

Twitter sentiment analysis in the context of Bitcoin price

www.sentcrypt.com



University of
Zurich^{UZH}

Contact Person :
Julien Donnet
Jjdonnet86@gmail.com
+41 78 964 00 16

Version: 1.0

Published:
Date: 09.12.2020



Table of contents

| | |
|---|-----------|
| Introduction | 3 |
| 1 a. Abstract | 3 |
| 1 b. Context | 3 |
| 1 What is bitcoin | 5 |
| 2 a. Bitcoin at a glance | 5 |
| 2 b. Bitcoin - Difference with traditional assets | 7 |
| 2 c. Other incentives to invest into crypto currencies and blockchain | 8 |
| 2 Bitcoin's price | 11 |
| 3 a. How is bitcoin priced? | 11 |
| 3 b. Is it an efficient market? | 13 |
| 3 b. Trends, Sentiment and social interactions | 14 |
| 3 Sentiment analysis | 16 |
| 4 a. Sentiment analysis in a nutshell | 16 |
| 4 c. Crypto currencies specifics and challenges | 17 |
| 4 e. What exists on the market | 18 |
| 4 Implementing a sentiment analyser | 20 |
| 5 a. Abstract SentCrypt | 20 |
| 5 b. Data collection | 20 |
| 5 c. Applying sentiment analysis | 22 |
| 5 d. Storage and Archiving | 23 |
| 5 g. Infrastructure and deployment | 24 |
| 5 e. Optimizing the data | 26 |
| 5 Results and conclusions | 28 |
| 6 Annex | 30 |
| 6.a Tweet example (Sample) | 30 |
| 6 b. Code prices | 31 |
| 6 c. Code tweets | 34 |
| 6 c. Code industrialization | 37 |
| 7 References | 39 |

1 Introduction

1 a. Abstract

Purpose – The purpose of this document is first to explore and present the state of the art concerning the valuation of Bitcoin and sentiment analysis in the context of Bitcoin price and then implement our own solution and framework to capture and compare the price of Bitcoin to the general sentiment of tweets.

Design – The collection of the data will be made with Python through APIs, the storing will be done in a PostgreSQL database, the infrastructure and the hosting will be done on Heroku. The data will be streamed live into charts and tables and then displayed on www.sentscrypt.com.

Findings – In general, we can identify a link and an interaction between twitter sentiment and the price of bitcoin.

Value – The work is done for educational purposes, as such it has limited financial or research value.

Keywords – Bitcoin, Bitcoin Price, Sentiment Analysis, Twitter, APIs, Python.

Paper type – Project report

1 b. Context

Being born in 2009 Blockchain and Bitcoin are rather young technologies, slowly maturing and attracting growing interest from the public, the media and the financial industry. The peak of the media coverage was observed in 2017 when the market grew from 14.589 B to 535.58 B in the span of 11 month (source: [link](#)). Suddenly Bitcoin was on the front page of every magazine of the planet and everybody was caught in the BTC frenzy.

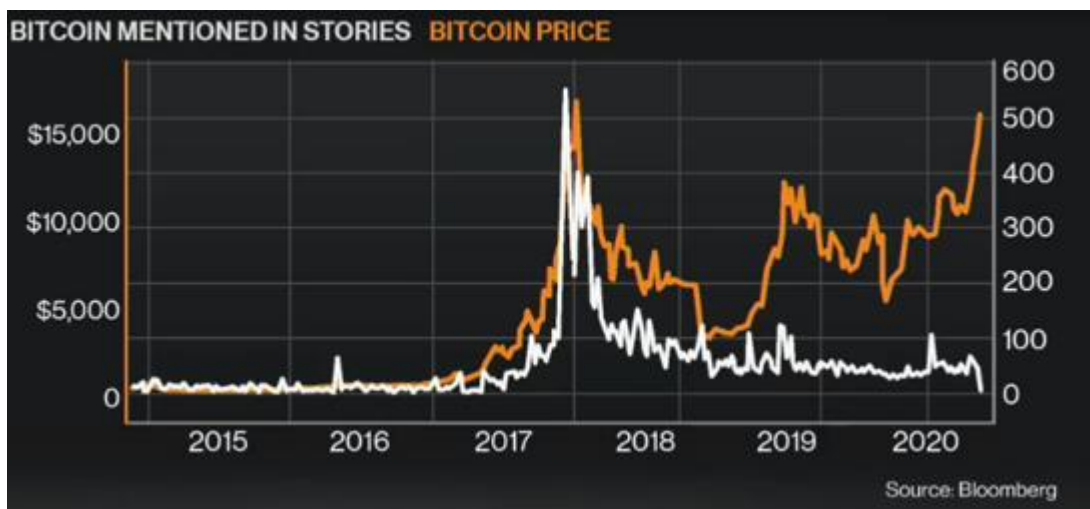


Figure 1 : Bitcoin price and stories

Not everyone was bullish about Bitcoin. The banking old guard was especially dubitative concerning this decentralized currency that could represent a threat to the establishment and to their business. In September 2017, James Dimon (CEO JPMorgan) was calling bitcoin a “fraud” used by [“Murderers, drug dealers and North Koreans”](#), the vast majority of banks were advising their clients to stay away from Bitcoin.

Since then the tide has turned and most of the big banks are experimenting with Blockchain, including JPMorgan. We are even seeing the birth of crypto banks in Switzerland, holding a banking license and regulated by the FINMA. Traditional payment players are also heavily investing in Blockchain, Alfred Kelly (VISA CEO) mentioned earlier this month (Nov 2020) that : “*Crypto is a developing part of payments in the*

world. It's in the very nascent state right now, and we're very interested in cryptocurrencies" ([link](#)). On their side PayPal has enabled their vast user base to buy, hold and sell cryptocurrencies in October 2020 ([link](#)), which represents a serious step towards mainstream adoption.

Although efforts have been made towards regulation and adoption, what remained from the early years is the noticeable volatility of the crypto currencies' prices. From the height of around \$20,089.00 USD on Dec 17, 2017 the price had fallen to \$3212.12 by Dec 14, 2018 and then raised back to \$19,312.39 by end of November 2020, offering a roller-coaster of emotions to the coin's holders.

But what exactly defines the price of this electronic asset and what drives the steep raises and falls of its price? In short, the price of bitcoin can be assessed by multiple ways, from the price dictated by market's offer and demand on the exchanges to the "fundamental" price that could be calculated by taking into account the cost of producing a bitcoin.

One of the established hypothesis is that the cryptocurrency market is heavily influenced by crowd trends, news and social medias. Big swings in the price can be observed when news are released on social medias concerning potential regulations or hacks. As a result, there is an increasing interest to develop and leverage machine learning techniques to detect in advance such swings and capitalize on those to make profit trading the cryptocurrency.

In this document we will first explore the nature of bitcoin and its price and then explain our approach concerning the building of our very own sentiment analysis solution.

2 What is bitcoin

2 a. Bitcoin at a glance

The best way to describe bitcoin would be probably to present it in the light of its white paper, released to the world by Satoshi Nakamoto in 2008 ([link](#)). Satoshi Nakamoto presented Bitcoin as “**a peer to peer electronic cash system**”. The main goal of the currency is to be an electronic version of “cash” that would allow online payments to be sent peer to peer directly, without sending it first through a financial institution. The rationale behind the paper was that as the transactions booked via a financial institution are reversible and since those financial institutions have to mediate disputes between actors in the network, the costs of those transactions are necessarily high. Thus, **limiting the minimum size of the transactions** and limiting the trust of each party involved in online transactions. Satoshi proposed a new electronic payment system, based on cryptographic proof instead of trust. Two willing parties would interact directly with each other without the need of a trusted party as a middle man. The transactions would be recorded in a permanent electronic ledger, without any possibility to reverse them, protecting the sellers from fraud. The solution would offer a solution to the double-spending problem by using a peer-to-peer distributed timestamp server to generate computational proof of the chronology of the transactions, as long as more than 50% of the nodes are honest and control more CPU power than a potential group of hostile actors.

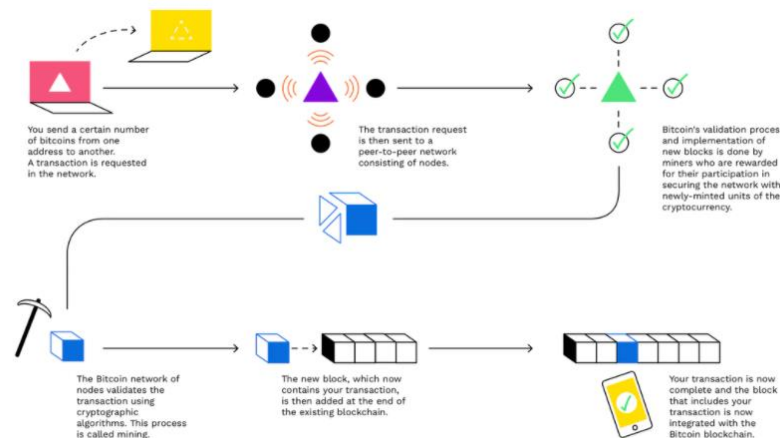


Figure 2 : [What is a blockchain ?](#)

Bitcoin offers three popular incentives to its users:

The first one is that Bitcoin is open-source, it's design is public and nobody owns or controls Bitcoin and everyone can take part in the eco system([link](#)). Privacy and decentralizations have been topics growing in importance in the last decades. Users data have been exposed and illegally accessed at countless iterations, whether it is through global corporations being hacked ([link](#)) or through global surveillance piloted by governments. According to Edward Snowden the NSA has been hacking into private and public communications for more than a decade, ([link](#)). In addition, we have also witnessed occurrences where centralized institutions simply seized the assets of the users or citizens, for instance in Cyprus where the funds of the citizens were simply confiscated by the government, no insured deposit of €100,000 or less were affected, though 47.5% of all bank deposits above €100,000 were seized ([link](#)). This issues have nurtured a growing distrust towards centralized institutions and favoured the emergence of decentralized alternatives, where no central government or private company holds absolute power and where the users are the driving force. In addition, bitcoin can offer a good alternatives to traditional currencies for citizens of countries experiencing high inflation, we have witnessed a raise of bitcoin's popularity in Venezuela lately, as the country is facing a heavy depreciation of the Venezuelan bolivar. Facing hyperinflation the country is turning to cryptocurrency adoption due to a combination of factors including migration, capital controls, risk of government seizure and exposure to petrol price, [link](#).

The second incentive is that Bitcoin and crypto currencies offers an alternative way of investing and earning money. Through mining, anyone with a functioning laptop, or any device capable of computing power ([link](#)), can plug in and become a miner. In a nutshell, miners are the validators of the transactions of the bitcoin network, they lend their computational power to solve cryptographic puzzles and validate transactions and in return they are rewarded with freshly minted bitcoins as an incentive. Today most of the market is dominated by a few big players, but anyone can virtually join of the pools and start earning bitcoin by mining.

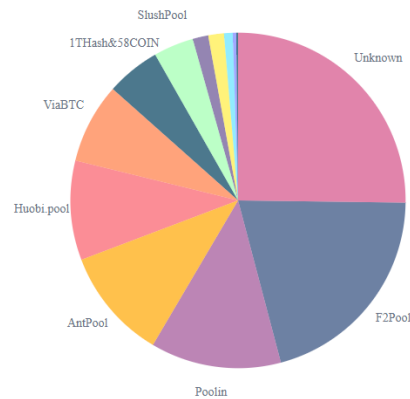


Figure 3: [Hashrate distribution](#)

New generations are also interested in alternative ways of investing their capital. The time where banks were all-mighty asset managers is slowly fading. New access to education through internet and the democratization of smaller players offering new innovative ways for investors to earn interest have offered viable alternatives to the banker. Especially with the raise of DeFi (decentralized finance), new options such as yield farming, crypto lending and staking have offered staggering returns in comparison to traditional asset classes. Very high volatility and a tremendous raise in the price of bitcoin has also offered good returns to bitcoin's investors. According to www.intotheblock.com, as of Today (30.11.2020) more than 95% of people who have bought BTC and held on it are in profit, as shown in the illustration below.

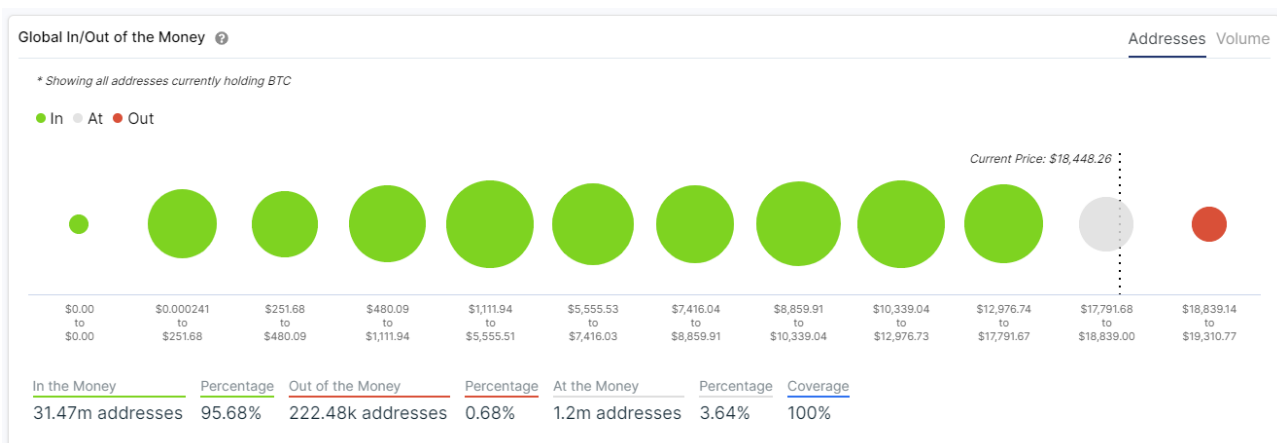


Figure 4: [Global In/Out of the Money](#)

The meteoritic climb of the price of bitcoin and natural prowess to communicate through social medias has given even more appeal to the currency for the newer generations, eager to find new digital ways to invest their money.

Finally, the third reason would be Financial inclusion. According to the World bank, 53% of the worlds' adult population is unbanked, [McKinsey & Company](#) agrees with this statement : "Half of the world is unbanked". This population does not have access to a regular checking account and even less to a payment card or a

credit card. Cryptocurrencies and Bitcoin offer an alternative to this population, everybody with a smartphone or even a regular phone can hold an account and operate peer to peer payments, without having to open a bank account with a regulated entity. Opening new ways to lend and finance your own business.

