute the smart contract (Dannen C., 2017).

Smart Contracts are also the foundations of Decentralized Applications (DApps), which are publicly available services running on the Ethereum network, made accessible through front-end programming languages by web browsers or mobile applications (Dannen C., 2017). DApps resemble regular applications but, as they are decentralized, have no ownership or authority over them, as they are executed and maintained by the Ethereum network rather than, as it would be the case for regular applications by the companies that own them and their centralized servers. DApps execute transactions through Smart Contracts (Dannen C., 2017), which offers the following advantages to their users: more privacy, as there is no longer a need to provide personal information to a central authority, as well as their developers, as Ethereum offers them a flexible platform for DApp creation. However, due to the code's immutability, DApps are challenging to scale.

Non-Fungible Tokens (NFTs) are unique tokens that cannot be copied or substituted and can be stored on the Ethereum network. Their properties make them undisputed proof of ownership of an online good (ethereum.org). NFTs have gained much media attention lately as they have been used to prove ownership of online art, which caused controversy. Since the NFTs are stored on the blockchain, they also constitute a public record of ownership, which means that, for instance, creators could request royalties for their work through smart contracts.

The Ethereum network's functionalities influence Ether's price; for instance, larger gas prices or the skyrocketing popularity of NFTs can significantly drive up the price of the cryptocurrency. However, they can also have a negative effect, such as the DAO fork incident, which resulted in Ether splitting into Ether and Ethereum classic (gemini.com). In September 2022, Ethereum switched its consensus mechanism from Proof-of-Work to Proof-Of-Stake (PoS). The PoS mechanism ensures high functionality of the blockchain due to validators staking their ETH tokens which act as collateral in case the validators do not fulfill their responsibilities or act maliciously (ethereum.org). The PoS mechanism is widely praised for requiring considerably less energy than PoW, thus being a much greener and more scalable solution for blockchain technology (time.com).

The evolution of Ether's price (Figure 3) resembles the evolution of Bitcoin, with the two following the same trends over time.

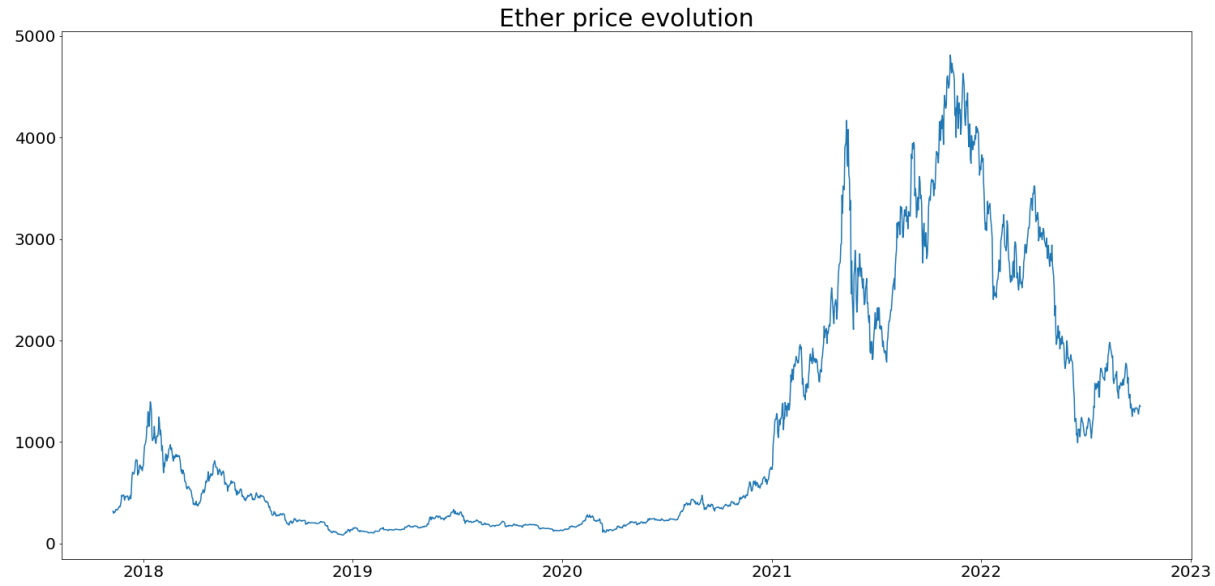


Figure 3 (The evolution of Ether prices)

2.4.3. Binance

Binance coin is the cryptocurrency issued by Binance cryptocurrency exchange (investopedia.com), the biggest cryptocurrency exchange in terms of trading volume of cryptocurrencies (coinmarketcap.com). The Binance exchange was funded in China in 2017 by Changpeng Zhao (history-computer.com) and is currently incorporated in the Cayman Islands (investopedia.com). The company quickly gained popularity; today, its U.S. branch alone is valued at 4.5 billion US Dollars (barrons.com). Binance exchange attracts customers mainly by offering low transaction fees. However, the platform also offers other services such as Binance Earn for earning interest on stablecoins, Binance Visa Card, which enables customers to pay with a cryptocurrency credit card by exchanging the cryptocurrencies for fiat currencies at the moment of purchase, Binance Smart Pool which helps miners optimize their efforts or Binance Labs and LaunchPad which focus on new, exciting cryptocurrency and blockchain-related projects (investopedia.com). Similarly to Ethereum, Binance also offers the possibility to create and trade NFTs (binance.com).

The popularity of the Binance exchange surely fostered the popularity of their native cryptocurrency: Binance Coin, especially as users receive discounts if they pay in the native tokens (investopedia.com). The Binance Coin (BNB) was initially based on the Ethereum blockchain but now has its own Binance chain (investopedia.com). Binance started with over 200 million tokens. However, to "maximize the BNB token's value and provide a sustainable and safe long-term growth plan for the BNB ecosystem" (cointelegraph.com), it undergoes quarterly burns; in fact, one-fifth of its profits are used for repurchasing and destroying the tokens (investopedia.com). Consequently, the Binance exchange buys back a certain number of Binance Coin tokens and destroys (burns) them. Thus, the total number of Binance Coin tokens available is 169,432,937, compared to the initial 200,000,000 (investopedia.com). The burn dates are announced on Twitter by the CEO of Binance (cointelegraph.com) and can influence the price of Binance Coin. It is important to note that Binance Coin also exists outside of the Binance exchange, as users can trade BNB on other platforms as well (investopedia.com).

Binance can offer this functionality as their BNB chain is a dual blockchain, consisting of BNB Beacon Chain and Binance Smart Chain (support.exodus.com). The Binance Smart Chain offers similar functionality to the Ethereum blockchain, like Smart Contracts and the ability to develop DApps; it works

on the Proof of Staked Authority (PoSA) consensus mechanism (academy.binance.com). PoSA consensus is more energy efficient than PoW used by Bitcoin's blockchain. Binance Coin is the native token of the BNB Beacon Chain, which focuses on fast transactions with small fees (academy.binance.com) and uses the PoW consensus mechanism (gizmodo.com.au). The main idea behind this solution is to ensure high speed, low transaction costs, and compatibility between the two chains and the Ethereum blockchain. However, the disadvantage is the centralization that comes with Binance being in control of the transactions and the criticism of the PoSA consensus mechanism favoring those with the most money to have control of the blockchain (ionos.com). Moreover, as the recent hacking attack (gizmodo.com.au) shows, using a dual blockchain may also cause security gaps that hackers could exploit.

The analysis refers to Binance Coin (BNB) and, for clarity reasons, from this point on, whenever the term 'Binance' is used, it relates to the cryptocurrency Binance Coin (BNB). The price evolution of Binance (Figure 4) shows that the price skyrocketed in early 2021 and has a similar pattern to Bitcoin and Ether.

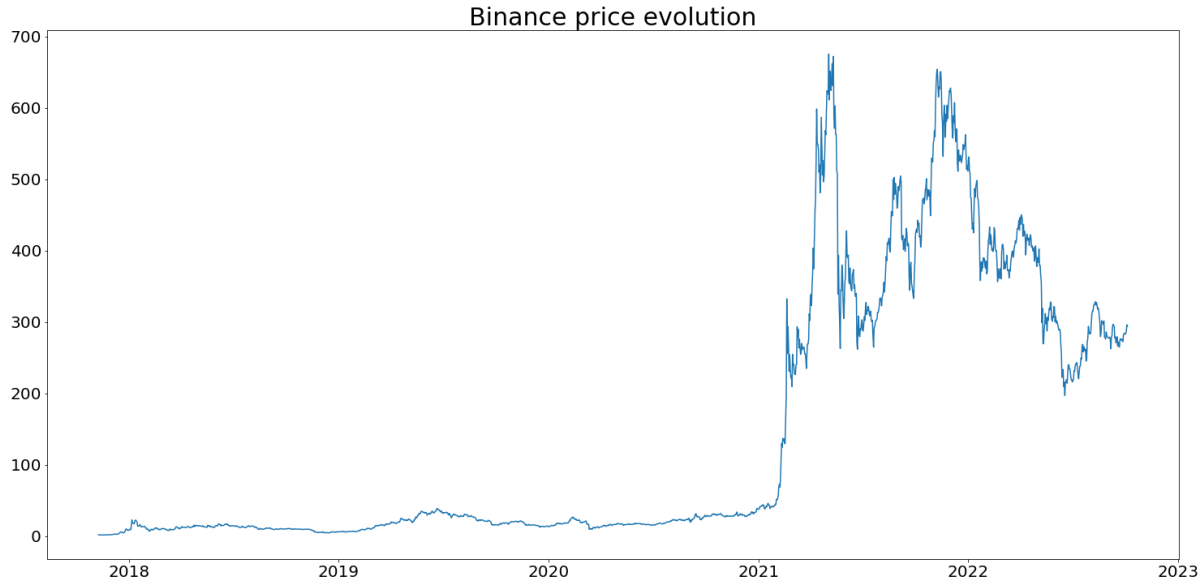


Figure 4 (The evolution of Binance prices)

2.4.4. Stablecoins

Stablecoins are a special type of altcoins; they address the high volatility, the main issue of cryptocurrencies as a payment method, by being pegged to a certain currency, commodity, or financial instruments (investopedia.com). By looking at the price evolution of Bitcoin, Ether, and Binance (Figures 2,3 and 4, respectively) it is easy to understand why transactions made in cryptocurrencies are risky; the most famous example of why this is an issue is Laszlo Haynecz, who paid 10,000 BTC for two pizzas in May 2010 (forbes.com), back then the transaction was worth about $41, while today it would be worth almost $21 million today (16th of January 2023). Stablecoins are meant to solve this problem by retaining a stable valuation.

There are three major mechanisms through which stablecoins can ensure their stability. Firstly, the stablecoins issuer can have a fiat currency reserve that acts as collateral, which helps ensure that the stablecoin maintains its peg. This solution helps close the gap between fiat currencies and cryptocurrencies, is scalable and easy to understand: However, it goes against the very premise of decentralization of cryptocurrencies, as the issuer is in control of operations and issuance of the coin, requires frequent and

independent audits, to ensure transparency, as well as acting as a trusted custodian to store the collateral fiat (medium.com).

Another way to ensure the stability of a stablecoin is to have another cryptocurrency as collateral. This solution is closer to the premise of Satoshi Nakamoto's white paper, as it offers more transparency and decentralization as all information is available on the blockchain. However, using another cryptocurrency as collateral transfers the issues of that cryptocurrency along with it, i.e., the high volatility of scalability issues (medium.com).

The third way to ensure a stablecoin's peg is by using an algorithm that regulates the supply of coins to maintain a given value. Thanks to this solution, collateral is not required, and transparency and independence of the stablecoin are ensured, as all information is on the blockchain, and the issuer cannot artificially control it. However, maintaining the peg through an algorithm is complicated to implement, usually using the premise of continuous future growth, and hence resembles a pyramid scheme, where the promise of future growth strengthens low prices. Moreover, problems arise during a bear market when the number of sell orders is too high for the system to absorb (medium.com).

Stablecoins were observed to be especially important during economic and geopolitical distress, especially during the Russian invasion of Ukraine (cryptonews.net), since people can exchange their fiat currency to a stablecoin in a fast and secure way, which in times of crisis, local banks cannot always ensure. Moreover, stablecoins are increasingly integrated into Web3 and traditional payment systems (cryptonews.net). The two stablecoins discussed in this paper are Tether (USDT) and USD Coin (USDC), as they are the biggest by market capitalization; both are discussed in the sections below.

2.4.5. Tether

Tether (USDT) was created in 2014 and began trading in 2015; first known as RealCoin but was later rebranded. Tether was one of the first stablecoins and is the biggest stablecoin by market capitalization; it is collateralized by large US Dollar reserves held in multiple banks (smartasset.com). Tether does not have its own blockchain infrastructure (coindesk.com); instead, the issuance of Tether tokens is viable on various blockchains such as Bitcoin and Ethereum, making it possible to users to transact Tether tokens through other blockchains (tether.to).

As explained in the previous section, having a fiat-collateralized stablecoin comes with certain disadvantages; since Tether does not have its own blockchain, the issuer, Tether Ltd., which has control over the issuance and destruction of USDT tokens to adjust to the supply and demand (coindesk.com). Furthermore, Tether comes under much scrutiny as it does not offer sufficient proof that the company does possess enough fiat collateral. Transparency is an issue for Tether as they only started publishing their assets after a court case that involved Bitfinex covering an $850 Million loss using Tether funds. Nevertheless, Tether Ltd. does not publish audits but quarterly attestations which are not verified by an independent body (coindesk.com). Tether is backed by US Dollar reserves and other assets such as corporate bonds, loans, or other investments, including cryptocurrencies.

Tether Ltd. is owned by iFinex, the parent company of the Bitfinex cryptocurrency exchange. In their study Griffin, J. M., & Shams, A. (2020) found that the 2017 Bitcoin boom was influenced by one entity that was using Tether and hypothesized that Tether might be an unbacked digital cryptocurrency inflating the prices. Tether Ltd. refuted the allegations of the study, saying that since the complete dataset had not been used, the study is in fact inconclusive (bitfinex.com). Despite its controversies and the lack of full transparency, Tether remains one of the most popular cryptocurrencies on the market. As shown in Figure 5. The price of Tether remains remarkably close to $1.

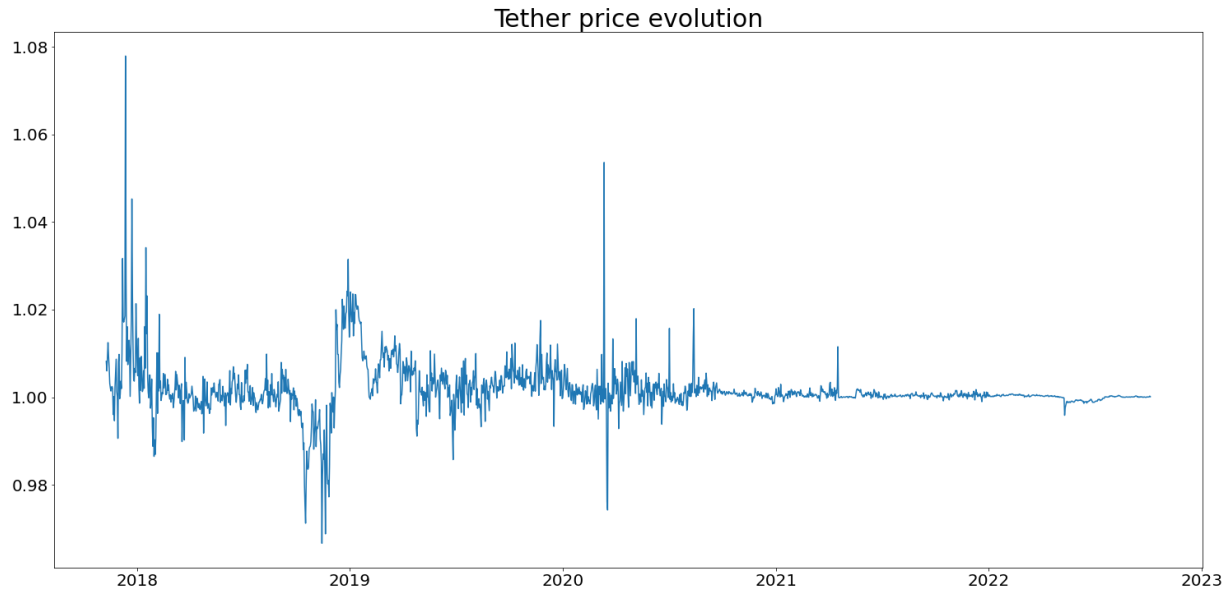


Figure 5 (The evolution of Tether prices)

2.4.6. USD Coin

USD Coin was launched in September 2018 by the Centre consortium, a collaboration between the Coinbase cryptocurrency exchange and Circle, a Money Transmitter backed by Goldman Sachs (cryptonews.com). Like Tether, USD Coin is a fiat-collateralized stablecoin pegged to the value of $1. USD Coin does not have its own blockchain; its tokens are available on other blockchains, such as Ethereum. USD Coin is a centralized stablecoin, with the issuer controlling the supply.

Circle ensures that each USDC token can be exchanged for US Dollars on a 1:1 ratio. Unlike Tether Ltd., Circle goes to great lengths to ensure transparency of its assets, publishing monthly audits and being a licensed money transmitter under U.S. law, having its financial statements audited annually by the SEC (circle.com). The tokens are issued when users send their US Dollars to the token issuer, who in turn creates an identical quantity of USDC by using a smart contract, thanks to which the created USDC tokens are simultaneously backed by the US Dollar reserves. An asymmetrical procedure is followed if a user wants to exchange their USDC for US Dollars. If the bank transfers work without issues, Circle USDC does not charge customers any feed for tokenizing and redeeming services (analyticsinsight.net).

Circle does a lot to minimize the issues that come with USD Coin being a fiat-collateralized stablecoin, as it ensures a superior level of transparency by being fully regulated and having its fiat reserves in one of the biggest banks. USD Coin is the second largest stablecoin by market capitalization, only behind Tether, which had a 4-year head start over USD Coin. Figure 6 shows that USD Coin successfully maintains its value close to $1.

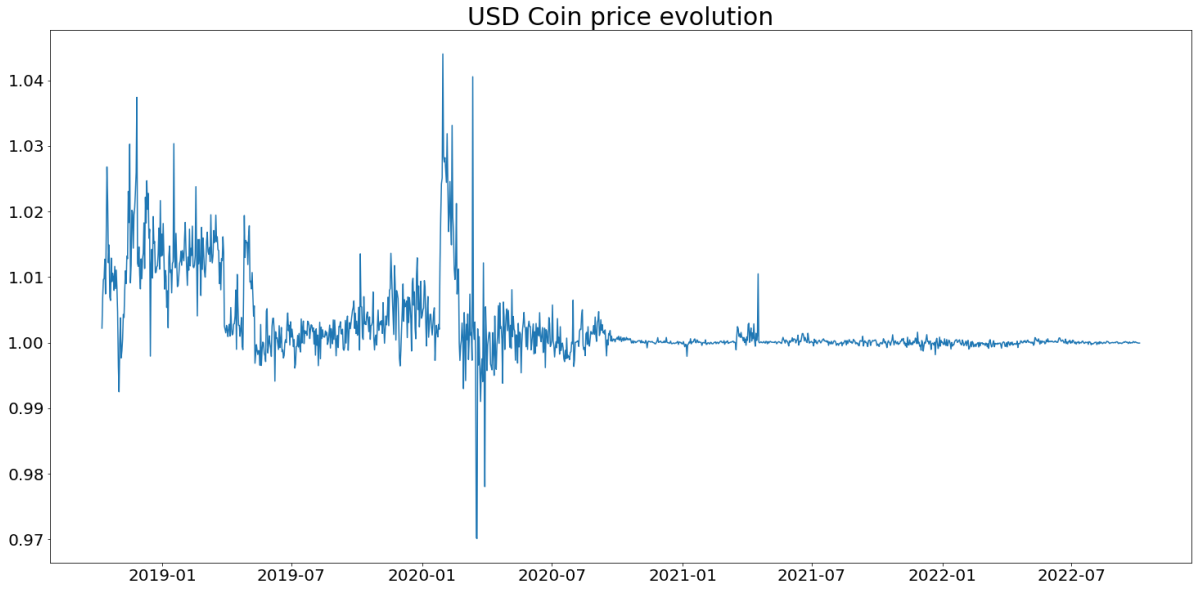


Figure 6 (The evolution of USD Coin prices)

2.5. Social media

Social media can be defined as websites and applications focusing on community-based input, interaction, content sharing, communication, and collaboration (techtarget.com). The inception of social media is said to be in the late 1990s. Today, almost 4.6 billion people use social media sites, more than half of the world's population (statista.com). Social media's rise in popularity can be addressed to technological innovations and widespread usage of personal computers and smartphones, as the social media application can be used on many electronic devices such as laptops, smartphones, tablets, and even certain electronic watches.

The rise of social media has led to the companies operating them becoming technological giants, the best example of which is Meta, the parent company of Facebook, which reached an over $1 billion US Dollar valuation in August 2021(macrotrends.net). Social media popularity has skyrocketed this millennium. Moreover, the users spend 2 hours and 29 minutes daily on various social media sites (smartinsights.com).

Social media offers unprecedented insight into people's thoughts, emotions, and opinions, which is more accessible than ever. This paper uses social media data and performs Social Media mining, defined by Zafarani et al. (2014) as "the process of representing, analyzing, and extracting actionable patterns from social media data". Thanks to Machine Learning and Natural Language Processing techniques, social big data can be used for opinion mining and, more specifically, sentiment analysis (Bello-Orgaza et al., 2016). The sentiment analysis aims at quantifying the subjectivity of a text, and, in the context of the big social media data, the aggregate subjectivity of social media posts can uncover people's general sentiment on a given subject. Social media mining spans many disciplines, such as computer science, statistics, as well as sociology, and ethnography (Zafarani et al., 2014).

This paper uses data from two social media in particular: Twitter and Reddit, as both are popular micro-blogging platforms that can be filtered through interests. Moreover, most of their content is in text format, which is easily extractable.

2.5.1. Twitter

Twitter is an online social networking service founded in 2006 by Jack Dorsey. Twitter's content consists of short messages called 'tweets' posted by users. A tweet can contain a maximum of 280 characters and include outside links, videos, or photos (lifewire.com). The users can like, reply, or retweet a tweet; the reply to a tweet is another tweet linked to the one it is in reply to, and a retweet is a reposting of the original tweet. The users can tag each other in tweets by using the @ symbol followed by the username and include hashtags following the symbol #. A hashtag is used to index keywords or topics on Twitter, which allows users to more easily navigate through topics that interest them (twitter.com).

Twitter has about 486 million users worldwide (datareportal.com), and over 500 million daily tweets are published (blog.hootsuite.com). Moreover, almost half of Twitter's audience consumes news regularly on the platform (blog.hootsuite.com). This also means that users may perceive the information on Twitter as being more official, with most of the world's leaders and international organizations having a Twitter account to communicate with their citizens. Thus, Twitter is a great candidate for opinion mining and sentiment analysis, as there are not only many tweets but since users might associate Twitter with more legitimate information, the tweets might have a bigger impact on peoples' actions when it comes to investing in cryptocurrencies. Moreover, as mentioned earlier, Elon Musks' tweets in 2021 had a near-immediate effect on Bitcoin prices (vox.com), which makes the analysis of tweets as a variable influencing cryptocurrency prices even more worthwhile.

2.5.2. Reddit

Reddit was created in 2005 by Steve Huffman and Alexis Ohanian and is valued at around 6 billion US Dollars (cnbc.com). Reddit is an online forum divided thematically into different discussion groups called 'subreddits'. A post on Reddit is called a 'submission'; a submission can be a text, image, video, or an internal or external link; other users can then upvote, comment or reply to other users' comments in the thread (medium.com). Thanks to the voting system, users know which content is the most valuable, as the posts with the most upvotes will appear higher on the page. Moreover, the post creators have a high incentive to produce quality content as they receive a 'Karma' score which reflects the quality of their submissions (medium.com).

Typically, a user will subscribe to many subreddits based on their interests; they will see the top posts of the subscribed subreddits on their front page. The front page and the individual subreddits page can be sorted in five diverse ways: hot – based on the voting algorithm and recency; new – based only on their recency; rising – gaining popularity, top – based only on the number of upvoting and controversial – showing posts with a high number of both up-votes and down-votes (medium.com). Furthermore, Reddit has a large community of volunteer moderators who may settle disputes and remove content they deem unsuitable or inappropriate (newscientist.com).

