Analysis Report (Chapter 3 & 4)

Investor Sentiment in the Cryptocurrency Market

Trung dang, 2020

Contents

1 Data Collection

1.1 CRIX - A Cryptocurrency Market Index

The *CRIX (CRyptocurrency IndeX)* was constructed by Trimborn & Härdle (2018) in order to track the entire cryptocurrency market performance as close as possible. The study adopts the usage of this index for broad representations of the cryptocurrency market as in a number of pioneering papers on cryptocurrencies, such as Hafner (2018), Chen et al. (2019), and Alexander & Dakos (2020). The index formula is given as:

$$CRIX_t = \frac{\sum_{i=1}^{n} MV_{it}}{Divisor_t} \tag{1.1}$$

in which MV_{it} is the market capitalization of the cryptocurrency i at time t, and n is the number of constituents of the index. $Divisor_t$ is the coefficient ensuring that the changes in the number of crypto tokens in circulation do not affect the value of the *CRIX*. At the starting point of the *CRIX, Divisor* is defined as:

$$Divisor_0 = \frac{\sum_{i=1}^{n} MV_{i0}}{1000} \tag{1.2}$$

where 1000 is the chosen starting value of the *CRIX*. Whenever the number of tokens of a cryptocurrency changes, the *Divisor* will be adjusted accordingly, thus assuring that the changes in *CRIX* are only caused by the price changes of the constituents.

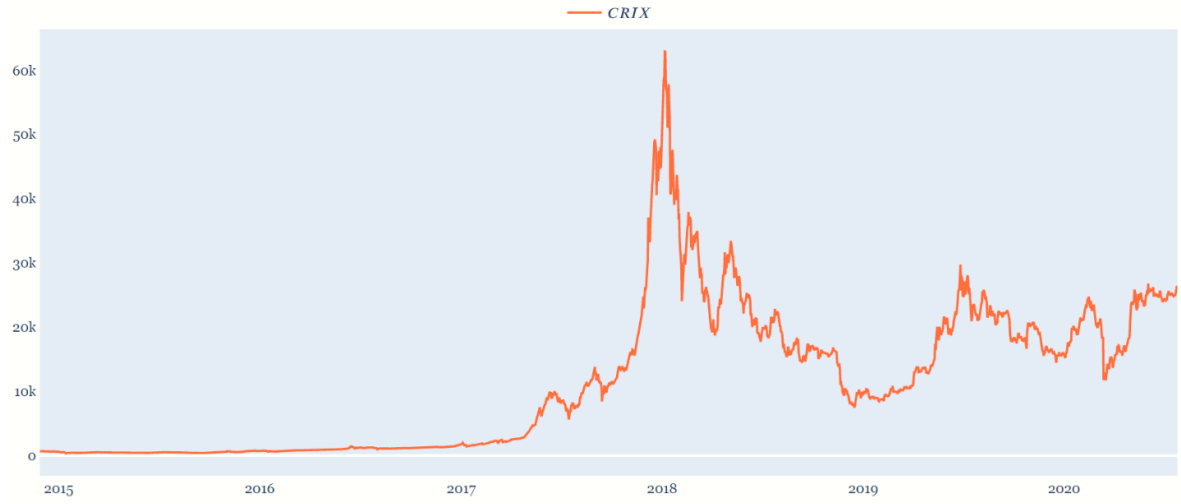

Figure 1.1: The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

In order to become eligible to be a constituent of the index, a cryptocurrency has to fulfill at least one of the two liquidity rules set out by its creators. The rules will not be mentioned explicitly in this study[1], but in principle, the components of *CRIX* have to be traded very frequently so that they can be easily converted to traditional fiat money or other cryptocurrencies. Also, the number of constituents n is not fixed as with the indices of relatively stable markets (such as the stock market with the *S&P500*) but actually allowed to variate to match a fast and dynamic market as cryptocurrencies. This number will be checked every 3 months to see if it still fits the market well, based on certain rules set out by the index's authors. As of 30 August 2020[2], the number of *CRIX*'s components is currently set at 30. In addition, the index

[1] For more details, see http://data.thecrix.de/#about

[2] See https://thecrix.de

is also reallocated monthly or more frequently in case the need arises. Lastly, the full data series is provided by the authors at the link http://data.thecrix.de/data/crix.json. In Figure 1.1, the reader could find a plot of the index for our analysis timeframe (from 28 November 2014 to 25 July 2020). It could be seen crucial moments of the cryptocurrency market, especially the unprecedented boom (and subsequent crash) occurred in late 2017 and early 2018, have been captured quite effectively by the *CRIX* series.

1.2 Direct Sentiment Measures

1.2.1 StockTwits Sentiment

This section begins with a detailed re-introduction of *StockTwits*. This platform is a social media similar to *Twitter*, but dedicated specifically to financial discussions where investors can express opinions on any financial assets supported by the platform by posting messages with a maximum length of 140 characters. *StockTwits* currently supports over 2,500 financial assets, including 532 different cryptocurrencies[3]. There are three features that make *StockTwits* a great candidate for performing sentiment analysis. Firstly, the site has a public API that allows automatic retrieval of users' historical messages. Secondly, conversations on *StockTwits* are ordered using a feature called "*cashtags*". For example, we have $BTC.X referring to Bitcoin on this platform. This feature enables one to download messages that are exclusively related to his or

[3] According to the site's API documentation, could be found at: https://api.stocktwits.com/symbol-sync/symbols.csv

her desired topics. Finally and most importantly, StockTwits has an optional feature for the users to tag their posts as "*Bullish"* or *"Bearish*". Conceptually, this feature makes applying supervised machine learning or similar algorithms very appealing to do because we have a very large dataset that is naturally labeled. Chen et al. (2019) exploited this to build a sophisticated lexicon that consists of nearly 10,000 terms frequently used by the investors of cryptocurrencies. According to the same study, this lexicon outperforms the financial lexicon by Loughran & McDonald by a substantial 32% of accuracy in classifying out-of-sample observations. Due to its superior performance, this lexicon will be employed to assess the sentiment of our retrieved messages at later stages.

The *StockTwits Public API*[4] is utilized for retrieving all messages posted between 28 November 2014 and 25 July 2020. Every retrieved message must contain at least one *cashtag* ending with ".X" (i.e, cashtag referring to a cryptocurrency on *StockTwits*). To avoid the domination of fake messages, possibly spammed by automatic bots, according to Pang et al. (2002), a maximum proportion of 1% of the dataset has to be imposed per user. The final dataset consists 2,045,322 messages (~287MB) from 48,648 distinct users and related to 519 cryptocurrencies. Among those, 999,720 messages are labeled by their posters as *"Bullish"* (818,452 messages ~ 40.02%) or *"Bearish"* (181,268 messages ~ 8.86%). The rest are unclassified messages. This imbalance is considered expected based on the previous claim on how investors are often optimistic when they share opinions on social media (see footnote **Error! Bookmark not defined.**). Table 1.1 displays 10 cryptocurrencies with the highest number of messages posted on the

[4] https://api.stocktwits.com/developers

platform during our analysis timeframe. It could be seen that messages about Bitcoin dominate nearly half of the total messages on *StockTwits*.