2 b. Bitcoin - Difference with traditional assets

Marc Andreessen, an internet pioneer and investor famously said that “Software is eating the world” in 2011. Since then we can only agree with Marc seeing how the world has digitized and how the companies that embraced software in 2011 are the current market leaders in their respective fields, no surprise that the top 5 market capitalization companies worldwide are all offering some type of software solutions.

We can see today that this maxim is also relevant for Bitcoin and crypto currencies as a whole. A share represents a partial ownership in a company, a bond is a debt but Bitcoin does not represent any underlying asset. So what is it ?

It is useful to first point out that as Bitcoin does not need any intervention from traditional financial institutions such as banks or clearing houses, the transactions can therefore be created and processed anywhere and anytime, providing business flexibility ([link](#)). Most brokers can only process transactions during working hours, making it difficult to book your trades on your favourite ticker during the weekend for instance. Bitcoin can be traded in the night and from any device.

It can be noted as well that Bitcoins are not being issued by a central authority, they are generated through a mining process and as such they have a predicable growth rate and are completely independent from the traditional eco systems. The limited supply of 21 million bitcoins can also be seen as an advantage to hedge against inflation. The predictability of the supply, the issuance and the cryptographic trust mechanism are all parameters that distinguish Bitcoin from traditional currencies or investment vehicles.

We can also observe a very high volatility in comparison to other traditional investment vehicles or currencies. This volatility can represent an additional risk for investors, and should be taken into account. An investor buying bitcoins must be prepared for big swings in the value of his portfolio. As mentioned in the previous section, the price of bitcoin has moved from \$20,000 to \$3,400 and then back to \$19,000 over the span of the last three years. It should be noted that according to Daniel Kahneman in his book “Thinking, fast and slow” humans are more sensible to losses than gains and that as a consequence such big dents in the portfolio can outweigh the gains in the perspective of the investor.

We can also cite the fact that as Bitcoin is not backed by a central authority, it is difficult to retrieve lost funds due to human error or hacks. In the case where the private address of an investor is compromised and all Bitcoins siphoned out of the wallet, it is simply impossible to reverse the transaction due to the very nature of Bitcoin. It is then quite difficult to identify the thief and to retrieve the Bitcoins. The maxim “not your keys, not your coins” is especially interesting in the subject of the security of internet exchanges, where your coins are stored in the wallets of the exchange and as such susceptible to hacks if security problems arise. We can offer the famous case of the hack of MtGox as an example to this point, [link](#).

Bitcoin is usually defined as a “crypto currency” or a “digital currency”. A currency is usually answering the following specific functions :

- A medium of exchange
- A unit of account
- Store of value

If we have a look at the three functions in regards to bitcoin, it becomes quite clear that Bitcoin is not a great medium of exchange. Data from [chainalysis](#) in 2018 indicated that most investors do not use Bitcoin as a medium of exchange but rather see it as an investment tool. The fees linked to every transaction, the fact that the transactions first need to be validated before being effective and the extreme volatility of the price are all factors limiting the usage of bitcoin as a real mainstream medium of exchange. The current pricing of bitcoin makes it also difficult to use as a unit of account. If we want to buy a pizza today with bitcoin, would the price 0.0012 BTC (CHF 20) be self-explanatory to the average consumer ?

In the end research are more or less in consensus to affirm that Bitcoin is more a store of value and as such rather an investment vehicle than a currency (Baur et al. 2015, Yermack 2013).

Integrating BTC into the asset portfolio can be interesting from multiple angles, according to Ria Bhutoria, director of research at fidelity digital assets, [link](#):

- **Potential** : *“Bitcoin is only at its beginning, with its \$197 billion market cap (October 7, 2020), it is only a drop in the bucket compared with markets bitcoin could disrupt.”* Bitcoin represents very high expected risk-adjusted returns
- **Correlation** : Fidelity’s report showcases that bitcoin’s behaviour is decoupled from other assets such as stocks or gold. *“Bitcoin is fundamentally less exposed to the prolonged economic headwinds that other assets will likely face in the next months and years. Combined with its multifaceted narratives and an interesting effect of persisting retail and growing institutional sentiment, it could be a potentially useful and uncorrelated addition to an investor’s portfolio toolkit.”* Indeed, as bitcoin is not issued by a central bank or backed by a government, therefore the monetary policy, inflation rates and economic growth measurements that typically influence the value of a currency simply do not apply to bitcoin.
- **Fees** : *“Alternative investments may be accompanied by fees that reduce the net returns investors receive such as management and performance fees”*. This is not the case with bitcoin as the only fees associated with it are the transaction fees linked to the execution of transactions and the eventual custody fees.
- Other noticeable advantages : Diversification, Liquidity, Accessibility.

Today, market leaders and public figures openly advise most investors to hold at least a small position in Bitcoin, as insurance. Fidelity advises its investors to hold 5% of their portfolio in Bitcoin ([link](#)), a Yale study advises that an optimal portfolio should include at least 6% of BTC ([link](#)). We can also cite Virgin galactic CEO, Chamath Palihapitiya who thinks investors should put a small percentage of their net worth in Bitcoin, around 1%, [link](#).

As a validation of this investment strategy we can witness today more and more asset managers including BTC positions into their portfolio, validating the reasons mentioned above. Here are a few examples :

- Stone Ridge Asset Management acquires \$115 Million of bitcoin, [link](#)
- MicroStrategy raises bitcoin holdings to \$425 Million after second purchase, [link](#)
- Square buys \$50 Million in bitcoin, [link](#)

2 c. Other incentives to invest into crypto currencies and blockchain

We have seen in the previous section how Bitcoin was slowly being accepted as a traditional investment vehicle as even traditional asset managers are slowly building up BTC positions.

When mentioning Bitcoin and cryptocurrencies in the context of investment. We should also at least dedicate a small section to the new investment opportunities offered by blockchain. We discussed already Bitcoin as a store of value and speculative investment vehicle, but what about DeFi ?

DeFi means “decentralized finance”, it is an umbrella term for a wide array of financial applications targeted at using Blockchain and DLT technologies to offer services traditionally offered by centralized financial institutions.

| | Traditional Financial System | Decentralized Finance |
|----------------------|--|---|
| Accessibility | Accessible to clients of the firm only | Accessible with anyone with an internet connection |
| Composability | Siloed products and services | Products can be mixed and matched without any friction and new products can natively be built on top of existing ones |
| Transparency | Products are developed in-house and proprietary. Customers have no ability to see the underlying code nor technology | Everything is open-source. Independent verification and driver of fast technological progress |
| Governance | Decisions taken by centralized structure | Decentralized and transparent decision making by the community (developers and users) |

Source : L1D presentation

The traction in DeFi is absolutely staggering, validating the appetite for alternative ways to use traditional financial tools in a decentralized context.

| DEFI PULSE | Name | Chain | Category | Locked (USD) ▼ | 1 Day % |
|------------|-----------------|----------|-------------|----------------|---------|
| 1. | Maker | Ethereum | Lending | \$2.70B | 4.94% |
| 2. | WBTC | Ethereum | Assets | \$2.26B | 1.84% |
| 3. | Compound | Ethereum | Lending | \$1.53B | 0.83% |
| 4. | Aave | Ethereum | Lending | \$1.43B | 4.01% |
| 5. | Uniswap | Ethereum | DEXes | \$1.20B | 0.28% |
| 6. | Curve Finance | Ethereum | DEXes | \$928.3M | 2.20% |
| 7. | Synthetic | Ethereum | Derivatives | \$760.8M | 8.26% |
| 8. | SushiSwap | Ethereum | DEXes | \$689.2M | 5.76% |
| 9. | Harvest Finance | Ethereum | Assets | \$674.6M | 0.93% |
| 10. | yearn.finance | Ethereum | Assets | \$449.6M | 3.29% |

Figure 5 top 10 Defi Projects

Total Value Locked (USD) in DeFi

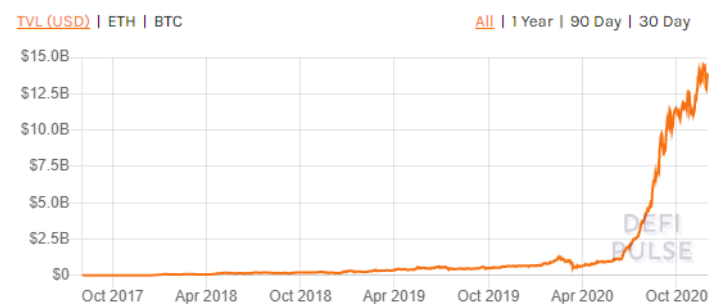


Figure 6 Total value locked in DeFi

In order to make things tangible, let's showcase a few projects in the Defi space that offer concrete applications :

Uniswap : "Uniswap is basically a set of computer programs that run on the Ethereum blockchain and allow for decentralized token swaps. Traders can exchange Ethereum tokens on Uniswap without having to trust anyone with their funds (i.e. exchanges). Meanwhile, anyone can lend their crypto to special reserves called "liquidity pools". In exchange for providing money to these pools, they earn fees (i.e. yield farming)" source : [link](#).

Estimated ETH and UNI Rewards for Eligible Uniswap Liquidity Pools

| Pool | Allocation | Contribution Amount | Est. Pool Share | Est. Daily ETH Rewards (USD) | Est. Daily UNI Rewards (USD) | Est. Total ETH Rewards (USD) | Est. Total UNI Rewards (USD) | Value at End (USD) | Yield | APY |
|--------------|-------------|---------------------|-----------------|------------------------------|------------------------------|------------------------------|------------------------------|-----------------------|--------------|---------------|
| ETH/USDC | 25% | \$250,000 | 0.07% | \$93.17 | \$284.62 | \$4,291.23 | \$13,109.47 | \$267,400.70 | 6.96% | 54.05% |
| ETH/USDT | 25% | \$250,000 | 0.06% | \$133.04 | \$267.64 | \$6,127.69 | \$12,327.65 | \$268,455.34 | 7.38% | 57.33% |
| ETH/DAI | 25% | \$250,000 | 0.08% | \$74.88 | \$343.12 | \$3,448.86 | \$15,804.12 | \$269,252.97 | 7.70% | 59.81% |
| ETH/WBTC | 25% | \$250,000 | 0.06% | \$34.02 | \$255.08 | \$1,566.85 | \$11,749.01 | \$263,315.86 | 5.33% | 41.36% |
| Total | 100% | \$1,000,000 | | \$335.10 | \$1,150.46 | \$15,434.63 | \$52,990.25 | \$1,068,424.88 | 6.84% | 53.14% |

Figure 7 Estimated ETH and UNI rewards

MakerDAO : MakerDAO is a decentralized organization built on Ethereum that allows lending and borrowing of cryptocurrencies without the need of an intermediary. MakerDAO is a set of smart contracts services that manage borrowing and lending through two currencies : DAI and MKR. By combining loans with a stable

currency (DAI), MakerDAO allows anyone to borrow money and reliably predict how much they have to pay back. Lenders are rewarded with interest and borrowers have access to new capital.

As we can see from the two examples mentioned above, the DeFi space is truly offering alternatives to services traditionally trusted by banks. In Fine, let's conclude by mentioning that according to the UBS global family office report from 2019, 57% [of the surveyed family offices] believe blockchain technology will fundamentally change the way we invest in the future.

3 Bitcoin's price

3 a. How is bitcoin priced?

After explaining what is Bitcoin, how does it compares to traditional assets and how blockchain can offer alternatives to traditional investment vehicles, let's focus on the actual valuation of bitcoin and see how can we attribute a value to this currency.

An interesting approach was proposed by Garcia and Al, 2014. The paper considers that the **fundamental value** of a bitcoin equals at least to the cost involved in its production (i.e. mining). This definition has the added value to be completely detached from the market price or speculation. The estimate is given by calculating the total number of SHA-256 hashes needed to mine one bitcoin and then using an approximation of the power requirements for mining those hashes and multiplying it by the price of power. This operation would yield a lower bound estimate of the fundamental value of a bitcoin. If we refer to calculation operated on [Investopedia](#) this amount would be around \$7,494 in Switzerland on 08.03.2018, if we trust [Forbes](#) it was around \$8206.64 in June this year (2020), following the halving.

Another way to approach the problem would be to calculate the **fair value** through the velocity of money. According to Campajola et al, preprint(2019) the standard theory of money velocity assumes that the following equation holds at any point in time :

$$MV = PQ$$

Where :

- M = total money supply in the system
- V = Velocity, the speed of money
- P = the overall price levels
- Q = Transactions, the physical quantity exchanges

In that case the price would be proportional to revenue and not profit, additionally as the price of bitcoin can be very volatile as the velocity changes, self-reinforcing feedback loop between hoarding and higher prices.

If we focus on the **market price** of Bitcoin, then we have to turn our eyes to the exchanges that are trading the crypto currency. The price of Bitcoin is then driven by the supply and demand mechanism of a classic market. When demand for bitcoin increases, the price increases, and when demand falls, the price falls. Let's see below what can impact supply and demand and as such the market price of bitcoin.

Supply :

As mentioned earlier the supply of bitcoin is limited to 21 million coins. New coins are minted at a predictable fixed rate when miners process blocks of transactions, the rate at which those new coins are mined is slowly decreasing with time:

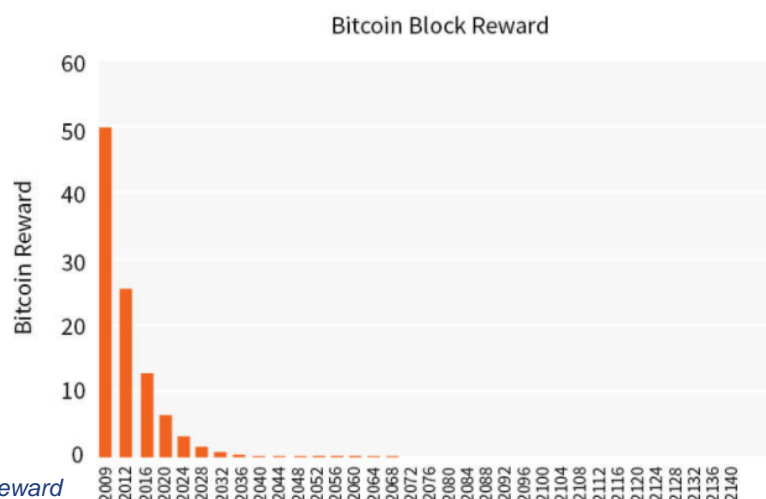


Figure 8 [bitcoin block reward](#)

With every halving (every 210,000 blocks), the reward to mine a full block (fresh bitcoins) is halved. As new coins are minted slower and slower, the demand can actually increase at a faster rate than the supply increases, rocketing the price up.