GameStop (theprint.in) example shows that Reddits' communities can impact the prices of stocks. The fact that Reddit users take investing advice from the site, along with the thematic breakdown of the content on Reddit, and most of the content is in written form, reviewed by moderators makes it a suitable place for opinion mining and sentiment analysis.

2.5.3. Wikipedia

The last data source for this paper will be Wikipedia trend. Wikipedia is a free, open-source, Internet-based encyclopedia, founded in 2001, and currently overseen by the nonprofit Wikimedia Foundation (britannica.com). Wikipedia offers many articles explaining a variety of topics and is one of the most visited sites on the internet. The main criticism of Wikipedia as a knowledge source is the fact that anyone can write or edit an article which makes the site not fully reliable. Wikipedia does have many volunteer

administrators who act as content moderators and have the power to, for instance, block and unblock user accounts, restrict or allow access to editing an article or delete and restore certain pages (en.wikipedia.org).

Despite its criticism, many people use Wikipedia to learn about new subjects, as Wikipedia pages get more than 2,500 views per second and over 6 million articles are available in English (webtribunal.net). Because of Wikipedia's popularity, the author has chosen to include the Wikipedia trend as a variable that could influence the prices of cryptocurrencies. Wikipedia trend represents the number of visits to a given Wikipedia article each day.

3. Literature Review

This paper aims to examine the relationship between online factors and cryptocurrency closing prices. In order to achieve that, the conducted literature review focuses on three aspects of cryptocurrency research.

Firstly, the cryptocurrency price prediction with only historical price values, which includes predictions using time series models and machine learning models, to establish which models are among the most successful at predicting closing prices. This aspect also includes the interlinkages between cryptocurrencies.

Secondly, the studies of online data related to cryptocurrencies, such as the sentiment analysis of social media data and other online factors. This is done for the purpose of examining which are the most studied online factors related to cryptocurrencies and what were the successful approaches in their analysis. This aspect includes Twitter, Reddit, and Wikipedia data and different sentiment analysis approaches.

Lastly, it focuses on the performance of machine learning models, which include online data as additional inputs to the models, to find out which online factors are deemed to be the most useful in predicting closing prices for cryptocurrencies.

3.1. Price prediction

Traditional time series models such as ARIMA produce excellent results for many economic data; yet, their performance with cryptocurrency price prediction is rather poor. A. Azari (2019) employed multiple ARIMA models to predict Bitcoin's future closing prices; however, due to Bitcoin's price vulnerability, the results produced large MSE values. Wirawan et al. (2019), on the other hand, found that ARIMA models work well for short-term predictions, producing very small MAPE. However, as the prediction period gets longer, the inaccuracy of the predictions increases drastically. Khedr et al. (2021) performed a survey of both traditional statistical as well as machine learning techniques most used in the literature for cryptocurrency price predictions. The most common traditional statistical methods are exponential smoothing, VAR-GARCH, and ARIMA. However, a common disadvantage of these models is that they require statistical assumptions such as the seasonality of the time series, and cryptocurrencies have no known seasonal factors. The most popular Machine Learning approaches for cryptocurrency price prediction are neural networks, especially ANN, RNN, LSTM, and GRU. Khedr et al. (2021) state that due to traditional statistical methods' drawbacks, machine learning techniques are of a better use when predicting cryptocurrency prices, even though that approach also has its own challenges. S. Biswas et al. (2021) have tested multiple neural network algorithms in order to predict the price of Litecoin and Monero. They suggest a pricing mechanism based on both GRU and LSTM. Hansun et al. (2022) have compared the performance of three recurrent neural network algorithms: LSTM, Bi-LSTM, and GRU, by using a multivariate approach to predict the prices of five cryptocurrencies (Bitcoin, Ether, Cardano, Tether, and Binance). Their findings suggest that Bi-LSTM and GRU have similar results, both better and more consistent than LSTM.

Apart from online factors, many scholars examine the relationship between different cryptocurrencies, as well as between cryptocurrencies and financial markets, for price prediction (S. Biswas et al., 2021) or effective diversification of cryptocurrency portfolios (I. Yousaf & S. Ali, 2020). Q. Ji et al. (2019) found that the connectedness via negative returns is considerably stronger than via positive ones as well as a cryptocurrency's market size is not necessarily related to its importance in return and volatility connectedness. Moreover, I. Yousaf and S. Ali (2020), in a study examining the interlinkages between Bitcoin, Ether, and Litecoin during the pre-Covid and Covid-19 period, found the conditional correlation between cryptocurrencies only became stronger during the latter. Sovbetov and Yhlas (2018) concluded that the crypto-market factors appear to be significant determinants both in the short- and long-run. In contrast, macro-financial factors such as the SP500 index seem to have a weak positive long-run impact on the studies of cryptocurrencies but a little negative in the short-run. Conlon et al. (2020) studied whether cryptocurrencies were a safe haven for investors during the Covid-19 period. They found that Bitcoin and Ethereum are not safe havens, and out of all the leading economies, only investors in the CSI 300 would benefit from diversifying their portfolio with those two cryptocurrencies. Tether, however, did act as a safe haven for all the indices throughout the Covid-19 period.

3.2. Online factors related to cryptocurrency prices

Cryptocurrencies are purely digital assets; hence, the online activity could heavily influence their price. One of the most popular online factors considered by scholars is the sentiment of the tweets, Lamon et al. (2017) created supervised machine learning models that were able to predict the days with the largest increases and decreases in the price of Bitcoin and Ethereum by using the sentiment of Tweets and news headlines. In a similar study, Valencia et al. (2019) used the VADER polarity score to establish sentiment for tweets concerning Bitcoin, Ethereum, Ripple, and Litecoin. They found that Twitter data could be beneficial in predicting the price movement of cryptocurrencies and that Neural Networks outperformed other supervised learning algorithms. Abraham et al. (2018) found that Tweet volume (the total amount of tweets, regardless of sentiment) and Google trends act as better predictors than the sentiment of the tweets, as the latter tends to remain positive regardless of price changes. Shen et al. (2018) have focused exclusively on Tweets volume and found that the previous day's tweets volume is a significant driver of the next day's trading volume and realized volatility but not returns.

Another very popular social media which gets a lot of scholars' attention is Reddit. Phillips et al. (2018) thanks to using the wavelet function have found that there exists a medium-term correlation between the popularity growth of a given cryptocurrency on Redditt forums as well as Wikipedia and Google trends and the price of a given cryptocurrency. ElBahrawy et al. (2019) have analyzed the evolution of Wikipedia pages and pageviews of 38 cryptocurrencies; their analysis has shown that Wikipedia data can benefit investors' decision-making. However, in the case of Reddit and Twitter, the most common way to examine the relationship is to examine the sentiment of submissions or submission titles. J. Bukovina and M. Marticek (2016) have found that the sentiment of Reddit's submission titles can explain a part of Bitcoin's total volatility. A more recent study by Seroyizhko et al. (2022) has found that using too much sentiment data from several subreddits deteriorates the performance of Neural Network prediction algorithms and points to evaluating the contribution of each set of sentiment features in future research.

3.3. Sentiment Analysis approaches

There are many ways in which the sentiment could be established; for instance, Rouhani & E. Abedin (2019) first used wordnet, an R package, to produce a test sample and then utilized several classification techniques to establish the sentiment of the remainder of the data. They have found that with SVM, one can predict the tweets' sentiment with high accuracy. T. Pano and R. Kashef (2020) conducted a thorough analysis of different pre-processing strategies that would yield the sentiment scores with the highest

correlation to Bitcoin prices in the COVID-19 period. They found that removing Twitter-specific tags tends to improve the correlation of sentiment scores with Bitcoin prices; however, even then, the sentiment correlates significantly with the prices only over shorter timespans. For the non-English social media sentiment, Huang et al. (2021) built a crypto dictionary for the sentiment analysis of Chinese social media, whose results yielded high accuracy. They recalled scores when used in predicting the price movements of Bitcoin, Ether, and Ripple. A. Burnie and E. Yilmaz (2019) decided to take a step further and analyze the price dynamics of given words from the Reddit submissions in the 2017-2018 period. Their Data-Driven Phasic Word Identification methodology concluded that the growing popularity of certain words follows the change in the price dynamics, for instance, the word 'ban', which referred to government regulations and internet companies banning cryptocurrency adverts.

3.4. Price prediction including online factors

A great advantage of machine learning algorithms is that most of them can include additional data that can help predict the price; for instance, K. Wołk (2020) included the Twitter sentiment in his models for short-term cryptocurrency prediction, he made a Python script which took actions automatically based on the predictions yielded by his models. After 30 days, the bot made a 14,82% profit. Just like Abraham et al. (2018), K. Wołk (2020) found that the sentiment tends to be positive no matter the price changes. However, a combination of Google trends and weighted sentiments is the most powerful predictor, with the negative being the predominant one. Eisen, A. M. (2018) employed an LSTM neural network and added Wikipedia pageviews to predict Bitcoin prices; the results have shown a strong relationship between Wikipedia Trend and the closing price of Bitcoin. Raju, S. M., & Tarif, A. M. (2020) concluded that a real-time LSTM model, including the sentiment from Twitter and Reddit posts, is considerably more effective at predicting future prices than an ARIMA model. One of the most recent studies by Critien et al. (2022), employed an ensemble method of various Neural Networks and included Twitter sentiment to perform a near-real-time price prediction for Bitcoin prices; their results showed an excellent MAE of 88.47%, better than the daily prediction.

3.5. This paper's contributions

The literature review shows that the most successful algorithms for cryptocurrency price prediction are LSTM, Bi-LSTM, and GRU, while time series models tend to be less successful. Moreover, the most studied online factors that can help explain the closing prices of cryptocurrencies are Twitter, Reddit, and Wikipedia trend. Additionally, VADER is the most widely used tool for sentiment analysis. Lastly, online factors such as social media sentiment greatly improve prediction models.

In order to answer the research questions, this paper will use data from Twitter, Reddit, and Wikipedia for the five chosen cryptocurrencies, four machine learning algorithms (RNN, LSTM, Bi-LSTM, and GRU), and two time series models (ARIMA and SARIMA), as well as the analysis of the covariance and random forest feature extraction results. Most of the relevant papers focused mainly on Wikipedia trend or Twitter or Reddit sentiment as potential factors influencing cryptocurrency prices; however, no paper has contained all those factors and provided a throughout comparison of each one's influence. This paper offers a comprehensive comparison of the influence of Wikipedia, Twitter, and Reddit variables on the five biggest cryptocurrencies.

The second contribution of this paper is the analysis of different variables of Twitter and Reddit data, not just sentiment and volume. The literature review section shows that, although the study of social media sentiment in terms of future price predictions makes the most intuitive sense, the research shows that other factors, such as Twitter volume or Wikipedia Trend tend to perform just as well or even better as future price predictors. Therefore, this paper also examines other variables of the social media data, such as the number of Likes, Retweets, or upvote ratio.

Literature Review Table

Paper	Objective	Model		Sentiment Analysis	Twitter	Reddit	Wiki	Studied cryptocurrencies
		Time Series	Machine Learning					
Seroyizhko et al. (2022)	Prediction, Sentiment Analysis	No	Yes	Yes, VADER and others	No	Yes	No	BTC
Critien et al. (2022)	Prediction	No	Yes	Yes	Yes	No	No	BTC
Hansun et al. (2022)	Prediction	No	Yes	No	No	No	No	BTC, ETH, ADA, USDT, BNB
S. Biswas et al. (2021)	Prediction	No	Yes	No	No	No	No	LTC, XMR
Huang et al. (2021)	Prediction	No	Yes	Yes	No	No	No	BTC, ETH, XPR
Khedr et al. (2021).	Survey on price prediction	Yes	Yes	No	No	No	No	BTC
K. Wołk (2020)	Prediction	No	Yes	Yes, VADER	Yes	No	No	BTC, ETH, ETN, XRP, ZEC, XMR
T. Pano & R. Kashef (2020)	Sentiment Analysis	No	No	Yes, VADER	Yes	No	No	BTC
Conlon et. al. (2020)	Risk Analysis	No	No	No	No	No	No	BTC, ETH, USDT
I. Yousaf & S. Ali (2020)	The return and volatility spillover analysis	Yes	No	No	No	No	No	BTC, ETH, LTC
Raju, S. M., & Tarif, A. M. (2020).	Prediction	Yes	Yes	Yes	Yes	Yes	No	BTC
ElBahrawy et al. (2019)	Prediction	No	No	No	No	No	Yes	Not stated
Q. Ji, et al. (2019)	Return and volatility spillover analysis	no	no	No	No	No	No	BTC, ETH, LTC, Dash, XRP, XML
A. Burnie & E. Yilmaz (2019)	Online factors influencing cryptocurrency prices	No	No	Yes, VADER	No	Yes	No	BTC
Valencia et al. (2019)	Prediction	no	Yes	Yes, VADER	Yes	No	no	BTC, ETH, XRP, LTC
S. Rouhani & E. Abedin (2019)	sentiment analysis	No	Yes	Yes	Yes	No	No	BTC, ADA, ETH, LTC, XRP
Amin Azari (2019)	Prediction	Yes	No	No	No	No	No	BTC
Wirawan et al. (2019)	Prediction	Yes	No	No	No	No	No	BTC
Abraham et al. (2018)	Prediction	No	No	Yes, VADER	Yes	No	No	BTC, ETH
Phillips, R. C., & Gorse, D. (2018)	Online factors influencing cryptocurrency prices	No	No	No	No	Yes	Yes	BTC, ETH, XMR, LTC
Sovbetov & Yhlas (2018)	Factors influencing the prices	No	No	No	No	No	No	BTC, ETH, Dash, LTC, XMR
Shen et al. (2018)	Online factors influencing cryptocurrency prices	Yes	No	No	Yes	No	No	BTC
Eisen, A. M. (2018)	Prediction	No	Yes	No	No	No	Yes	BTC, Dash, DOGE, ETH, LTC, XMR, NEM, XRP, XLM
Lamon et al. (2017)	Prediction	No	Yes	Yes	Yes	No	No	BTC, ETH, LTC
J. Bukovina & M. Marticek (2016)	Volatility and sentiment analysis	No	No	Yes	No	Yes	No	BTC

Table 1 (Literature review table)

4. Methodology

This section introduces the methodology used throughout the paper, which provides a structured approach that helps the author answer the research questions and examines the relevant topics.

The Research Onion model created by Saunders et al. (2007) represents the methodological process essential for conducting quality research. The study will hence go through each of the Research Onion model's layers presented in Figure 7, starting with the research philosophy, followed by the research approach. Afterward, the methodological choice, research strategy, purpose, and time horizon will be discussed in the research design section.

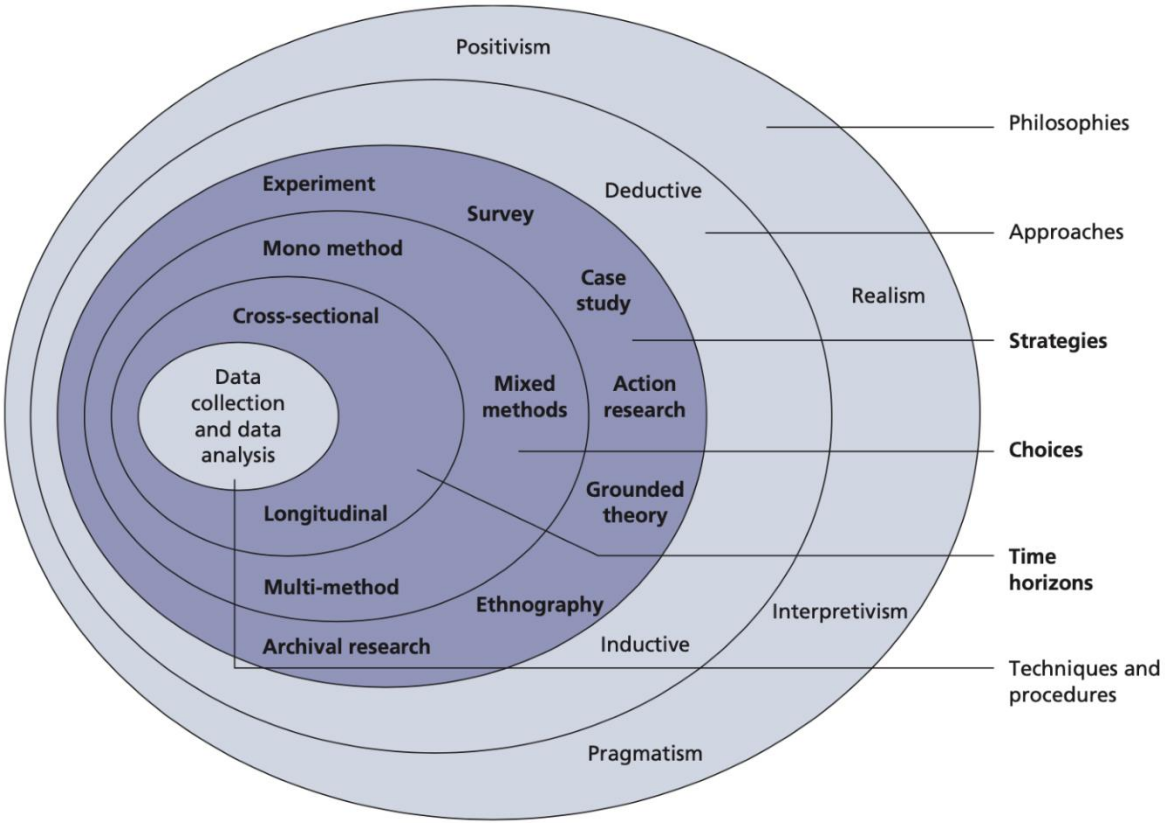


Figure 7 (Research onion from Saunders et al., 2007)

4.1. Research philosophy

Research philosophy, the first layer of the research onion, is "a system of beliefs and assumptions about the development of knowledge" Saunders et al. (2007). The choice of the underlying research philosophy is important, as it impacts how the research is conducted and how the author understands their findings and research process. Hence, the research philosophy must be aligned with the other research onion layers. Saunders et al. (2007) list four major research philosophies: pragmatism, positivism, realism, and interpretivism. In order to properly argue the choice of the underlying research philosophy, the philosophies mentioned above will be discussed below.

Research philosophies can be best distinguished by their assumptions in ontology, epistemology, and axiology, which in turn vary depending on whether one takes an objective or subjective stance; hence all these concepts will be explained first.

Ontology is a set of assumptions about the nature of reality. From a closer perspective, it answers the question of what one constitutes as a fact (Saunders et al., 2007). Therefore, the author's ontological assumptions establish the research's perception of studies phenomena and whether they exist independently of human understanding and interpretation.

Saunders et al. (2007) state, "Epistemology concerns what constitutes acceptable knowledge in a field of study.". Hence, epistemology refers to what the researcher believes to be acceptable knowledge and the acceptable way to convey this knowledge to others. Epistemological choices are essential as they influence the latter layers of the research onion.

Axiology is a discipline of philosophy that examines the concepts of values and judgment (Saunders et al., 2007). It refers to the role of researchers' values and ethics, reflected in their research design choices, collected data, or how it was analyzed and presented.

Lee & Lings (2008) explains how these concepts are strictly interconnected: ontology establishes how the researchers perceive reality, and epistemology establishes the knowledge one can learn about reality and is thus strictly dependent on how we define reality to be. Axiology establishes the research's aim, whether the researcher is trying to predict, explain or understand the world. Lastly, Lee & Lings (2008) mentions that Methodology establishes how the researcher will go about their research and fully depends on the chosen ontological, epistemological, and axiological assumptions.

Ontology, Epistemology, and Axiology might adopt two extreme stances; objectivism and subjectivism. Objectivism is the belief that social entities exist in reality which is external to social factors (Saunders et al., 2007). Consequently, social actors do not affect the social world, and, at its most extreme stance, it assumes that there exists only one social reality in which all the social actors exist (Saunders et al., 2007).

Subjectivism resides on the other side of the spectrum; its core beliefs are that social actors' views and subsequent actions result in social phenomena. Consequently, social phenomena are not static but continuous and under constant revision, as they are influenced by the process of social interactions (Saunders et al., 2007). Hence, for a subjectivist, the reality is socially constructed and under continuous revision rather than existing independently of social actors.

Pragmatism treats research questions as the best determinant of epistemological, ontological, and axiological assumptions (Saunders et al., 2007). Consequently, this approach makes the researcher choose the right assumptions and work with different variations of assumptions and approaches from all three philosophical assumptions. In their comparison of the four research philosophies table, Saunders et al. (2007) (p.119) also mention that researchers' values are significant while interpreting the results, as the researcher takes both a subjective and objective approach.

The positivist stance is the stance of a natural scientist; this research philosophy focuses mostly on observable data, as that is the only credible data from a positivist's researchers' perspective (Saunders et al., 2007). Consequently, to adopt a positivist stance, the researcher should collect their data with an emphasis on large samples and measurements of quantitative data, as well as is likely to focus on already existing theories to develop hypotheses which in turn will be tested in their research (Saunders et al., 2007). The positivist approach should be conducted value-free so that the researcher becomes independent of the collected data (Saunders et al., 2007).

Realism treats reality as what the researcher perceives with their senses so that the objects exist independently of the human mind (Saunders et al., 2007). Two main schools of thought within realism are direct realism which states that data can be easily misinterpreted due to insufficient information, and critical

realism, which refutes that claim. The researcher's values, as well as background, are hence essential, as their senses and biases are unavoidable for the results' interpretation (Saunders et al., 2007).

Interpretivism focuses on understanding the differences between humans as social actors since the interpretivist belief is that reality is socially constructed and constantly in motion (Saunders et al., 2007). To analyze the data in interpretivist research, one must understand the subjective reality behind the details of a situation and the motivations of the affected social actors. Moreover, as the researcher is inherently subjective and is, therefore, part of the research, their values greatly matter to the interpretation of the results (Saunders et al., 2007).

Thanks to the understanding of the four major research philosophies and their assumptions that will further determine the latter research layers, the researcher has decided to follow the positivist stance. The reasoning behind this choice is the following, from an ontological perspective, the relationship between cryptocurrency prices and online factors exists independently of this paper. Epistemological choices are crucial for this paper, as it studies the relationship between online factors, including online sentiment and cryptocurrency price. Although people's sentiment is a social phenomenon, in this paper, it is represented by numbers from the preprocessed data, which makes it quantitative and independent of the researcher. Lastly, to conduct positivist research, the researcher must detach themselves from their values and beliefs to attain objective results (Saunders et al., 2007).

4.2. Research approach

The two major research approaches that could be employed in a study are deduction and induction. Even though the deductive approach is mostly linked to positivist research and the inductive approach is more related to the interpretivist research philosophy, the choice of a given research philosophy does not rule out the use of each of the two research approaches (Saunders et al., 2007).

The deductive approach focuses on deducting hypotheses and expressing them in operational terms, afterward testing those hypotheses, critically examining the outcomes, and, if necessary, changing the underlying theory based on the yielded results (Saunders et al., 2007). This approach is mostly related to natural sciences and the use of quantitative data; the meaning of the hypotheses to be operational is to make it possible for different tests can be applied and measurements can be taken to accept or rule out the hypotheses, as well as clearly describe that process (Saunders et al., 2007).

The inductive approach process starts with the data collection rather than a hypothesis; the goal is to formulate a theory through the exploration of the study's phenomenon. Hence, the inductive approach focuses on the context in which the events occur and often focuses on smaller samples of qualitative data (Saunders et al., 2007).

Considering the available data and the positivist stance, the deduction approach was chosen for this paper. Furthermore, the paper aims to explain the relationship between online factors and cryptocurrency prices using a large amount of quantitative data, which fits Saunders et al. (2007) description of a deductive approach. Therefore, the research questions have been formulated based on existing theories regarding the relationship between online factors and cryptocurrency prices, specified in the Literature Review section. The chosen models and variables are the result of what was mentioned in the found research papers and the data exploration conducted by the author.