Table 1.1: 10 cryptocurrencies with the highest number of messages on *StockTwits* (28/11/2014 - 25/07/2020)

Cryptocurrency	**Number of Messages**
Bitcoin ($BTC.X)	964,167
Ethereum ($ETH.X)	169,481
Litecoin ($LTC.X)	147,808
TRON ($TRX.X)	140,884
Ripple ($XRP.X)	128,281
Dogecoin ($DOGE.X)	41,293
Bitcoin Cash ($BCH.X)	32,779
Ethereum Classic ($ETC.X)	24,342
Stellar ($XLM.X)	23,962
Bitcoin SV ($BSV.X)	23,472

Next, a *text processing* procedure is applied to the collected messages so that only

relevant information remains. The main idea is to remove all words that are unnecessary for classifying the sentiment of sentences. The first step is to lowercase all messages. For some reason, messages collected using *StockTwits API* have all punctuations represented by their HTML codes instead of the Unicode (human-readable) characters. For example, "!" was expressed as "!". All of these HTML codes are converted back to Unicode format. All words with more than 3 repeated letters, such as *"hellooooooooooo*", which appear frequently as a form of emphasized sentiment on online discussion platforms[5], are shrunk to a maximum length of 3 (i.e. *"hellooo"*). Currencies hashtags (such as *"$BTC.X"*, *"$ETH.X"*), money values (such as *"$10k"*, *"€200"*), hyperlinks (such as *"http://stocktwits.com"*), numbers (such as 10,000 or 5k), username (such as *"@username"*) are replaced respectively by the word *"cashtag"*, *"moneytag"*, *"linktag"*, *"numbertag"*, and *"usertag"*. Next, all stop words such as *"I"*, *"he"*, *"she"*, *"it"*, *"herself"*, *"himself"*, *"the"*, *"an"*, *"a"*, etc.[6] are also removed. All punctuations are also removed except *"!"* and *"?"*. Lastly, the prefix *"negtag_"* is added to any word consecutive to negative words such as *"no"*, *"nor"*, *"isn't"*, etc.[7] For example, *"can't fly"* is converted into *"negtag_fly"*. The *text processing* methodology adopted by this study is inspired largely by Renault (2017) and Chen et al. (2019). However, this study modifies their procedures by adding some extra steps, namely escaping HTML symbols, or providing a more detailed list of stop words and negative words. Examples of messages before and after processing are given

[5] See Brody & Diakopoulos (2011)

[6] The complete list of 136 stop words to be removed could be found in the file *stopwords.csv* in the GitHub repository of this project, which will be linked in the conclusion chapter.

[7] The complete list of 43 negative words could be found in the file *negative.csv* in the GitHub repository of this project.

in Table 1.2.

Table 1.2: Text processing examples

Before Processing	After Processing
$BTC.X wow this is way ahead of binance!	cashtag wow way ahead binance!
$BTC.X needs to hold above 10,200.. I really don't want to be bored by a 2 day consolidation between 10,000-10,200	cashtag needs hold numbertag really don't want bored numbertag day consolidation numbertag numbertag
$1ST.X Got fouled, this is a POS. need to hodl cos I'm fucked at the moment with this one. Will take 5 years to break even	cashtag got fouled pos need hodl cos fucked moment one take numbertag years break even
$SPY wallstreet is piling into bitcoin and crypto. Smart money leaving before the crash $BTC.X	cashtag wallstreet piling bitcoin crypto smart money leaving crash cashtag
$BTC.X to all the bitcoin bears out there.	cashtag bitcoin bears there

At this stage, we utilize the crypto-specific lexicon created by Chen et al. (2019)[8] to measure sentiment in each processed message. To make this happen, the messages must be transformed into vectors of words that could be unigrams (i.e. single word) or bigrams (i.e. two words side by side) in a procedure called *text vectorization*. For example, suppose there is a sentence such as *"I have a lovely dog"*, then the vectorizer will split that sentence into a vector of unigrams:

$$[\text{“}I\text{”} \ , \text{“}have\text{”}, \text{“}a\text{”} \ , \text{“}lovely\text{”}, \text{“}dog\text{”}]$$

and a vector of bigrams:

$$[\text{“}I \ have\text{”} \ , \text{“}have \ a\text{”}, \text{“}a \ lovely\text{”} \ , \text{“}lovely \ dog\text{”}]$$

The reason why we have to convert the messages into vectors of unigrams and bigrams is that the lexicon of Chen et al. (2019) consists of both types of words. Moreover, bigrams even account for the vast majority of the words inside their lexicon with 7,329 out of 9,613 words are bigrams. At this moment, we could proceed to compute the sentiment score of each individual message in the exact order as follows. First, sentiment of the vector of bigrams is measured, then any bigram that has been matched with a word in the lexicon is removed. Next, sentiment of the vector of unigrams is measured. The sentiment scores of both unigrams and bigrams are summed up and divided by the total number of terms that have been matched with the lexicon. For instance, suppose we have a processed message such as *"moneytag pulls back pump"*. The bigram *"pulls back"* will be first taken into account with a sentiment weight of 0.88. The unigram *"pulls"* with a sentiment score of 0.54 will be skipped according to

[8] The crypto-specific lexicon is published by the main author (Professor Cathy Y. Chen) on her personal website: https://sites.google.com/site/professorcathychen/resume

the previous rules. However, the unigram *"pump"* will be considered with a weight of -0.30. Since there are 2 terms that have been matched with the lexicon, the sentiment score of the message is evaluated at $(0.88 - 0.30)/2 = 0.29$. After every individual message has been computed for its sentiment, the sentiment score for each day is derived by averaging the sentiment scores of all messages published on that day. The series of the daily sentiment of StockTwits users will be denoted as $SENT_{StockTwits}$ from this point. Figure 1.2 plots this data series versus the *CRIX* index. For some reason, the sentiment score on *StockTwits* fluctuates strongly during the early days of our sampling period (pre-2017), then becomes more stable after that. Interestingly, the sentiment score seems to be almost always positive after June 2017. We will discuss this phenomenon in more detail when we examine the descriptive characteristics of the dataset in chapter 0.

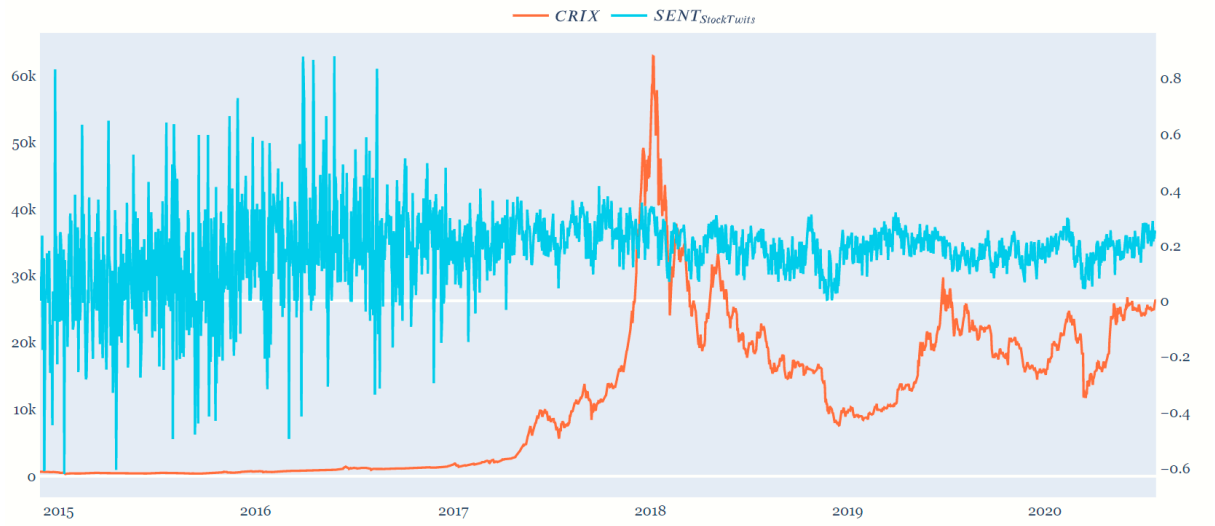

Figure 1.2: The $SENT_{StockTwits}$ & The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

1.2.2 Reddit Sentiment

Previously in section **Error! Reference source not found.**, we learned that *Reddit* is also an excellent source for topic-specialized textual data. Speaking more about *Reddit*, this is an American online discussion website that has been ranked as the 18th-most visited website globally (as of 30 August 2020) by the popular web traffic analysis company *Alexa Internet*[9]. Registered members (also known as *"Redditors"*) can submit content to the site such as links, text posts, images, videos, etc. (called *"submissions"*). These submissions are then voted up or down and discussed by other members. Posts are organized by subject into user-created boards (called *"subreddits"*), which cover a variety of topics, from politics, science, sports, fitness to video games, cooking, pets, investing, etc. Each subreddit is started with the prefix *"/r/"*.