Demand :

Bitcoin is not the only cryptocurrency out there and as such experiences the heavy competition from other “alt coins”. It is clear that so far Bitcoin is king and is the most well-known and popular crypto currency. Most of the mainstream funds investing in crypto currencies are taking huge positions in Bitcoin for instance. And as such it still enjoys a very heavy dominance, around 63.17% on 30.11.2020 as shown in the graph below.



Figure 9 : [Bitcoin dominance Index](#)

Competition from alternative coins can and has proven to impact the price of Bitcoin, especially during the ICO frenzy of 2017 where the bitcoin dominance has melted from a staggering 90% to less than 40%. What is clear is that after the dust has settled following the crypto crash of early 2018 and we entered the crypto winter, Bitcoin has slowly crawled back to a more dominant position. While tokens were dying left and right, the Bitcoin was claiming back its throne, cementing that in the end Bitcoin was indeed the pillar of crypto.

Demand is also impacted by the availability of the coin on public exchanges, as those are the main channel for mainstream aficionados to acquire the precious coin. The user experience and the overall “easiness” to book purchase transactions and trade bitcoin on those exchanges can represent a barrier to entry to some users. The more it is easy to open an account, transfer FIAT money from your regular bank account to the exchange and buy crypto, the more an exchange can draw additional participants to create a network effect. We can only but acknowledge the efforts that the main exchanges have put in motion in the last years to democratize the access to bitcoins and other cryptocurrencies. We can now buy cryptocurrencies instantly using a credit card or wire money overnight to exchanges to buy bitcoins with fiat in an unprecedented smooth process.

In order to propose bitcoins and crypto currencies to the masses, we cannot oversee the regulations and legal matters impact on the demand. In the last years we have witnessed over and over again how news concerning potential bans of cryptocurrencies had a visible impact on the price of bitcoin. As most regulators are still wrapping their head around the blockchain space and have troubles to come up with a viable and universal regulation it is quite clear that nothing is set in stone concerning the availability of the crypto currencies globally. While it is starting to get quite clear that no country will simply ban blockchain or access to cryptocurrencies in order to avoid a technological lockout, we can fathom how making it more difficult to trade or acquire bitcoin might have an impact on demand.

Finally, it can be assumed from past academic research that there is a link of causality between media headlines, activity on social medias and the price of bitcoin. We will abord this subject more in detail in the following sections so let’s leave it at that for the moment.

As of today, 31.11.2020, the bitcoin is currently priced at USD 19,335.50. Quite near it’s all time high from December 2017.

So in the end, why does bitcoin has any value ? Well, let's cite www.bitcoin.org on that one :

"Bitcoins have value because they are useful as a form of money. Bitcoin has the characteristics of money (durability, portability, fungibility, scarcity, divisibility, and recognizability) based on the properties of mathematics rather than relying on physical properties (like gold and silver) or trust in central authorities (like fiat currencies). In short, Bitcoin is backed by mathematics. With these attributes, all that is required for a form of money to hold value is trust and adoption. In the case of Bitcoin, this can be measured by its growing base of users, merchants, and start-ups. As with all currency, bitcoin's value comes only and directly from people willing to accept them as payment."

3 b. Is it an efficient market?

According to Wikipedia, the efficient-market hypothesis is a hypothesis in financial economics that states that asset prices reflect all available information at an instant t. A direct implication from that is that it is impossible to beat the market consistently on a risk-adjusted basis since market prices have already all known parameters "priced in" and as such should only react to new information.

After reviewing the literature it seems clear that bitcoin and cryptocurrencies should not be considered as an efficient market. Olivier Kraaijeveld and Johannes De Smedt provided a fine overview concerning the nature of the cryptocurrency markets in regards to the efficient market hypothesis :

Ciaian and al. (2018) find that bitcoin's market and other alt coins are heavily interdependent and that the correlation is stronger during the short-term than the long term. Whereas, by definition, in an efficient market successive price changes are completely independent (Fama, 1970).

Additionally, in an efficient market investors are assumed to be rational and give value to an asset based on its fundamental value. However in an article by Silverman and al (2017), an economist states that due to the lack of intrinsic value and the impact of speculation on cryptocurrencies price, there is no possibility for cryptocurrencies to be valued fundamentally, and thus making the market irrational.

We can also observe varying exchanges for Bitcoin from one exchange to another.

At the time that we speak :

- Kraken BTC/USD : \$19,317.1
- Binance BTC/USD : \$19,339.33
- Coinbase BTC/USD : \$19,321.57
- Crypto.com BTC/USD : \$19,332.67

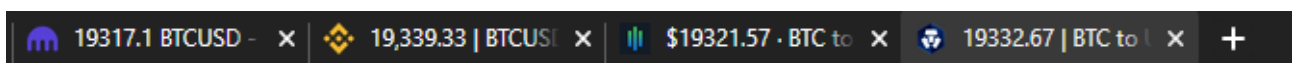


Figure 10 Price of bitcoin on different exchanges

This of course encourages operating arbitrage by simultaneously selling at high and buying at lower prices on different exchanges to profit from the difference in price. Arbitrage is of course a characteristic of inefficient markets. In practice Arbitrage is not always feasible by taking into consideration the transactional costs, delays in filling the orders and the default risk (i.e. MtGox).

Urquhart (2016) concludes his research by affirming that the bitcoin market is inefficient but might be progressively transitioning into a more efficient market after analysing the price of bitcoin over the period 2010 to 2016.

Mensi and al. (2019) also find price patterns between Ethereum and Bitcoin, implying that both currency's markets are inefficient.

Hanlin Yang finds strong price momentum in cryptocurrency prices, which should not happen in an efficient market.

In a nutshell, the consensus still seems to tend towards the fact that bitcoin and cryptocurrencies are not an efficient market.

3 b. Trends, Sentiment and social interactions

Behavioural psychology approaches to the stock market analysis and by analogy, to bitcoin trading, offers a plausible and promising alternative to the EMH. Daniel Kahneman is also sceptic concerning our ability to beat the market *“they are not going to do it. It’s just not going to happen”*, even by using behavioural psychology. But what about Bitcoin? Bitcoin is surely a less mature asset class than traditional finance and as such may be more prone to trends, social interactions and sentiment. In general, behavioural bias leads investors to invest non-rationally and as such can have an impact on the pricing of an asset, especially in a young market where institutional investments are less dominant. Lo, (2012) agrees with this affirmation, stating that *“a relatively new market is likely to be less efficient than a market that has been in existence for decades”*.

The adaptive markets hypothesis, showcased by Lo (2004) is probably more pertinent in the context of cryptocurrencies. In Lo’s opinion EMH is not completely wrong but probably incomplete as it does not fully explain market behaviour and the concept of rationality. Lo states that *“markets are not always efficient, but are usually competitive and adaptive, varying in their degree of efficiency as the environment and investor population change over time”* Lo (2014).

In regards to our document and our work, what is of interest to us is to determine if we can identify a strong correlation between the price of bitcoin and the activity on social medias, social interactions and social factors.

In terms of Media attention, Yannick Zjörjen from the university of Zurich has touched the subject in his work *“Predicting Bitcoin – Gauging the market for bitcoin using web search queries”*.

From an empirical perspective we can see in the quick screenshot below that we can find indeed a pattern in google trends for bitcoin and the price of bitcoin. To be noted that this observation is less visible today as showcased in the introduction section of this document.

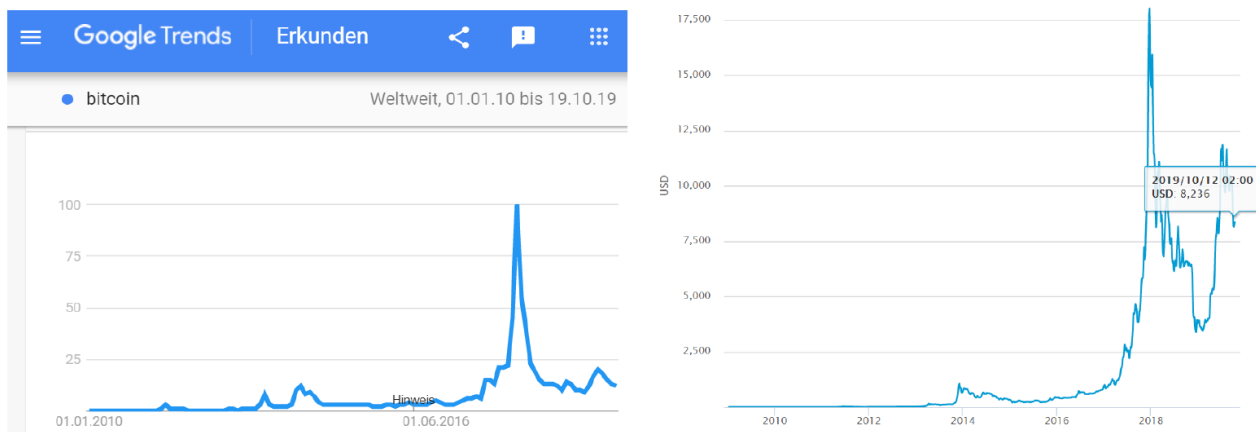


Figure 11 Google trends

Nevertheless, this topic was already heavily studied in the literature.

Engelberg and Parsons (2011) have found evidences that media sentiment has an impact on trading by identifying a link between the sentiment and the trading activity when this link is broken following natural disasters.

Kristoufek (2013) has showed that there exists a strong correlation between the price of bitcoin, daily Wikipedia and weekly search volumes.

Following Kristoufek's study we can cite David Garcia and Al. (2014) that has identified two positive feedback loops that lead to price bubbles in the absence of exogenous stimuli: one driven by word of mouth and the other by new bitcoin adopters. The research group has observed that:

Social cycle: search volume increases with price, word of mouth increases with search volume and price increases with word of mouth. This represents the feedback cycle between social dynamics and price in the bitcoin economy.

User adoption cycle : search volume increases with price, amount of new users increases with search interest and price increases with increases in user adoption.

What is quite surprising is that the team also identified a negative relation between search and price. Indeed 3 of 4 largest daily price drops at that time (2014) were preceded by the 1st, 4th and 8th largest increases in google search volume the day before. It would be interesting to see if this phenomena has continued in the following years.

In the following paper, Garcia and Schweitzer (2015) combined in their observations economic signals of volume and price of Bitcoin, adoption of the bitcoin technology, transaction volume with social signals related to information search, word of mouth volume, emotional valence and opinion polarization from tweets related to bitcoin in order to predict increases or decreases in the price of bitcoin. They then applied these insights and predictions to design and test algorithmic trading strategies to finally confirm the hypothesis that trading based social media sentiment has the potential to yield positive returns on investment.

Kaminsiki and Gloor (2014) affirmed through their study that negative tweets and tweets that express uncertainty correlate moderately positive with the bitcoin trading volume and negatively with the bitcoin price, they finally concluded that Twitter sentiment mirrors but does not predict the market price of bitcoin.

Vytautas Karalevicius and Al. (2017) identified that interaction between media sentiment and the bitcoin price exists, and that there is a tendency for investors to overreact on news in a short period of time through their study.

Olivier Kraaijeveld and Al. (2018) also found that twitter sentiment has predictive power for the returns of bitcoin, bitcoin cash and Litecoin.

In fine, after reviewing the academic literature on the subject, analysing the global sentiment from social medias seems to be a good option to try to predict the price of bitcoin. Twitter seemed to be the most relevant social media source as twitter users often post sentiment infused posts to the community, looking to discuss and to broadcast opinions. In addition twitter possesses a good API in order to retrieve streaming tweets.

4 Sentiment analysis

4 a. Sentiment analysis in a nutshell

As mentioned in the previous section, we have established that there is some kind of correlation between the news and social media and the cryptocurrency market. Probably due to the very nature of the cryptocurrency market and its young age, events and sentiment in news or social media have a tendency to impact positively or negatively the price of bitcoin. It is quite clear then that having the ability to analyse on the fly the global sentiment and key indicators can offer an edge to an intelligent investor. Of course this task is not trivial and requires good knowledge in natural language processing, trained and operational models, custom made for specific fields of finance. This field of research is of course not only applicable to crypto currencies, in the regular financial markets most of the big investment firms and asset managers developed some kind of automatic trading based on AI powered algorithms, by crawling through the earnings reports of companies or acting on specific alerts and indicators. It is quite an interesting field to explore.

Anyway, at its roots sentiment analysis is part of natural language processing, focusing on identifying the sentiment or emotion or “affective state” in textual communication, be it tweets, Facebook posts or whole news articles. Sentiment analysis (SA) is a subdiscipline of deep learning and is itself composed of a multitude of tools and techniques in order to find the right classification for the right data set. In the context of cryptocurrencies we can identify a few types of SA that could provide us with interesting findings.

Polarity: Analysis of a text and determining whether the sentiment is rather negative, neutral or positive. In most of the SA tools that we reviewed the sentiment is classified on a discrete range of values that can for instance go from -1 (very negative) to 0 (neutral) to 1 (positive). As an example, we can take a few tweets from our database as this exactly the type of sentiment analysis that we have used.

- Tweet 1 : "AVOID THIS SCAM!! Bitcoin mining scam will take your money. Avoid working with the following at all costs;... “
- Polarity : -0.9306
- Tweet 2 : " Best performing asset class Bitcoin. Best performing stock Tesla. 2020"
- Polarity: 0.8979

Emotion: The classifier has a set of different emotions such as “happiness”, “sadness”, “anger”, “despair”, “rage” and will try to associate those emotions to a text, giving an idea about the state of emotion the author is in.

Aspect Sentiment: This type of SA goes a bit further and will try to assign a sentiment to specific features or topics in the text. It will break down the text into chunks and will permit to generate more granular insights into it. As such, we can for instance get the general sentiment towards Trump in an article concerning American politics for instance.