4.3. Research design

The research design is to be considered as the general plan of how the author will answer their research question(s); it is to contain clear objectives, specify the sources of the collected data as well as state the limitations and ethics of the study, all in relation to the research question(s) (Saunders et al., 2007). The

author guided their choices in the research design into what would be the most effective approach to answering the research questions as well as what would be feasible for time constraints.

Hence, the following section will present the methodological choice, research purpose, strategy, time horizon, data collection, analysis, and quality. The vast amount of quantitative data from various sources, which is used in this paper, necessitates the data collection and preprocessing to be a separate section.

4.4. Research choice

Methodology enables the researcher to accomplish study objectives and answer research questions properly (Saunders et al., 2007). The methodological choice is represented as the third layer of the research onion (Figure 7), and its choice is naturally influenced by the choices made regarding the first two layers, the research philosophy and research approach. The first decision regarding the methodological choice concerns the quantitative, qualitative, or mixed research design methods.

The qualitative method is usually more appropriate for the interpretivist research philosophy as it focuses on gathering data about people and social and behavioral aspects; however, it is also possible to quantify the qualitative data by, for instance, counting the frequency of certain events (Saunders et al., 2007).

The quantitative method is usually more appropriate for the positivist philosophy and deductive approach, as it focuses on collecting substantial amounts of structured data, which is represented numerically (Saunders et al., 2007). That process makes it easier to conduct tests and measurements necessary for hypothesis testing, the core of the deductive approach, and freeing the researcher of biases, which is necessary for the positivist research philosophy.

The mixed method, in accordance with its name, combines the qualitative and quantitative methods in a manner that is most appropriate for the specific research.

The paper will take the quantitative approach for two reasons. Firstly, the gathered data, cryptocurrencies prices, Twitter's and Reddit's data and sentiment, and Wikipedia trend can all be represented numerically; the quantitative method is the most appropriate to examine the relationship between these variables. Secondly, the quantitative method complements the chosen research philosophy of positivism and the deductive research approach.

4.5. Research purpose

The next step in the research design refers to the determination of the research purpose. Saunders et. al. (2007) presents three types of research purpose; exploratory, descriptive, and explanatory. Research may, however, have more than one purpose, moreover, the purpose of the research can change over time, all depending on the research question (Saunders et. al., 2007). The exploratory purpose refers to the study of a topic which is oftentimes new and understudied, the direction of the research may change depending on the data and evidence collected through the research process (Saunders et. al., 2007). The explanatory purpose refers to studies focusing on a situation or problem to explain the causal relationship between certain variables (Saunders et. al., 2007). The descriptive purpose, in accordance with its name refers to studies that describe a given phenomenon, however, Saunders et. al. (2007) points out that oftentimes description alone is not enough, and many descriptive studies tend to become descripto-explanatory studies.

This paper's fundamental purpose is explanatory, as to answer the research questions the causal relationship between the prices of the chosen five cryptocurrencies and the online factors will be investigated. The process of determining the relationship between a set of variables and one main variable, the price falls very well within the explanatory purposes' definition. The literature review section concludes that despite many research papers being published about this topic, the results are not always conclusive, hence this

paper, which investigates not only five biggest cryptocurrencies but also a multitude of variables from Twitter, Reddit and Wikipedia adds to the understanding of the subject, and is, therefore, worth studying.

4.6. Research strategy

The following research onion's layer is the research strategy, which is to be considered a general plan on how one will approach and answer the research questions. (Saunders et al., 2007). The research strategies can be applied to any research philosophy or approach, but some combinations are more appropriate than others. The research questions and objectives should guide the choice of a research strategy, and the extent of existing knowledge, the amount of time, and other available resources (Saunders et al., 2007).

Saunders et al. (2007) state that numerous research strategies are available such as experiments, surveys, case studies, action research, grounded theory, ethnography, or archival research. In order to answer the research question, two methods seem appropriate an experiment and a case study.

The experiment research strategy focuses on studying causal links between different variables; the experiment research strategy is also concerned with the size and relative importance of that relationship; they tend to answer the 'how' and 'why' questions (Saunders et al., 2007).

The case study research strategy focuses on conducting an empirical investigation of a particular and contemporary phenomenon (Yin, 2003). Moreover, it also focuses on the context of the studied phenomenon since, within a case study, the boundaries between the phenomenon and its context are ambiguous (Yin, 2003). Saunders et al. (2007) state that the case study research strategy answers questions like 'why?' as well as the 'what?' and 'how?'.

Both research strategies are appropriate to answer the research questions and are highly appropriate for explanatory studies. However, the case study research strategy would fulfill the research objectives better. The core difference between the two research strategies lies within their understanding of the context of the phenomena, for the experiment research strategy requires a highly controlled context. In contrast, a case study acknowledges its ambiguity from the studies phenomena (Saunders et al., 2007).

Moreover, this paper uses triangulation, different data collection techniques from multiple sources, and focuses on multiple cryptocurrencies in hopes of achieving generalizable results. This makes this study a multiple cases study, as although one could consider cryptocurrencies as a single phenomenon, different variables may influence the prices of different cryptocurrencies in various manners. The choice of the five cryptocurrencies, Bitcoin, Ether, Binance, Tether, and USD Coin, as stated before, is due to them being the largest cryptocurrencies by market capitalization (coinmarketcap.com).

Considering the provided evidence, the case study research strategy seems the most appropriate to provide coherence throughout the research design, which will lead to answering the research questions.

4.7. Time Horizon

The fifth layer of the research onion is the time horizon of the study. Saunders et al. (2007) state that there are two types of time horizons; cross-sectional and longitudinal. Moreover, the time horizons are independent of the choices made in the previous onion layers, research strategy, and research method (Saunders et al., 2007).

Saunders et al. (2007) refer to a cross-sectional research time horizon as a 'snapshot', a study of a particular phenomenon at a particular time. Moreover, Malhotra, N., & Birks, D. F. (2007) define a cross-sectional study as a "type of research design involving the collection of information from any given sample of population elements only once." (p.74). The cross-sectional studies can be further subdivided into single and multiple cross-sectional studies; in the first case, there is only one sample of the information, which is

obtained only once; in the latter, there is more than one sample, but the information from each is only obtained once (Malhotra, N., & Birks, D. F., 2007).

The longitudinal research time horizon aims at examining the change and development of a certain phenomenon over time (Saunders et al., 2007). Therefore, the longitudinal time horizon study involves a fixed sample whose evolution over time is investigated (Malhotra, N., & Birks, D. F., 2007).

This paper's analysis spans over a considerable period of time, in the most extreme case of Bitcoin, as long as ten years of data. The purpose of the study is not to examine the evolution of the relationship between online factors and cryptocurrencies' prices over time but rather to examine the relationship in the given moment and during the Covid-19 pandemic period. Additionally, because the data was only gathered once, no consideration was given to changes over time. Therefore, the study has a cross-sectional research time horizon, as it is more appropriate to answer the research question and fulfill the research objectives.

4.8. Analysis Plan

This section will discuss the techniques and tools used to answer the research questions and fulfill the research objectives. As presented in the literature review section, most research focuses on the relationship between cryptocurrency price and Twitter sentiment. While less common Reddit sentiment and Wikipedia trend produce encouraging results (Phillips, R. C., & Gorse, D., 2018). Hence, this paper will include Twitter data acquired via *snscrape,* Reddit data acquired via *Pushshift.io,* and the Wikipedia Trend acquired through *pageviews.vmcould.org*.

The literature review section shows that VADER is one of the most common tools implemented for the sentiment of social media posts. Therefore, the text from the acquired tweets and Reddit submissions will undergo NLP preprocessing to conduct a sentiment analysis in python using the VADER sentiment analyzer, whose functioning and output will be further discussed in the following subsection.

The closing prices of each cryptocurrency will also be analyzed based on their correlation and covariance with the closing prices of the other cryptocurrencies. The two stablecoins, Tether and USD Coin, will be analyzed separately in relation to each other as they are different from the other three chosen cryptocurrencies, with their goal being to have a value as close to $1 as possible.

In order to establish the relationship between the online variables and closing prices of cryptocurrencies, the four datasets; cryptocurrency price, Wikipedia trend, as well as Reddit and Twitter datasets will be merged, a process which will group the outputs daily by date and hence remove the raw text and leave the sentiment instead.

As each cryptocurrency has many online variables to consider, the next step of the analysis will be the feature selection, i.e., choosing the features to be included in later models. In order to perform the feature selection, firstly, each variable will be compared with the price of the corresponding cryptocurrency by using Pearsons' r value of correlation in relation to the price. Secondly, Random Forest feature extraction will be used to establish the importance of each variable in explaining the relationship with the price.

Afterward, the online variables' relationship with the closing price of the corresponding cryptocurrency will be tested by observing whether they can improve cryptocurrency price predictions. The chosen models will be run firstly just on the historical closing prices, with no additional data; these results will be compared to the results produced by models, which include each of the online variables selected through the feature selection process.

Based on Khedr et al. (2021) survey, the most employed machine learning algorithms for cryptocurrency price prediction are: RNN, LSTM, and GRU, the author decided to use those exact algorithms as well as

the Bi-LSTM to observe whether the bidirectional flow of information will increase the models' accuracy. Despite the drawbacks of traditional statistical models when it comes to cryptocurrency price prediction pointed out by Khedr et al. (2021), the author decided to use the time series models of ARIMA and SARIMA first to observe whether some seasonal effects that can improve the prediction and secondly, as these are autoregressive models and additional data cannot be added like for the neural network algorithms, they will act as a second baseline for the machine learning models. The general outline of the analysis plan is presented in the diagram below (Figure 8.)

The entire code for the paper was done in Python and is attached to the thesis as separate appendices. There are three code files; *Data_download.ipynb*, *Sentiment.ipynb* and *MT_analysis.ipynb*. Data_download.ipynb file contains the Python script used to download Twitter and Reddit data. Sentiment.ipynb file contains the Python script used to perform the sentiment analysis on Tweets and Reddit submissions. MT_analysis.ipynb file contains the Python script used to perform the Time Series and Machine Learning models.

There are two .zip folders containing data: *'Data_for_MT_analysis.zip'* and *'Data_sentiment_analysis.zip'*. Data_for_MT_analysis.zip contains five datasets, one per cryptocurrency. Each data file contains daily prices, wikipedia trend as well as all Twitter and Reddit variables. The Data_sentiment_analysis.zip folder contains 14 data files. Five files for cryptocurrency prices, five for Twitter data about each cryptocurrency, three for Reddit data about Bitcoin, Ether and Binance, and one file containing Wikipedia trend for all five cryptocurrencies. These files can be used to run the Sentiment.ipynb Python script.

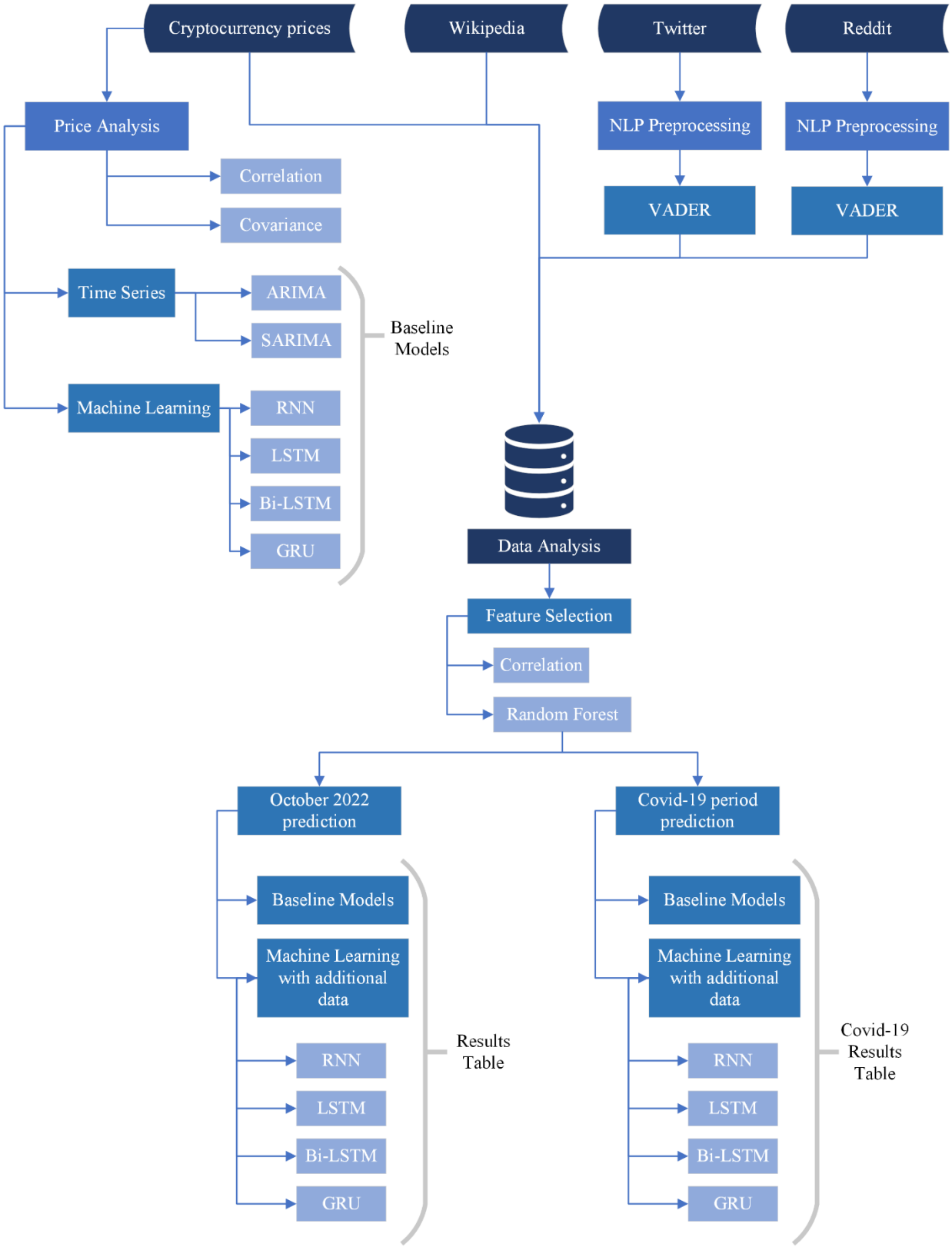


Figure 8 (Analysis plan)

5. Data Collection and Preprocessing

The previous sections established that the paper will be a cross-sectional case study, which follows a positivist research philosophy, with a deductive research approach and a quantitative research method and with an explanatory research purpose.

This section will provide a description and explanation of the data and the rationale behind selecting the chosen variables included in the models. Firstly, the data type will be specified, followed by a description of the data acquisition process. Subsequent sections will present the data preprocessing and the sentiment analysis using the VADER sentiment analyzer, followed by the section on merging the data into one dataset ready for the chosen models. The general outline of the section is presented in Figure 9.

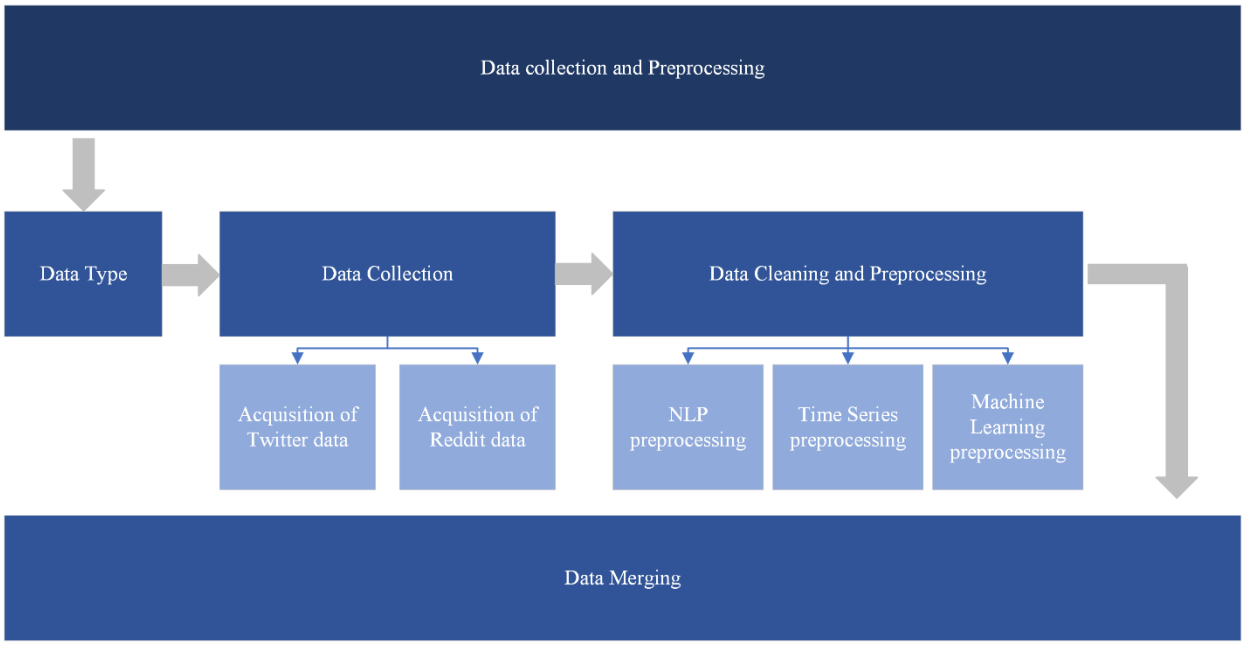


Figure 9 (Structure of Data collection and preprocessing section)

5.1. Data Type

There are two types of data that can be gathered for research; primary and secondary data. The primary data relates to new data created by the researcher tailored to address the research problem, making the researchers the first users of the data (Veal et al., 2017).

Secondary data is the data that is already collected for other purposes; it is often omitted but can be extremely helpful (Saunders et al., 2007). Most importantly, secondary data may but does not have to come from someone else's research but may be available from sources such as government departments conducting surveys, newspapers, or other organizations (Saunders et al., 2007).

Malhotra, N., & Birks, D. F. (2007) state that prior to collecting primary data, one should exhaust the possibilities that the secondary data has to offer and focus on the primary data only if the secondary data provides marginal results. Following Malhotra, N., & Birks, D. F. (2007) statement, the secondary data collected from publicly available sources yielded satisfying results, answered the research questions, and fulfilled the research objectives; the author believes there was no need to collect primary data.

The data gathered for this paper, despite being mostly in the format of raw tweets and Reddit submissions, has been gathered from existing databases, i.e., Twitter, Reddit, Wikipedia, and Yahoo! Finance. Therefore, following the definition of primary and secondary data discussed above, all data used in this paper is of secondary nature, as it has been collected from publicly available sources, and the author is not the first to use it for research purposes.

5.2. Data Collection

This paper has four data sources; Twitter from which the raw tweets about each cryptocurrency were downloaded, Reddit from which the raw Reddit submissions were obtained, pageviews from which the Wikipedia trend for the Wikipedia page of each cryptocurrency was downloaded and Yahoo! Finance from which the historical prices of each cryptocurrency were downloaded.

The acquisition of Wikipedia trends and historical cryptocurrency prices was very straightforward. Both websites provide a download function for the chosen datasets, which downloads the chosen dataset in a .csv format. The daily historical prices for Bitcoin were acquired from the 17th of September 2014 till the 5th of October 2022, the prices of Ether, Binance, and Tether were from the 9th of November 2017 to the 5th of October 2022, and the USD Coin from the 8th of October 2019 till 5th of October 2022.

The acquisition of Twitter and Reddit data required multiple steps; hence the process of collection of each one will be specified in the two subsections below.

5.2.1. The collection of Reddit Data

Reddit offers access to their API freely to the users who want to download the raw submissions from their site. However, Pushshift, a data platform with a track record in peer-reviewed publications and an active community of several hundred users, is easier to use and offers a larger limit than Reddit's API (Baumgartner et al., 2020).

In order to retrieve all the Reddit submissions, PMAW, a multithread Pushshift.io API Wrapper for reddit.com comment and submission searches, was used (github.com/mattpodolak/pmaw). As discussed in the theoretical background section, Reddit can be navigated through different subreddits, each being a separate discussion forum. The goal was to retrieve all the submissions from the most popular subreddits about each of the five chosen cryptocurrencies. Three subreddits were considered 'r/bitcoin' with 4.7 million members, 'r/EthTrader' with 2.2 million members, and 'r/binance' with 0.88 million members as they were the subreddits concerning the chosen cryptocurrencies with the most members (Reddit.com). The reason for not including the subreddits concerning Tether and USD Coin is that they have about 10,300 and 3,100 members, respectively (Reddit.com). The small audience these Reddit submissions can reach was the primary reason for not including data from Tether's and USD Coin's subreddits.

The output was a data frame of all submissions from each respective subreddit along with multiple variables, out of which the following six were chosen to be included for the analysis: upvote ratio, awards, score, number of comments as well as the 'selftext' which is the actual text of the submission and 'created utc'. The UTC timestamp had to be converted into a standard date format that does not include hours. The selftext included several submissions which were either blank, including only pictures, or had '[removed]' and '[deleted]' as their values. These rows were removed as they provided no value to the sentiment analysis. The preprocessing of the Reddit submissions' text was identical to the preprocessing of the text of the tweets and will be discussed in the latter section, text preprocessing. 252,568 valid submissions were pulled from the r/Bitcoin, 40,861 from r/EthTrader and 53,005 from r/Binance.

5.2.2. The acquisition of Twitter Data

Twitter data was acquired through scraping Twitter thanks to the use of snscrape. Snscrape is a scraper for social networking services (SNS); it allows users to efficiently collect data from social networks such as Twitter, Facebook, or Instagram through filtering based on, e.g., hashtags, user profiles, or searches (github.com/JustAnotherArchivist/snscrape).

Twitter, unlike Reddit, does not have clearly defined discussion forums; user posts are visible to one's followers or accessible thanks to a search for certain hashtags. Therefore, the relevant tweets were identified by searching for the hashtag symbol ('#') followed by specific words. Therefore, the chosen hashtags were symbols and names of the chosen cryptocurrencies: '#Bitcoin', '#btc', '#Ethereum', '#eth', '#Binance', '#bnb', '#Tether', '#usdt', '#USDCoin' and '#usdc'.

In order to collect only the tweets that may have influenced the price, the author also decided to restrict the tweets to those that have only had at least one like and one retweet and were written in the English language. Thanks to that, it is certain that the collected tweets had at least the minimum engagement and are applicable for sentiment analysis using VADER.

The dataset of tweets concerning Bitcoin with hashtags '#Bitcoin' and '#btc' has a total of 3,853,299 tweets, dating from the 7th of January 2010 to the 11th of October 2022. The dataset of tweets concerning Ether with hashtags '#Ethereum' and '#eth' has a total of 1,640,045 tweets, dating from the 1st of August 2015 till the 11th of October 2022. The dataset of tweets concerning Binance with hashtags '#Binance' and '#bnb' has a total of 359,855 tweets, dating from the 1st of June 2017 to the 11th of October 2022. The dataset concerning Tether with hashtags '#Tether' and '#usdt' has 68,169 tweets dating from the 3rd of January 2017 till the 11th of October 2022. Lastly, the dataset concerning USD Coin with hashtags '#USDCoin' and '#usdc' has 12,496 tweets dating from the 19th of January 2018 till the 11th of October 2022. Therefore, a total of 5,933,837 tweets were scraped from Twitter.

The Twitter dataset contains a total of 15 columns; however, only six were kept: 'Datetime', the date of the creation of the tweet, 'Replies Count', the number of replies to a tweet, 'Retweet Count', the number of retweets, 'Like Count', the number of likes, 'Quote Count', the numbers of quotes as well as 'Text', representing the raw text of the tweet. The reason for keeping those variables is that the date needs to be kept for the prediction models, the text will be analyzed based on the sentiment, and the rest is numeric, which means it will remain while grouping the dataset by date in the merging process.

5.3. Data Cleaning and Preprocessing

This section details the preprocessing of the data; firstly, for the NLP preprocessing for the sentiment analysis for VADER, both Reddit submissions and Tweets will be subject to identical preprocessing. Secondly, the process of making the time series stationery for the ARIMA model will be presented. The preprocessing necessary for the Neural Networks models (RNN, LSTM, Bi-LSTM, and GRU) will be discussed.