This study employs the *Pushshift Reddit API* designed by the *"/r/datasets"* moderators to extract all submissions and comments posted in two subreddits strictly related to cryptocurrencies having more than 1 million subscribers, namely *"r/Bitcoin"* (~1.6m members), and *"r/CryptoCurrency"* (~1.1m members). Those submissions and comments are collected, once again, from 28 November 2014 to 25 July 2020. It is noteworthy that although *Reddit* also has its own official public API, the *Pushshift Reddit API* provides access to *Reddit* data in a much faster and more efficient way. Thus, it is chosen for the analysis in this study. After imposing a similar maximum proportion of 1% of the dataset on each poster, the final dataset contains 1,290,318

[9] https://www.alexa.com/siteinfo/reddit.com

submissions (~327MB) from 291,009 distinct users, and an astonishing 13,015,447 comments (~2.61GB) from 498,460 distinct users. Since Reddit is also an informal source of information where people also use emojis, slangs, and short sentences, the same procedures of text processing, vectorization, and sentiment analysis are applied to extract sentiment information from Reddit submissions and comments as with StockTwits messages. Also, because comments are usually shorter and more informal than submissions, the analysis is done separately for both sources of data, resulting in two unique sentiment indices, denoted as $SENT_{\text{RedSub}}$ and $SENT_{\text{RedCom}}$. The reader could find below those series plotted versus the *CRIX* index in Figure 1.3 and Figure 1.4. Both of the series appears to vary in ranges relatively narrower compared to the one of the *StockTwits* messages' sentiment score. Over the sampling period, the $SENT_{\text{RedSub}}$ series reaches its highest point during the market downturn in late 2018, while the $SENT_{\text{RedCom}}$ peaks during the famous boom of late 2017.

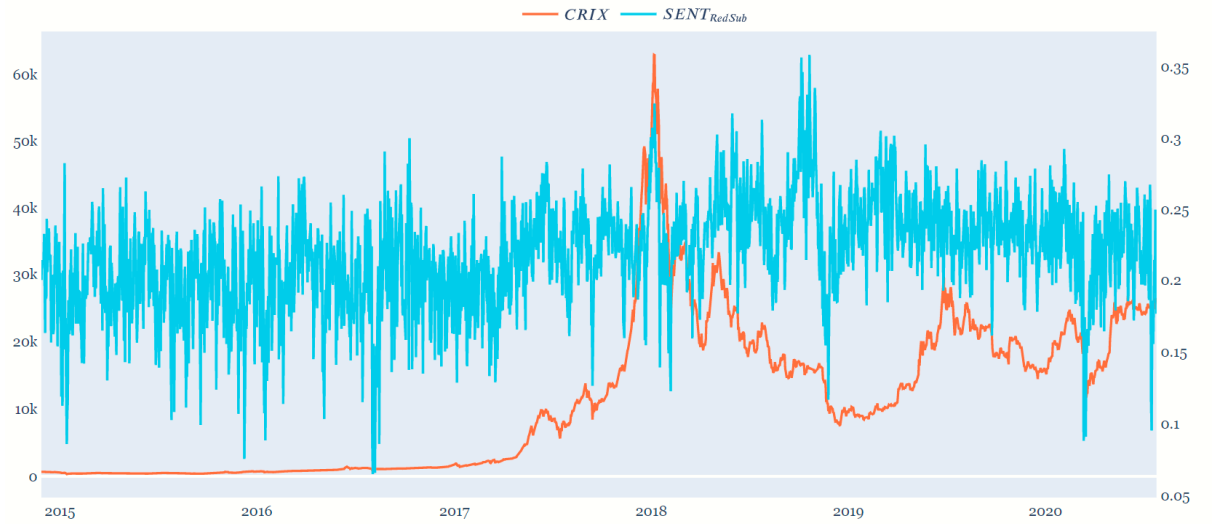

Figure 1.3: The $SENT_{RedSub}$ & the *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on

collected data.

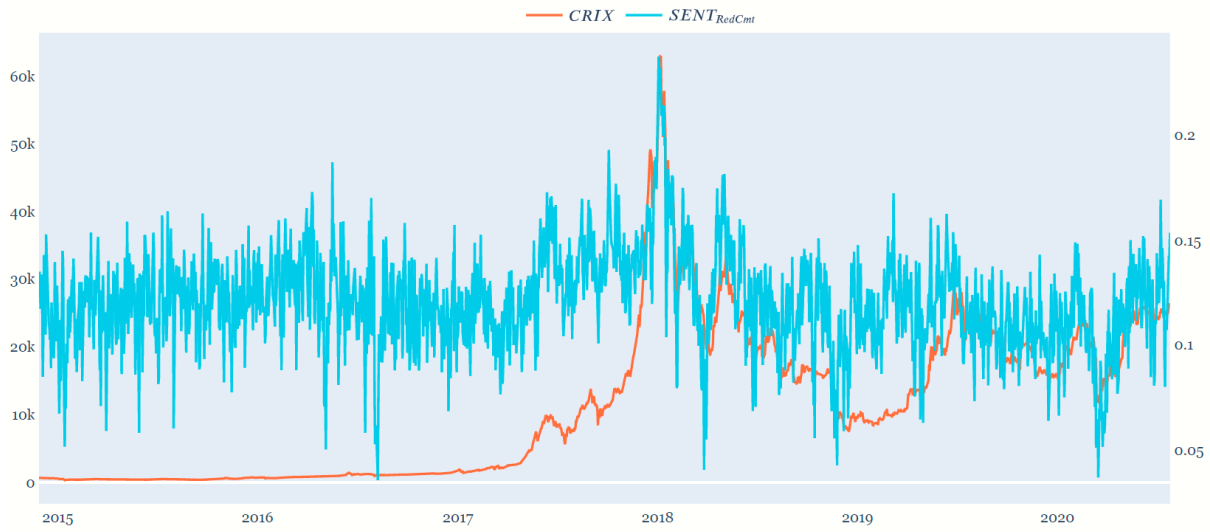

Figure 1.4: The $SENT_{RedCmt}$ & The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

1.3 Indirect Sentiment Measures

1.3.1 Financial Indicators

Following the rationales given in the closing part of section **Error! Reference source not found.**, we attempt to collect data of two potential sentiment proxies that are measures of market performance.

VCRIX - A Cryptocurrency Volatility Market Index

The first candidate is the *VCRIX (Volatility CRypto IndeX)*, created by Kolesnikova (2018) to grasp the risk induced by the cryptocurrency market. This index plays similar

roles in the market of crypto as the *VIX* in the US stock market, or the German implied volatility index *VDAX* in the German stock market. Normally, computing the implied volatility of a financial market requires the presence of derivative instruments inside that market. However, despite the lack of a developed cryptocurrency derivatives market, Kolesnikova has successfully simulated the *VCRIX* on the basis of the *CRIX* market index. The data series could be collected at the author's website[10]. As usual, in Figure 1.5, one could find a plot of the *VCRIX* versus the *CRIX* during our analysis timeframe. It could be seen over our sampling period, the *VCRIX* index reaches its peak of 2324 points very recently, on 14 March 2020, and hits its bottom being 113 points on 3 September 2016.