For the sake of simplicity and as it seems to be a good trade-off we chose to go with Polarity for this exercise. In order to apply sentiment analysis in the context of cryptocurrencies we need first to identify where to find the data. At first glance the usual sources for this type of exercise are either Reddit or Twitter. As an avid Redditor myself I was first keen to try it but after further investigations I decided to go with Twitter. First because it was a new field for me, as I am not tweeting, secondly because it seemed to be a more relevant source to get valuable personal messages. [Brandwatch](#) reported that there are 330 million monthly active users, 500 million people visit the website each month without logging in, 500 million tweets are sent each day (6,000 tweets per second). Additionally 24.6% of verified accounts are held by journalists and the dominant age bracket of tweeter users is the 25-34 segment as mentioned on [Statista](#). The facts that the majority of the users are young, informed and prone to technology are all parameters that pushed me to take the decision to go with twitter as my data source. If you also add the fact that Twitter proposes a native API for developers to plug in and collect streaming tweets, then it seemed as the go-to solution in order to collect some sentiment on Bitcoin! To further convince us to use twitter as the data source we can cite Tafti et al. (2016) and Li and al. (2017) that affirm that micro blogging platforms such as Twitter are well suited to provide a broad and general live market snapshot. Twitter is able to spread content, sentiment, news before mainstream media have the time to react and publish articles, which makes it even better concerning the analysis of the effect of sentiment in the context of crypto markets.

We will collect the tweets from twitter with a filter to gather only tweets relative to bitcoin and then attempt to assign a polarity rating to each of them.

4 b. Crypto currencies specifics and challenges

We have seen above that in general crypto currencies can be considered as partly irrational, prone to behavioural bias and an inefficient market. In this context it seems clear that it would be possible to use sentiment analysis in order to predict price swings in the price of bitcoin. But in reality implementing a fully functional sentiment analysis classifier and especially in the context of crypto currencies represents a set of challenges and difficulties.

The first one would be probably the fact that even if today it is quite easy to use premade SA libraries without any knowledge in deep learning, those are rather generic et do not extrapolate well into context specific jargon. They can be used fine to get a first glimpse on the general sentiment on simple regular sentences like “I love watching a good movie on a Sunday night” but they are not specialized in particular fields. Specific market terminology is important when trying to get accurate sentiment from text treating of a very specific niche subject. As an example we can cite the CRIX index that provides also a daily tweet sentiment and who has incorporated specific crypto jargon into the calculation of the sentiment to make it more precise

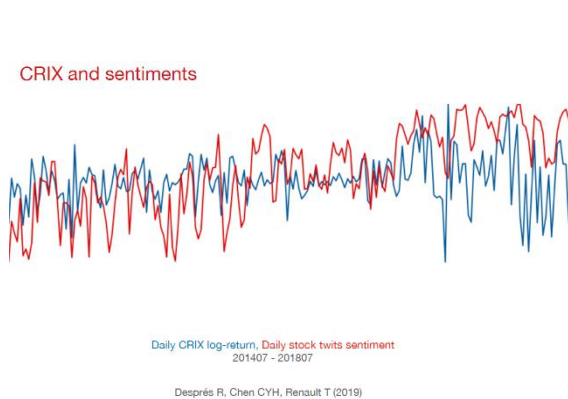




Figure 11 Crix and sentiments

| Term | Sentiment weight |
|---|------------------|
|  hodl | 0.90 |
| hodl ! | 0.54 |
| hackers | 0.85 |
| miner | -0.74 |
| bitcoin  | 0.62 |
| scam | -0.73 |
| fixing scam | -0.77 |
| | -0.86 |

Crypto specific terms

Figure 12 Crix and sentiment

In the literature we can of course cite Loughran and McDonald (2011) who have observed that the performance of a sentiment analysis classifier improves when using a specific dictionary or lexicon.

A second challenge to use sentiment analysis in the context of Bitcoin and Twitter would be the noisiness and the contextualization. Indeed, well written, informative news should be written in a neutral tone and as such should generate a roughly neutral sentiment, on the other hand social media and especially Twitter contain sentiment and heavy polarization but they usually react to already known events. In addition, social medias can produce misleading polarization as they are noisy and highly subjective. The crypto space is rather dynamic and full of unexpected events that polarizes even more the general sentiment pushing it into the extremes and further away from objectivity.

Finally, as Bitcoin and cryptocurrencies are still considered as young technologies, they evolve very fast and as such the jargon and lexical field also evolves rapidly. Terms that were used yesterday might be irrelevant tomorrow and new terms are introduced every day. Who was using “Airdrops” 4 years ago? What is “moon” referring to? Is it the reddit’s crypto currency or the raise in price for another currency? An evolving lexicon is obviously a nightmare for a specialized sentiment analysis machine learning algorithm.

If the goal is to predict the price of bitcoin then using solely sentiment analysis on tweets seems to be too narrow, a more holistic approach encapsulating multiple data sources, capturing all three areas of sentiment analysis described in the previous section would probably grant better and exploitable results. Luckily, we are merely doing an education project and for that purpose sentiment analysis on a twitter feed will be considered as a realistic and interesting goal.

4 c. What exists on the market

In this section, let's have a look at what exists on internet in the field of sentiment analysis for crypto currencies and bitcoin.

Crypto Fear & Greed Index: [Link](#)

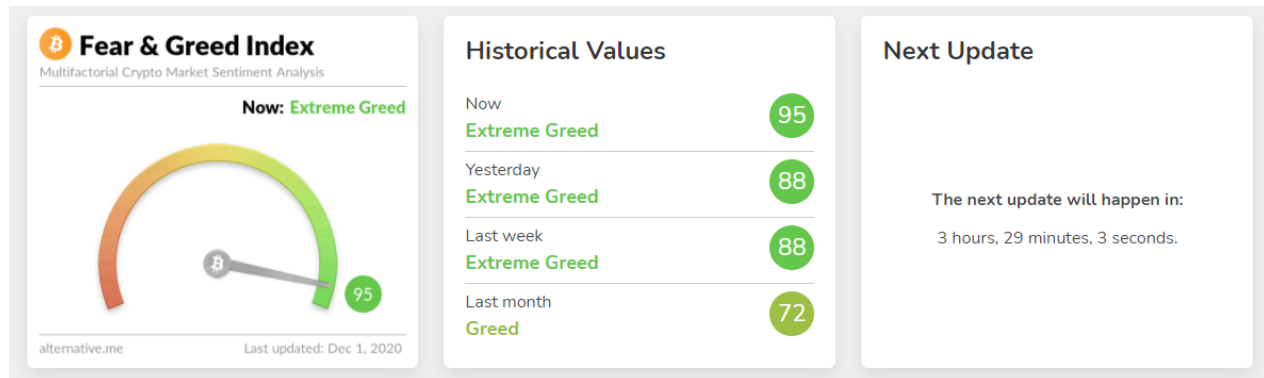


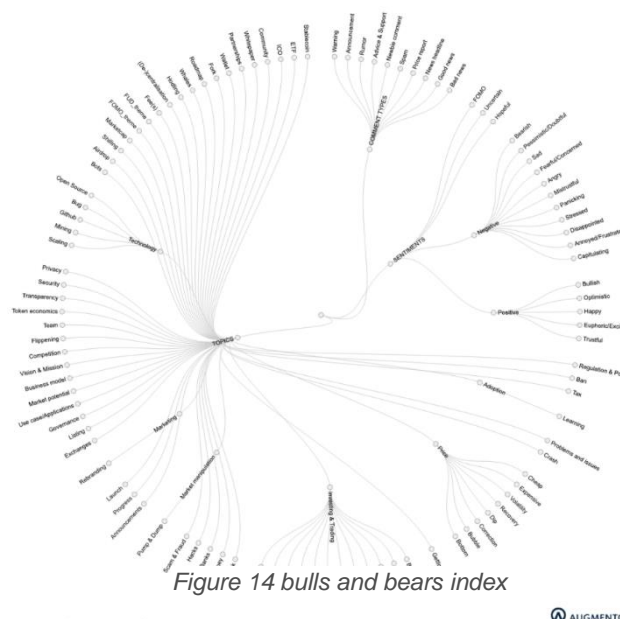
Figure 13 Crypto Fear And Greed index

The crypto fear and greed index is pretty popular on internet, it is quite similar to the [CNN's fear & greed index](#) but measures the sentiment in the crypto market rather than in the stock market. They crunch the numbers from a multitude of data sources into a single score. According to the website, the crypto market is very emotional and as such people tend to get greedy when the price is rising with increasing FOMO (fear of missing out) effect and over-sell coins when the market is in a down trend. According to the meter, in case of extreme fear it can be a good opportunity to buy coins, when the greed is high than that means that a correction is coming. As we can see from the screen shot, we can expect a correction coming soon according to the website.

The data sources used by the crypto fear & Greed index are : volatility (25%), Market momentum/volume (25%), Social Media (15%), Surveys (15%), Dominance (10%) and trends (10%).

Bulls & Bears Index: [Link](#)

The bulls & bears index is measuring social media sentiment to showcase how bullish or bearish social interactions are in the context of bitcoins. The index is collecting data from Twitter, Reddit and Bitcoin talk for



instance and crunch it using a classifier trained on crypto-specific dictionaries. The algorithm used is taking into account 93 different sentiments and topics in order to determine the mood of the market.

Bitcoin Sentiment Index: [Link](#)

The bitcoin sentiment index also crunches posts and interactions on social media platform in order to calculate a bitcoin sentiment index that seems to correlate only moderately to the price of bitcoin if we refer to the charts available on the website. In addition to the bitcoin sentiment the index offers as well the volume of bitcoin's mentions on social media and an indicator concerning the "buzz" around specific cryptocurrencies. Unfortunately, the website does not disclose any explanation on the methodology used to calculate the different indicators and requires a paid plan in order to retrieve live values.



Figure 15 Bitcoin sentiment Index

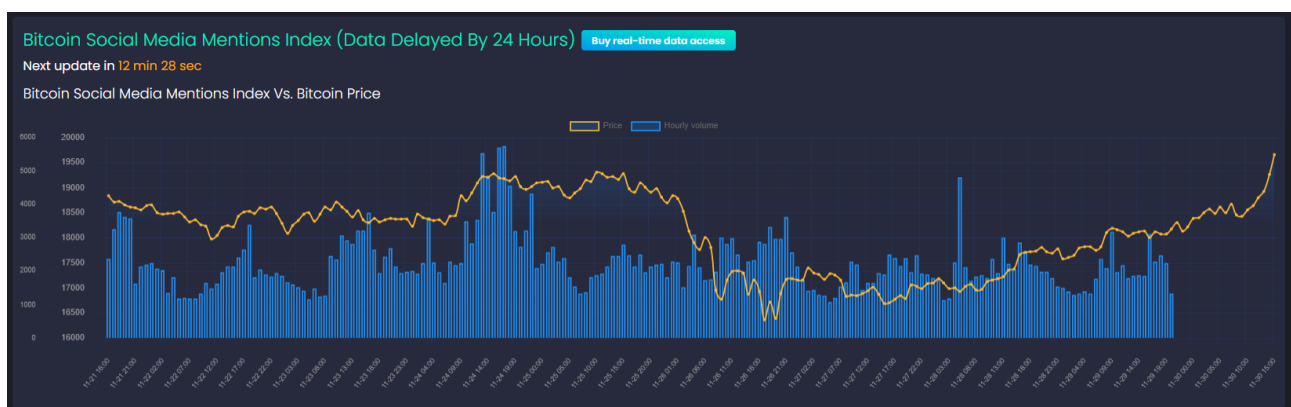


Figure 16 Bitcoin Social media mentions

5 Implementing a sentiment analyser

5 a. Abstract SentCrypt

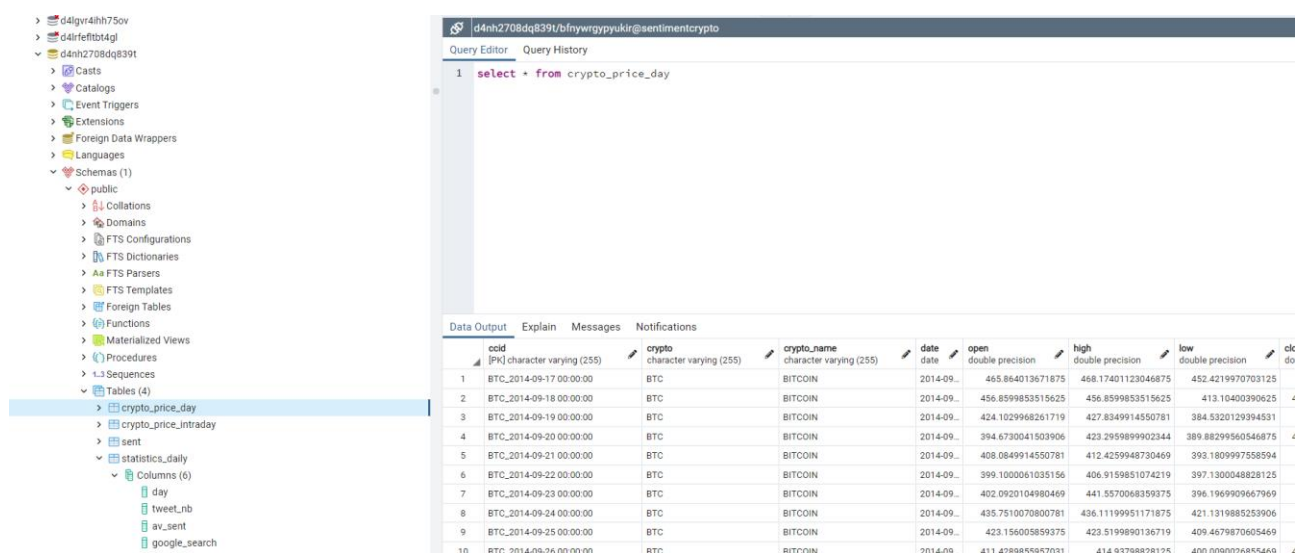
My main objective for the closing project of the CAS in Blockchain at University of Zürich was to build something real and tangible combining two of my hobbies: machine learning and crypto currencies. I wanted to experiment with a fun and real-life project in order to familiarize myself with sentiment analysis and NLP technologies while producing a service or a good related to crypto currencies. I quickly realized that building a website from scratch while experimenting with those new technologies that passionate me was actually an interesting project and would provide me with valuable knowledge that I could eventually leverage as well in my portfolio in the future. I settled for sentiment analysis using the twitter API and the collection of bitcoin prices from APIs to plot them both into a single webpage, everything else is bonus.