5.3.1. NLP preprocessing

In order to perform the sentiment analysis on Tweets and Reddit submissions, the raw text had to be preprocessed. For most NLP tasks, the pre-processing would include steps such as lowercasing the text, removing the punctuation, stopwords, and emojis, along with text normalization techniques lemmatizing, and stemming. VADER can deal with uft-8 encoded emojis and treats punctuation and capitalization as important metrics of the sentiment polarity score; hence many popular NLP pre-processing techniques remove much value for VADER information (towardsdatascience.com).

Therefore, this paper will aim to remove the parts of the text that provide no value to the sentiment analysis and keep everything that it does. The text will not be lowercased, and the punctuation and emojis will be kept as they may be of great value while analyzing social media data with the VADER sentiment analyzer. However, the stopwords, URLs, and other user mentions will be removed as they do not provide much value to the sentiment analysis. The ampersand symbol (&) was represented as '&amp' in the acquired tweets, had also been removed as it is considered a stopword.

The text will subsequently be normalized—more specifically, it will be first stemmed and then lemmatized. The reason for text normalization is the fact that VADER is a rule and lexicon-based sentiment analyzer; this means that some variations of words may not be in the built-in lexicon; the process of finding the root forms of each word should make it more plausible that they do exist in the VADER lexicon.

5.3.2. Time Series Preprocessing

In order for an ARIMA model to produce a forecast, the time series needs to be firstly stationary, meaning that their properties cannot depend on the time at which the series is observed (Hyndman, R. J., & Athanasopoulos, G. (2018), chapter 8.). Hence, the data cannot contain any trend or seasonality. If a time series is non-stationary, it needs to be transformed; the transformations used in this paper are logarithmic transformation, which means that a logarithm of the original values is taken, and differencing, which is subtracting a given value of a time series with a past value of the time series (Hyndman, R. J., & Athanasopoulos, G. (2018), chapter 8).

In order to determine whether the data is stationary, the Augmented Dickey-Fuller (ADF) test was performed. ADF test assumes that the data is non-stationary; hence the goal is to reject the null hypothesis and achieve p-values below 0.05 (analyticsvidhya.com).

Since the project requires five ARIMA and five SARIMA models (one for each cryptocurrency), a custom-made function was made to establish how many transformations were required for the given data and to plot the time series decomposition for each cryptocurrency.

From the analysis, it was established that in all scenarios, the cryptocurrencies: Bitcoin, Ether, and Binance needed both logarithmic and differencing transformations, whereas Tether and USD Coin did not require any transformations as both series were already stationary.

As per pmdarima's documentation (alkaline-ml.com/pmdarima/tips_and_tricks) for time series with daily frequency, the m hyperparameter corresponds to the number of observations per seasonal cycle, will be set to 7 for all cryptocurrencies and time horizons.

5.3.3. Machine Learning Preprocessing

As mentioned in the previous section, all chosen neural network algorithms: RNN, LSTM, Bi-LSTM, and GRU, are distinct recurrent neural networks; hence they require the same type of preprocessing. Neural Networks does not accept null values, although it is not an issue for most of the variables; in some cases, it does bear some importance; hence all 'NaN' values will be replaced with '0' by using pd.fillna(0) function.

Most importantly, however, the input data must be normalized, i.e., all features must be presented on the same scale. This practice is necessary as thanks to both positive and negative values used as inputs, it makes learning more flexible, and the weights assigned to some inputs would be updated much faster, hurting the learning process (towardsdatascience.com). In order to do that, the MinMaxScaler() function from the sklearn package was used.

5.4. Data Merging

This section will first explain the process of merging Twitter, Reddit, and Wikipedia trend and cryptocurrency prices.

Firstly, the price dataset will need to be trimmed to just the closing price; hence, the variables: 'Open', 'High', 'Low', 'Adj Close', and 'Volume' will be dropped. Furthermore, as the sentiment will be based on the compound score, the sentiment analysis variables ‘pos’, ‘neg’, and ‘neu’ for both Twitter and Reddit will be dropped.

The dataset containing prices and the Wikipedia trend have the same indices (Date) and a similar number of observations, hence merging them is easy with the *pd.merge(df, how=’left’)* function, which keeps the price index. However, the datasets containing Tweets and Reddit submissions have multiple data points per date and need to be further preprocessed.

Firstly, the sentiment of each tweet needs to be classified, as mentioned in the earlier sections, this paper will follow Pano, T., & Kashef, R. (2020) classification of any tweet or submission having a compound sentiment of above 0.05 as positive, below –0.05 as negative and the rest as neutral. Furthermore, following Abraham et al. (2018) findings that the tweet volume may be of great importance, the volume of tweets and Reddit submissions will be established by adding 1 to each tweet and submission.

Afterward, the Twitter and Reddit datasets were made to have the same indices as the price dataset thanks to *pd.groupby().sum()* function. Therefore, all the variables represent the sum of the values of that variable from all the Tweets or Reddit submissions made each day. Therefore, for instance, the daily number of Twitter likes represents the total number of likes received by all tweets made on a given day. It is important to mention that the compound sentiment, as aforementioned, is the output of the VADER sentiment analyzer; thanks to the use of *groupby(‘Date’).sum()* function, the scores from all Tweets from a given day were added up, providing the daily compound sentiment for Twitter. In contrast, all Reddit submissions from a given day were added up, providing the daily compound sentiment for Reddit submissions.

This made it possible for these datasets to be merged with the price dataset by the same method as the Wikipedia dataset. Furthermore, another custom-made function was made to highlight the general sentiment in percentage terms by dividing the given sentiment volume of the posts on the chosen social media by the total volume of the posts on that chosen social media.

Therefore, the Bitcoin, Ether, and Binance datasets have 25 variables each, while the Tether and USD Coin, since they do not contain the Reddit data, have 13 variables each. Naturally, those are too many variables. In the following section, Feature selection will address this issue and provide the rationale for dropping some variables.

6. Results

The structure of the results section is presented in Figure 8 from the section '4.8 Analysis plan'. The results of machine learning models with additional data are broken down into further subsections, one for cryptocurrency. Hence, the following section is split into two significant subsections; price analysis and data analysis. Price analysis presents the correlation and covariance analysis and the results of the chosen time series and machine learning models for the prices of cryptocurrencies alone. The Data Analysis part presents the results of the correlation and random forest feature importance of the online data and the results of the machine learning model, which includes the online data. Lastly, the additional analysis of the online data throughout the Covid-19 period is presented.

The time series and machine learning models' performance will be judged by implementing two accuracy measures; RMSE and MAPE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Root Mean Squared Error (RMSE) is defined as the square root of the quotient of the squared sum of errors and the length of the prediction. Importantly, RMSE is scale dependent; therefore, for a time series with large values, like Bitcoin, an RMSE of, e.g., 200 is remarkably good, whereas the same RMSE for a time series with small values like Binance would be considered very poor.

$$MAPE = \frac{1}{n} \times \sum \left| \frac{actual\ value - forecast\ value}{actual\ value} \right|$$

The second accuracy measure, Mean Absolute Percentage Error (MAPE), is calculated by summing the quotients of errors and actual values and dividing that sum by the length of the prediction, and lastly, multiplying it by 100 in order to get the percentage terms. MAPE is not scale dependent; hence a MAPE score of, e.g., 1.5 is an excellent result regardless of the time series.

6.1. Price analysis

6.1.1. Correlation and Covariance

In order to examine the relationship between the cryptocurrencies, covariance and correlation between the cryptocurrencies will be examined. Covariance measures the direction of the relationship, while Pearsons' r correlation coefficient measures the strength of the relationship.

The correlation and covariance between the closing prices of BTC, ETH, BNB, USDT, and USDC

Cryptocurrency pairs	Covariance	Correlation	
		p-value	Pearson's r
BTC- ETH	19335684.418	0	0.925
BTC- BNB	2886263.234	0	0.909
ETH- BNB	215236.134	0	0.959
USDT- USDC	8.39E-06	7.79E-25	0.265

Table 2 (The correlation and covariance between the closing prices of BTC, ETH, BNB, USDT, and USDC)

In order to establish the correlation between the closing prices, Pearson's correlation coefficient was used from the *scipy* package in Python. Hence, the correlation is on the spectrum from +1, representing a perfect positive correlation, to –1, representing a perfect negative correlation.

The covariance table between the five cryptocurrencies, presented in Table 2 shows that Bitcoin has a positive relationship with Ether and Binance; unsurprisingly, the relationship between Ether and Binance is also strongly positive. The two stablecoins, Tether and USD Coin, have a positive relationship.

The correlation coefficient between the Bitcoin price and the Ether price is 0.925; between Bitcoin and Binance, 0.908 and lastly, between Ether and Binance, it is 0.959. All three coefficients are positive and very high, indicating a strong positive correlation between the prices. The two stablecoins, Tether and USD Coin, have a correlation coefficient of 0.265, showing a small positive correlation. None of the p-values are significant, so one can reject the null hypothesis of observing the correlations of this strength under the assumption that the two variables are uncorrelated.

6.1.2. Time series models

Table 3 compares the performance of ARIMA and Seasonal ARIMA for each cryptocurrency and time horizon. As aforementioned, the choice of the order and seasonal order were made using *model_selection* of the *pmdarima* package in Python.

Time Series models results

Cryptocurrency	Days	Order	RMSE	MAPE
Bitcoin	1 day	(3,0,3)	415.70	2.06
		(1,0,0)(2,0,0)[7]	78.08	0.39
	2 days	(3,0,2)	677.95	3.01
		(1,0,0)(2,0,0)[7]	1074.55	3.92
	3 days	(3,0,2)	572.17	2.43
		(1,0,0)(2,0,0)[7]	884.06	2.94
Ether	1 day	(2,0,2)	10.10	0.75
		(3,0,0)(2,0,0)[7]	6.23	0.46
	2 days	(2,0,2)	33.36	2.06
		(3,0,0)(2,0,0)[7]	59.59	3.32
	3 days	(2,0,2)	27.63	1.58
		(3,0,0)(2,0,0)[7]	51.12	2.89
Binance	1 day	(2,0,3)	3.34	1.14
		(3,0,0)(2,0,0)[7]	4.62	1.57
	2 days	(2,0,3)	13.53	3.75
		(3,0,0)(2,0,0)[7]	12.80	3.74
	3 days	(3,0,2)	9.67	2.89
		(3,0,0)(2,0,0)[7]	10.47	2.62
Tether	1 day	(2,0,1)	1.65E-05	1.65E-03
		(3,0,3)(0,0,0)[7]	2.92E-05	2.92E-03
	2 days	(2,0,1)	1.43E-04	1.09E-02
		(3,0,3)(0,0,0)[7]	1.36E-04	1.10E-02
	3 days	(2,0,1)	1.21E-04	9.12E-03
		(3,0,3)(0,0,0)[7]	1.14E-04	8.80E-03
USD Coin	1 day	(1,0,2)	2.62E-05	2.62E-03
		(3,0,2)(0,0,1)[7]	8.67E-06	8.67E-04
	2 days	(1,0,2)	3.81E-05	3.67E-03
		(3,0,2)(0,0,1)[7]	2.69E-05	2.29E-03
	3 days	(1,0,2)	6.09E-05	5.47E-03
		(3,0,2)(0,0,0)[7]	5.64E-05	4.52E-03

Table 3 (Time Series models results)

Out of 15 models in total, in 8 cases, the SARIMA performed better, whereas in 7 cases ARIMA was superior. For both Bitcoin and Ether, SARIMA performed better for the 1-day time horizon but not for the 2- and 3-day time horizons; this is especially interesting considering how much better the SARIMA performance is over the ARIMA performance for the 1-day horizon. For USD Coin, SARIMA performed

the best in all time horizons, indicating that the time series has strong seasonality. Hence, apart from the USD Coin, it is unclear whether seasonality significantly influences the price.

6.1.3. Machine learning models

Table 4 shows the results of the neural network models, including only historical prices for all five cryptocurrencies. In the case of Bitcoin, Ether, and Binance, the neural network algorithms outperformed the seasonal and non-seasonal ARIMA models. However, the opposite is true for the stablecoins, where the ARIMA and SARIMA models significantly outperformed the neural network models.

For Bitcoin, the GRU model performed best for the 1-day and 3-day prediction intervals, whereas LSTM performed best for the 2-day prediction interval. Neural Network algorithms perform better than ARIMA and SARIMA models, with GRU outperforming them for each time horizon. Overall, the best performance was the GRU for a 1-day horizon of a MAPE of only 0.04, whereas the worst prediction was made by the LSTM for a 3-day prediction interval with a MAPE of 3.68.

For Ether, the Bi-LSTM model perfumed best for the 1-day and 2-day prediction interval, whereas RNN performed best for the 3-day time horizon. The best predictions made by the neural networks have better accuracy than seasonal and non-seasonal ARIMA, although the performance is very similar. Interestingly, in Ether's case, GRU and LSTM had the worst performances, while the opposite is true for Bitcoin.

Binance's GRU model had the best performance for 1-day prediction, LSTM for 2-day prediction, and Bi-LSTM for 3-day prediction. RNN had the worst performance, with a MAPE close to or above 5, while the other neural network algorithms had a MAPE of below two across all time horizons. The neural networks also slightly outperformed SARIMA and ARIMA models.

Tether's GRU model had the best performance in the 1-day prediction, while Bi-LSTM had the best performance in the 2 and 3-day prediction. For Tether, seasonal and non-seasonal ARIMA performed much better than the Neural Network algorithms, with MAPE seven times smaller for all prediction intervals.

For USD Coin, RNN outperformed all neural network algorithms, with a MAPE of below 0.04 in all the prediction intervals. However, the SARIMA models performed significantly better than Neural Networks with a MAPE 30 times smaller for a 1-day prediction interval, 13 times smaller for a 2-day interval, and two times smaller for a 3-day interval.

The baseline neural network models result

Cryptocurrency	Model	1 day		2 days		3 days	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Bitcoin	RNN	380.76	1.89	532.21	2.61	435.08	1.62
	LSTM	146.87	0.73	106.69	0.47	415.24	1.50
	Bi-LSTM	92.74	0.46	615.17	2.97	823.18	3.86
	GRU	8.89	0.04	281.16	1.00	285.52	1.15
Ether	RNN	22.15	1.64	25.86	1.89	31.83	1.93
	LSTM	57.45	4.25	59.92	4.41	49.69	3.33
	Bi-LSTM	2.69	0.199	7.73	0.49	32.33	2.02
	GRU	87.33	6.46	77.06	5.62	71.65	5.19
Binance	RNN	16.30	5.54	18.94	6.36	14.62	4.81
	LSTM	4.64	1.58	3.38	0.99	6.24	1.80
	Bi-LSTM	3.87	1.316	6.50	2.07	3.52	1.01
	GRU	2.51	0.85	3.02	1.01	5.38	1.55
Tether	RNN	6.95E-04	0.070	7.95E-04	0.079	9.93E-04	0.099
	LSTM	1.67E-03	0.167	1.60E-03	0.160	1.63E-03	0.163
	Bi-LSTM	9.66E-04	0.097	8.93E-04	0.089	9.70E-04	0.097
	GRU	1.04E-03	0.104	9.58E-04	0.095	9.88E-04	0.099
USD Coin	RNN	2.63E-04	0.026	3.44E-04	0.034	1.04E-04	0.010
	LSTM	1.87E-03	0.187	1.88E-03	0.188	1.89E-03	0.189
	Bi-LSTM	1.11E-03	0.111	1.10E-03	0.110	1.12E-03	0.112
	GRU	1.80E-03	0.180	1.76E-03	0.176	1.82E-03	0.182

Table 4 (The baseline neural network models result)

6.2. Data Analysis

6.2.1. Correlation

The most correlated online variables to their respective cryptocurrency closing price

	Twitter's top 5 features			Reddit's top 5 features			Wikipedia Trend	
	variable	p-value	Pearson's r	variable	p-value	Pearson's r	p-value	Pearson's r
Bitcoin	compound	0	0.878	upvote ratio	0	0.722	8.38E-09	0.106
	Pos_vol	0	0.861	number of comments	1.06E-130	0.427		
	Volume	0	0.837	compound	5.41E-94	0.366		
	Likes	0	0.829	Pos_vol	5.54E-75	0.329		
	Retweets	0	0.815	Neu_vol	5.31E-68	0.313		
Ether	Neu_vol	6.58E-297	0.729	upvote ratio	2.28E-125	0.521	2.23E-107	0.487
	Likes	7.03E-277	0.712	number of comments	1.71E-37	0.296		
	Neg_vol	3.68E-274	0.709	compound	1.88E-28	0.257		
	Volume	9.75E-274	0.709	Pos_vol	5.62E-22	0.225		
	Pos_vol	7.51E-246	0.682	Volume	6.49E-18	0.202		
Binance	Likes	1.75E-301	0.733	compound	5.11E-144	0.553	0	0.867
	Retweets	2.57E-281	0.716	upvote ratio	1.58E-137	0.542		
	Neg_vol	2.64E-269	0.705	Pos_vol	5.16E-118	0.508		
	Volume	1.43E-266	0.702	Volume	1.47E-109	0.491		
	Pos_vol	2.17E-264	0.700	Neu_vol	2.00E-103	0.479		
Tether	Volume	3.10E-18	-0.204	*(No Reddit data for Tether)*			2.9E-07	-0.121
	Neu_vol	1.62E-17	-0.199					
	Pos_vol	8.47E-15	-0.182					
	Neg_vol	1.81E-13	-0.173					
	compound	3.20E-10	-0.148					
USD Coin	Pos_vol	7.43E-26	-0.270	*(No Reddit data for USD Coin)*			1.74E-36	-0.322
	Volume	9.66E-26	-0.270					
	compound	7.38E-24	-0.259					
	Neu_vol	2.35E-21	-0.245					
	Neg_vol	3.05E-12	-0.181					

Table 5 (The most correlated online variables to their respective cryptocurrency closing price)

Table 5 shows the five most correlated Twitter and Reddit variables and the Wikipedia trend. The correlation between online variables and the closing price was calculated similarly to the correlation between closing prices using Pearson's correlation coefficient. One can observe that none of the p-values are significant, i.e., all are below 0.05, which means that the null hypothesis of uncorrelated variables can be rejected. Tether and USD Coin do not have any Reddit variables since, as mentioned in the previous sections, their respective subreddits do not have enough engagement.

For the three cryptocurrencies that include Twitter and Reddit data, Bitcoin, Ether, and Binance, the Twitter data seems much more correlated with the closing prices than Reddit data. Interestingly, for neither of the three cryptocurrencies, does the negative sentiment have a negative correlation, but rather a significant positive correlation, both for Twitter and Reddit. For most cryptocurrencies, the online variables have significantly lower correlation coefficients than the ones between the cryptocurrencies' closing prices. Interestingly, the daily sentiments and volume are not the only online variables significantly correlated with the closing price.

For Bitcoin, the most significant Twitter variables are; the Twitter daily compound sentiment with a coefficient of 0.878, followed by the number of positive tweets with a coefficient of 0.861, and the Volume of all tweets with a coefficient of 0.837. The most significant Reddit variables are; the daily upvote ratio, with a coefficient of 0.722, the daily number of comments, with a coefficient of 0.427, and the daily compound sentiment for Reddit submissions, with a coefficient of 0.366. Lastly, the daily Wikipedia trend has a very low correlation of 0.106. The five Twitter variables have a very strong correlation with the price, whereas only one Reddit variable, the upvote ratio, correlates with similar strength. The correlation coefficient of 0.878 for the Twitter daily compound sentiment is slightly smaller than the one between Bitcoin's and Binance's closing prices of 0.909; however substantially lower than the correlation coefficient between Bitcoin's and Ether's closing prices of 0.959.

Ether's most significant Twitter variables are the daily number of neutral tweets with a coefficient of 0.729, the daily number of Twitter likes with a coefficient of 0.712, followed the daily number of negative Tweets and the daily Tweet volume, both with a coefficient of 0.709 and lastly, the number daily of positive Tweets with a coefficient of 0.682. Reddit variables for Ether are significantly less correlated with the closing price than the Twitter variables, with coefficients of 0.521, 0.296, and 0.257 for the daily upvote ratio, daily number of comments, and daily Reddit compound sentiment, respectively. In fact, in comparison with Bitcoin and Binance, the Reddit data for Ether is the least correlated with its closing price. Lastly, the correlation between Wikipedia's trend and Ether's closing price, with a coefficient of 0.487, is of more significance than in Bitcoin's case. Therefore, Ether's online variables correlate less with Ether's closing price than Bitcoin's and Binance's closing prices.

For Binance, the most significant Twitter variables are; the daily number of likes with a correlation coefficient of 0.733, the daily number of Retweets with a correlation coefficient of 0.716, and the daily number of negative Tweets with a correlation coefficient of 0.705. The most significant Reddit variables are the daily compound sentiment with a correlation coefficient of 0.553, the daily upvote ratio with a correlation coefficient of 0.542, and the daily number of positive Reddit submissions with a correlation coefficient of 0.508. The most significant variable overall is the Wikipedia trend with a correlation coefficient of 0.867, which is the only correlation coefficient comparable with the correlation coefficients between the closing price of Binance and Ether, 0.925, and the closing price of Binance and Bitcoin, 0.909.

Tether showcases a negative correlation between its online variables and its closing price for the Twitter data; the most significant is the daily volume of Tweets with a correlation of –0.204, followed by the daily number of neutral, positive, and negative Tweets with correlation coefficients of –0.199, -0.182 and –0.173 respectively. The correlation coefficient between the Wikipedia trend of Tether and its closing price is –0.121. Therefore, for Tether, none of its online variables have a correlation coefficient equal to or higher than the correlation coefficient between Tether's and USD Coin's price

USD Coin also shows a negative correlation between its online variables and its closing price, however a slightly more significant than Tether's. The most significant correlation coefficients for Twitter are; the daily number of positive Tweets (-0.270), the daily volume of Tweets (-0.270), and the daily Twitter compound sentiment (-0.259). The most significant correlation coefficient for the closing price of USD

Coin is for the Wikipedia trend (-0.322). Therefore, the Wikipedia trend, the daily number of positive Tweets, and the daily volume of Tweets correlate more with USD Coin's closing price than Tether's closing price.

6.2.2. Random Forest Feature Extraction

Random Forest regressor is a regression algorithm that combines the ensemble learning method, which is the process of using multiple models trained over the same data and averaging the results in hopes that the errors of each model are independent and the decision tree framework, which is a process of splitting the data based on variable's values. Hence, the Random Forest Regressor, in accordance with its name, creates a multitude of random decision trees, which aim at predicting the values of the target variables and average the results of those decision trees to yield a set of predictions (towardsdatascience.com).

The Random Forest Regressor from the nltk package in python can, however, do more than just predict future values; thanks to the *RandomForestRegressor.feature_importances_* attribute, it is possible to obtain information on how useful each variable in the regression process was. The method to obtain that information is the Gini importance measure, which measures how much impurity reduction (i.e., information gain) each variable provides on average among all the decision trees produced by the random forest. This method ensures fast computation even on large datasets; however, it does tend to give more importance to features heavily correlated with the target variable (mljar.com). The raw output of the function was on a scale from 0 to 1 in terms of importance so that all variables' importance's summed up to 1. Table 6 presents the three most important variables for each cryptocurrency in percentage terms.

The variables which will be considered are: 'Wiki_trend', 'Replies_tw', 'Retweets_tw', 'Likes_tw', 'Quote_tw', 'compound_tw', along with the volume of each sentiment and volume of all Tweets, as well as the daily percentage of each sentiment. Moreover, Bitcoin, Ether, and Binance will also include the variables from the Reddit dataset: 'n_com_red', 'score_red', 'awards_red', 'upvote_ratio_red', along with the volume of each sentiment and volume of all subreddit submissions, as well as the daily percentage of each sentiment. This totals 25 variables for Bitcoin, Ether, and Binance and 13 for Tether and USD Coin.