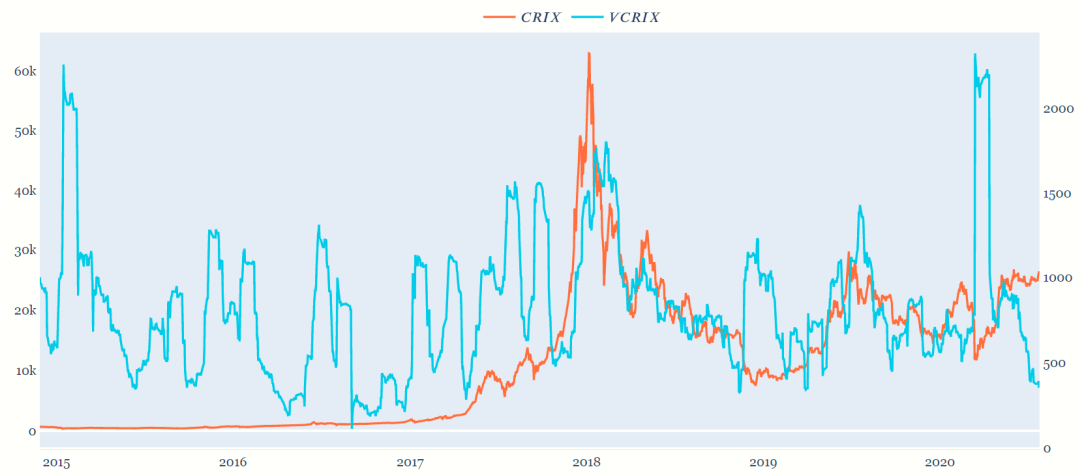

Figure 1.5: The *VCRIX* series & The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

[10] http://data.thecrix.de/data/crix11.json

Market Trading Volume

Another possible candidate in this family of financial indicators is the market's aggregated trading volume. The data series has been downloaded using *Nomics Public API*[11]. *Nomics* is a price-tracking website for crypto assets that is comparable to websites such as *CoinMarketCap* or *CoinGecko*. In terms of popularity among crypto investors, *Nomics* is behind *CoinMarketCap* and *CoinGecko*, however, this site is employed since only its API provides the information required by our analysis. This data series, denoted as VOL_{Trade}, is quoted in USD and available in a daily frequency (Figure 1.6). Initially, the trading volume series appears to co-move very well with the market index series, but stops to do so until the early of this year (2020) when it bursts significantly with a peaked trading volume that equals to approximately 4 times of the highest point reached during the bubble of late 2017.

[11] https://docs.nomics.com

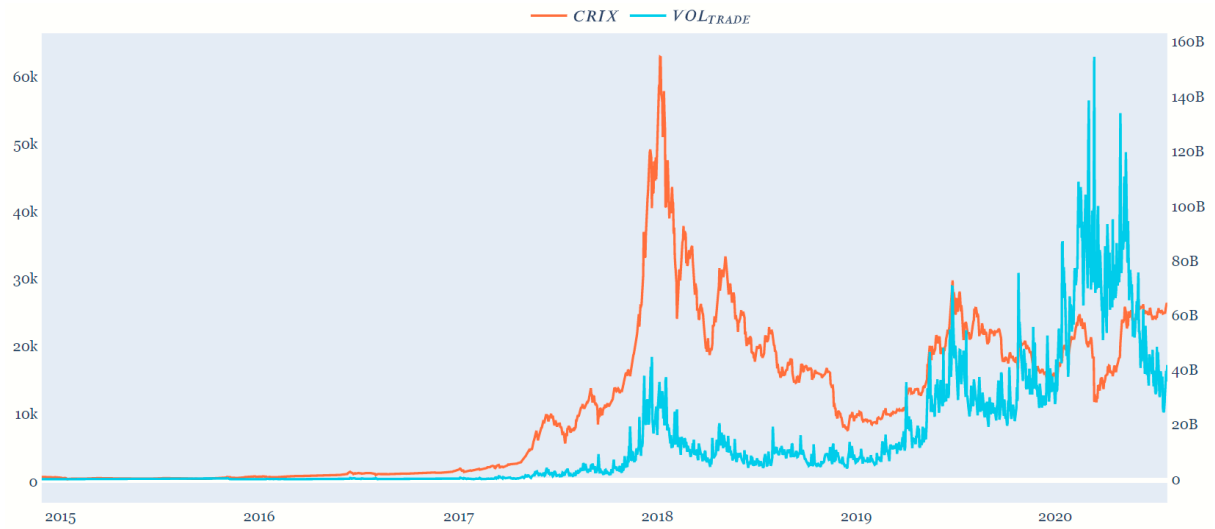

Figure 1.6: The VOL_{Trade} series & the CRIX series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

1.3.2 Online Platforms' Metrics

Google Search Volume

Based on the important recommendations of Abraham et al. (2018) that have been discussed in section **Error! Reference source not found.**, *Google Search Volume*, a search index provided by *Google Trends*, is employed as an indirect sentiment measure. At this stage, the *Python* package *Pytrends*[12] is implemented to collect the *Search Volume* of the keyword *"bitcoin"* from 28 November 2014 to 25 July 2020. The reason why this keyword is selected for our analysis is simply that according to *Google Trends*, it is the crypto-related keyword that is most searched for on *Google*. The

[12] The Pseudo (Unofficial) API for *Google Trends* designed for *Python* users: https://pypi.org/project/pytrends

search volume is scaled automatically across the time period by the *Pytrends* package. Once again, this sentiment measure, denoted in our analysis as VOL_{Google}, is plotted agaisnt the *CRIX* in Figure 1.7. It is not surprising that the *Google Search Volume* peaks during December 2017 and January 2018. Moreover, there seem to be more observers paying attention to the crypto market after the market boom of 2018 than before the incident.

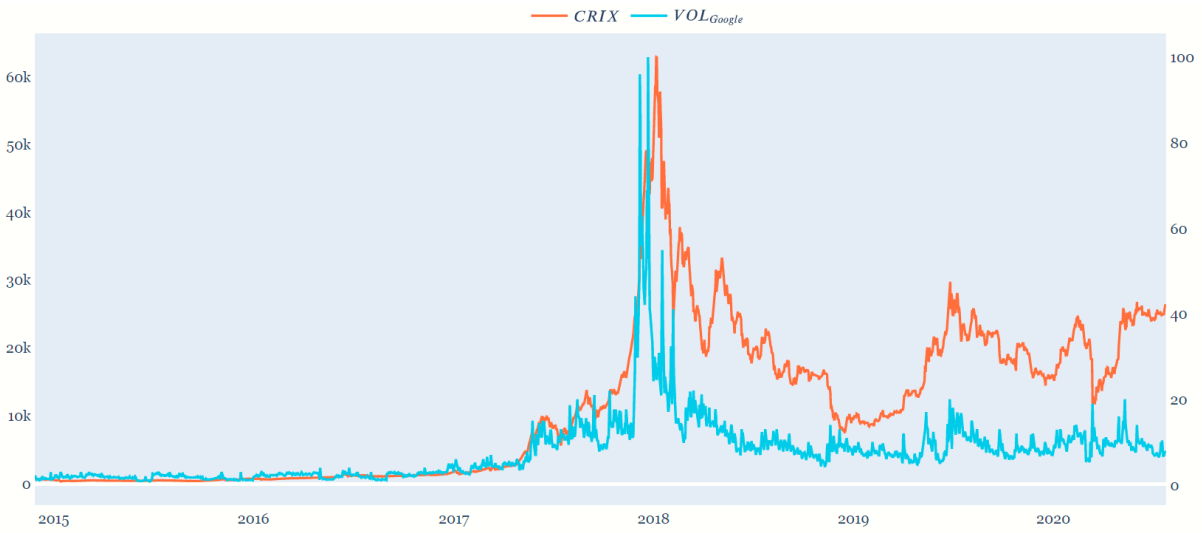

Figure 1.7: The VOL_{Google} & The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

Metrics of StockTwits and Reddit

Also, since a large number of messages and discussions on *StockTwits* and *Reddit* have been collected in previous sections, it is quite straightforward to also compute the metrics of these online platforms and utilize those as proxies for sentiment. The data

series of *StockTwits* message volume, the number of *Reddit* submissions, and the number of *Reddit* comments are denoted respectively as $VOL_{StockTwits}$, VOL_{RedSub}, and VOL_{RedCmt}. The reader might also find their plots versus the *CRIX* in the figures below. The submission and comment volumes on *Reddit* appear to co-move extremely closely with each other and the market index. There exist some point where their increases (decreases) even precede the movement of the market index. In contrast, the message volume on *StockTwits* seems to only spike during crucial events of the market. A possible explanation might be that discussions on *Reddit* take place regularly every day since there are not only investors but also tech enthusiasts getting involve on this platform. In contrast, a vast majority of *StockTwits* users are probably investors and they will mostly talk about the market when there are some large movements ongoing.