I usually learn by doing so I logged in to my GoDaddy account and bought a domain name that was simple, easy to remember and actually quite relevant for the nature of the project: www.sentcrypt.com.

Once I was settled on the name, I had to think about the actual content that was going to be displayed on the website, the technology to be used, the infrastructure and all the data aspects. I was adamant about using python for the backend scripts and the data collection and processing, I decided to give a shot to Heroku concerning the infrastructure and the hosting and finally benchmarked a few different APIs and feeds in the data collection process.

5 b. Data collection

Once the domain name was secured and the minimum infrastructure deployed on Heroku I first created a database. Heroku provides a free DB on PostgreSQL natively to experiment with. The problem is that the free DB is limited to 10,000 rows and 1 GB in storage space. While the storage space is not a big problem, the number of rows is a serious limitation as I am planning on collecting hundreds of thousands of Bitcoin prices and tweets. I started with the free plan just to setup the basics, connect to the DB via the PostgreSQL GUI, create a table, insert a few lines, connect then via a python script and insert lines from there. Everything was working as expected so I upgraded to the Hobby basic plan that offers a comfortable 10,000,000 row limit.



The screenshot shows the PostgreSQL GUI interface. On the left, the database structure is visible, including a table named 'crypto_price_day'. The main window displays a query result for the query 'select * from crypto_price_day'. The result is a table with 10 rows and 10 columns: 'oid', 'crypto', 'crypto_name', 'date', 'open', 'high', 'low', 'close', 'tweet_nb', and 'av_sent'.

| oid | crypto | crypto_name | date | open | high | low | close | tweet_nb | av_sent |
|-----|-------------------------|-------------|------------|-------------------|--------------------|--------------------|-------------------|----------|---------|
| 1 | BTC_2014-09-17 00:00:00 | BTC | 2014-09-17 | 465.864013671875 | 468.17401123046875 | 452.4219970703125 | 465.864013671875 | 0 | 0 |
| 2 | BTC_2014-09-18 00:00:00 | BTC | 2014-09-18 | 456.8599853515625 | 456.8599853515625 | 413.10400390625 | 456.8599853515625 | 0 | 0 |
| 3 | BTC_2014-09-19 00:00:00 | BTC | 2014-09-19 | 424.1029968261719 | 427.8349914550781 | 384.5320129394531 | 424.1029968261719 | 0 | 0 |
| 4 | BTC_2014-09-20 00:00:00 | BTC | 2014-09-20 | 394.6730041503096 | 423.2959899902344 | 389.88299560546875 | 394.6730041503096 | 0 | 0 |
| 5 | BTC_2014-09-21 00:00:00 | BTC | 2014-09-21 | 408.0849914550781 | 412.4259948730469 | 392.1809997558594 | 408.0849914550781 | 0 | 0 |
| 6 | BTC_2014-09-22 00:00:00 | BTC | 2014-09-22 | 399.1000061035156 | 406.9159851074219 | 397.1300048828125 | 399.1000061035156 | 0 | 0 |
| 7 | BTC_2014-09-23 00:00:00 | BTC | 2014-09-23 | 402.0920104980469 | 441.5570068359375 | 396.1969909667969 | 402.0920104980469 | 0 | 0 |
| 8 | BTC_2014-09-24 00:00:00 | BTC | 2014-09-24 | 435.7510070800781 | 436.11199951171875 | 421.1319885253906 | 435.7510070800781 | 0 | 0 |
| 9 | BTC_2014-09-25 00:00:00 | BTC | 2014-09-25 | 423.156005859375 | 423.5199890136719 | 409.4679870605469 | 423.156005859375 | 0 | 0 |
| 10 | BTC_2014-09-26 00:00:00 | BTC | 2014-09-26 | 411.4289855957031 | 414.93798828125 | 400.0090026855469 | 411.4289855957031 | 0 | 0 |

Figure 17 Sentcrypt DB

The DB being installed I then focused on retrieving the prices of Bitcoins. Even here the choice is quite wide, do I want to retrieve only daily prices? Or intraday prices with more granularity, every 5 minutes? 10 minutes? Which API to use and how to make it scalable and robust? And how do I want to retrieve the prices, via an API or a web crawler that would scrap the precious data from the internet? While I was first interested to use beautiful soup to do some scrapping, I then decided to go against it. I wanted the website to

be robust and maintainable, web scrappers might work one day and be patched the next one, potentially requiring manual workarounds and code adaptation. In the process of benchmarking the data acquisition process I actually tried to scrap the data from coinmarketcap.com at first, the code was working the first day but then was not able to retrieve the data the next one, cementing my decision to settle for a more perennial approach using an API.

The first Data I acquired was the closing price of Bitcoin. I first used the coinmarketcap.com API and was surprised to see that the free API does not offer any historical prices. Not even the closure price. So I first collected the closing price from the website directly as it is quite easy to collect it manually from the historical data page for any cryptocurrency, [here](#). While the historical data goes back quite far, what if I miss one day through the API, should I collect the price myself manually and insert it into the DB? Out of the question. So, I had a look at the paid API subscriptions. The price for a hobbyist feed from coinmarketcap.com for personal projects is \$29 per month, for 1 month of historical data. A scandalous price in my opinion.

I then looked for alternatives to collect the closing price of Bitcoin and settled for nomics which offered a free API permitting to swiftly retrieve the daily prices with historical prices as well. The API documentation was clear, the API key was sent right away and I then experimented with Nomics to retrieve the prices and inject them into my PostgreSQL DB on Heroku via python scripts.

The next problem was the fact that even if nomics offered historical daily prices, it did not offer intraday historical prices. I was then entering the phase of the project where I actually was trying to decide what did I want to plot precisely on the website. I realized that I want to plot daily prices, that's fine, but if I want to calculate sentiment on the fly and plot it along the price of bitcoin then I need also a more granular feed called "intraday". Nomics could offer this partly but I wanted to also retrieve past values and secure the access to at least a few days back of intraday values to palliate the risk of a deficient script.

I was also experimenting in parallel the retrieval and plotting of stock tickers such as MSFT, TSL or the SP500 in order to build a custom dashboard to follow my own stocks and was actually also including BTC and ETH in the dashboard:

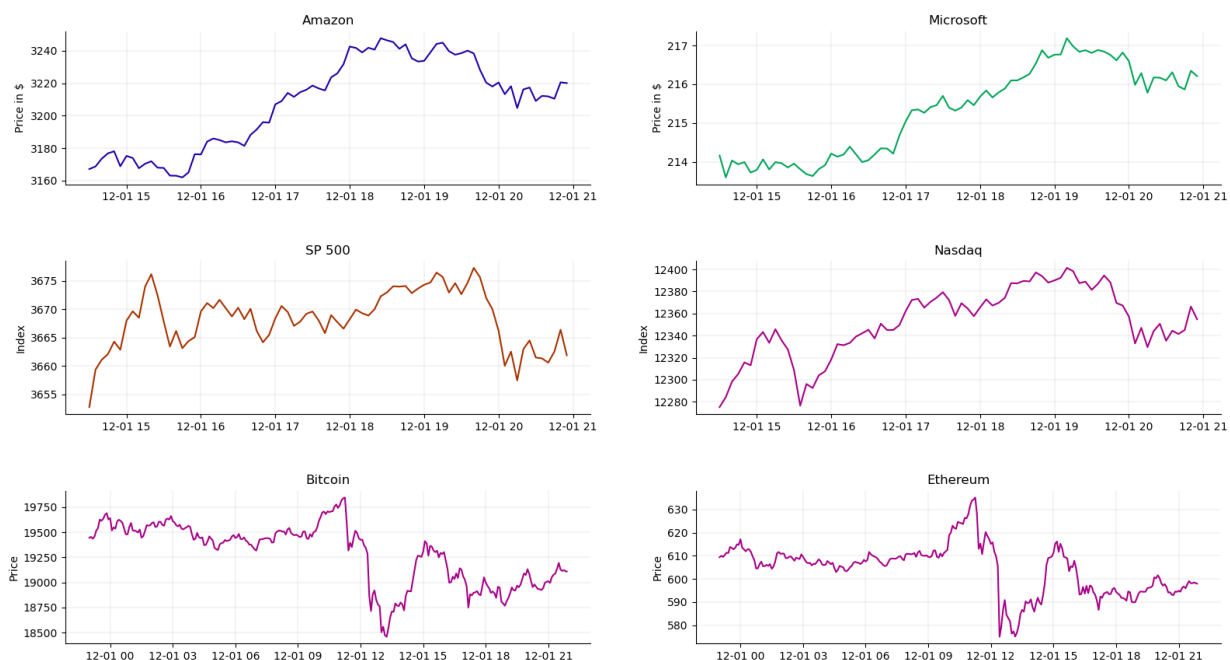


Figure 18 Tickers

I then realized that I was actually using the yahoo finance API to retrieve the data for my tickers, which as a matter of fact contains intraday and best of all contain the last 30 days of intraday data for bitcoin, free of charge. The API is easy to use and requires only to pip install the yfinance library. Retrieving intraday prices for bitcoin is literally done in two lines of code:

```
bitcoin = yf.Ticker("BTC-USD")
```

```
hist_bitcoin = bitcoin.history(start=date_today, interval="5m" )
```

yahoo finance was a revelation and I quickly modified my closing + high + low daily price code to retrieve data from this source. I now had also a clean feed for my intraday data. With an available backlog of 30 days it is quite easy to retrieve eventual missing data in case of service interruption or network problems. But how to retrieve the data older than 30 days? I could of course only settle to go back 30 days in time and start to gather data from that point in time. But I wasn't really comfortable with that, I wanted to build a fully functional and complete database containing closing, low, high and intraday prices of Bitcoin. I searched on internet for days to retrieve the data and finally bought it from fiverr from a user named "data_dealer". This gentleman sold me 7 years of clean intraday data for Bitcoin for CHF 9. I was curious to know how did he manage to get it, and he was kind enough to provide the explanation. He was scrapping it directly from the graphs on coinmarketcap.com. In any case, I loaded the data and picked up from there with the yahoo finance feed to go forward. Here is the link to the yahoo finance API: [link](#).

My script could run and just pick up the last prices available with a timer, querying the API every x seconds and then sleeping. But running a script indefinitely is not really a best practice. So, I settled for script that will be launched every 5 minutes, retrieve the data that was published since last run and insert it into the DB. It would also check at every occurrence if a price for the day before is present in the DB and if not would query it and insert it as well to make sure that we maintain both the daily prices and the intraday prices.

Concerning the tweets there is a limited amount of options. I can either try to scrap the data via specific searches or I can request a Twitter developer API and start from there. I was interested to try the streaming API from twitter so I decided to go that way. The developer keys are not granted instantly so I had to wait a few days after making the request here: [link](#). Once the keys were received I searched for a clean way to query the API and retrieve the tweets. The market solution seems to be [tweepy](#), which is simply described as "An easy-to-use Python library for accessing the Twitter API".

In order to start receiving the tweets, I had to fill the API credentials from Twitter, create a listener class in order to receive in input the streaming tweets and then open the connection with our credentials. The tweets are received in the form of a json file and have to be opened. An example of such a tweet is provided in annex. I also signified a tracking options to only retrieve tweets with the word "bitcoin" included in order to only retrieve relevant tweets.

5 c. Applying sentiment analysis

Once the tweets are being correctly parsed in our listener and streamed live from the twitter API. What is left to do is actually to retrieve the text of the tweet itself and apply sentiment analysis on it. I could build our own sentiment analyser using one the following technics:

- Supervised machine learning
- Lexicon based
- Hybrid (ML+LB)
- Graph based approach

For the sake of time I decided to go rather with an out of the box solution leveraging one of the available library solutions in python. While researching the subject two libraries were standing out and appeared to be the standard:

1) [TextBlob](#) : Simplified Text Processing

TextBlob is a library used for the processing of textual data. It offers a simple and straight forward API to dive into common NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation. Here is the complete list of features natively available in the library:

Features

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis

- Classification (Naive Bayes, Decision Tree)
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- n-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions
- WordNet integration

2) [Vader sentiment](#): Valence Aware Dictionary and sentiment Reasoner

This is a lexicon and rule-based sentiment analysis library that is especially sensible to sentiments expressed in social media. From the GitHub repo we can read. *“It Implements the grammatical and syntactical rules described in the paper, incorporating empirically derived quantifications for the impact of each rule on the perceived intensity of sentiment in sentence-level text. Importantly, these heuristics go beyond what would normally be captured in a typical bag-of-words model. They incorporate **word-order sensitive relationships** between terms. For example, degree modifiers (also called intensifiers, booster words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity.”*

The [GitHub](#) offers as well a lot of examples and a good explanation concerning the scoring mechanism.

Vader outputs the sentiment in the following way:

```
World:{'neg': 0.026, 'neu': 0.492, 'pos': 0.482, 'compound': 0.9798}
```

The compound score is obtained by summing the valence scores of each word in the lexicon and then adjusted to the rules and then finally normalized to be between -1 (extreme negative) to +1 (extreme positive). It is the most straight forward metric to use for a text if we want a unidimensional output.

In general, to be usable, the consensus is to classify as such the score:

- positive sentiment: compound score ≥ 0.05
- neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
- negative sentiment: compound score ≤ -0.05

At first sight it seems quite natural to go with Vader as I will be exclusively calculating sentiment analysis on social media content. After conducting additional research, Vader sentiment appears to be also better at tackling slang and emojis. Cementing my first opinion. It would be interesting to benchmark both libraries on our dataset of tweets.

In practice using Vader sentiment is extremely simple. After installing the library via pip install and importing it, the only thing to do is call it on a text:

```
vs = analyzer.polarity_scores(tweet)
sentiment = vs['compound']
```

And that's it! Retrieval of the sentiment and storage done.

5 d. Storage and Archiving

Now that we have tackled the data retrieval and the application of the sentiment analysis library the challenge is to store it in a smart way and industrialize it to limit the manual work involved. With our Heroku database we are limited to 10'000'000 records, and while it can appear to be a high number, it is not that high.