Random Forest feature extraction results

Cryptocurrency	Top 3 features	Importance
Bitcoin	Likes_tw	55,61%
	Replies_tw	26,43%
	Wiki_trend	7,6%
Ether	Volume_tw	39,14%
	Neu_vol_tw	34,45%
	Wiki_trend	7,8%
Binance	Volume_tw	44,21%
	Pos_vol_tw	23,06%
	compound_tw	11,34%
Tether	Wiki_trend	20,8%
	compound_tw	12,6%
	Likes_tw	9,52%
USD Coin	Wiki_trend	32,46%
	compound_tw	12,38%
	Likes_tw	11,75%

Table 6 (Random Forest feature extraction results)

The three most important features for Bitcoin are the daily number of Twitter likes (55,61%), Twitter Replies (26,43%), and the Wikipedia trend (7,6%). Both the daily number of Twitter likes and Replies have a significant correlation with the closing Bitcoin price (0.829 and 0.763); what is very surprising is that Bitcoin had the least significant correlation coefficient with Wikipedia trend as compared to other cryptocurrencies; despite that, Wikipedia trend had been very useful for the Random Forest Regressor.

For Ether, the three most important features chosen by the Random Forest Regressor are the daily Tweet Volume (39,14%), the daily number of neutral tweets (34,45%), and the Wikipedia trend (7,8%). Similarly, to Bitcoin, the Wikipedia trend stands out by having significant feature importance but low correlation (0.487).

Interestingly, the variable with the highest correlation coefficient for Binance, the Wikipedia trend, is not present among these most important variables.

For Tether, the three most important features chosen by the Random Forest Regressor were the Wikipedia trend (20,8%), the daily compound sentiment (12,6%), and the daily number of likes on Twitter (9,52%). Wikipedia trend is not significantly correlated with Tether's closing price; nonetheless, the Random Forest Regressor found it to be the most important feature for the prediction.

USD Coin's most important features are the Wikipedia trend (32,46%), the daily compound Twitter sentiment (12,38%), and the daily number of likes on Twitter (11.75%). Interestingly, for both stablecoins, the chosen features are the same and even in the same order, but with slightly different proportions.

Interestingly, the top 3 features combined are extremely important for Bitcoin, Ether, and Binance, having 89.64%, 81,39%, and 78,61% significance. Moreover, in all three cases, the third feature is significantly less important than the first two. However, this is not the case for stablecoins. For Tether, the sum of the three most important features is 42,92%, and for USD Coin, it is 56,59%. Moreover, the features seem to

be similarly important. All the results from the Random Forest feature extraction can be seen in Appendix 5.

6.2.3. Feature Selection

It can be observed that both the correlation table, as well as the Random Forest feature importance, show that the compound sentiment, the daily sentiment volumes, as well as the overall social media posts volume are important for all the cryptocurrencies and should therefore be included in all the models. The Random Forest feature importance also shows that the daily number of Likes, Retweets, and Replies on Twitter can be important for most cryptocurrencies.

Even though the Reddit variables seemed to have a less significant correlation and were not deemed the most important ones, they will also be included in the neural network models. The Wikipedia trend will also be included in neural network models. Lastly, the percentage of each sentiment for each social media will also be included in the models. The models for Bitcoin, Ether, and Binance have 25 variables, while the models for Tether and USD Coin have 13 variables to consider, as they do not include Reddit data.

6.2.4. Machine Learning models' results

6.2.4.1. Bitcoin

Table of results for the neural network models including additional variables for Bitcoin

Days	Model	Twitter's top 3 features			Reddit's top 3 features			Baselines & Wiki		
		variable	RMSE	MAPE	variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	pct_neu	1403.95	6.96	compound	188.5	0.93	Base	380.76	1.89
		Quote	1813.57	9.00	Pos_vol	219.3	1.09	ARIMA	78.08	0.39
		Neg_vol	2070.46	10.27	Volume	233.3	1.16	Wiki_trend	1593.13	7.90
	LSTM	pct_pos	650.51	3.23	pct_neu	294.4	1.46	Base	146.87	0.73
		Likes	909.76	4.51	Neu_vol	534.4	2.65	ARIMA	78.08	0.39
		Volume	1024.94	5.08	score	554.3	2.75	Wiki_trend	998.55	4.95
	Bi - LSTM	Pos_vol	79.08	0.39	Neg_vol	13.1	0.06	Base	92.74	0.46
		compound	80.57	0.40	pct_pos	36.9	0.18	ARIMA	78.08	0.39
		pct_pos	87.24	0.43	Neu_vol	208.4	1.03	Wiki_trend	53.89	0.27
	GRU	pct_neg	6.94	0.03	n_comments	134.7	0.67	Base	8.89	0.04
		Pos_vol	121.41	0.60	Pos_vol	377.9	1.87	ARIMA	78.08	0.39
		Volume	429.90	2.13	pct_neg	422.0	2.09	Wiki_trend	1221.11	6.06
2 days	RNN	pct_neu	1544.48	7.59	Pos_vol	224.7	1.05	Base	532.21	2.61
		Likes	1596.87	7.88	Neg_vol	368.1	1.67	ARIMA	677.95	3.01
		compound	2081.93	10.28	Volume	454.9	2.14	Wiki_trend	1891.89	9.32
	LSTM	Likes	284.78	1.31	Neu_vol	287.2	1.33	Base	106.69	0.47
		pct_neu	363.51	1.71	Neg_vol	444.2	2.15	ARIMA	677.95	3.01
		Retweets	670.10	3.24	Volume	594.8	2.90	Wiki_trend	1417.65	6.98
	Bi - LSTM	Quote	198.71	0.83	score	181.0	0.83	Base	615.17	2.97
		Pos_vol	204.88	0.97	awards	374.7	1.66	ARIMA	677.95	3.01
		compound	255.06	1.21	pct_neu	747.5	3.44	Wiki_trend	412.79	1.86
	GRU	pct_neu	338.91	1.42	Pos_vol	195.9	0.86	Base	281.16	1.00
		Retweets	515.79	2.52	n_comments	247.6	0.88	ARIMA	677.95	3.01
		pct_neg	581.13	2.71	awards	365.1	1.55	Wiki_trend	812.00	3.91
3 days	RNN	pct_neu	742.10	3.41	Pos_vol	262.5	1.17	Base	435.08	1.62
		Retweets	1046.15	4.07	compound	301.7	1.43	ARIMA	572.17	2.43
		compound	1284.40	5.87	Volume	348.6	1.44	Wiki_trend	707.94	3.18
	LSTM	Likes	860.75	3.99	awards	303.8	1.42	Base	415.24	1.50
		Quote	1092.49	5.22	score	361.6	1.70	ARIMA	572.17	2.43
		pct_pos	1361.96	6.61	Neg_vol	1114.7	5.32	Wiki_trend	1339.47	6.49
	Bi - LSTM	pct_neg	245.92	0.99	Neg_vol	428.4	1.65	Base	823.18	3.86
		pct_pos	307.97	1.30	pct_pos	738.2	3.21	ARIMA	572.17	2.43
		Neu_vol	432.77	1.89	pct_neg	740.4	3.49	Wiki_trend	777.13	3.61
	GRU	Neu_vol	340.55	1.37	Neg_vol	911.3	4.30	Base	285.52	1.15
		Neg_vol	357.23	1.60	compound	912.2	4.30	ARIMA	572.17	2.43
		Likes	474.76	1.89	Volume	1064.4	5.09	Wiki_trend	661.89	2.96

Table 7 (Table of results for the neural network models including additional variables for Bitcoin)

From Table 7, one can observe that Bitcoin's 1-day prediction interval the RNN had not been improved by any Twitter variables, as they had an RMSE of 1400 or above and a MAPE of 7 or above, in comparison to the baselines RMSE of 380 and MAPE of 1.89. The Reddit variables, however, did significantly improve the model; the most significant improvement was provided by the compound Reddit sentiment, which has an RMSE and MAPE twice as small as the baseline. None of the models, however, performed better than SARIMA (1,0,0)(2,0,0)[7]. In the LSTM model for a 1-day prediction interval, none of the additional data improved the model's accuracy.

For the Bi-LSTM model 1 day prediction interval, the three best models, including Twitter variables, performed similarly to each other, all slightly better than the baseline. Whereas the number of Negative Reddit submissions and the daily positive Reddit submission percentage, with the MAPE of 0.06 and 0.18, respectively, both significantly outperformed not only the Bi-LSTM baseline model as well as the SARIMA (1,0,0)(2,0,0)[7] model which had a MAPE of 0.39. The Wikipedia trend with a MAPE of 0.27 also outperformed both the Bi-LSTM baseline as well as the SARIMA (1,0,0)(2,0,0)[7] model.

The GRU produced the best 1-day prediction, with the model including the daily percentage of negative tweets with an RMSE of 6.94 and MAPE of 0.03; it was also the only model to outperform the GRU baseline model, which had an excellent RMSE and MAPE score of 8.89 and 0.04 respectively. Unlike the other three neural network algorithms, the Reddit and Twitter data performed very similarly.

For Bitcoin's 2-day prediction interval, the RNN had not been improved by any Twitter variables or the Wikipedia trend. However, the Reddit variables, especially the daily number of positive Reddit submissions, significantly improved the baseline RNN model yielding an RMSE of 244.7 and MAPE of 1.05, less than half of the RNN baseline and about a third of the ARIMA (3,0,2) model.

For LSTM, none of the variables outperformed the model's baseline RMSE of 106.69 and MAPE of 0.47. The three best-performing Twitter and Reddit variables performed similarly to each other, all outperforming the ARIMA (3,0,2) model.

Five Bi-LSTM models outperformed the baseline model and ARIMA(3,0,2) model, the daily number of Twitter Quotes, the number of positive Tweets, the compound Twitter sentiment, the Reddit score, the Reddit awards, and the Wikipedia trend. The overall best model was the daily Reddit score with an RMSE of 181 and MAPE of 0.83, compared to the Bi-LSTM baseline of RMSE of 615.17 and MAPE of 2.97.

For GRU, none of the Twitter variables improved the baseline model; however, the daily number of Reddit positive submissions and the number of Reddit comments outperformed both the GRU baseline and the ARIMA (3,0,2) model.

For Bitcoin's 3-day prediction, the RNN model had not been improved by any Twitter variable. In contrast, all three best-performing Reddit variables improved the RNN baseline and outperformed the ARIMA (3,0,2) model. The best-performing variable, the daily number of positive Reddit submissions, had an RMSE of 172.58, better than the baseline, and a MAPE of 0.45, better than the baseline.

The LSTM model only had one variable improving the baseline, the daily number of Reddit awards, which had an RMSE and MAPE of 303.8 and 1.42, respectively, compared to the baseline's RMSE of 415.24 and MAPE of 1.5. Apart from the daily number of Reddit awards, only the daily Reddit score outperformed the ARIMA (3,0,2) model.

The Bi-LSTM model was improved by the three best Twitter and Reddit variables and the Wikipedia trend. The best performance overall was the daily percentage of negative Twitter sentiment, which had an RMSE of 245.92 and MAPE of 0.99 compared to the baselines RMSE of 823.18 and MAPE of 3.86. Apart from the daily percentage of negative Twitter sentiment, three other variables outperformed the ARIMA (3,0,2) model, the daily percentage of positive Twitter sentiment, the daily number of neutral Tweets, and the daily number of Negative Reddit submissions.

For the GRU, none of the variables improved the baseline model; however, the three best-performing Twitter variables: the daily number of neutral Tweets, the daily number of negative Tweets, and the daily number of Twitter Likes outperformed the ARIMA (3,0,2) model.

6.2.4.2. Ether

Table of results for the neural network models including additional variables for Ether

Days	Model	Twitter's top 3 features			Reddit's top 3 features			Baselines & Wiki		
		variable	RMSE	MAPE	variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	Replies	49.83	3.68	n_comments	36.74	2.72	Base	22.15	1.64
		pct_neu	140.82	10.41	Volume	98.12	7.25	ARIMA	6.23	0.46
		compound	155.04	11.46	Neg_vol	102.34	7.57	Wiki_trend	251.51	18.59
	LSTM	Quote	3.09	0.23	upvote_ratio	3.05	0.23	Base	57.45	4.25
		compound	28.14	2.08	awards	3.87	0.29	ARIMA	6.23	0.46
		Replies	47.77	3.53	score	23.48	1.74	Wiki_trend	55.35	4.09
	Bi - LSTM	pct_neu	35.70	2.64	n_comments	3.39	0.25	Base	2.69	0.20
		pct_neg	53.90	3.98	Volume	5.45	0.40	ARIMA	6.23	0.46
		Quote	54.85	4.05	Neg_vol	9.92	0.73	Wiki_trend	17.58	1.30
	GRU	compound	4.05	0.30	pct_pos	2.12	0.16	Base	87.33	6.46
		Quote	14.30	1.06	upvote_ratio	2.63	0.19	ARIMA	6.23	0.46
		pct_neu	27.44	2.03	Neu_vol	9.36	0.69	Wiki_trend	7.69	0.57
2 days	RNN	Replies	55.52	4.07	n_comments	51.45	3.66	Base	25.86	1.89
		pct_neu	134.65	9.91	Volume	115.03	8.39	ARIMA	33.36	2.06
		compound	152.42	11.23	Neu_vol	125.93	9.16	Wiki_trend	256.00	18.85
	LSTM	Quote	3.56	0.26	upvote_ratio	2.67	0.195	Base	59.92	4.41
		compound	31.50	2.31	awards	3.18	0.23	ARIMA	33.36	2.06
		Replies	51.62	3.79	score	20.61	1.50	Wiki_trend	58.55	4.31
	Bi - LSTM	pct_neu	42.89	3.12	Neg_vol	8.12	0.58	Base	7.73	0.49
		pct_neg	44.78	3.21	n_comments	13.39	0.81	ARIMA	33.36	2.06
		Quote	63.18	4.62	Volume	15.44	0.98	Wiki_trend	26.48	1.86
	GRU	compound	23.11	1.34	pct_pos	7.28	0.45	Base	77.06	5.62
		Quote	26.24	1.79	upvote_ratio	12.05	0.72	ARIMA	33.36	2.06
		pct_neu	36.34	2.61	Neu_vol	19.24	1.28	Wiki_trend	18.88	1.22
3 days	RNN	pct_neu	35.54	1.74	n_comments	24.97	1.40	Base	31.83	1.93
		compound	135.00	9.25	awards	27.06	1.65	ARIMA	27.63	1.58
		Replies	190.23	13.86	Pos_vol	39.45	2.35	Wiki_trend	152.10	11.21
	LSTM	Quote	44.27	2.95	pct_neu	14.97	0.98	Base	49.69	3.33
		Replies	49.35	3.54	score	20.48	1.49	ARIMA	27.63	1.58
		pct_neu	69.63	4.94	n_comments	25.98	1.81	Wiki_trend	64.70	4.58
	Bi - LSTM	pct_neg	21.26	1.30	pct_neg	15.026	0.92	Base	32.33	2.02
		Quote	35.03	2.30	upvote_ratio	17.06	1.12	ARIMA	27.63	1.58
		Neg_vol	54.50	3.61	score	17.71	1.09	Wiki_trend	28.87	1.71
	GRU	compound	41.96	2.58	pct_neg	38.73	2.62	Base	71.65	5.19
		pct_neu	52.53	3.54	Neg_vol	45.29	3.12	ARIMA	27.63	1.58
		Volume	58.22	4.10	pct_pos	69.24	4.51	Wiki_trend	90.84	6.61

Table 8 (Table of results for the neural network models including additional variables for Ether)

The 1-day prediction interval for RNN and Bi-LSTM with the online variables, offer no improvement of the baseline model; on the contrary, the models' accuracy measures worsen considerably. However, both the three Reddit and three Twitter variables improve the baseline model; out of those six variables, only three perform better than SARIMA(3,0,0)(2,0,0)[7] model. Namely the daily number of Quotes on Twitter, the number of awards on Reddit, and the upvote ratio on Reddit. The best model's performance is the daily upvote ratio, which has a MAPE and RMSE of more than 18 times smaller than the baseline LSTM. Moreover, the model also considerably outperforms the SARIMA model since it has a MAPE of 0.23 and RMSE of 3.05 compared to SARIMA's 0.46 and 6.23.

The biggest improvement, however, must be attributed to the inclusion of the daily positive sentiment percentage of Reddit submissions for the GRU model. The RMSE improved from 87.33 to 2.12, and MAPE from 6.46 to 0.16. Moreover, the model, which included the daily positive percentage of Reddit submissions, had accuracy measures nearly three times smaller than SARIMA and was, overall, the best model for Ether's 1-day interval prediction. The other Reddit and Twitter variables and the Wikipedia trend also improved the baseline model.

Similarly to the 1-day prediction interval, in the 2-day prediction interval, both RNN and Bi-LSTM showed no improvement once the online variables were included in the models. Like the previous prediction interval, for 2-days, the Reddit upvote ratio yielded the best results, with an RMSE of 2.67 and MAPE of 0.195 compared to the baselines 59.92 and 4.41, while also performing considerably better than the ARIMA(2,0,2). The other Twitter and Reddit variables also offered considerable improvements to the baseline. The GRU results are consistent with the previous time horizon, with the daily positive percentage of Reddit submissions producing the best model. In this case, however, the three Reddit variables, two Twitter variables, and Wikipedia trend outperformed the ARIMA(2,0,2) model.

In the 3-day prediction interval, all the neural network models were improved by the online variables. RNN was improved the most by including the daily number of comments on Reddit. In contrast, the LSTM model, which included the daily neutral sentiment percentage of Reddit submissions, had an RMSE and MAPE more than three times smaller than its corresponding baseline. The Bi-LSTM model was most improved by including the daily negative sentiment percentage of Reddit submissions. In contrast, the GRU model was improved the most by including the daily Twitter compound sentiment.

6.2.4.3. Binance

Table of results for the neural network models including additional variables for Binance

Days	Model	Twitter's top 3 features			Reddit's top 3 features			Baselines & Wiki		
		variable	RMSE	MAPE	variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	Neu_vol	8.13	2.76	pct_neg	11.47	3.90	Base	16.30	5.54
		pct_neg	19.44	6.61	Neu_vol	24.71	8.40	ARIMA	3.34	1.14
		compound	24.95	8.48	n_comments	27.07	9.20	Wikipedia	44.81	15.23
	LSTM	pct_pos	4.94	1.68	pct_pos	0.24	0.08	Base	4.64	1.58
		pct_neg	6.34	2.16	score	6.09	2.07	ARIMA	3.34	1.14
		pct_neu	7.52	2.56	n_comments	6.87	2.34	Wikipedia	36.23	12.32
	Bi - LSTM	Retweets	6.00	2.04	pct_pos	4.41	1.50	Base	3.87	1.32
		Likes	7.32	2.49	awards	11.27	3.83	ARIMA	3.34	1.14
		pct_pos	8.35	2.84	score	12.16	4.13	Wikipedia	18.24	6.20
	GRU	pct_neg	10.95	3.72	pct_pos	14.03	4.77	Base	2.51	0.85
		Quote	13.99	4.76	awards	14.22	4.83	ARIMA	3.34	1.14
		pct_pos	16.48	5.60	score	16.50	5.61	Wikipedia	16.03	5.45
2 days	RNN	Neu_vol	6.96	2.32	Neu_vol	28.27	9.50	Base	18.94	6.36
		pct_neg	22.37	7.52	n_comments	29.83	10.06	ARIMA	12.8	3.74
		compound	32.41	10.73	pct_neg	33.09	9.60	Wikipedia	46.99	15.90
	LSTM	pct_pos	7.47	2.42	score	8.02	2.65	Base	3.38	0.99
		pct_neg	8.21	2.72	pct_pos	8.28	2.02	ARIMA	12.8	3.74
		pct_neu	8.97	3.00	n_comments	8.79	2.92	Wikipedia	37.43	12.67
	Bi - LSTM	Retweets	4.26	1.11	pct_pos	9.32	2.85	Base	6.50	2.07
		Likes	5.18	1.25	awards	13.66	4.56	ARIMA	12.8	3.74
		compound	11.01	2.95	score	14.53	4.86	Wikipedia	19.94	6.73
	GRU	pct_neg	14.12	4.68	pct_pos	12.63	4.25	Base	3.02	1.01
		Quote	15.35	5.18	awards	17.23	5.76	ARIMA	12.8	3.74
		pct_pos	19.15	6.43	score	19.27	6.47	Wikipedia	17.79	6.00
3 days	RNN	pct_neg	6.37	1.78	awards	8.45	2.66	Base	14.62	4.81
		pct_pos	9.31	2.96	score	8.71	2.73	ARIMA	10.47	2.62
		pct_neu	42.14	14.27	Pos_vol	9.74	2.78	Wikipedia	30.81	10.38
	LSTM	pct_pos	7.51	2.03	upvote_ratio	7.70	2.30	Base	6.24	1.80
		pct_neu	8.75	2.76	Neu_vol	8.23	2.55	ARIMA	10.47	2.62
		pct_neg	15.53	5.21	n_comments	10.19	3.26	Wikipedia	36.88	12.53
	Bi - LSTM	Replies	5.17	1.56	upvote_ratio	19.71	6.62	Base	3.52	1.01
		Volume	8.34	2.32	pct_pos	21.73	7.30	ARIMA	10.47	2.62
		Retweets	8.65	2.67	awards	21.85	7.37	Wikipedia	33.98	11.54
	GRU	Quote	9.08	2.86	pct_pos	7.90	2.24	Base	5.38	1.55
		Replies	15.43	5.06	compound	7.92	2.43	ARIMA	10.47	2.62
		pct_pos	22.18	7.49	score	8.11	2.49	Wikipedia	30.81	10.38

Table 9 (Table of results for the neural network models including additional variables for Binance)

For the 1-day prediction horizon for Binance, the RNN models worked worse than other neural network algorithms as none was close to the ARIMA(2,0,3). The most improved RNN model included the daily

volume of Tweets with neutral sentiment and had an RMSE of 8.13 and MAPE of 2.76 compared to the baselines of 16.3 and 5.54. The LSTM models were much closer to their LSTM baseline; however, only one model improved the model; interestingly, it was also the best model overall for Binance, it included the daily positive sentiment percentage of Reddit submissions and had an RMSE of only 0.24 and MAPE of 0.08, its baseline, and ARIMA models had an RMSE of 4.64 and 3.34 and MAPE of 1.58 and 1.14 respectively. This incredibly accurate prediction was an outlier as for both Bi-LSTM and GRU, no variable outperformed their respective Neural Network baselines or the ARIMA (2,0,3).

For the 2-day prediction interval, the RNN model had a disappointing performance, with most MAPE scores above 5. Only one variable, the daily number of Tweets with a neutral sentiment, improved the baseline. Moreover, the LSTMs, nor GRU's performance only worsened after including online variables. For Bi-LSTM, two models outperformed both the baseline of RMSE of 6.5 and MAPE of 2.07, as well as SARIMA(3,0,0)(2,0,0)[7] with an RMSE equal to 12.8 and MAPE of 3.74. The first model included the daily number of Retweets, and the second model included Twitter Likes; the former yielded the best results with an RMSE of 4.26 and MAPE of 1.11.

The 3-day prediction interval had the least number of variables that improved the baseline model, as neither LSTM, Bi-LSTM, nor GRU was improved by including the online variables. For RNN, the model, which included the daily percentage of negative Tweets, had the best performance and improved the baselines RMSE from 14.62 to 6.37 and MAPE from 4.81 to 1.78. The RNN model, which included the daily percentage of negative Tweets, also outperformed the SARIMA(3,0,0)(2,0,0)[7], which had an RMSE of 10.47 and MAPE of 2.62.

6.2.4.4. Stablecoins

The table representing the results for Tether can be seen in Appendix 1. The models yield exponentially small accuracy measures, which makes them challenging to compare. Generally, very few models, including online variables, outperformed the SARIMA/ARIMA models; interestingly, the Wikipedia trend consistently outperformed the corresponding baseline model across all neural network algorithms and prediction intervals. Twitter's negative sentiment bears the most importance for Tether, as both the daily percentage of negative Tweets and the daily volume of negative tweets consistently improve the baseline and are in the top 3 best models.

The table representing USD Coin can be seen in Appendix 2. USD Coin's predictions have the same issues as Tether's; the RMSE and MAPE are exponentially small and need to be represented with a scientific notation. For USD Coin, not a single neural network model outperformed the SARIMA models, even though the models including online variables consistently improved the corresponding baseline.