To conclude this chapter, we have collected the daily time series of 9 different individual measures for sentiment, namely $SENT_{StockTwits}$, $SENT_{RedSub}$, $SENT_{RedCmt}$, *VCRIX*, VOL_{Trade}, VOL_{Google}, $VOL_{StockTwits}$, VOL_{RedSub}, and VOL_{RedCmt}. The sample period is from 28 November 2014 to 25 July 2020. Besides from that, the index chosen to become the aggregated measure of the crypto market is the *CRIX* time series, also obtained for the pre-mentioned period. At this point, the data collection process has been completed. In the next chapter, we will move forward to the actual analysis where our true measure of sentiment in the crypto market is finally computed, according to the plan proposed at the end of chapter 3. This sentiment measure will also be utilized to test the suspicion set out in section **Error! Reference source not found.** that public expectations on cryptocurrencies (evaluated by the market prices of those digital

assets) are currently suffering from irrational hypes.

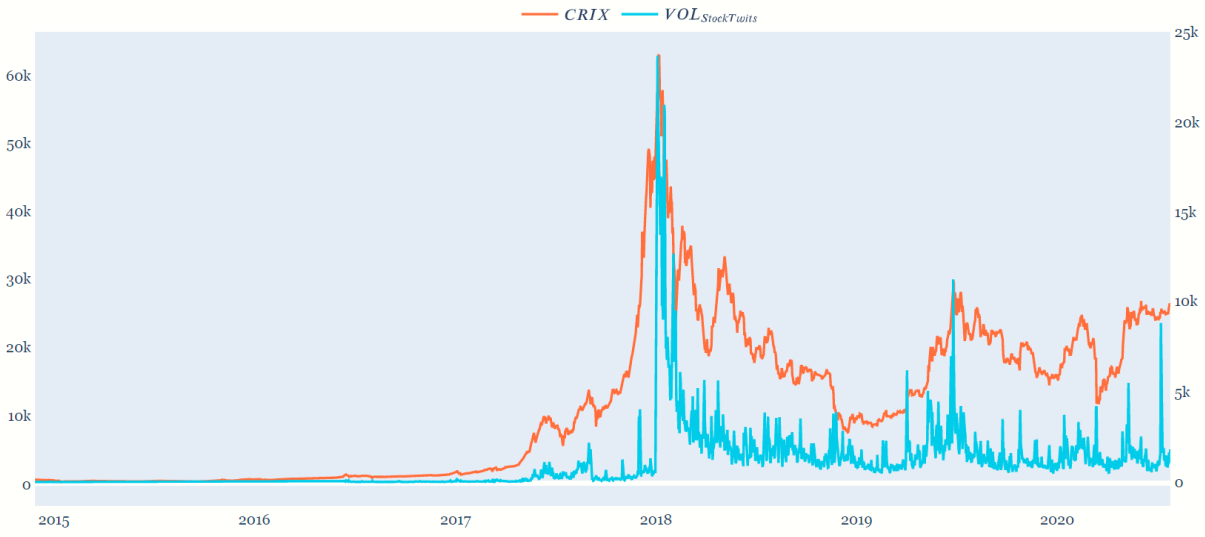

Figure 1.8: The $VOL_{StockTwits}$ series & The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

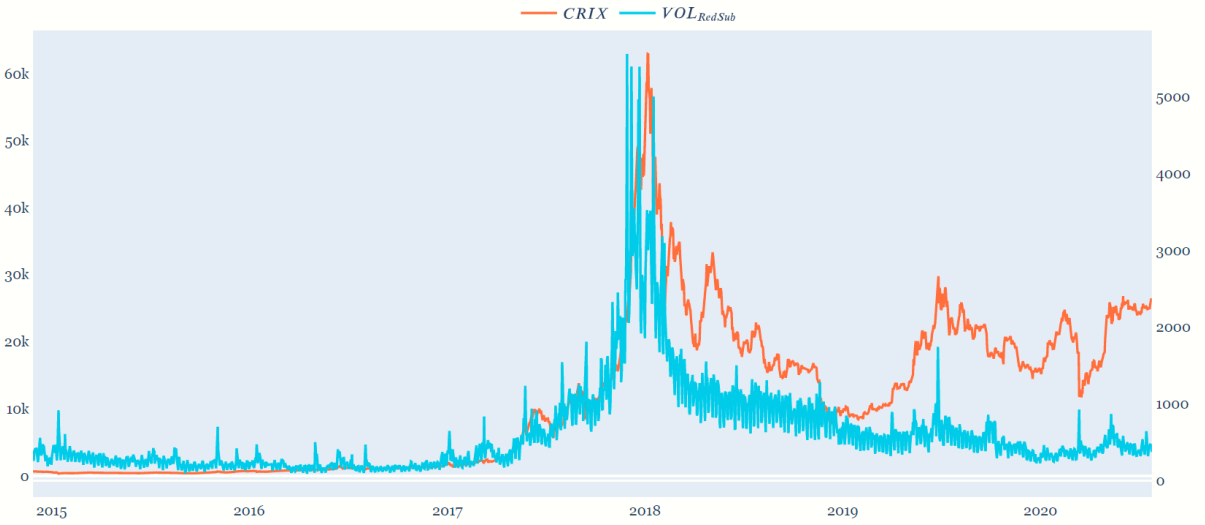

Figure 1.9: The VOL_{RedSub} series & The *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

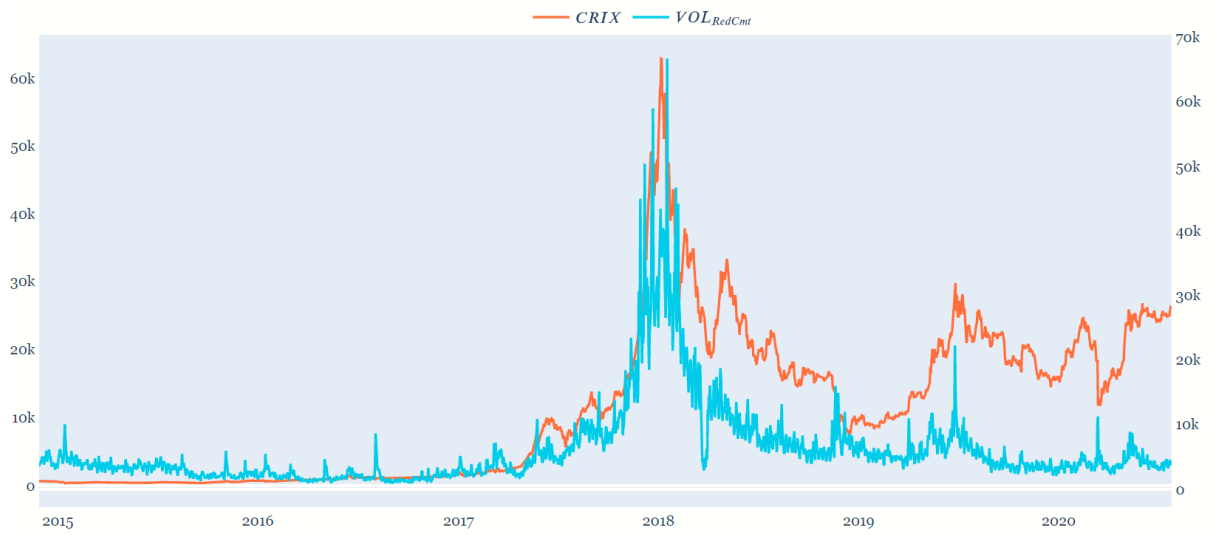

Figure 1.10: The VOL_{RedCmt} series & the *CRIX* series (28/11/2014 - 25/07/2020). Source: Own elaboration based on collected data.