We are already using 800'000 rows for our intraday historical bitcoin prices, and we are storing 288 additional rows per day. We also have also 2500 records of daily prices and growing (albeit slowly). So, the Bitcoin part seems to be under control and we do not have to worry about it for the next two years if we settle for a limit of around 1 million rows for our crypto currency.

But what about the tweets? Based on a quick benchmark I am receiving around 2500 tweets per hour, that's 60'000 tweets per day. How am I supposed to first store them in a reliable way and plot them on the website? It seems quite tedious to do.

The way I want to plot the data on the website is also a parameter to take into account while designing the way we are going to store it. And vice versa. Both are quite interconnected and a global solution and vision has to be established before releasing anything into production. After careful consideration, here is the solution that appeared to be the most pragmatic:

Display:

- 1) Daily price of bitcoin with high, low and closure values for every day for the last few years. Updated daily to incorporate the values for the value of the previous day. Interval of one day between two values
- 2) Intraday price of bitcoin showing the daily price of bitcoin with a granularity of 5 minutes intervals
- 3) Daily sentiment aggregate, showing for every past day the average daily sentiment, no history available at the moment but the historical values are built with our scripts daily
- 4) Intraday average sentiment on rolling window of 5 minutes. The graph is updated every 5 minutes with an average value of the tweets streamed in this interval
- 5) A table displaying the last streaming tweets with the polarity and a colour code to identify the negative and positive ones

Storing:

- 1) One daily table containing the daily prices of bitcoin, one row per day
- 2) One intraday table of prices for bitcoin, one row every 5 minutes
- 3) One intraday tweets table, storing all the tweets streamed into the API + the calculated sentiment. Deletion of all the entries older than 30 days on a daily basis
- 4) One daily table containing an average of the sentiment of the past day, calculated daily

This way the data is clearly segmented, we can query the tables to retrieve current tweets being streamed, we can calculate and display the rolling average sentiment and we can as well plot the daily and intraday prices of bitcoin.

Automatizing and scheduling all those scripts is a challenge but once in place everything should run smoothly and with minimal supervision.

5 g. Infrastructure and deployment

Concerning the deployment and the tools to be used by the website I chose the following options :

Hosting/Deployment: [Heroku](#)

The name comes from "Heroic" and "Haiku", offering a shout out to "Matz" the Japanese creator of the Ruby language. Heroku is a cloud platform as a service, it is a container-based cloud platform built for developers to deploy, manage and scale modern applications. It simply offers an easy to use framework to deploy webapps quickly. It is supposed to be beginner friendly, and while I had no big problems using it, I found some of the processes quite complex. I was for instance expecting the upgrade of a database to be a one button operation, whereas it actually required bash/shell scripts to be run on my side and manual copy of the data from the initial free DB to the new paid DB. I hesitated to try to deploy the apps on a regular cloud provider, but Heroku seemed more "out-of-the box" oriented and seemed more adapted for personal projects. I am the developer, the change management and the project manager, I cannot handle a DevOps hat in addition!

The products offered by Heroku are:

- Platform
- Postgres
- Redis
- Teams
- Enterprise

- Connect
- Elements

I used Platform to deploy our Python scripts and Postgres to store the data.

Web Development: [Flask](#)

Flask is a micro web framework in python, it is called a microframework because it does not require any additional libraries or technologies to build a web application. The application can take the form of web pages similar to the what we are trying to build at www.sentcrypt.com or a blog, a wiki or even web based calendar or project management tools. The main features of flask are :

- Built-in development server, fast debugger
- Integrated support for unit testing
- RESTful request dispatching
- Jinja2 Templating
- Support of secure cookies
- Lightweight and modular design allows for a flexible framework

A possible alternative to Flask would have been Django. But as I am a beginner in the domain I preferred to stick to something simpler. Django is a full featured framework and as such it comes with tons of features out-of-the box, I prefer a more iterative approach here. Starting with a bare metal skeleton and expanding on top of it with additional libraries when needed. It is nice to have all features included, but especially in “simple” project as ours we don’t need all those fancy tools. Flask offered us simplicity and flexibility.

Plotting: [Bokeh](#)

Bokeh is a visualization library in python, it is especially adapted for building data applications, it offers a wide range of visualization options and chart possibilities. As we will be plotting data concerning prices and time series about sentiment, having a performant way to show it is paramount. I was considering using dash as this is another dashboard-building popular option.

All the coding was done in Python with a vast array of libraries that you can access in the code overview in annex.

Handling authentication

An interesting point to raise is the management of the credentials to handle the connections to the different APIs and to the database and how to industrialize it without hardcoding the keys into the code. When you are working with Heroku or with Github it can be quite tempting to upload your code with hardcoded access keys in order to make it work directly, it is just faster. But this approach has two direct flaws :

The first one is that from a confidentiality perspective this is just plain bad. There are crawlers on Github scanning all the public repos to gather exactly this, API keys, that are then sold on the black market for other people to use. In general it is best practice to decouple your code and your credentials, both should be stored in two different points and your code should access those centralized credentials via a config parser.

The second reason is that it makes your code more flexible and maintainable, in the future if you have hundreds of files accessing those credentials you don’t want to have to change those in every single file in case those credentials are updated. You want to be able to do it in one single file that all your code is accessing.

You can do that by using a `ConfigParser` class which implements a basic configuration language which provides a structure similar to what’s found in Microsoft Windows INI files. You can use this to write Python programs which can be customized by end users easily.

Here is an example of my credentials.ini file (without real keys obviously) :

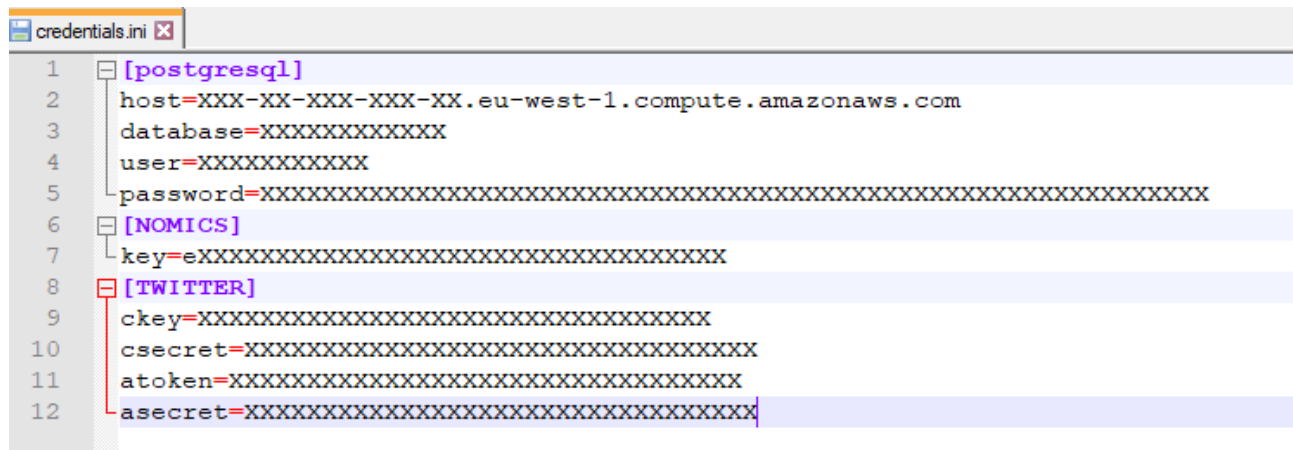


Figure 12 Credentials.ini example

And here is a sample of code accessing the credentials :

```
#Get Connection details in .ini file for POSTGRES
def config_POSTGRES(filename='credentials.ini', section='postgresql'):
    # create a parser
    parser = ConfigParser()
    # read config file
    parser.read(filename)
    # get section, default to postgresql
    db = {}
    if parser.has_section(section):
        params = parser.items(section)
        for param in params:
            db[param[0]] = param[1]
    else:
        raise Exception('Section {0} not found in the {1} file'.format(section,
filename))
    return db
```

Using the configparser I simply access the credentials.ini file and then access the section that is relevant, in that case the credentials for the database, and retrieve the values !

5 e. Optimizing the data

Once the data pipeline is running smoothly and collecting all the information that I needed it was time to start to work on the data itself. No additional effort was to be done on the bitcoin prices for now, we might expand at a later stage to gather as well other metrics concerning bitcoin such as :

- Number of wallets
- Number of transactions
- Wikipedia reads
- Google searches
- Mentions in mainstream media

We might as well expand the scope of our project to include other crypto currencies but for now we are ok on that side.

It is less the case concerning the tweets. We are calculating a sentiment on text, the calculation is quite subjective and highly depends on the quality of the data and on the processing. After reviewing samples of the tweets in the database it was very clear that the calculation of the sentiment was polluted by various factors and that as such the sentiment was not reliable enough.

| | id_tweet [PK] character varying (255) | date timestamp without time zone | tweet character varying (500) | sentiment character varying (255) |
|----|--|-------------------------------------|--|--------------------------------------|
| 1 | 1332249247765303297 | 2020-11-27 10:05:47.318 | Great project | 0.6249 |
| 2 | 1332249261832949761 | 2020-11-27 10:05:50.672 | Good project | 0.4404 |
| 3 | 1332249302408646656 | 2020-11-27 10:06:00.346 | 2) Top 10 Underlying Tokens for | 0.6037 |
| 4 | 1332249304115916807 | 2020-11-27 10:06:00.753 | How's everyone feeling about #bitcoin today? | 0.128 |
| 5 | 1332249319009824769 | 2020-11-27 10:06:04.304 | Yup, LTF scenario in play. Missed the move. Onto the next one :) | 0.4939 |
| 6 | 1332249345782124544 | 2020-11-27 10:06:10.687 | Right. Have you considered hard money as a solution? Gold has failed in the past at doi... | -0.34 |
| 7 | 1332249346666991618 | 2020-11-27 10:06:10.898 | Indeed : | 0.1779 |
| 8 | 1332249368150196225 | 2020-11-27 10:06:16.02 | At first I thought started accepting Bitcoin | 0.3818 |
| 9 | 1332249392955310081 | 2020-11-27 10:06:21.934 | best project on 2020,must join in this legit project | 0.7506 |
| 10 | 1332249407610368002 | 2020-11-27 10:06:25.428 | people interested in #crypto should definitely watch this! nice conversation between an... | 0.8172 |
| 11 | 1332249407610368002 | 2020-11-27 10:06:44.607 | Dollar cost averaging (DCA) is the best way to invest in #Bitcoin. Why? | 0.6260 |

Figure 13 Tweet extract from the database

My first observation was that it was hard for me to benchmark the quality of the sentiment analysis with tweets in languages I did not understand. Vader sentiment is “supposed” to work alright with other languages than English but to avoid any problems I preferred to filter out all the tweets which are not in English. This can be done quite easily as the language of the tweet is part of the API. It was only a matter of identifying the right field.

After filtering the language I also decided to filter out all the tweets by accounts with less than 50 followers. The threshold of 50 is completely arbitrary on my part but I figured that this would filter out at least part of the bots and noise. As we are trying to do sentiment analysis in the context of bitcoin’s price I also considered that smaller accounts with less followers do not hold the same reach as bigger accounts and that as such the sentiment of their tweets is probably less impactful.

By reviewing the tweets streaming into the database I also observed that I was getting a high number of retweets. Especially in the crypto space twitter is full of contests and people trying to win some obscure alt coins by retweeting. This noise is particularly lacking any value and hence I also filtered out all the retweets. I did this by taking out all the tweets where the first two letters are “RT”.

Finally I also implemented changes in the tweets themselves. I decide to remove all mentions of other users. On twitter you can tag another user with “@”. It is self-explanatory that user names are not relevant in the context of sentiment analysis so I just deleted the strings starting with “@”. I then also deleted all the URLs from the tweets as those did not add much to the content and finally I also removed all the special characters (except #) and punctuations in order to further reduce the noise. Hashtags do hold value in the context of this study so I decided to keep those in.

Let’s have a look at a few tweets and the calculated sentiment :

"#BTC could drop to \$16,000 or possibly to the \$13,800 level. However, the probability for a drop to the \$13,000 r... " **-0.4939**

"I'm buying bitcoin in bulk 😊 At the best rates 🚀 You will be amazed by the awesome offers...I pay fast #dontmissout " **0.9022**

"What are you afraid of? Invest today and see your life changing!! 100% guarantee of profit earning! DM ME NOW! 📞 📧 g... " **0.7243**

"NOO WAY!!!! BITCOIN READY TO SHOCK US AGAIN!!!! [WATCH BEFORE MONDAY] AL... via " **-0.3111**

"Congrats you have won \$33 in Bitcoin #BSV 🎉 Pickaw seed: Cd55adhDu3RxEYow" **0.7964**