6.2.5. Covid-19 period Results

SARS-Cov-2 outbreak was declared a pandemic on the 11th of March 2020 (who.org); since then, the disease has claimed the lives of over 6 million people worldwide, causing countries to take unprecedented measures to contain the spread. Many countries have entered 'Covid lockdowns', meaning that citizens were heavily encouraged to stay at home to avoid spreading the virus, while restaurants, gyms, and many other workplaces were closed. The Covid-19 pandemic had many consequences on health regulations, the economy, and society. Although the pandemic is not over, its spread has slowed, and many countries reopened their economies due to increased vaccination efforts. This paper defines the Covid-19 period from the 11th of March 2020 to the 26th of January 2021. This section examines the results from the Machine Learning models run for the Covid-19 period.

6.2.5.1. Bitcoin

Table of results for the neural network models during the Covid-19 period for Bitcoin

Days	Model	Twitter's top 3 features			Reddit's top 3 features			Baselines & Wikipedia		
		variable	RMSE	MAPE	variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	Neu_vol	3307.4	10.87	Volume	22.3	0.07	Baseline	4719.5	15.51
		pct_neu	4061.0	13.34	Pos_vol	279.8	0.92	ARIMA	933.02	2.72
		Neg_vol	4327.9	14.22	n_com	546.0	1.79	Wikipedia	962.9	3.16
	LSTM	pct_pos	2536.2	8.33	pct_neu	4095.8	13.46	Baseline	15051.2	49.46
		pct_neg	3072.8	10.10	pct_neg	5196.6	17.08	ARIMA	933.02	2.72
		Neg_vol	16729.2	54.97	pct_pos	5306.0	17.44	Wikipedia	17484.8	57.45
	Bi - LSTM	pct_pos	1659.0	5.45	pct_neu	527.2	1.73	Baseline	13398.9	44.03
		pct_neg	1777.0	5.84	pct_pos	3222.2	10.59	ARIMA	933.02	2.72
		Likes	14407.0	47.34	awards	13628.8	44.78	Wikipedia	13704.0	45.0
	GRU	compound	1527.8	5.02	pct_neg	343.4	1.13	Baseline	3733.2	12.27
		pct_pos	2495.9	8.20	awards	1628.8	5.35	ARIMA	933.02	2.72
		Likes	15776.4	51.84	pct_pos	2515.9	8.27	Wikipedia	2936.8	9.65
2 days	RNN	Neu_vol	5873.1	16.82	upvote ratio	826.2	2.59	Baseline	6284.1	19
		Replies	6016.0	18.38	n_com	1137.0	3.16	ARIMA	577.6	1.34
		pct_pos	6033.3	18.19	Volume	1374.4	2.94	Wikipedia	2264.7	6.15
	LSTM	pct_pos	1938.1	5.72	pct_neu	3020.2	8.54	Baseline	16689.6	51.89
		pct_neg	4414.3	13.17	pct_pos	3802.4	10.02	ARIMA	577.6	1.34
		Neg_vol	18215.1	56.75	pct_neg	4014.8	11.96	Wikipedia	19029.3	59.29
	Bi - LSTM	pct_pos	1791.3	5.59	pct_neu	2194.5	5.44	Baseline	15072.9	46.78
		pct_neg	1793.0	5.62	pct_pos	2295.8	5.89	ARIMA	577.6	1.34
		Likes	16072.8	49.94	awards	15282.8	47.45	Wikipedia	15375.7	47.74
	GRU	compound	1734.6	5.38	pct_neg	2294.4	5.38	Baseline	5668.6	16.73
		pct_pos	4402.8	12.62	awards	3743.4	10.20	ARIMA	577.6	1.34
		Likes	17411.3	54.17	pct_pos	4438.7	12.73	Wikipedia	4945.0	14.31
3 days	RNN	Likes	4451.1	12.24	upvote ratio	1530.8	4.04	Baseline	7500.6	21.87
		Neu_vol	4491.4	11.49	Neu_vol	2076.6	5.74	ARIMA	1809.72	5.18
		Retweets	5587.6	15.34	Volume	2517.0	5.75	Wikipedia	2988.6	8.05
	LSTM	pct_pos	2033.0	5.96	pct_neu	2473.1	6.01	Baseline	17498.3	53.06
		pct_neg	5235.5	15.17	pct_pos	3107.6	6.91	ARIMA	1809.72	5.18
		Likes	18567.9	56.36	pct_neg	3483.4	9.95	Wikipedia	19824.0	60.24
	Bi - LSTM	pct_neg	2060.9	6.19	pct_pos	1929.0	4.69	Baseline	15870.0	48.05
		pct_pos	2110.2	6.28	pct_neu	2715.6	7.06	ARIMA	1809.72	5.18
		Likes	16465.7	49.92	awards	16051.9	48.62	Wikipedia	16152.5	48.93
	GRU	pct_pos	5326.2	15.03	pct_neg	3023.7	7.58	Baseline	17498.3	53.06
		Likes	17898.4	54.30	awards	4510.1	12.38	ARIMA	1809.72	5.18
		Retweets	18305.7	55.54	pct_pos	5198.5	14.75	Wikipedia	5720.9	16.36

Table 10 (Table of results for the neural network models during the Covid-19 period for Bitcoin)

For RNN, the top 3 best models from Twitter and Reddit, and Wikipedia consistently improved the corresponding baseline models across all the prediction intervals. The Reddit models generally had much better accuracy than Twitter or Wikipedia trend models. The best Twitter models differ among the three prediction intervals; however, in each one, the daily volume of neutral sentiment Tweets was among the best three models for Twitter. Similarly for Reddit, although the best prediction models differed from one prediction interval to another, however, the daily volume of Reddit submissions was among the best in all three prediction intervals; it also yielded the best model overall with an RMSE of only 22.3 and MAPE of 0.07 as compared to its baseline of RMSE of 4719.5 and MAPE of 15.5.

The results of the LSTM and Bi-LSTM models are very similar and highly consistent among the time prediction intervals. For both LSTM and Bi-LSTM, across the three time prediction intervals, the best model is yielded by including the daily percentage of tweets with positive sentiment; for LSTM during the 1-day prediction interval, it improves the baseline accuracy almost six times, for 2-day prediction interval and 3-day prediction interval nine times. In contrast, for Bi-LSTM, the accuracy measures are consistently about eight times better. Moreover, for both LSTM and Bi-LSTM, the second-best twitter variable is the daily percentage of tweets with negative sentiment, while the third-best model is either the daily compound Twitter sentiment or the number of Likes; however, it does to improve the baseline in any case. The best Reddit models for both LSTM and Bi-LSTM, across the three prediction intervals, are the daily percentage of Reddit submissions with neutral, negative, and positive sentiment, and all three improve the baseline. Lastly, the Wikipedia trend does not improve the baseline model in any prediction intervals for either LSTM or Bi-LSTM.

For GRU, the results are very consistent for Reddit variables and Wikipedia trend; there is less consistency among Twitter variables. For Reddit, the best model across all prediction intervals is the daily percentage of negative sentiment Reddit submissions, the second-best is the daily number of awards on Twitter, and the third-best is the daily percentage of positive Reddit submissions; all the Reddit models significantly improve the baseline model. The Wikipedia trend also slightly outperforms the baseline model in all prediction interval horizons. The Twitter results for GRU are less consistent among the three prediction intervals; the variable which always outperforms the baseline model is the daily percentage of positive sentiment Tweets; for 1-day and 2-day prediction intervals, the best model from Twitter is the compound sentiment, although it is not present among the best three Twitter models in the 3-day prediction interval.

The time series models were observed to outperform the machine learning models significantly. Only a few models with additional variables outperformed the time series baseline. There was, however, no pattern as to which variable or machine learning algorithm outperformed the corresponding time series model. The results for the ARIMA and SARIMA orders for the Covid-19 period for all cryptocurrencies can be seen in Appendix 6.

6.2.5.2. Ether

Table of results for the neural network models during the Covid-19 period for Ether

Days	Model	Twitter's top 3 features			Reddit's top 3 features			Baselines & Wikipedia		
		variable	RMSE	MAPE	variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	Neg_vol	150.2	11.99	pct_neu	118.2	9.43	Baseline	458.2	36.57
		Neu_vol	183.8	14.67	upvote_ratio	168.2	13.42	ARIMA	71.9	5.23
		Volume	197.3	15.75	pct_pos	183.8	14.66	Wikipedia	188.1	15.01
	LSTM	Likes	27.9	2.23	score	21.3	1.70	Baseline	139.1	11.10
		Retweets	40.6	3.24	upvote_ratio	26.5	2.11	ARIMA	71.9	5.23
		pct_pos	60.3	4.81	pct_neg	27.1	2.16	Wikipedia	90.8	7.25
	Bi - LSTM	Neg_vol	42.5	3.39	Pos_vol	7.1	0.57	Baseline	234.3	18.69
		Pos_vol	69.2	5.52	Volume	10.3	0.82	ARIMA	71.9	5.23
		compound	85.0	6.78	compound	12.1	0.97	Wikipedia	66.5	5.3
	GRU	compound	258.8	20.65	pct_neu	118.2	9.43	Baseline	306.6	24.46
		pct_neu	271.4	21.66	upvote_ratio	168.2	13.42	ARIMA	71.9	5.23
		Quote	274.3	21.89	pct_pos	183.8	14.66	Wikipedia	128.0	10.21
2 days	RNN	Neg_vol	198.5	14.90	score	303.0	22.98	Baseline	468.8	36
		Volume	263.6	19.74	n_com	315.1	24.15	ARIMA	68.96	4.14
		Neu_vol	275.9	20.25	Neg_vol	338.7	26.06	Wikipedia	214.4	16.43
	LSTM	pct_pos	42.7	2.50	upvote_ratio	22.2	1.69	Baseline	170.0	12.91
		Likes	64.1	4.35	score	25.8	1.96	ARIMA	68.96	4.14
		pct_neg	75.8	5.60	awards	35.8	2.07	Wikipedia	122.0	9.13
	Bi - LSTM	Neg_vol	90.6	6.23	Neg_vol	52.5	3.89	Baseline	275.7	21.04
		Pos_vol	117.0	8.40	Pos_vol	63.0	3.62	ARIMA	68.96	4.14
		compound	135.1	9.81	Volume	63.2	3.74	Wikipedia	112.8	8.09
	GRU	compound	302.9	23.13	pct_neu	165.5	12.29	Baseline	350.1	26.82
		pct_neu	310.1	23.76	upvote_ratio	222.2	16.67	ARIMA	68.96	4.14
		Quote	317.6	24.29	pct_pos	225.5	17.11	Wikipedia	173.4	12.95
3 days	RNN	Neg_vol	249.8	17.87	score	333.5	24.66	Baseline	486.4	36.71
		Volume	278.9	20.57	upvote_ratio	351.8	26.45	ARIMA	48.6	3.41
		Neu_vol	288.7	21.04	n_com	357.6	26.47	Wikipedia	242.2	17.94
	LSTM	pct_pos	37.4	2.23	upvote_ratio	24.2	1.76	Baseline	188.4	13.93
		pct_neg	64.4	4.48	awards	30.3	1.72	ARIMA	48.6	3.41
		Likes	82.6	5.57	score	34.9	2.47	Wikipedia	141.9	10.31
	Bi - LSTM	Neg_vol	117.6	7.97	Neg_vol	71.5	4.98	Baseline	303.1	22.50
		Pos_vol	144.9	10.15	Neu_vol	74.6	5.01	ARIMA	48.6	3.41
		compound	164.5	11.64	Volume	86.8	5.41	Wikipedia	138.1	9.69
	GRU	compound	333.5	24.77	pct_neu	195.1	14.07	Baseline	380.8	28.39
		pct_neu	339.5	25.28	pct_pos	254.8	18.76	ARIMA	48.6	3.41
		Quote	347.9	25.88	Neg_vol	289.2	21.28	Wikipedia	203.1	14.72

Table 11 (Table of results for the neural network models during the Covid-19 period for Ether)

For Ether's RNN 1-day prediction interval, the Reddit data slightly outperformed the Twitter data; however, for the remaining two prediction intervals, Twitter data was superior. Moreover, the three best models from Twitter and the three best models for Reddit and the Wikipedia trend improved their corresponding baseline in each prediction interval. Moreover, the results for models for Twitter variables were highly consistent for all three prediction intervals; the best model was the daily volume of negative tweets, the second and

third best were the volume of all tweets, and the volume of tweets with a neutral sentiment. In comparison to Twitter, the results for Reddit were less consistent, with the daily percentage of Reddit submissions with neutral sentiment having the best accuracy measures for the 1-day prediction interval but not being among the three best prediction intervals for the remaining two prediction intervals. No Reddit variable was among the best three Reddit models for all three prediction intervals; however, the Reddit score, upvote ratio, and the number of Reddit comments were twice among the best three models. The Wikipedia trend was consistently improving the baseline performing slightly worse than the Twitter variables but better than the Reddit variables.

LSTM's results were far more consistent among the three prediction intervals; the models with Reddit variables had better accuracy measures than models with Twitter variables. The Wikipedia trend had the worse accuracy measures but still outperformed all the corresponding baseline models. The order among the best Twitter models varies depending on the prediction interval; however, the best three Twitter models for all prediction intervals consist of the daily number of Twitter Likes and the daily percentage of tweets with positive and negative sentiment. Similarly, for Reddit, the best models for Reddit were obtained by the Reddit score, upvote ratio, and the daily number of Reddit awards. The model with the best scores overall was the daily Reddit upvote ratio for a 2-day prediction interval, with accuracy measures more than seven times lower than the corresponding baseline.

For Bi-LSTM, for all the prediction intervals the Reddit results were consistently better than the Twitter results and Wikipedia trend, although all predictions had consistently worse accuracy measures with each consecutive prediction interval. Twitter had the same three variables yielding the best three models for all three prediction intervals; the daily number of positive and negative sentiment tweets and the daily compound Twitter sentiment. The results for Reddit were less consistent in terms of variables yielding the best results; only the daily volume of Reddit submissions was among the best three results in each time horizon; not once, however, had it produced the best-performing model. The other variables consisted of the volume of positive, neutral, or negative sentiment Tweets and the compound Twitter sentiment.

GRU's predictions for Ether during the Covid-19 period were highly consistent among the three prediction intervals, with the Reddit data and Wikipedia trend performing better than Twitter data. For Twitter for all prediction intervals, the best model was yielded by including the compound Twitter sentiment, the second-best model was obtained by including the daily percentage of Tweets with the neutral sentiment, and the third best with the daily number of Twitter quotes. The Reddit models were slightly less consistent; however, the daily percentage of Tweets with neutral sentiment was always the best-performing model, and the daily percentage of Tweets with positive sentiment was present in the best three models for all three prediction intervals.

The time series models have again significantly outperformed the baseline machine learning models; however, the LSTM and Bi-LSTM models, with additional data, have outperformed the time series models across all prediction intervals.

6.2.5.3. Binance

Table of results for the neural network models during the Covid-19 period for Binance

Days	Model	Twitter's top 3 features			Reddit's top 3 features			Baselines & Wikipedia		
		variable	RMSE	MAPE	variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	pct_pos	4.1	10.05	pct_neg	5.4	13.21	Baseline	7.5	18.37
		Neg_vol	5.4	13.21	pct_pos	6.2	15.05	ARIMA	0.67	1.51
		Quote	6.1	14.93	awards	7.3	17.78	Wikipedia	19.9	48.66
	LSTM	pct_pos	5.0	12.29	pct_pos	3.6	8.85	Baseline	12.0	29.18
		pct_neu	10.2	24.86	pct_neu	12.1	29.60	ARIMA	0.67	1.51
		pct_neg	12.2	29.73	pct_neg	12.4	30.19	Wikipedia	14.5	35.48
	Bi - LSTM	pct_neu	3.9	9.56	upvote_ratio	10.7	26.07	Baseline	11.4	27.77
		Neu_vol	11.1	27.15	pct_neu	11.0	26.74	ARIMA	0.67	1.51
		Replies	11.4	27.72	Pos_vol	11.0	26.78	Wikipedia	10.5	25.7
	GRU	Likes	12.2	29.88	Pos_vol	13.2	32.09	Baseline	13.4	32.68
		Replies	12.5	30.39	awards	13.6	33.11	ARIMA	0.67	1.51
		pct_neu	13.5	32.88	compound	13.7	33.49	Wikipedia	13.7	33.50
2 days	RNN	Neg_vol	4.5	10.64	pct_neg	5.9	14.08	Baseline	8.2	20
		pct_pos	4.6	10.94	pct_pos	7.3	17.32	ARIMA	0.55	1.21
		compound	7.2	17.16	awards	8.0	18.99	Wikipedia	20.4	48.74
	LSTM	pct_pos	4.6	10.92	pct_pos	2.9	6.68	Baseline	12.7	30.39
		pct_neu	11.1	26.39	pct_neu	12.9	30.86	ARIMA	0.55	1.21
		pct_neg	12.9	30.86	pct_neg	13.1	31.22	Wikipedia	15.3	36.48
	Bi - LSTM	pct_neu	4.9	11.47	upvote_ratio	11.6	27.61	Baseline	12.2	29.06
		Neu_vol	11.9	28.49	pct_neu	11.7	28.03	ARIMA	0.55	1.21
		Retweets	12.0	28.65	Pos_vol	11.9	28.29	Wikipedia	11.3	27.07
	GRU	Likes	13.2	31.40	Pos_vol	13.8	33.06	Baseline	14.2	33.92
		Replies	13.3	31.75	awards	14.4	34.34	ARIMA	0.55	1.21
		pct_neu	14.3	34.17	compound	14.4	34.41	Wikipedia	14.5	34.68
3 days	RNN	Neg_vol	5.1	11.76	pct_neg	6.4	14.96	Baseline	8.6	20.21
		pct_pos	5.1	11.93	pct_pos	7.7	18.00	ARIMA	1.09	2.51
		compound	7.3	17.26	awards	8.3	19.63	Wikipedia	14.3	33.76
	LSTM	pct_pos	4.5	10.62	pct_pos	2.8	6.56	Baseline	13.0	30.86
		pct_neu	11.5	27.10	pct_neg	13.4	31.65	ARIMA	1.09	2.51
		pct_neg	13.2	31.25	pct_neu	13.4	31.61	Wikipedia	14.0	33.24
	Bi - LSTM	pct_neu	5.2	12.20	pct_neu	12.0	28.52	Baseline	12.5	29.55
		Retweets	12.2	28.90	compound	12.4	29.30	ARIMA	1.09	2.51
		Replies	12.3	29.15	upvote_ratio	12.4	29.36	Wikipedia	12.4	29.34
	GRU	Likes	13.6	32.19	Pos_vol	13.8	32.58	Baseline	14.5	34.42
		Replies	13.7	32.45	compound	14.6	34.67	ARIMA	1.09	2.51
		pct_neu	14.7	34.74	Neu_vol	14.7	34.74	Wikipedia	14.0	33.10

Table 12 (Table of results for the neural network models during the Covid-19 period for Binance)

For RNN, in all prediction intervals, the Twitter models were the best-performing ones; Reddit models were worse; however still outperformed the corresponding baseline models, and the Wikipedia trend models, however, performed significantly worse than the corresponding baseline in each prediction interval. The same variables for both social media yielded the best results. The three best models for Twitter were obtained by adding the daily percentage of Tweets with positive sentiment, the daily volume of negative

Tweets, and the daily compound Twitter sentiment. The three best Reddit models were obtained by adding the daily percentage of Reddit submissions with positive sentiment, the daily number of Reddit awards, and the daily percentage of Reddit submissions with positive sentiment.

For LSTM, significantly fewer models were able to improve their corresponding baseline. The best model in each prediction interval included the daily percentage of Reddit submissions with a positive sentiment; however, it was also the only Reddit variable that improved the corresponding baseline in all prediction intervals. The best-performing Twitter model in each prediction interval included the daily percentage of Tweets with a positive sentiment. The Wikipedia trend in each prediction interval performed worse than its corresponding baseline model.

For Bi-LSTM, in each prediction interval, there was only one model which significantly improved the baseline was the daily percentage of Tweets with the neutral sentiment. The remaining models, including other variables, were performing very close to the baseline; what is interesting to note is that the daily percentage of Reddit submissions with neutral sentiment had also been among the three best Reddit models; however, its results were only slightly more accurate.

For GRU in each prediction interval, the results were very similar, with all variables performing slightly better or worse than the corresponding baseline model. The best model in each prediction interval was the daily number of Twitter Likes; even that model only improved the accuracy measures very slightly compared to the other neural network algorithms. The second-best model was the daily number of Twitter Replies, and the third-best model, which managed to outperform the baseline, was the daily volume of Reddit submissions with positive sentiment. Wikipedia trend outperformed the baseline only for the 3-day prediction interval.

The time series models have performed significantly better than both machine learning models with and without additional data.

6.2.5.4. Stablecoins

Similarly to the previous results for the October 2022 prediction, the prediction for the Covid-19 period the results for the stablecoins yield exponentially small accuracy measures, which makes them challenging to compare. The table of results for Tether can be seen in Appendix 3, for USD Coin in Appendix 4, the results will be briefly presented below.

For Tether, the results for each neural network algorithm were consistent across the prediction intervals. The best-performing neural network algorithm was GRU, for which the models which included positive and compound sentiments from Twitter produced the best forecasts. For RNN, the models which included the negative sentiment and the number of quotes and replies performed the best. For LSTM, the models that included the negative and neutral sentiments had the best accuracy measures. The Bi-LSTM was the only neural network algorithm for which the Wikipedia trend improved the baseline model; apart from that, the best-performing models included the volume of all Tweets and either positive, negative, or neutral sentiments. The time series models have performed very poorly for Tether, having accuracy measures 100 times larger than their machine learning counterparts.

For USD Coin, the RNN had by far the worst performance as the 1-day prediction interval baseline was the only one being improved; for the rest of the neural network algorithms, all three of the best models from Twitter outperformed the corresponding baseline. The LSTM baseline was most improved by the daily percentage of Tweets with a neutral sentiment, followed by the models with the daily percentage of Tweets with a positive sentiment, and lastly by the daily number of Twitter Replies. The Bi-LSTM baseline was outperformed by models including the daily volume of Tweets and daily percentage of Tweets with positive sentiment; moreover, the Wikipedia trend also improved the models for all three prediction intervals. For

GRU, the models which included the positive or compound sentiment yielded the best results. The time series models have performed incredibly well for the 1- and 2-day prediction interval, outperforming all other models significantly. However, for the 3-day time horizon, their results have been very poor.

6.3. Value at Risk analysis

In order to showcase just how risky an investment in cryptocurrencies is, the Value at Risk of each cryptocurrency was calculated.

Value at Risk (VaR) is a statistical technique used to measure potential losses to a given stock or portfolio over a given period (corporatefinanceinstitute.com). Although there are three major methods to calculate VaR, this paper will use the historical method, for which the VaR value is calculated by creating a histogram of historical returns and choosing the confidence interval from there. For this paper, the 1-day VaR was calculated with a Python script.

Value at Risk for the five cryptocurrencies

Confidence level	Bitcoin	Ether	Binance	Tether	USD Coin
90%	-3.83%	-5.36%	-5.52%	-0.34%	-0.34%
95%	-6.01%	-7.69%	-7.67%	-0.55%	-0.52%
99%	-10.56%	-13.63%	-13.68%	-1.32%	-1.01%

Table 13 (VaR values for the five cryptocurrencies)

The results show that an investor in Bitcoin has a 90% confidence level that their losses will not exceed 3.83%, 95% confidence that their losses will not exceed 6.01%, and 99% confidence that their losses will not exceed 10.56% for a 1-day investment in Bitcoin. An Ether investor, for a 1-day investment in Ether, has a 90% confidence level that their losses will not exceed 5.36%, a 95% confidence level that their losses will not exceed 7.69% and a 99% confidence level that their losses will not exceed 13.63%. A Binance investor has a 90% confidence level that their losses will not exceed 5.52%, a 95% confidence level that their losses will not exceed 7.67% and a 99% confidence level that their losses will not exceed 13.68% for a 1-day investment in Binance.