2 Sentiment and Cryptocurrency Return

2.1 A Composite Sentiment Index

2.1.1 Descriptive Statistics

We begin the analysis by taking a look at some of the important characteristics of the collected dataset. In Table 2.1, one could find the summary statistics reported for the ten variables acquired in the last chapter. In fact, there are not too many points that could be put into comparison between those variables since they are mostly measured on different scales, except the ones that are derived using sentiment analysis, namely $SENT_{StockTwits}$, $SENT_{RedSub}$, and $SENT_{RedCmt}$. While these measures are generally positive (which is in line with the statement above about the optimism of online investors), the sentiment score of *StockTwits* seems to display greater volatility than the ones coming from *Reddit*. This phenomenon is exactly similar to in the study of Chen et al. (2019) where the crypto-specific lexicon is also applied on *StockTwits* and *Reddit* textual data.

Table 2.1: Summary Statistics (Before Standardization). Source: Own elaboration.

Variable	Mean	Std. Dev
CRIX	11391.7	11269.3
$SENT_{StockTwits}$	0.17860	0.14028
$SENT_{RedSub}$	0.21820	0.03836
$SENT_{RedCmt}$	0.12006	0.02369
VCRIX	820.852	386.930

VOL_{Trade}	1.37E+10	2.19E+10
VOL_{Google}	7.75068	7.87399
$VOL_{StockTwits}$	1002.12	1847.41
VOL_{RedSub}	609.919	602.834
VOL_{RedCmt}	6095.96	6565.66
Number of Observations	2041	

Next, all sentiment measures (except *CRIX*) are standardized for zero-mean and unit-variance. This is because later we want to apply *PCA* on those variables. In *PCA*, we are interested in the components that maximize the joint-variation between variables. If the variables are left unscaled, some (e.g. $SENT_{RedCmt}$) will vary less than others (e.g. VOL_{Trade}) because of their respective scales and *PCA* might determine that the direction of maximum variance more closely corresponds with the variable with high relative variation. The summary statistics after standardization are presented in Table 2.2 below.

Table 2.2: Summary Statistics (After Standardization). Source: Own elaboration

Variable	**Kurtosis**	**Skewness**	**Minimum**	**Maximum**
$SENT_{StockTwits}$	5.28	-0.68	-5.70	5.00
$SENT_{RedSub}$	0.77	-0.33	-4.00	3.65
$SENT_{RedCmt}$	2.06	0.29	-3.55	4.93
VCRIX	1.98	1.17	-1.83	3.89

VOL_{Trade}	5.45	2.25	-0.63	6.43
VOL_{Google}	32.02	4.30	-0.88	11.72
$VOL_{StockTwits}$	47.27	5.67	-0.54	12.29
VOL_{RedSub}	14.85	3.21	-0.87	8.20
VOL_{RedCmt}	17.47	3.65	-0.79	9.23
Number of Observations	2041			

One might also examine the correlations between standardized sentiment measures in Table 2.3. The sentiment indicators appear to be highly correlated with one another with 34 out of 36 pairs showing statistically significant relationships (with p-values less than 0.01). Two pairs that do not show significant correlations (colored in red) both involve the volatility index *VCRIX*. The results further reveal that most relationships are positive (which is in line with conventional expectations), except the ones that entail *VCRIX* and the market trading volume. This result is still understandable, to a certain degree, as both of these variables are not very straightforward measures of sentiment. Usually, market volatility indices could only become meaningful during the time when investors feel extremely bearish about the market. In addition, while it is true that investors tend to trade more as they feel more confident about future prospects, the trading volumes could also rise in the episodes of market crashes where people panic sell their assets. In other words, while financial-based indicators could sometimes turn into valid proxies of sentiment during special moments of the market, their performance (on measuring sentiment) will not stay consistent over time. Therefore, they are recommended to be used together with other indicators for the

purpose of measuring sentiment. Back to the discussion on the relationship between sentiment individual indicators, another noteworthy detail is that the highest correlations (around 0.8 ~ 0.9) are present among the metrics of online platforms, namely VOL_{Google}, $VOL_{StockTwits}$, VOL_{RedSub}, and VOL_{RedCmt}, meaning that they seem to be measuring the same phenomenon (possibly sentiment). Put this differently, one might guess that those variables could be among the most valid individual measures for the sentiment of crypto investors. It should also be noted that this speculation is in line with the results of Abraham et al. (2018) (mentioned previously in section **Error! Reference source not found.**).

According to the arguments of Brown & Cliff (2004) and Baker & Wurgler (2007) given in section **Error! Reference source not found.**, significant correlations between sentiment indicators make it conceptually appealing to construct a composite index that can represent sentiment in a clearer way and hope that it could stay valid for a prolonged time. In order to isolate as much information as possible into the final index, the *PCA* method used by Baker & Wurgler (2007) will be adopted to combine the sentiment indicators. The next section will describe this process in more detail.

Table 2.3: Contemporanous Correlations. Source: Own elaboration

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
(1)	1.000								
(2)	0.217	1.000							
(3)	0.220	0.374	1.000						
(4)	-0.044	-0.046	0.069	1.000					

(5)	0.058	0.131	-0.130	0.160	1.000				
(6)	0.158	0.184	0.243	0.276	0.327	1.000			
(7)	0.089	0.217	0.244	0.259	0.339	0.497	1.000		
(8)	0.137	0.199	0.340	0.290	0.107	0.878	0.613	1.000	
(9)	0.106	0.124	0.348	0.342	0.118	0.839	0.693	0.955	1.000

(1): $SENT_{StockTwits}$ (4): $VCRIX$ (7): $VOL_{StockTwits}$
(2): $SENT_{RedSub}$ (5): VOL_{Trade} (8): VOL_{RedSub}
(3): $SENT_{RedCmt}$ (6): VOL_{Google} (9): VOL_{RedCmt}

2.1.2 Principal Component Analysis

Before applying *PCA* on the standardized sentiment indicators, Bartlett's test of sphericity is employed to check if their correlation matrix diverges significantly from the identity matrix. Since the p-value is less than 1% (Table 2.4), the null hypothesis of orthogonal variables can be rejected, implying that *PCA* can be used to derive the compression of all available sentiment indicators.

Table 2.4: Bartlett's Test of Sphericity. Source: Own elaboration

Bartlett's Test of Sphericity	Approx. Chi-Square	12125.57
	df	36
	Sig.	.00

Similar to Baker & Wurgler (2007) and Khan & Ahmad (2018), the composite sentiment index is the first principal component of the nine individual indicators, which

is simply a linear combination of those variables with the coefficients chosen to capture the most joint-variation across all nine series. All indicators' coefficients, also called *component loadings*, are presented in the following equation, portraying that the indicators have the signs as expected.

$$\begin{aligned} SENT_9 = \; & 0.116SENT_{StockTwits} + 0.166SENT_{RedSub} + 0.226SENT_{RedCmt} + 0.207VCRIX \\ & + \; 0.162VOL_{Trade} + 0.460VOL_{Google} + 0.394VOL_{StockTwits} + 0.484VOL_{RedSub} \\ & + 0.487VOL_{RedCmt} \end{aligned} \tag{2.1}$$

It could be seen that five indicators have *component loadings* less than 0.3. Based on a rule of thumb set by Hair et al. (2009), we can attempt to eliminate every variable with a loading below 0.3 and construct another composite index from the four remaining variables. The new index (SENT_4) has the *loadings* as follows:

$$\begin{aligned} SENT_4 = \; & 0.499VOL_{Google} + 0.419VOL_{StockTwits} + 0.534VOL_{RedSub} \\ & + 0.539VOL_{RedCmt} \end{aligned} \tag{2.2}$$

The equation of SENT_4 shows that all indicators have the same signs and almost the same (but larger) coefficients as in the original index ($SENT_9$). The correlation between $SENT_9$ and $SENT_4$ is extremely high (0.976), indicating that eliminating four sentiment indicators did not cause any significant change. This can also be seen in Figure 2.1 where we plot the two series against each other. The only crucial difference between the series perhaps occurs in the early years (from 2015 to 2017) where the original index $SENT_9$ seems to vary much stronger than the compact version $SENT_4$. This is easy to understand since the five indicators removed to form $SENT_4$ fluctuate very widely during those years, as one could see in the previous Figure 1.2, Figure 1.3, and Figure 1.4. To avoid losing early variation, the study proceeds with the sentiment

index computed using all nine indicators for the final analysis. From this point, the index will be denoted solely as $SENT$ instead of $SENT_9$.