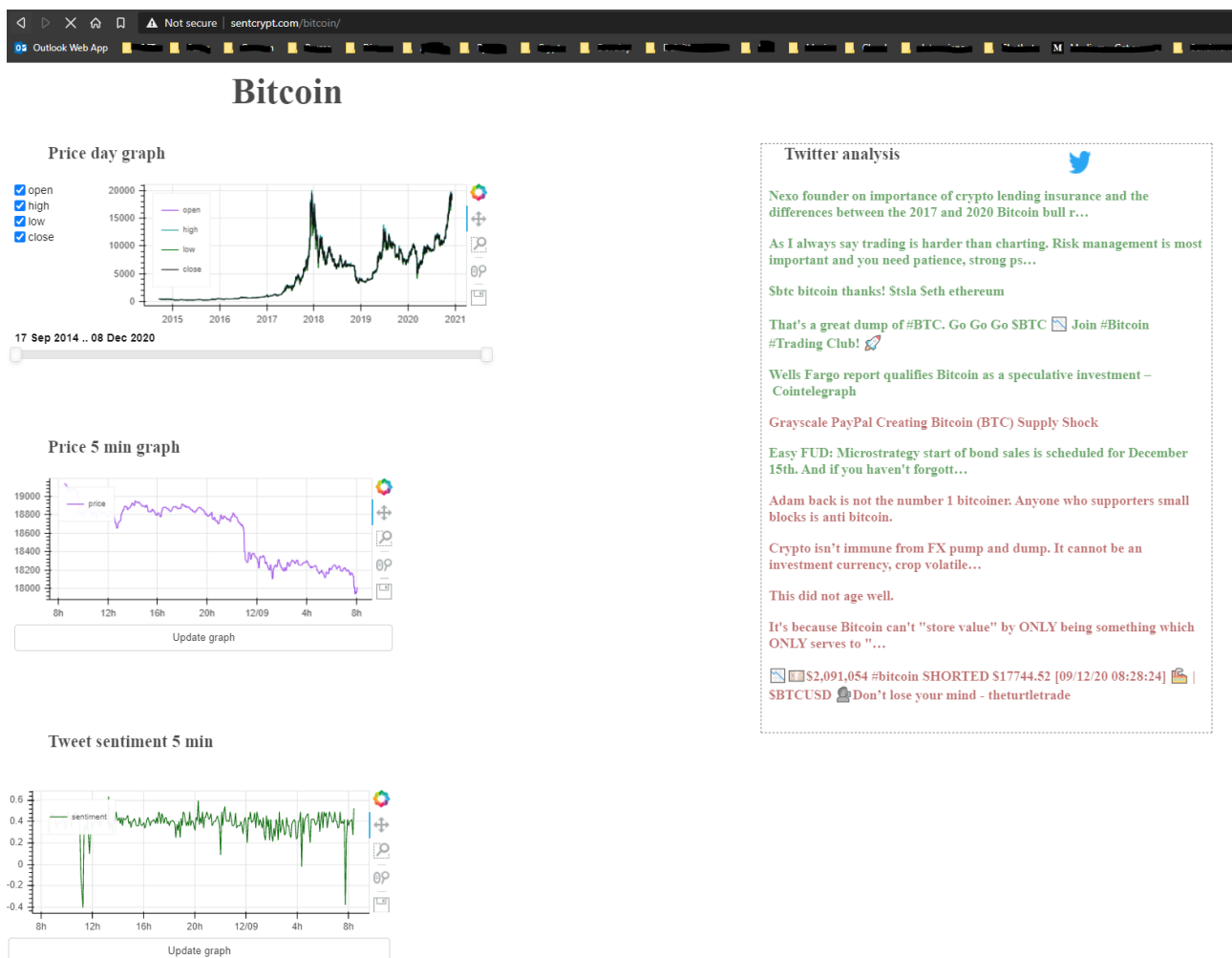
As we can observe easily, the sentiment analysis is working alright, the general sentiment is correctly estimated in my opinion. But the tweets are rarely relevant. On this small sample that I took randomly from the table, 3 tweets are obviously self-interested scams or shills. It is therefore fair to ask ourselves the question if the analysis of such tweets has any added value to determine the future price of bitcoin or if this noise is only bad luck in our random walk.

6 Results and conclusions

The website is still under construction but is already capable of displaying most of the information that I want it to display. It is showcasing:

- Historical daily closing, high and low price of bitcoin (updated daily)
- Historical intraday price of bitcoin (updated every 5 minutes)
- Sentiment of today on a rolling window of 5 minutes, (updated every 5 minutes)
- Sample of tweets being streamed with the calculated sentiment (updated every refresh). Green tweets are positive and red tweets are negative
- (future work) : historical daily sentiment aggregate
- (future work) : additional metrics
- (future work) : Live calculation of VAR and correlation between all values

Here is a snapshot from the current state of the website :



We demonstrated in the first sections of this work that in general the literature has been going towards the idea that there is a clear link and correlation between tweets and other social medias indicators and the price of bitcoin. In our work we were not able to go as far as to benchmark the nature of this relationship. Indeed we have been more focused on the building of a website, the infrastructure and the industrialization of the whole process rather than analysing the data and trying to get insights from it. But now that all the foundations have been led and that the data is being streamed and acquired in a 100% automatic way, we will be able to use it and try to see if we can indeed confirm the hypothetical link between the price of Bitcoin and twitter sentiment.

From a technical perspective we were able to solve all our problems and challenges, from the hosting, data pipelines, DB admin, continuous deployment and integration. But we have clearly identified limitation factors concerning the quality of the data. We identified some problems in the previous section of this document, let's elaborate further.

The twitter API only enables to reach 1% of the total volume of tweets going through twitter. This in itself is a limitation concerning what we can retrieve for the famous micro blogging website. By only reaching 1% of the volume we miss a lot of the content and it is quite clear that we are subject to RNG and there is a possibility that we get highly skewed data. In addition, tweets are only 280 char long and usually contains hashtags, tagging, URLs, emojis and various "fillers" that further limit the number of available relevant words for analysis.

We observed the fact that in terms of quality tweets have limited value. Bots (estimated to be around 1-14% of total volume according to Olivier Kraaijeveld 2018), spam, contests are representing significant noise.

Specific to sentiment analysis, a fast evolving jargon and lexicon, consequence of the lack of maturity of the technology and the fact that it is sometimes difficult to quantify affective states and subjective information without proper interpretation of emojis represent challenges to efficiently extract an accurate sentiment benchmark.

We will cover this part of the analysis as a next step of this work.

7 Annex

7.a Tweet example (Sample)

```
{
'created_at': 'Tue Nov 17 09:15:35 +0000 2020',
'id': 1328627836475871232,
'id_str': '1328627836475871232',
'text': 'Bitcoin Marking New Highs, Game of Thrones Actor Attracted Towards Crypto
Space\n\nhttps://t.co/NozKNwjk4i',
'source': '<a href="https://mobile.twitter.com" rel="nofollow">Twitter Web App</a>',
'truncated': False,
'in_reply_to_status_id': None,
'in_reply_to_status_id_str': None,
'in_reply_to_user_id': None,
'in_reply_to_user_id_str': None,
'in_reply_to_screen_name': None,
'user': {
'id': 2235329532,
'id_str': '2235329532',
'name': 'John Morgan',
'screen_name': 'johnmorganFL',
'location': 'For now, Earth',
'url': None,
'description': 'Even when he has a 50/50 shot, the odds are 80/20 in his favor 🚀',
'translator_type': 'none',
'protected': False,
'verified': False,
'followers_count': 18360,
'friends_count': 17112,
'listed_count': 48,
'favourites_count': 123,
'statuses_count': 63811,
'created_at': 'Sun Dec 08 02:06:19 +0000 2013',
'utc_offset': None,
'time_zone': None,
'geo_enabled': False,
'lang': None,
'contributors_enabled': False,
```

```
'is_translator': False,
'profile_background_color': '000000',
'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme9/bg.gif',
'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme9/bg.gif',
'profile_background_tile': False,
'profile_link_color': '2FC2EF',
'profile_sidebar_border_color': '000000',
'profile_sidebar_fill_color': '000000',
'profile_text_color': '000000',
'profile_use_background_image': False,
'profile_image_url': 'http://pbs.twimg.com/profile_images/1326671287641665537/BfJb673P_normal.jpg',
'profile_image_url_https':
'https://pbs.twimg.com/profile_images/1326671287641665537/BfJb673P_normal.jpg',
'profile_banner_url': 'https://pbs.twimg.com/profile_banners/2235329532/1388183287',
'default_profile': False,
'default_profile_image': False,
'following': None,
'follow_request_sent': None,
'notifications': None},
'geo': None,
'coordinates': None,
'place': None,
'contributors': None,
'is_quote_status': False,
'quote_count': 0,
'reply_count': 0,
'retweet_count': 0,
'favorite_count': 0,
'entities': {'hashtags': [], 'urls': [{'url': 'https://t.co/NozKNWjk4i', 'expanded_url':
'https://coinpedia.org/news/game-of-thrones-actor-attracted-towards-bitcoin/', 'display_url':
'coinpedia.org/news/game-of-t...', 'indices': [81, 104]}], 'user_mentions': [], 'symbols': [], 'favorited': False,
'retweeted': False, 'possibly_sensitive': False, 'filter_level': 'low', 'lang': 'en', 'timestamp_ms':
'1605604535584'}
```

1605604535584 Bitcoin Marking New Highs, Game of Thrones Actor Attracted Towards Crypto Space

7 b. Code prices

```
#/*****
#Nom ..... : Insert Streaming BTC Price Into postgres_v02.py
#Context .....: Natural language processing and Crypto Prices
#Role .....: Get BTC prices and insert in DB
```

```
#Auteur ..... : JDO
#Version ..... : V1
#Date ..... : 09.12.2020
#Language : Python
#Version : 3.7.8
#*****/
#*****/

from datetime import date, timedelta, datetime
import time
import json
from unicode import unicode
import time
import sqlite3
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import psychopg2
from configparser import ConfigParser
import yfinance as yf
from pytz import timezone
import pytz

#Dates Variables
date_today = date.today()
date_yesterday = date.today() - timedelta(days=1)
#Get BITCOIN TICKER from yahoo finance
bitcoin = yf.Ticker("BTC-USD")
#Update table with missing daily prices for BITCOIN
COUNTER_INSERT = 0

#Function to get Connection details in .ini file for POSTGRES DB
def config_POSTGRES(filename='credentials.ini', section='postgresql'):
    # create a parser
    parser = ConfigParser()
    # read config file
    parser.read(filename)
    # get section, default to postgresql
    db = {}
    if parser.has_section(section):
        params = parser.items(section)
        for param in params:
            db[param[0]] = param[1]
    else:
        raise Exception('Section {0} not found in the {1} file'.format(section,
filename))
    return db

#Get Connection details in .ini file for POSTGRES DB and open connection
conn = None
params = config_POSTGRES()
print('Connecting to the PostgreSQL database...')
conn = psychopg2.connect(**params)
cur = conn.cursor()

#----- UPDATE CRYPTO_PRICE_DAY TABLE -----
#if we need all history we can use period = max
hist_bitcoin_daily = bitcoin.history(period="max")

#If not let's get the last record from the table
cur.execute("SELECT MAX(date) FROM crypto_price_day;")
x=cur.fetchone()

#We query the potential missing data from yahoo finance
```



```

hist_bitcoin_daily = bitcoin.history(start=x[0],end=date_yesterday)

#Then we loop over the data retrieved from yahoo finance
for index, row in hist_bitcoin_daily.iterrows():
    ccid='BTC'+ '_' +str(index)
    cur.execute("SELECT * FROM crypto_price_day where ccid=%s;",(ccid,))
    x=cur.fetchone()
    #If no record concerning the retrieved row then we insert
    if x is None:
        cur.execute("INSERT INTO crypto_price_day (ccid, crypto, crypto_name,
date, open, high, low, close) VALUES (%s, %s, %s, %s, %s, %s, %s,
%s)",(ccid,'BTC','BITCOIN',index,row['Open'],row['High'],row['Low'],row['Close']
))
        conn.commit()
        COUNTER_INSERT += 1
    else:
        continue

#Print metrics
print('-----')
print('FINISHED CRYPTO_PRICE_DAY UPDATE. INSERTED %s rows'%(COUNTER_INSERT))

#----- UPDATE CRYPTO_PRICE_INTRADAY TABLE -----
-
#Reinitialize counter
COUNTER_INSERT = 0

#let's get the last record from the table
cur.execute("SELECT MAX(date) FROM crypto_price_intraday;")
x=cur.fetchone()

#We query the potential missing data from yahoo finance
hist_bitcoin_intraday = bitcoin.history(start=x[0],interval="5m")

#Then we loop over the data retrieved from yahoo finance
for index, row in hist_bitcoin_intraday.iterrows():
    ccid='BTC'+ '_' +str(index)
    cur.execute("SELECT * FROM crypto_price_intraday where ccid=%s;",(ccid,))
    x=cur.fetchone()
    #If no record we insert into the table
    if x is None:
        cur.execute("INSERT INTO crypto_price_intraday (ccid, crypto,
crypto_name, date, price) VALUES (%s, %s, %s, %s,
%s)",(ccid,'BTC','BITCOIN',index,row['Close']))
        conn.commit()
        COUNTER_INSERT += 1
    else:
        continue

#Print metrics
print('-----')
print('FINISHED CRYPTO_PRICE_INTRADAY UPDATE. INSERTED %s
rows'%(COUNTER_INSERT))

#----- STREAMING UPDATES -----
print('-----')
print('ENTERING STREAMING')

while True:
    #now = datetime.now()- timedelta(hours=1)
    now = datetime.now()
    current_time = now.strftime("%H:%M:%S")
    #CRYPTO_PRICE_INTRADAY Streaming update

```

```

hist_bitcoin_intraday = bitcoin.history(start=date_today,interval="5m" )
ccid='BTC'+ '_' +str(hist_bitcoin_intraday.index[-1])
Date = hist_bitcoin_intraday.index[-1]
price=hist_bitcoin_intraday.iloc[-1]['Close']
cur.execute("SELECT * FROM CRYPTO_PRICE_INTRADAY where ccid = %s", (ccid,))
x=cur.fetchone()
if x is None:
    cur.execute("INSERT INTO CRYPTO_PRICE_INTRADAY (ccid, crypto,
crypto_name, date, Price) VALUES (%s, %s, %s, %s,
%s)", (ccid, 'BTC', 'BITCOIN', Date, price))
    conn.commit()
    print("%s : Value for BTC INTRADAY inserted PRICE = %s for DATE =
%s"%(current_time, price, Date))
else:
    print("%s : No new BTC INTRADAY price"%(current_time))

#CRYPTO_PRICE_DAY Streaming update
hist_bitcoin_daily = bitcoin.history(start=date_yesterday,
end=date_yesterday)
ccid='BTC'+ '_' +str(hist_bitcoin_daily.index[-1])
Date = hist_bitcoin_daily.index[-1]
Open=hist_bitcoin_daily.iloc[-1]['Open']
High=hist_bitcoin_daily.iloc[-1]['High']
Low=hist_bitcoin_daily.iloc[-1]['Low']
Close=hist_bitcoin_daily.iloc[-1]['Close']

cur.execute("SELECT * FROM CRYPTO_PRICE_DAY where ccid = %s", (ccid,))
x=cur.fetchone()
if x is None:
    cur.execute("INSERT INTO CRYPTO_PRICE_DAY (ccid, crypto, crypto_name,
date, Open, High, Low, Close) VALUES (%s, %s, %s, %s, %s, %s, %s,
%s)", (ccid, 'BTC', 'BITCOIN', Date, Open, High, Low, Close))
    conn.commit()
    print("%s : Value for BTC DAY inserted PRICE = %s for DATE =
%s"%(current_time, Close, Date))
else:
    print("%s : No new BTC DAY closing price"%(current_time))

time.sleep(120)