One's risk is significantly reduced if one chooses to invest in either of the two stablecoins. Tether shows a 90% confidence level of losses not exceeding 0.34%, a 95% confidence level of losses not exceeding 0.55%, and a 99% confidence level of losses not exceeding 1.32% for a 1-day investment. USD Coins show the best results as an investment in that cryptocurrency shows a 90% confidence level of losses not exceeding 0.34%, a 95% confidence level of losses not exceeding 0.52%, and a 99% confidence level of losses not exceeding 1.01%,

7. Discussion

In order to address the research questions, this study examined the influence of Twitter, Reddit, and Wikipedia data on the prices of the five biggest cryptocurrencies, as well as how important each social media and the sentiment within each social media is. Additionally, the Covid-19 period was examined to observe how the relationship changed. Lastly, the Value at Risk of each cryptocurrency was calculated to give a perspective on how risky investing in cryptocurrencies is.

7.1. Feature selection

The results for the sentiment analysis were consistent among the five cryptocurrencies and social media. As the literature presented in the earlier section suggests, positive sentiment is the most common, both on Twitter and Reddit for all the five cryptocurrencies. Interestingly the second most common is the neutral sentiment, this may be due to the NLP preprocessing strategy chosen by the author, or bots and advertisement that the author has failed to remove. The negative sentiment is the least common one. The proportions vary little despite the price fluctuations, which shows that the sentiment remains mostly positive, despite the price movements, which is consistent with the findings from the literature presented in the earlier section.

The correlation table shows that the volume of each sentiment is far more correlated to the closing price than the relative sentiment; however, the positive sentiment does not always have the most significant correlation coefficient compared to the other sentiments. Thus, negative and neutral sentiments are also important despite being less common. Moreover, the findings also show that many non-sentiment variables are heavily correlated with the closing price. The variables in question are variables, such as the daily number of likes, retweets, the daily number of all Tweets, the upvote ratio, and the number of comments on Reddit. The Random Forest feature extraction shows comparable results suggesting that these variables and Wikipedia trend bear more importance in price prediction than the sentiments. In the author's view, the variables in question measure user engagement on social media; hence, they will be referred to as *engagement measures*. The importance of the engagement measures explains why the volume of a sentiment performs better than the relative sentiment, as the volume of a sentiment contains both the sentiment and the volume, one of the engagement measures.

7.2. Bitcoin

October 2022 and Covid-19 period prediction results for Bitcoin indicate that the Reddit variables performed better on average than Twitter variables. However, a few models which included Twitter variables produced the best results overall. For the October 2022 prediction, despite many models outperforming the baseline model, the results show no patterns. Even within the same neural network algorithm, the variables which produced the best models were inconsistent among the prediction intervals. The Covid-19 pandemic period prediction results had worse accuracy metrics than the results for the October 2022 prediction, which is to be expected considering the price fluctuations were more significant. Similarly, the results did not produce a clear pattern that would indicate the best overall sentiment or engagement metric from either social media to be the best predictor of the closing price. However, the results were highly consistent within each Neural Network algorithm, meaning that for each neural network algorithm, the same variables produced the best models for each prediction interval.

Therefore, the results suggest that while the online data, in general, does have a relationship with the closing price of Bitcoin, as some of the variables are successful with their predictions, however, it is impossible to point out which variables would be the most helpful with predicting the future price. The main finding is that the 'r/BTC' subreddit is more reliable in producing forecasts which outperform the baseline model.

Furthermore, the finding suggests that Reddit's subreddit forum structure with active content moderators is more efficient at gathering users who are more likely to invest and act on the advice or opinion posted on the forum, as compared to Twitter, where anyone can post their opinion and have their content become a trending topic. Hence, the subreddit data should not be discarded, as despite having less significant correlation coefficients and not being picked by the Random Forest feature extraction can produce excellent predictions.

Moreover, the Time Series models have produced very encouraging results. However, they have been mostly outperformed by the machine learning models with additional data for the October 2022 prediction; they have been much more successful for the Covid-19 period prediction. The Covid-19 Time Series results may be due to the prediction intervals being very short, and as Wirawan et al. (2019) found, ARIMA models perform very well for short-term cryptocurrency price prediction. Nonetheless, this indicated that the Covid-19 period may have had some seasonal effects, possibly correlating with important events related to the pandemic.

7.3. Ether

Ether's results for the October 2022 prediction were similar to those of Bitcoin, with the Reddit variables outperforming the Twitter variables on average, while the Twitter variables had few models with the best accuracy measures overall. In October 2022 prediction for Ether, the sentiment variables from both Twitter and Reddit produced few models which outperformed the baseline and the engagement metrics. The number of Replies on Twitter and comments on Reddit produced more accurate models than any of the sentiments. Interestingly, for the Covid-19 period prediction, all the three best models from each social media and the Wikipedia trend have improved their corresponding baseline model, sometimes improving the accuracy more than tenfold. The results are consistent within each neural network algorithm; for each neural network algorithm, the same variables produced the best models for each prediction interval. Similarly to the October 2022 prediction for the Covid-19 period prediction, the engagement metrics have also produced most of the best models. The machine learning models consistently outperformed the time series models for the October 2022 and Covid-19 predictions. The aspect of engagement metrics as a predictor of the cryptocurrency price is often overlooked in the literature. However, as the results show, the engagement metrics may hold valuable data, which explains the behavior of the closing price. The results suggest that for Ether, the engagement metrics, rather than the sentiment of social media posts about the cryptocurrency influences the price the most.

7.4. Binance

Interestingly, Binance's results for the October 2022 predictions show that the relationship between the price and the social media data is not as strong as the correlation coefficients and Random Forest suggested. Only a handful of models, which included the online variables, outperformed the corresponding baseline. From the models that outperformed their corresponding baseline, Twitter data performed better than Reddit; however, the results show no pattern regarding which variable or sentiment would be the best predictor. For Covid-19 period predictions, many more models with online variables outperformed their corresponding baseline. Despite inconsistent results, both Twitter and Reddit had a similar performance; for Reddit, the positive sentiment had been the best predictor, whereas, for Twitter, the engagement metrics were the best predictors. However, for the Covid-19 predictions, the time series models performed significantly better than any models with online data. The results from the predictions contradict to some degree the results from the correlation and Random Forest, which indicated that there is a strong relationship between the online variables and the closing price of Binance. A possible explanation for this is that Binance is mainly known as a cryptocurrency exchange platform, and most online mentions are about the platform rather than the currency itself.

7.5. USD Coin

USD Coin results for the October 2022 prediction show that the engagement metrics from Twitter produced the best models; the Wikipedia trend outperformed the corresponding baseline model more often than for the other coins. The results for the Covid-19 period show that fewer online variables outperformed the baseline model, and both positive sentiment and engagement metrics from Twitter produced the best predictions. The best predictions overall were produced by the models, which included the volume of tweets. However, the results are inconsistent, as different variables yield the best results within each prediction interval and each Neural Network algorithm. Moreover, as mentioned in the earlier section, the exponentially small results are due to the tiny price fluctuations of USD Coin. Therefore, the results show that USD Coin is very successful at maintaining its value close to $1. However, the fact that the online factors influence the price of a stablecoin is interesting, as USD Coin's goal is to retain a value equal to $1 and should therefore be independent of any other factors.

7.6. Tether

For the October 2022 predictions, Tether is the cryptocurrency that had the most models outperforming the corresponding baseline; moreover, the results show a clear trend. The best results were achieved by the models, which included the daily negative percentage of the tweets. The results were also consistent across the different neural network algorithms and prediction intervals. The daily negative percentage of Tether's tweets is one of the least correlated variables with Tether's closing price, and the Random Forest feature extraction also did not assign much importance to the daily negative percentage of Tweets. The negative sentiment may be this significant because cryptocurrencies are now in the bear market, mostly losing value. Tether, as a stablecoin, is subject to less speculation than Bitcoin, Ether, or Binance, which makes the fact that the social media sentiment is this successful in predicting the price especially intriguing. This may also be the sign that some investors, fearing the bearish market, convert their Tether holdings into US Dollars. The results for the Covid-19 period for Tether were similar to the ones from the October 2022 predictions' results. However, they had worse accuracy metrics, and the Wikipedia trend did not outperform the corresponding baseline. Similarly to the October 2022 prediction results, the Covid-19 period results show the importance of the negative sentiment but not to the same extent; moreover, the engagement metrics also seem to be excellent predictors.

7.7. Value at Risk

The VaR results show that Bitcoin, Ether, and Binance are very volatile assets with a risk of substantial losses even for short-term investments. While Bitcoin is the least risky of the three, it may be because it is the most established cryptocurrency, which has been present for over a decade. Binance and Ether have very similar values, where, even at a 90% confidence level, one risks losing over 5%, and almost 14% at the 99% confidence level.

Interestingly, the VaR results suggest that Tether and USD Coin are far less risky than the other three cryptocurrencies; Bitcoin, Ether and Binance. The VaR values for both stablecoins are below 1% for the 90% and 95% confidence intervals and slightly above 1% for the 99% confidence interval. The significantly low VaR values show that Tether and USD Coin fulfill their objective of having a value as close to $1 as possible.

7.8. Limitations and Lessons Learned

The most time-consuming part of the project has been running the machine learning models; a single neural network algorithm took about 1 to 3 minutes to run on the author's personal machine, depending on whether an additional variable was included in the model, in comparison the time series models took a couple of seconds to run. Running all the baseline models for the data analysis and the Covid-19 part alone took about 6 hours. Bitcoin, Ether, and Binance had 25 variables each, and Tether and USD Coin had 13 variables each. This resulted in running 2,424 models with additional data, which took over 80 hours to compute.

This is to be considered as both a limitation and a lesson learned. Tuning in the hyperparameters of a neural network is a challenging task, which this paper could not elaborate on due to the above-described constraints, making it a limitation. It is a lesson learned, as the substantial number of models made it even more challenging to choose the best hyperparameters. Certain variables performed better or worse in various configurations, and tuning each model separately was unfeasible.

The long computing time of the neural network models was also a constraint for the NLP preprocessing. Had the computing time been considerably shorter, the sentiment analysis could have been repeated with multiple NLP preprocessing techniques observing which one was the most effective. Lastly, the results could have been expanded by including more than one online variable in a model so that all engagement metrics from a given social media or the sentiment from both social would be bundled together.

8. Conclusion

Cryptocurrencies are decentralized virtual assets, which are very volatile, and present an opportunity for large profits and the risk of large losses. Cryptocurrencies are typically not backed by any physical assets, which, along with their volatility, makes it especially challenging to predict their prices; this led to scholars trying to find factors that may explain the behavior and nature of cryptocurrency prices. This study focused on the five biggest cryptocurrencies by market capitalization (as of the 16th of January 2023, coinmarketcap.com): Bitcoin, Ether, Binance, Tether, and USD Coin.

Firstly, the closing prices of the five cryptocurrencies were analyzed in relation to each other; stablecoins were found to be different from Bitcoin, Ether, and Binance; hence, these two groups were analyzed separately. Through calculating the covariance and correlation of the closing prices, it was found that Bitcoin, Ether, and Binance prices have a strong relationship and are strongly correlated. In contrast, USD Coin and Tether are considerably less correlated. Following the existing literature, it was found that for cryptocurrency price predictions, the main two approaches are; traditional statistical methods and machine learning models (Khedr et al., 2021). The paper implemented the most popular methods from Khedr et al. (2021) survey for the price prediction of the five chosen cryptocurrencies for 1-, 2- and 3-day prediction intervals. The chosen traditional statistical method was ARIMA and SARIMA models, whereas the chosen machine learning models were RNN, LSTM, Bi-LSTM, and GRU models. The results were consistent with the existing literature, as the Machine learning methods outperformed the traditional statistical models. Furthermore, the seasonal factor of ARIMA model was not always the best traditional statistical model, which proved that cryptocurrencies do not have significant seasonal trends.

In order to examine the relationship between online factors and the closing price of cryptocurrencies, data from Twitter, Reddit and Wikipedia were acquired. In order to obtain data that had a significant engagement, only Tweets with a minimum of 1 like or 1 retweet were collected; moreover, only data from subreddits related to Bitcoin, Ether, and Binance was collected as the subreddits for Tether and USD Coin did not have enough members.

Social media sentiment is one of the most often analyzed online factors in cryptocurrency literature; for this reason, this paper uses VADER, an open-source, rule and lexicon-based sentiment analysis tool designed especially for social media posts to perform the sentiment analysis. Afterward, the data from Twitter, Reddit, and Wikipedia were merged with the daily cryptocurrency closing prices for each respective cryptocurrency.

The variables were then analyzed based on their correlation with the closing price. It was found that for Bitcoin, Ether, and Binance, Twitter data is very strongly and positively correlated with the closing price, much more than Reddit data. Furthermore, all three sentiments, positive, negative, and neutral, are among the most correlated features among all five cryptocurrencies. The Tether and USD Coin results show a much less significant negative correlation with the closing price. The online variable's importance was further examined by the Random Forest feature extraction, which showed that apart from the online sentiment, features such as Twitter Likes, the volume of Tweets, and Wikipedia trend have much importance in explaining the prices of cryptocurrencies.

The features chosen through the correlation analysis and Random Forest feature extraction were afterward used as additional data for the machine learning models, RNN, LSTM, Bi-LSTM, and GRU, to observe which variables will improve the baseline models. The results show that despite being less correlated and given little importance in the Random Forest feature extraction, the Reddit data tends to produce the most accurate predictions for Bitcoin, Ether, and Binance, whereas the Wikipedia trend rarely improves the baseline model despite being given much importance by the Random Forest algorithm. The accuracy

measures for stablecoin's results are exponentially small, most probably due to the tiny price fluctuations of the two stablecoins, Tether and USD Coin. The results for Bitcoin, Ether, and USD Coin do not show a specific variable or sentiment that would consistently improve the baseline model. However, in many cases, the non-sentiment variables, such as the number of Twitter likes or the Reddit upvote ratio, tend to be the best models. Tether's results show a clear pattern of negative sentiment consistently performing as the best model, regardless of the prediction interval or the chosen algorithm. Binance's results show that very few models have managed to improve the baseline model at all.

In order to observe whether the results would be consistent, the models were afterward repeated for the Covid-19 period, which was defined as starting on the 11th of March 2020 and ending on the 26th of January 2021. This time the results were highly consistent among the same machine learning algorithm but not within the same prediction interval. Moreover, the overall accuracy of the models was worse than before, most probably due to larger price fluctuations during the Covid-19 period; this resulted in many models improving the baseline more than tenfold, suggesting that online variables were more important during the Covid-19 period. Bitcoin's results, just like previously, show that Reddit variables produce more accurate models than Twitter variables, but no specific variable or sentiment is consistently the best model. Ether's results show that most of the best models were produced by engagement metrics rather than social media sentiment and that Reddit data produces better models than Twitter data. For Binance, in contrast to before, many variables outperformed the baseline, with Twitter and Reddit data performing very similarly. Tether's results are similar, with the negative sentiment still consistently producing some of the best models, however not to the same extent as before. USD Coin's results do not show a clear pattern as to which variable or sentiment would be the best one. An additional VaR analysis proved that cryptocurrencies present a considerable risk for investors, with Bitcoin, Ether, and Binance having VaR values of over -10% for a 99% confidence level. The stablecoins, however, were found to have VaR values significantly lower at about 1% for a 99% confidence level.

The findings, therefore, prove that online variables have a strong relationship with the closing prices of Bitcoin, Ether, Binance, Tether and USD Coin, with Reddit variables having a better predictive ability despite being less correlated than Twitter data or the Wikipedia trend. It is difficult to tell which sentiment has the most overall significance. However, the less-examined engagement metrics such as the number of Likes, Retweets, or Replies on Twitter or upvote ratio, number of comments, or the score on Reddit deliver outcomes comparable to or better than those produced by the sentiments. Lastly, the machine learning models perform better than the traditional statistical models and can be further improved by including online data.

9. References

Academic papers:

Tredinnick, L. (2019). Cryptocurrencies and the blockchain. Business Information Review, 36(1), 39-44.

Van Alstyne, M. (2014). Why Bitcoin has value. Communications of the ACM, 57(5), 30-32.

Berentsen, A., & Schär, F. (2018). A short introduction to the world of cryptocurrencies. FRB of St. Louis Working Review.

Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). Bitcoin and cryptocurrency technologies: a comprehensive introduction. Princeton University Press.

Schinckus, C. (2021). Proof-of-work based blockchain technology and Anthropocene: An undermined situation?. Renewable and Sustainable Energy Reviews, 152, 111682.

Dannen, C. (2017). Introducing Ethereum and solidity (Vol. 1, pp. 159-160). Berkeley: Apress.

Griffin, J. M., & Shams, A. (2020). Is Bitcoin really untethered?. The Journal of Finance, 75(4), 1913-1964.

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

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: an introduction. Cambridge University Press.

Phillips, R. C., & Gorse, D. (2018). Cryptocurrency price drivers: Wavelet coherence analysis revisited. PloS one, 13(4), e0195200., Chicago,

Ji, Q., Bouri, E., Lau, C. K. M., & Roubaud, D. (2019). Dynamic connectedness and integration in cryptocurrency markets. International Review of Financial Analysis, 63, 257-272.

Sovbetov, Y. (2018). Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. Journal of Economics and Financial Analysis, 2(2), 1-27.

Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., ... & Zhang, J. (2021, April). Lstm based sentiment analysis for cryptocurrency prediction. In International Conference on Database Systems for Advanced Applications (pp. 617-621). Springer, Cham.

Biswas, S., Pawar, M., Badole, S., Galande, N., & Rathod, S. (2021, March). Cryptocurrency price prediction using neural networks and deep learning. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 408-413). IEEE.

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. SMU Data Sci. Rev, 1(3), 1-22.

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy, 21(6), 589.

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), 1.

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Systems, 37(2), e12493.

Yousaf, I., & Ali, S. (2020). Discovering interlinkages between major cryptocurrencies using high-frequency data: new evidence from COVID-19 pandemic. Financial Innovation, 6(1), 1-18.

Conlon, T., Corbet, S., & McGee, R. J. (2020). Are cryptocurrencies a safe haven for equity markets? An international perspective from the COVID-19 pandemic. Research in International Business and Finance, 54, 101248.

Bukovina, J., & Marticek, M. (2016). Sentiment and bitcoin volatility. University of Brno.

Burnie, A., & Yilmaz, E. (2019). Social media and bitcoin metrics: which words matter. Royal Society open science, 6(10), 191068.

Pano, T., & Kashef, R. (2020). A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. Big Data and Cognitive Computing, 4(4), 33.

Shen, D., Urquhart, A., & Wang, P. (2019). Does twitter predict Bitcoin?. Economics Letters, 174, 118-122.

Rouhani, S., & Abedin, E. (2019). Crypto-currencies narrated on tweets: a sentiment analysis approach. International Journal of Ethics and Systems, 36(1), 58-72.

Khedr, A. M., Arif, I., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Intelligent Systems in Accounting, Finance and Management, 28(1), 3-34.

Azari, A. (2019). Bitcoin price prediction: An ARIMA approach. arXiv preprint arXiv:1904.05315.

Wirawan, I. M., Widiyaningtyas, T., & Hasan, M. M. (2019, September). Short term prediction on bitcoin price using ARIMA method. In 2019 International Seminar on Application for Technology of Information and Communication (iSemantic)(pp. 260-265). IEEE.

Eisen, A. M. (2018). Prediction of Cryptocurrency Price Using Wikipedia Page Views (Doctoral dissertation, Ben-Gurion University of the Negev).

ElBahrawy, A., Alessandretti, L., & Baronchelli, A. (2019). Wikipedia and cryptocurrencies: Interplay between collective attention and market performance. Frontiers in Blockchain, 2, 12.

Seroyizhko, P., Zhexenova, Z., Shafiq, M. Z., Merizzi, F., Galassi, A., & Ruggeri, F. (2022). A Sentiment and Emotion Annotated Dataset for Bitcoin Price Forecasting Based on Reddit Posts. In Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (pp. 50-56).

Hansun, S., Wicaksana, A., & Khaliq, A. Q. (2022). Multivariate cryptocurrency prediction: comparative analysis of three recurrent neural networks approaches. Journal of Big Data, 9(1), 1-15.

Raju, S. M., & Tarif, A. M. (2020). Real-time prediction of BITCOIN price using machine learning techniques and public sentiment analysis. arXiv preprint arXiv:2006.14473.

Critien, J. V., Gatt, A., & Ellul, J. (2022). Bitcoin price change and trend prediction through twitter sentiment and data volume. Financial Innovation, 8(1), 1-20.

Lee, N., & Lings, I. (2008). Doing business research: a guide to theory and practice. Sage.

Saunders, M., Lewis, P. H. I. L. I. P., & Thornhill, A. D. R. I. A. N. (2007). Research methods. Business Students 4th edition Pearson Education Limited, England.

Yin, R. K. (2003). Case study research design and methods third edition. Applied social research methods series, 5.

Malhotra, N., & Birks, D. F. (2007). An applied approach. Marketing research. London: Prentice Hall.

Veal, A. J. (2017). Research methods for leisure and tourism. Pearson UK.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In Proceedings of the international AAAI conference on web and social media (Vol. 14, pp. 830-839).

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 3285-3292). IEEE.

Jurafsky, D., & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2018

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Websites:

https://investingnews.com/daily/tech-investing/blockchain-investing/bitcoin-price-history/#toggle-gdpr

https://finance.yahoo.com/quote/BTC-USD/

https://explodingtopics.com/blog/number-of-cryptocurrencies

https://www.investopedia.com/articles/investing/031416/bitcoin-vs-ethereum-driven-different-purposes.asp

https://www.simplilearn.com/bitcoin-mining-explained-article

https://buybitcoinworldwide.com/volatility-index/

https://www.euromoney.com/learning/blockchain-explained/what-is-blockchain

https://theprint.in/theprint-essential/the-gamestop-story-how-a-group-of-investors-on-reddit-gave-wall-street-a-wild-week/595181/

https://www.vox.com/recode/2021/5/18/22441831/elon-musk-bitcoin-dogecoin-crypto-prices-tesla

https://www.gemini.com/cryptopedia/the-dao-hack-makerdao

https://www.investopedia.com/terms/b/binance-coin-bnb.asp

https://coinmarketcap.com/rankings/exchanges/

https://history-computer.com/binance-history/

https://www.barrons.com/articles/u-s-arm-of-crypto-exchange-binance-is-valued-at-4-5-billion-watch-for-an-ipo-51649253200

https://www.investopedia.com/terms/b/binance-exchange.asp

https://cointelegraph.com/news/what-is-bnb-auto-burn-and-how-does-it-work

https://www.binance.com/en/nft/home

https://support.exodus.com/article/1820-bnb-beacon-chain-and-smart-chain

https://academy.binance.com/en/articles/an-introduction-to-binance-smart-chain-bsc

https://www.gizmodo.com.au/2022/10/binance-says-us100-139-million-of-crypto-produced-out-of-thin-air-by-hackers/

https://www.ionos.com/digitalguide/online-marketing/online-sales/binance-smart-chain/

https://www.investopedia.com/terms/s/stablecoin.asp

https://www.forbes.com/sites/rufaskamau/2022/05/09/what-is-bitcoin-pizza-day-and-why-does-the-community-celebrate-on-may-22/
https://medium.com/@MakerDAO/stablecoins-strengths-weaknesses-62cd47bb7fbf
https://cryptonews.net/editorial/investments/stablecoins-their-types-advantages-disadvantages/
https://smartasset.com/financial-advisor/tether-what-is-it
https://www.coindesk.com/learn/2022/06/01/what-is-tether-how-usdt-works-and-what-backs-its-value/
https://tether.to/en/how-it-works/
https://www.bitfinex.com/posts/432
https://cryptonews.com/coins/usd-coin/
https://www.circle.com/en/usdc
https://www.analyticsinsight.net/usd-coin-everything-you-should-know-about-the-second-largest-stablecoin/
https://www.techtarget.com/whatis/definition/social-media
https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
https://www.macrotrends.net/stocks/charts/META/meta-platforms/net-worth
https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
https://www.oberlo.com/statistics/what-age-group-uses-social-media-the-most
https://www.lifewire.com/what-exactly-is-twitter-2483331
https://help.twitter.com/en/using-twitter/how-to-use-hashtags
https://blog.hootsuite.com/twitter-statistics/
https://datareportal.com/essential-twitter-stats
https://www.cnbc.com/2020/11/03/alexis-ohanian-reflects-on-selling-reddit-for-10-million.html
https://medium.com/@michaelmaieli2019/reddit-organizing-communities-for-every-topic-on-the-internet-d784ba2b140b
https://www.newscientist.com/article/2325828-reddit-moderators-do-3-4-million-worth-of-unpaid-work-each-year/
https://www.britannica.com/topic/Wikipedia
https://www.techtarget.com/whatis/definition/Wikipedia
https://en.wikipedia.org/wiki/Wikipedia:Administrators
https://webtribunal.net/blog/how-many-people-use-wikipedia/#gref
https://ethereum.org/en/developers/docs/consensus-mechanisms/pos/
https://time.com/nextadvisor/investing/cryptocurrency/proof-of-work-vs-proof-of-stake/
https://resources.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.html
https://github.com/mattpodolak/pmaw
https://github.com/JustAnotherArchivist/snscrape
https://github.com/cjhutto/vaderSentiment
https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9
https://medium.com/fintechexplained/understanding-auto-regressive-model-arima-4bd463b7a1bb
http://alkaline-ml.com/pmdarima/

https://builtin.com/data-science/recurrent-neural-networks-and-lstm

https://analyticsindiamag.com/lstm-vs-gru-in-recurrent-neural-network-a-comparative-study/

https://towardsdatascience.com/are-you-scared-vader-understanding-how-nlp-pre-processing-impacts-vader-scoring-4f4edadbc91d

https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/

https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d

https://towardsdatascience.com/random-forest-regression-5f605132d19d

https://mljar.com/blog/feature-importance-in-random-forest/

https://alkaline-ml.com/pmdarima/tips_and_tricks.html#period

https://corporatefinanceinstitute.com/resources/risk-management/value-at-risk-var/

10. Appendix

Appendix 1.