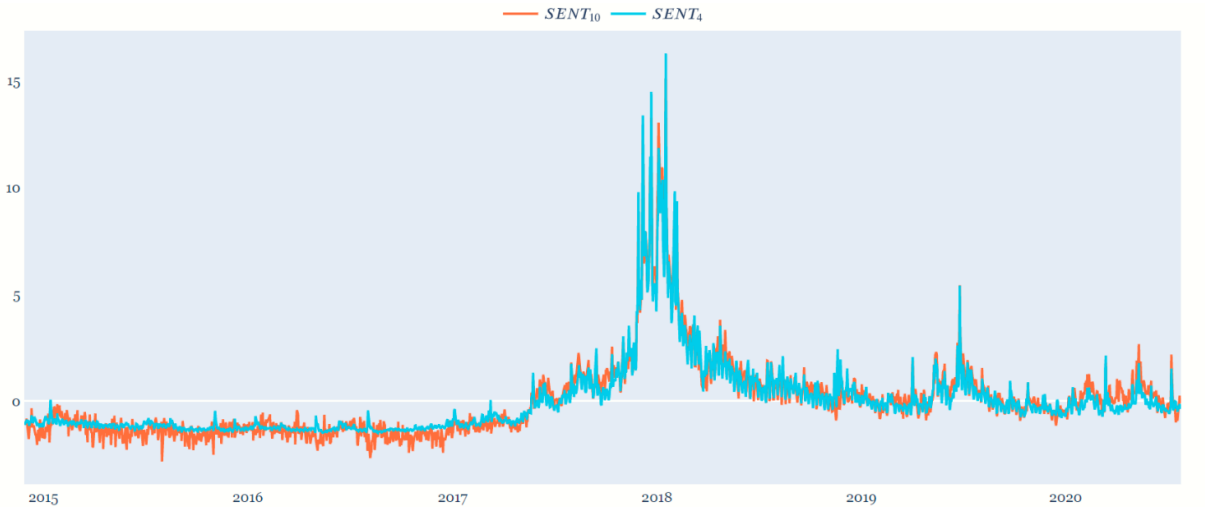

Figure 2.1: $SENT_9$ & $SENT_4$. Source: Own elaboration.

The correlation vector of nine individual sentiment indicators and the composite sentiment index, given in Table 2.5, shows that all sentiment proxies are directly related to the index. To summarize, it appears that we have successfully isolated the unobserved factor of sentiment from various individual indicators.

Table 2.5[13]: Correlation Vector of Individual Sentiment Measures with the Composite Index. Source: Own elaboration.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
SENT	0.222	0.319	0.443	0.401	0.302	0.883	0.761	0.934	0.940

[13] Values in parenthesis represent significance levels of the corresponding correlation coefficients.

(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
(1): $SENT_{StockTwits}$			(4): $VCRIX$			(7): $VOL_{StockTwits}$		
(2): $SENT_{RedSub}$			(5): VOL_{Trade}			(8): VOL_{RedSub}		
(3): $SENT_{RedCmt}$			(6): VOL_{Google}			(9): VOL_{RedCmt}		

2.2 Vector Autoregressive Analysis

With the composite index in hand, we now turn to test the key hypotheses about how sentiment affects the market of cryptocurrency. Some of the earlier discussion in section **Error! Reference source not found.** suggests that an increase (decrease) in sentiment may lead to a temporary price change in the same direction in the short term. In an initial plot (Figure 2.2), the new sentiment index *SENT* appears to be highly promising as it seems to precede most of the trends happening in the market index *CRIX*. A *Vector Autoregressive (VAR)* model will be employed to see the interaction between the two series in a formal way. The analysis will also attempt to identify any possible causality between the sentiment index and the crypto market index.

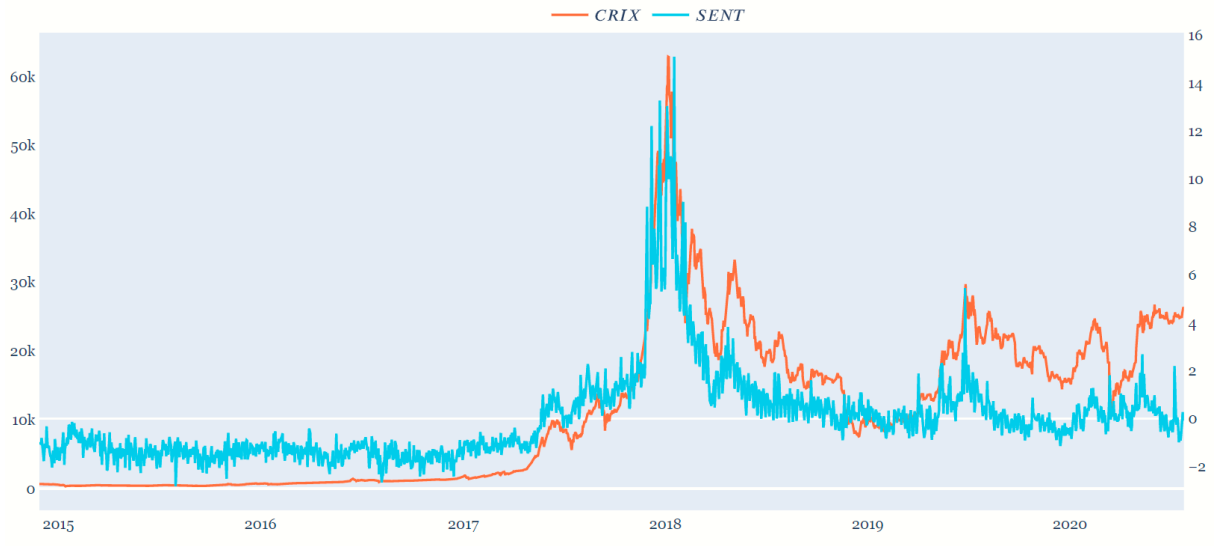

Figure 2.2: The *SENT* index & the *CRIX series*. Source: Own elaboration.

2.2.1 Stationarity Test

To cover various combinations of relationships, the market return series (denoted as RM) and the first difference of the sentiment index (denoted as $\Delta SENT$) will also be studied. The formulas of these series (at time t) are given as follows:

$$RM_t = \frac{CRIX_t - CRIX_{t-1}}{CRIX_{t-1}} \tag{2.3}$$

&

$$\Delta SENT_t = SENT_t - SENT_{t-1} \tag{2.4}$$

Before applying VAR, it is necessary to test the series' stationarity using the KPSS test (Kwiatkowski et al., 1992) and the ADF test (Dickey & Fuller, 1979). It is noted that the two tests have opposite null and alternative hypotheses where the KPSS test has the null of stationarity. The test results are summarized in Table 2.6. The $\Delta SENT$

and the *RM* series are found to be stationary and not contain any unit root, as could also be seen from their graphs in Figure 2.3.