```

7 c. Code tweets

```

#/******
#Nom ..... : Insert Streaming Tweets Into postgres_v03.py
#Context .....: Natural language processing and Crypto Prices
#Role .....: Get tweets, apply sentiment analysis and store in DB
#Auteur ..... : JDO
#Version ..... : V1.1
#Date ..... : 09.12.2020
#Language : Python
#Version : 3.7.8
#*****/
#*****/

from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json
from unicode import unicode
import time

```

```
import sqlite3
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import psycpg2
from configparser import ConfigParser
import yfinance as yf
import random
import numpy as np
import re
from datetime import date, timedelta, datetime
import datetime

#Get Connection details in .ini file for POSTGRES
def config_POSTGRES(filename='credentials.ini', section='postgresql'):
    # create a parser
    parser = ConfigParser()
    # read config file
    parser.read(filename)
    # get section, default to postgresql
    db = {}
    if parser.has_section(section):
        params = parser.items(section)
        for param in params:
            db[param[0]] = param[1]
    else:
        raise Exception('Section {0} not found in the {1} file'.format(section,
filename))
    return db

#Opening connection to postgres database
conn = None
params = config_POSTGRES()
print('Connecting to the PostgreSQL database...')
conn = psycpg2.connect(**params)
cur = conn.cursor()

#Import the sentiment analyzer from VADER
analyzer = SentimentIntensityAnalyzer()

#Get Connection details in .ini file for TWITTER
def config_TWITTER(filename='credentials.ini', section='TWITTER'):
    # create a parser
    parser = ConfigParser()
    # read config file
    parser.read(filename)
    # get section, default to postgresql
    twit_config = {}
    if parser.has_section(section):
        params = parser.items(section)
        for param in params:
            twit_config[param[0]] = param[1]
    else:
        raise Exception('Section {0} not found in the {1} file'.format(section,
filename))
    return twit_config

#Get the connection details
param_twitter=config_TWITTER()

def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
```

```

        input_txt = re.sub(i, '', input_txt)
    return input_txt

def clean_tweets(tweets):
    #remove twitter Return handles (RT @xxx:)
    tweets = np.vectorize(remove_pattern)(tweets, "RT @[\\w]*:")
    #remove twitter handles (@xxx)
    tweets = np.vectorize(remove_pattern)(tweets, "@[\\w]*")
    #remove URL links (httpxxx)
    tweets = np.vectorize(remove_pattern)(tweets, "https?:/[A-Za-z0-9./]*")
    #remove special characters, numbers, punctuations (except for #)
    tweets = np.core.defchararray.replace(tweets, "[^a-zA-Z]", " ")

    return tweets

class listener(StreamListener):
    def on_data(self, data):
        try:
            data = json.loads(data)
            print("-----STARTING-----")
            #print(data)
            #Cleaning the tweets
            if data['text'][:2] != 'RT' and data['user']['followers_count'] > 50
and data['lang']=='en':
                id_tweet=data['id']
                #print("-----PRINT RAW-----")
                #print(data['id_str'])
                #print(data['text'])
                tweet_list = []
                tweet_list.append(data['text'])
                tweet = clean_tweets(tweet_list)
                #print("-----PRINT CLEANED-----")
                #print(tweet[0])
                time_ms = int(data['timestamp_ms']/1000)
                time_dt=datetime.datetime.fromtimestamp(time_ms).isoformat()
                vs = analyzer.polarity_scores(tweet)
                sentiment = vs['compound']
                #print("-----PRINT TIME-----")
                print(time_dt)
                #print("-----PRINT SENTIMENT-----")
                #print( sentiment)
                if sentiment != 0.0:
                    print("Tweet Insertion started")
                    cur.execute("INSERT INTO sent (id_tweet, date, tweet,
sentiment) VALUES (%s,%s, %s, %s)",(id_tweet,time_dt, tweet[0], sentiment))
                    conn.commit()
                    print("Tweet Inserted")
                else:
                    print("Sentiment unclear")
            else:
                #print(data['text'])
                print("Tweet does not look relevant")
                #print("Language =s"%data['lang'])
                #print("Follower count =s"%data['user']['followers_count'])
                #print("RT Status = %s"%data['text'][:2])
        except KeyError as e:
            print(str(e))
        return(True)

    def on_error(self, status):
        print(status)

```

```
while True:
    try:
        auth = OAuthHandler(param_twitter['ckey'],param_twitter['csecret'])
        auth.set_access_token(param_twitter['atoken'], param_twitter['asecret'])
        twitterStream = Stream(auth, listener())
        twitterStream.filter(track=["Bitcoin"])
    except Exception as e:
        print(str(e))
        time.sleep(5)
```

6 c. Code industrialization

```
#!/*****
#Nom ..... : CleanUp and Insert History_v01.py
#Context .....: Natural language processing and Crypto Prices
#Role .....: clean up of daily tweets and archiving of sentiment
#Auteur ..... : JDO
#Version ..... : V1.1
#Date ..... : 09.12.2020
#Language : Python
#Version : 3.7.8
#*****/

#*****/

from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json
from unicode import unicode
import time
import sqlite3
from configparser import ConfigParser
import yfinance as yf
import random
import numpy as np
import re
from datetime import date, timedelta, datetime
import datetime
import pandas as pd
import pandas.io.sql as sqlio
import psycopg2

#Get Connection details in .ini file for POSTGRES
def config_POSTGRES(filename='credentials.ini', section='postgresql'):
    # create a parser
    parser = ConfigParser()
    # read config file
    parser.read(filename)
    # get section, default to postgresql
    db = {}
    if parser.has_section(section):
        params = parser.items(section)
        for param in params:
            db[param[0]] = param[1]
    else:
        raise Exception('Section {0} not found in the {1} file'.format(section,
filename))
```

```

return db

#Opening connection to postgres database
conn = None
params = config_POSTGRES()
print('Connecting to the PostgreSQL database...')
conn = psycopg2.connect(**params)
cur = conn.cursor()
date_yesterday = date.today() - timedelta(days=1)
date_beforeyesterday = date.today() - timedelta(days=2)
date_Month = date.today() - timedelta(days=30)

#Get all the sentiments from the sent table for the last day
print(date.today())
print(date_beforeyesterday)
dat = pd.read_sql_query("SELECT * FROM sent where sentiment is not NULL and
CAST(date AS DATE)<%s and CAST(date AS DATE)>%s ;",conn,
params=(date.today(),date_beforeyesterday,))
Number_tweets = len(dat.index)
Average_sentiment = 0

#calculate sum of sentiments
for index, row in dat.iterrows():
    Average_sentiment += float(row['sentiment'])

#calculate the average sentiment
if Number_tweets > 0:
    Average_sentiment=Average_sentiment/Number_tweets

    #Check if there is already a line for yesterday
    cur.execute("SELECT * FROM statistics_daily where
day=%s;",(date_yesterday,))
    x=cur.fetchone()

    #insert the aggregated sentiment for yesterday
    if x is None:
        cur.execute("INSERT INTO statistics_daily (day, tweet_nb, av_sent,
google_search, transactions_nb, wallet_nb) VALUES (%s, %s, %s, %s, %s,
%s)",(date_yesterday,Number_tweets,Average_sentiment,0,0,0))
        conn.commit()
    else:
        print('Already in the table bro')

    #Delete values older than a month
    cur.execute("DELETE FROM sent where date < %s;",(date_Month,))
    conn.commit()

```


8 References

- Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- www.sentcrypt.com
- <https://www.tradingview.com/markets/cryptocurrencies/global-charts/>
- <https://www.vice.com/en/article/nedqgw/jpmorgan-ceo-says-bitcoin-is-for-murderers-drug-dealers-and-north-korea>
- <https://www.cryptoglobe.com/latest/2020/11/paypal-ceo-alfred-kelly-we-are-very-interested-in-cryptocurrencies/>
- <https://newsroom.paypal-corp.com/2020-10-21-PayPal-Launches-New-Service-Enabling-Users-to-Buy-Hold-and-Sell-Cryptocurrency>
- <https://bitcoin.org/en/bitcoin-paper>
- <https://bitcoin.org/en/>
- https://en.wikipedia.org/wiki/List_of_security_hacking_incidents
- https://en.wikipedia.org/wiki/Edward_Snowden
- https://en.wikipedia.org/wiki/2012%E2%80%932013_Cypriot_financial_crisis
- <https://www.coindesk.com/bitcoin-adoption-venezuela-research>
- <https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/by-the-numbers-are-your-smart-home-devices-being-used-as-cryptocurrency-miners>
- www.intotheblock.com
- https://theses.ubn.ru.nl/bitstream/handle/123456789/4434/MTHEC_RU_Sjoerd_Klabbers_s4384458.pdf?sequence=1
- Thinking fast and slow Daniel Kahneman
- <https://www.mckinsey.com/industries/public-and-social-sector/our-insights/half-the-world-is-unbanked#:~:text=Research%20reveals%20that%20more%20than,to%20save%20or%20borrow%20money.>
- https://en.wikipedia.org/wiki/Mt._Gox
- <https://blockchain.news/news/bitcoin-should-make-5-investment-portfolio-fidelity>
- <https://www.investopedia.com/news/every-portfolio-should-have-6-bitcoin-yale-study/>
- <https://cointelegraph.com/news/virgin-galactic-ceo-everyone-should-have-1-of-their-assets-in-bitcoin>
- <https://blockchain.news/news/115-million-bitcoin-asset-manager-primary-treasury-reserve-asset>
- <https://news.bitcoin.com/nasdaq-microstrategy-bitcoin-425-million/>
- <https://www.cnbc.com/2020/10/08/square-buys-50-million-in-bitcoin-says-cryptocurrency-aligns-with-companys-purpose.html#:~:text=Payment%20company%20Square%20is%20buying,the%20second%20quarter%20of%202020.>
- <https://defipulse.com/>
- <https://academy.binance.com/en/articles/what-is-uniswap-and-how-does-it-work>
- <https://www.investopedia.com/news/how-much-does-it-cost-mine-bitcoin-around-world/>
- <https://www.forbes.com/sites/youngjoseph/2020/06/07/why-the-actual-cost-of-mining-bitcoin-can-leave-it-vulnerable-to-a-deep-correction/?sh=1abf18e26067>
- The evolving liaisons between the transaction networks of Bitcoin and its price dynamics - Carlo Campajola - The evolving liaisons between the transaction networks of Bitcoin and its price dynamics (2019)
- <https://www.forextime.com/eu/education/forex-tutorials/bitcoin-block-rewards>
- <https://www.tradingview.com/symbols/CRYPTOCAP-BTC.D/>
- <https://bitcoin.org/en/>
- The predictive power of public Twitter sentiment for forecasting cryptocurrency prices : Olivier Kraaijeveld, Johannes De Smedt (2018)
- Using sentiment analysis to predict interday Bitcoin price movements, Vytautas Karalevicius, Niels Degrande and Jochen De Weerd (2017)
- Ciaian, P., Rajcaniova, M., Kanacs, D., 2016. The economics of Bitcoin price formation. Appl. Econ. 48 (19), 1799–1815.

- Ciaian, P., Rajcaniova, M., Kanacs, D., 2018. Virtual relationships: short- and long-run evidence from Bitcoin and altcoin markets. *J. Int. Financ. Markets Inst. Money* 52, 173–195.
- Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work. *J. Financ.* 25 (2), 383–417.
- Silverman, G., Murphy, H., Authers, J., 2017. Bitcoin: an investment mania for the fake news era.
- The inefficiency of Bitcoin - A Urquhart (2016)
- Mensi, W., Lee, Y.-J., Al-Yahyaee, K.H., Sensoy, A., Yoon, S.-M., 2019. Intraday downward/upward multifractality and long memory in Bitcoin and Ethereum markets: An asymmetric multifractal detrended fluctuation analysis. *Financ. Res. Lett.* 31, 19–25.
- Behavioral anomalies in cryptocurrency markets - Hanlin Yang (2019)
- Lo, A.W., 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective.
- Lo, A.W., 2012. Adaptive markets and the new world order (corrected May 2012). *Financ. Anal. J.* 68 (2), 18–29.
- Predicting Bitcoin – Gauging the market for bitcoin using web search queries - Yannick Zjörjen (2017)
- Engelberg, E.J. and Parsons, A.C. (2011), “The causal impact of media in financial markets”, Vol. 66 No. 1, pp. 67-97.
- Kristoufek, L. (2013), “Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era”, *Scientific Reports* 3, 3415 EP, available at: <http://dx.doi.org/10.1038/srep03415>
- The digital traces of bubbles: feedback cycles be-tween socio-economic signals in the Bitcoin economy - David Garcia and AI (2014)
- Social signals and algorithmic trading of Bitcoin - David Garcia, Frank Schweitzer (2015)
- Nowcasting the Bitcoin Market with Twitter Signals - Jermain Kaminski (2016)
- <https://www.brandwatch.com/blog/twitter-stats-and-statistics/>
- <https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/>
- Tafti, A., Zotti, R., Jank, W., 2016. Real-time diffusion of information on Twitter and the financial markets. *PloS ONE* 11 (8).
- Li, X., Wang, C.A., 2017. The technology and economic determinants of cryptocurrency exchange rates: the case of Bitcoin. *Decis. Support Syst.* 95, 49–60.
- Li, X., Xie, H., Chen, L., Wang, J., Deng, X., 2014. News impact on stock price return via sentiment analysis. *Knowl.-Based Syst.* 69 (1), 14–23.
- Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Financ.* 66 (1), 35–65.
- <https://alternative.me/crypto/fear-and-greed-index/>
- <https://money.cnn.com/data/fear-and-greed/>
- <https://www.augmento.ai/bitcoin-sentiment/>
- <https://www.bittsanalytics.com/sentiment-index/BTC>
- <https://coinmarketcap.com/currencies/bitcoin/historical-data/>
- <https://p.nomics.com/cryptocurrency-bitcoin-api>
- <https://pypi.org/project/yahoo-finance/>
- <https://developer.twitter.com/en>
- <https://www.tweepy.org/>
- <https://textblob.readthedocs.io/en/dev/>
- <https://pypi.org/project/vaderSentiment/#:~:text=VADER%20Sentiment%20Analysis, on%20texts%20from%20other%20domains.>
- <https://github.com/cjhutto/vaderSentiment>
- <https://dashboard.heroku.com/>
- <https://flask.palletsprojects.com/en/1.1.x/>
- <https://bokeh.org/>