Table of results for Tether including additional variables for the October 2022 prediction

Days	Model	Twitter's top 5 features			Baselines & Wiki		
		variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	pct_neg	2.75E-04	0.0275	Base	6.95E-04	0.0695
		Pos_vol	4.02E-04	0.0402	ARIMA	1.65E-05	0.0016
		Neg_vol	1.10E-03	0.1095	Wikipedia	3.84E-04	0.0384
	LSTM	Neg_vol	4.86E-04	0.0486	Base	1.67E-03	0.1671
		pct_neu	5.21E-04	0.0521	ARIMA	1.65E-05	0.0016
		pct_neg	7.35E-04	0.0735	Wikipedia	9.52E-04	0.0952
	Bi -LSTM	pct_neg	2.05E-05	0.0021	Base	9.66E-04	0.0966
		Neg_vol	5.02E-05	0.0050	ARIMA	1.65E-05	0.0016
		Pos_vol	2.46E-04	0.0246	Wikipedia	4.28E-04	0.0428
	GRU	pct_neg	2.29E-04	0.0229	Base	1.04E-03	0.1035
		Neg_vol	4.30E-04	0.0430	ARIMA	1.65E-05	0.0016
		pct_neu	7.43E-04	0.0743	Wikipedia	6.20E-04	0.0619
2 days	RNN	Pos_vol	2.91E-04	0.0244	Base	7.95E-04	0.0789
		pct_neg	3.96E-04	0.0381	ARIMA	1.43E-04	0.0109
		Neg_vol	1.00E-03	0.0997	Wikipedia	4.34E-04	0.0431
	LSTM	pct_neu	4.05E-04	0.0379	Base	1.60E-03	0.1600
		Neg_vol	5.44E-04	0.0541	ARIMA	1.43E-04	0.0109
		pct_neg	7.29E-04	0.0729	Wikipedia	8.77E-04	0.0874
	Bi -LSTM	pct_neg	4.36E-05	0.0039	Base	8.93E-04	0.0890
		Neg_vol	4.77E-05	0.0048	ARIMA	1.43E-04	0.0109
		pct_neu	2.28E-04	0.0203	Wikipedia	5.31E-04	0.0522
	GRU	pct_neg	1.92E-04	0.0188	Base	9.58E-04	0.0955
		Neg_vol	3.15E-04	0.0274	ARIMA	1.43E-04	0.0109
		pct_neu	5.84E-04	0.0552	Wikipedia	6.90E-04	0.0686
3 days	RNN	pct_neu	3.57E-04	0.0338	Base	9.93E-04	0.0989
		Neg_vol	8.25E-04	0.0737	ARIMA	1.14E-04	0.0088
		pct_neg	8.63E-04	0.0835	Wikipedia	9.29E-04	0.0922
	LSTM	Neg_vol	3.92E-04	0.0367	Base	1.63E-03	0.1629
		pct_pos	7.86E-04	0.0782	ARIMA	1.14E-04	0.0088
		Pos_vol	8.02E-04	0.0799	Wikipedia	6.48E-05	0.0057
	Bi -LSTM	pct_neg	3.10E-04	0.0296	Base	9.70E-04	0.0967
		pct_neu	1.06E-03	0.1053	ARIMA	1.14E-04	0.0088
		Neg_vol	1.09E-03	0.1084	Wikipedia	1.21E-03	0.1211
	GRU	pct_neg	3.88E-04	0.0363	Base	9.88E-04	0.0985
		Quote	7.81E-04	0.0733	ARIMA	1.14E-04	0.0088
		Volume	9.06E-04	0.0838	Wikipedia	8.67E-04	0.0863

Appendix 2.

Table of results for USD Coin including additional variables for the October 2022 prediction

Days	Model	Twitter's top 5 features			Baselines & Wiki		
		variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	pct_neg	8.70E-04	0.0870	Base	2.63E-04	0.0263
		compound	1.49E-03	0.1486	ARIMA	8.67E-06	0.0009
		Neg_vol	2.26E-03	0.2261	Wikipedia	2.68E-03	0.2679
	LSTM	pct_pos	3.92E-04	0.0392	Base	1.87E-03	0.1869
		Likes	6.28E-04	0.0628	ARIMA	8.67E-06	0.0009
		pct_neu	6.35E-04	0.0635	Wikipedia	1.38E-03	0.1379
	Bi -LSTM	pct_pos	2.96E-04	0.0296	Base	1.11E-03	0.1108
		pct_neu	4.46E-04	0.0446	ARIMA	8.67E-06	0.0009
		pct_neg	4.88E-04	0.0488	Wikipedia	1.27E-03	0.1269
	GRU	Neg_vol	1.43E-04	0.0143	Base	1.80E-03	0.1798
		Likes	5.06E-04	0.0506	ARIMA	8.67E-06	0.0009
		Retweets	5.22E-04	0.0522	Wikipedia	1.39E-03	0.1394
2 days	RNN	Pos_vol	4.62E-04	0.0418	Base	3.44E-04	0.0344
		compound	9.03E-04	0.0858	ARIMA	2.69E-05	0.0023
		Likes	1.99E-03	0.1945	Wikipedia	6.20E-04	0.0618
	LSTM	Replies	1.10E-04	0.0097	Base	1.88E-03	0.1877
		pct_pos	1.86E-04	0.0162	ARIMA	2.69E-05	0.0023
		Quote	1.97E-04	0.0143	Wikipedia	9.72E-04	0.0972
	Bi -LSTM	pct_neu	1.04E-04	0.0099	Base	1.10E-03	0.1096
		pct_neg	5.61E-04	0.0561	ARIMA	2.69E-05	0.0023
		pct_pos	6.03E-04	0.0599	Wikipedia	8.65E-04	0.0865
	GRU	pct_pos	2.15E-04	0.0212	Base	1.76E-03	0.1762
		Likes	3.07E-04	0.0299	ARIMA	2.69E-05	0.0023
		Neg_vol	4.90E-04	0.0484	Wikipedia	1.16E-03	0.1155
3 days	RNN	Pos_vol	9.30E-04	0.0746	Base	1.04E-04	0.0096
		Neu_vol	1.40E-03	0.1345	ARIMA	5.64E-05	0.0045
		Likes	1.50E-03	0.1428	Wikipedia	5.42E-04	0.0541
	LSTM	pct_neu	3.68E-04	0.0354	Base	1.89E-03	0.1886
		pct_neg	3.77E-04	0.0372	ARIMA	5.64E-05	0.0045
		Retweets	5.18E-04	0.0513	Wikipedia	5.30E-04	0.0528
	Bi -LSTM	pct_neg	5.35E-04	0.0534	Base	1.12E-03	0.1124
		Retweets	5.60E-04	0.0559	ARIMA	5.64E-05	0.0045
		Likes	6.07E-04	0.0511	Wikipedia	1.15E-03	0.1150
	GRU	pct_neg	1.68E-04	0.0150	Base	1.82E-03	0.1817
		Retweets	2.86E-04	0.0216	ARIMA	5.64E-05	0.0045
		pct_neu	5.03E-04	0.0480	Wikipedia	1.84E-03	0.1841

Appendix 3.

Table of results for Tether including additional variables for Covid- 19 prediction

Days	Model	Twitter's top 5 features			Baselines & Wikipedia		
		variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	Replies	1.45E-03	0.14	Baseline	3.74E-03	0.37
		Quote	2.33E-03	0.23	ARIMA	3.03E-01	30.21
		Retweets	3.15E-03	0.31	Wikipedia	2.46E-02	2.45
	LSTM	pct_neg	1.12E-03	0.11	Baseline	3.64E-03	0.36
		pct_neu	2.16E-03	0.22	ARIMA	3.03E-01	30.21
		Neg_vol	2.21E-03	0.22	Wikipedia	5.78E-03	0.58
	Bi -LSTM	Volume	5.56E-04	0.06	Baseline	1.83E-03	0.18
		Neu_vol	1.09E-03	0.11	ARIMA	3.03E-01	30.21
		Neg_vol	1.51E-03	0.15	Wikipedia	1.26E-03	0.13
	GRU	pct_pos	2.10E-04	0.02	Baseline	2.15E-03	0.22
		compound	6.60E-04	0.07	ARIMA	3.03E-01	30.21
		Pos_vol	7.46E-04	0.07	Wikipedia	5.50E-03	0.55
2 days	RNN	Replies	1.45E-03	0.15	Baseline	3.56E-03	0.36
		Quote	2.13E-03	0.21	ARIMA	1.00	100
		Neg_vol	2.97E-03	0.25	Wikipedia	2.39E-02	2.38
	LSTM	pct_neg	9.31E-04	0.09	Baseline	3.54E-03	0.35
		Neg_vol	1.99E-03	0.20	ARIMA	1.00	100
		pct_neu	2.07E-03	0.21	Wikipedia	5.55E-03	0.55
	Bi -LSTM	Volume	7.47E-04	0.07	Baseline	1.71E-03	0.17
		Neu_vol	9.55E-04	0.09	ARIMA	1.00	100
		Neg_vol	1.38E-03	0.14	Wikipedia	1.24E-03	0.12
	GRU	pct_pos	1.49E-04	0.01	Baseline	2.02E-03	0.20
		compound	5.56E-04	0.05	ARIMA	1.00	100
		Pos_vol	6.47E-04	0.06	Wikipedia	5.29E-03	0.53
3 days	RNN	Replies	1.23E-03	0.12	Baseline	3.17E-03	0.31
		Quote	1.81E-03	0.17	ARIMA	8.17E-01	66.74
		Neg_vol	2.42E-03	0.17	Wikipedia	2.33E-02	2.32
	LSTM	pct_neg	8.11E-04	0.08	Baseline	3.19E-03	0.31
		Neg_vol	1.66E-03	0.15	ARIMA	8.17E-01	66.74
		pct_neu	1.77E-03	0.17	Wikipedia	5.13E-03	0.51
	Bi -LSTM	Neu_vol	7.89E-04	0.07	Baseline	1.42E-03	0.13
		Neg_vol	1.14E-03	0.10	ARIMA	8.17E-01	66.74
		Volume	1.32E-03	0.12	Wikipedia	1.65E-03	0.16
	GRU	compound	6.17E-04	0.06	Baseline	1.71E-03	0.16
		Pos_vol	6.25E-04	0.06	ARIMA	8.17E-01	66.74
		Volume	6.52E-04	0.06	Wikipedia	4.88E-03	0.48

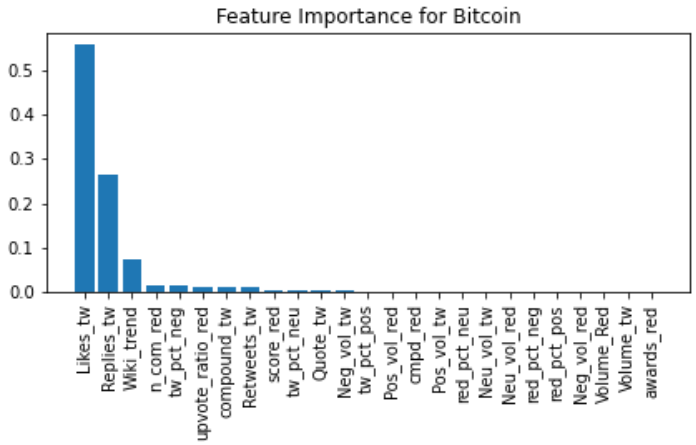
Appendix 4.

Table of results for USD Coin including additional variables for Covid- 19 prediction

Days	Model	Twitter's top 5 features			Baselines & Wikipedia		
		variable	RMSE	MAPE	variable	RMSE	MAPE
1 day	RNN	pct_neg	2.14E-04	0.02	Baseline	3.04E-03	0.30
		Neg_vol	2.93E-03	0.29	ARIMA	8.45E-06	0.00084
		Replies	4.26E-03	0.43	Wikipedia	2.66E-02	2.66
	LSTM	pct_neu	1.85E-03	0.19	Baseline	2.51E-03	0.25
		pct_pos	2.06E-03	0.21	ARIMA	8.45E-06	0.00084
		Replies	2.48E-03	0.25	Wikipedia	5.76E-03	0.58
	Bi -LSTM	Volume	5.89E-05	0.01	Baseline	2.78E-03	0.28
		pct_pos	6.34E-04	0.06	ARIMA	8.45E-06	0.00084
		compound	8.18E-04	0.08	Wikipedia	1.30E-03	0.13
	GRU	pct_pos	1.21E-03	0.12	Baseline	2.22E-03	0.22
		compound	1.22E-03	0.12	ARIMA	8.45E-06	0.00084
		Pos_vol	1.89E-03	0.19	Wikipedia	5.03E-03	0.50
2 days	RNN	Replies	3.37E-03	0.32	Baseline	2.86E-03	0.29
		pct_neu	4.39E-03	0.42	ARIMA	3.40E-05	0.0028
		Quote	4.56E-03	0.46	Wikipedia	2.58E-02	2.58
	LSTM	pct_neu	1.45E-03	0.14	Baseline	2.31E-03	0.23
		pct_pos	1.90E-03	0.19	ARIMA	3.40E-05	0.0028
		Replies	2.22E-03	0.22	Wikipedia	5.54E-03	0.55
	Bi -LSTM	Volume	3.98E-04	0.03	Baseline	2.57E-03	0.26
		pct_pos	4.55E-04	0.04	ARIMA	3.40E-05	0.0028
		Likes	7.49E-04	0.07	Wikipedia	1.50E-03	0.15
	GRU	compound	1.03E-03	0.10	Baseline	2.02E-03	0.20
		pct_pos	1.05E-03	0.10	ARIMA	3.40E-05	0.0028
		Pos_vol	1.65E-03	0.16	Wikipedia	4.79E-03	0.48
3 days	RNN	Replies	3.10E-03	0.30	Baseline	2.86E-03	0.29
		Quote	4.53E-03	0.45	ARIMA	5.00E-01	50.05
		pct_neu	5.04E-03	0.49	Wikipedia	2.59E-02	2.59
	LSTM	pct_neu	1.56E-03	0.15	Baseline	2.37E-03	0.24
		pct_pos	2.06E-03	0.20	ARIMA	5.00E-01	50.05
		Replies	2.27E-03	0.23	Wikipedia	5.60E-03	0.56
	Bi -LSTM	Volume	3.85E-04	0.03	Baseline	2.64E-03	0.26
		pct_pos	4.69E-04	0.04	ARIMA	5.00E-01	50.05
		Neu_vol	7.02E-04	0.07	Wikipedia	1.42E-03	0.14
	GRU	compound	1.10E-03	0.11	Baseline	2.09E-03	0.21
		pct_pos	1.22E-03	0.12	ARIMA	5.00E-01	50.05
		Pos_vol	1.68E-03	0.17	Wikipedia	4.84E-03	0.48

Appendix 5.

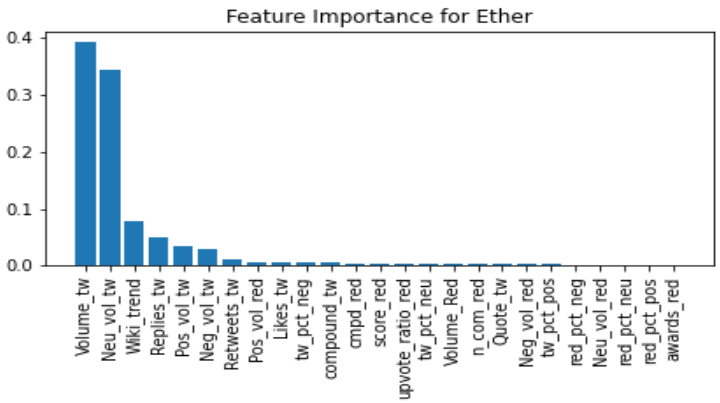
Random Forest feature extraction Results for Bitcoin



Feature	Importance
Likes_tw	55,61%
Replies_tw	26,43%
Wiki_trend	7,6%
n_com_red	1,71%
tw_pct_neg	1,43%
upvote_ratio_red	1,33%
compound_tw	1,08%
Retweets_tw	1,05%
score_red	0,38%
tw_pct_neu	0.37%

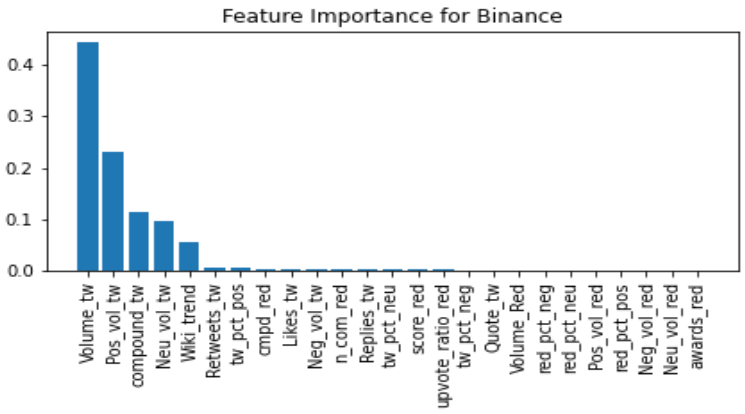
.

Random Forest feature extraction Results for Ether



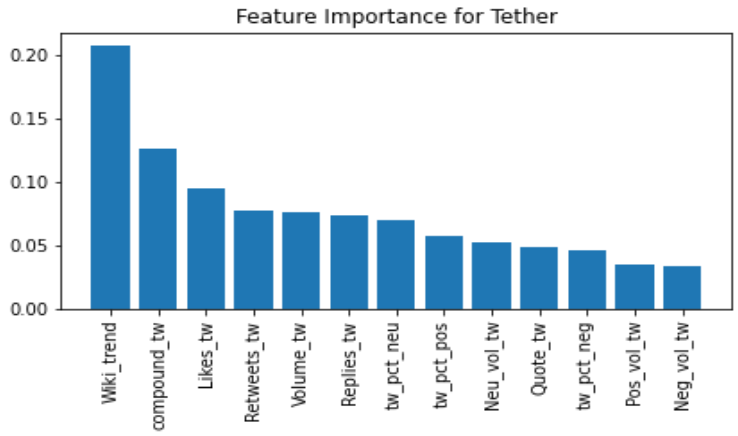
Feature	Importance
Volume_tw	39,14%
Neu_vol_tw	34,45%
Wiki_trend	7,8%
Replies_tw	5,08%
Pos_vol_tw	3,51%
Neg_vol_tw	3,01%
Retweets_tw	1,16%
Pos_vol_red	0,65%
Likes_tw	0,61%
tw_pct_neg	0,55%

Random Forest feature extraction Results for Binance



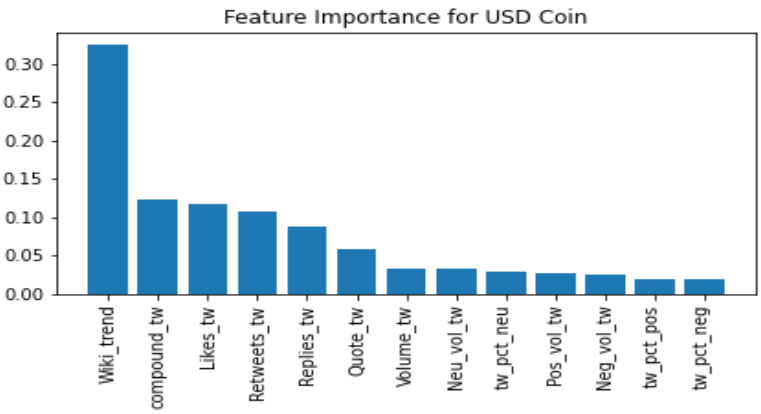
Feature	Importance
Volume_tw	44,21%
Pos_vol_tw	23,06%
compound_tw	11,34%
Neu_vol_tw	9,76%
Wiki_trend	5,7%
Retweets_tw	0,78%
tw_pct_pos	0,57%
cmpd_red	0,5%
Likes_tw	0,5%
Neg_vol_tw	0,45%

Random Forest feature extraction Results for Tether



Feature	Importance
Wiki_trend	20,8%
compound_tw	12,6%
Likes_tw	9,52%
Retweets_tw	7,72%
Volume_tw	7,68%
Replies_tw	7,41%
tw_pct_neu	6,94%
tw_pct_pos	5,73%
Neu_vol_tw	5,19%
Quote_tw	4,88%

Random Forest feature extraction Results for USD Coin



Feature	Importance
Wiki_trend	32,46%
compound_tw	12,38%
Likes_tw	11,75%
Retweets_tw	10,76%
Replies_tw	8,73%
Quote_tw	5,73%
Volume_tw	3,32%
Neu_vol_tw	3,25%
tw_pct_neu	2,83%
Pos_vol_tw	2,67%

Appendix 6.

Time series results for Covid-19 period prediction

Cryptocurrency	Days	Order	RMSE	MAPE
Bitcoin	1 day	(3, 0, 2)	933.02	2.72
		(2, 0, 1) (2, 0, 1) [7]	1311.54	3.83
	2 days	(2, 0, 2)	577.55	1.34
		(3, 0, 0) (1, 0, 1) [7]	1507.58	3.54
	3 days	(3, 0, 2)	3459.45	7.35
		(2, 0, 1) (2, 0, 1) [7]	1809.72	5.18
Ether	1 day	(3, 0, 1)	87.96	6.39
		(2, 0, 1) (2, 0, 1) [7]	71.93	5.23
	2 days	(3, 0, 2)	68.96	4.14
		(2, 0, 1), (2, 0, 1) [7]	93.93	6.65
	3 days	(3, 0, 1)	163.05	8.93
		(1, 0, 0), (2, 0, 1) [7]	48.62	3.41
Binance	1 day	(3, 0, 2)	0.67	1.51
		(2, 0, 2), (1, 0, 1) [7]	0.85	1.89
	2 days	(3, 0, 2)	0.55	1.21
		(2, 0, 1), (1, 0, 1) [7]	1.10	2.48
	3 days	(3, 0, 1)	2.43	4.06
		(2, 0, 1), (2, 0, 1) [7]	1.09	2.51
Tether	1 day	(3, 0, 0)	1	100
		(0, 0, 1), (0, 0, 1, 7)	0.3	30.21
	2 days	(3, 0, 0)	1	100
		(3, 0, 0), (0, 0, 0) [7]	1	100
	3 days	(3, 0, 0)	0.82	66.74
		(3, 0, 0) (0, 0, 0) [7]	0.82	66.74
USD Coin	1 day	(1, 0, 1)	8.446E-06	0.00084
		(0, 0, 0), (0, 0, 0) [7]	1	100
	2 days	(1, 0, 1)	3.4E-05	0.0028
		(0, 0, 0), (0, 0, 0) [7]	1	100
	3 days	(0, 0, 0)	1	100
		(0, 0, 0), (0, 0, 1) [7]	0.5	50.05