Table 2.6: Stationarity Tests. Source: Own elaboration

	KPSS	**p-value**	**ADF**	**p-value**
SENT	4.3250	<0.01	-2.3810	0.1472
CRIX	1.9274	<0.01	-2.1351	0.2306
$\Delta SENT$	0.1561	>0.1	-11.0189	0.0000
RM	0.0342	>0.1	-16.3843	0.0000

This also implies that the *SENT* and the *CRIX* series are non-stationary time series with an order of integration of 1[14]. The later *VAR* analysis will be applied to the two stationary series, namely $\Delta SENT$ and *RM*. A plot of these two series could be found in Figure 2.3. A side note is that it is also possible to apply *VAR* on the pre-transformed variables (*SENT* and *CRIX*) if we could verify whether they are co-integrated. If this is the case, our *VAR* model will become a *Vector Error Correction Model (VECM)*, which could be thought of as a restricted version of *VAR*. However, since the relationship between the two stationary variables ($\Delta SENT$ and *RM*) is already straightforward to be interpreted in from a financial perspective (and also commonly investigated in studies concerning market sentiment), the author decides to proceed with these two variables and does not check for the co-integration between the *SENT*

[14] Non-stationary series that takes 1 time of differencing to become stationary.

and *CRIX* series.

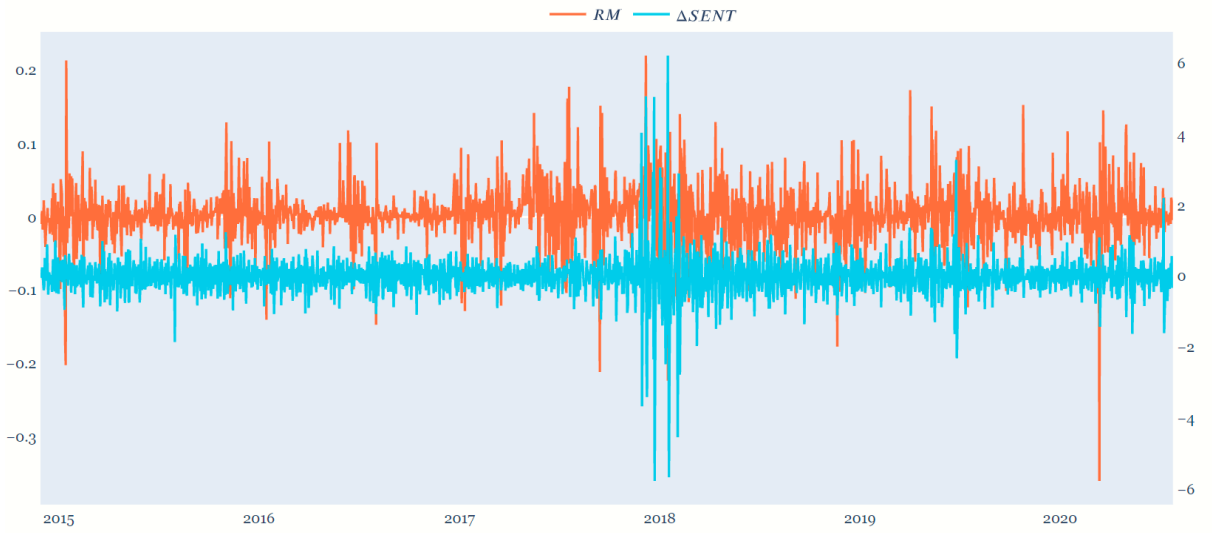

Figure 2.3: The $\Delta SENT$ index & the *CRIX* Return (RM). Source: Own elaboration.

2.2.2 VAR Results

The general *VAR(p)* model specification is given as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \varepsilon_t \tag{2.5}$$

where $Y_t = [RM_t, \Delta SENT_t]^T$, α is the vector of intercepts, β_i is the time-invariant (2x2)-matrix of coefficients, and ε_t is the vector of error terms. Goodness-of-fit criteria *BIC* suggests $p = 5$.

Table 2.7 reports the result from estimating the *VAR* model using daily market return (RM) and daily changes in the sentiment index ($\Delta SENT$). The blocks of rows indicate the contribution of each independent variable at lags 1 - 5. As we want to find out the influences of sentiment on market return, the first block of rows and the first column

should be our primary interest.

The first block of rows shows that $\Delta SENT$ is a powerful predictor of itself. All lagged values of the series from 1 to 5 day(s) are negative and significant at 1%. This outcome is very similar to what is observed by Brown & Cliff (2004) in the stock market where their composite index also displays strongly negative autocorrelations.

However, at the part showing from the relationship between the market return variable RM and lagged values of $\Delta SENT$, we acquire different results compared to Brown & Cliff (2004). Our market return variable RM is significantly positively correlated with $\Delta SENT$ at lag 1, 3, 4 (at 1%), and lag 5 (at 5%), while their sentiment index shows no correlation at all with their market variables. One possible explanation for this phenomenon has been discussed in section **Error! Reference source not found.**, where we anticipate that sentiment influences should be stronger (and consequently, easier to be captured) in a young and innovation market like crypto than a well-established market like stocks. Moreover, the positive correlation of RM and $\Delta SENT$ at all lags is also consistent with the expectation that rising (falling) waves of sentiment lead to temporary increase (decrease) in cryptocurrency prices in the short horizon.

On a side note, changes in the sentiment index are also significantly correlated with the lagged market return (at 1% level with lag 1, 4; at 5% level with lag 2). Thus, one might also suspect that returns could drive the changes in sentiment in the opposite direction.

Table 2.7: VAR(5) Results (RM & $\Delta SENT$). Source: Own elaboration

Independent Variable	**Lag**	**Dependent Variable**	
		RM	$\Delta SENT$
$\Delta SENT$	1	0.0088***	-0.2349***
	2	0.0004	-0.3481***
	3	0.0045***	-0.2748***
	4	0.0044***	-0.2132***
	5	0.0038**	-0.2154***
RM	1	-0.0427*	0.9456***
	2	0.0198	0.6894**
	3	0.0152	0.2291
	4	0.0317	1.3735***
	5	0.0034	0.2589

* Indicate significance at the 10% level
** Indicate significance at the 5% level
*** Indicate significance at the 1% level

There is a popular mantra in the world of statisticians that goes *"Correlation is not causation"*, referring to the inability to formally deduce a cause-and-effect relationship between two variables solely on the basis of an observed correlation between them. Thus, it is not quite straightforward to interpret the significant estimates of the previous VAR model into a statement of predictability between the sentiment index and the market return. To determine whether the composite sentiment index is useful

in forecasting the market returns series, one has to rely on a different statistical test called *Granger-causality*, which we will turn our attention to in the next section.

2.2.3 Granger-causality Analysis

The *Granger-causality* test, first proposed by the Nobel laureate Clive W. J. Granger (1969), is a statistical test for assessing whether adding a time series leads to significantly better *forecasts*[15] of another one. A time series X is said to *Granger-cause* (i.e., have a predictive power of) another series Y if it can be shown that the past values of X provide statistically significant information about the values of Y. By applying the *Granger-causality* test on the changes in sentiment index and the market return, we retrieve the results as presented in Table 2.8. Since the null hypotheses of no *Granger-causality* are rejected in both test directions with very small p-values, it could be seen that the two variables, namely $\Delta SENT$ and RM, are both strong predictors of the remaining one. Interestingly, our results coincide with the results of similar papers (but done in different markets) such as the ones by Alrabadi & Waleed (2015) (who studied the stock market of Jordan), Khan & Ahmad (2018) (who did a nearly identical analysis in the stock market of Pakistan). It could be seen that the markets where sentiment index is documented to have a predictive power of returns (using the above approach) are often young, emerging, and volatile ones. This further confirms our previous anticipation of how investors in novel markets often get swayed by waves of sentiment. In contrast, studies conducted in a well-established market,

[15] *Forecast* is the term used by Hamilton (1994) rather than the original author of the test.

such as Brown & Cliff (2004), show no short-term return predictability in their sentiment index.

Table 2.8: Granger-causality Test Results. Source: Own elaboration.

Null Hypothesis	**Test Statistics**	**Critical Value**	**p-value**
$\Delta SENT$ does not Granger-cause RM	7.597	2.216	0.000
RM does not Granger-cause $\Delta SENT$	7.305	2.216	0.000

2.3 A Simple Trading Strategy

Since the short-run return predictability of the composite sentiment index has been proved to exist, intuitively, one could think about the possibility of profitable trading strategies that are implemented based on this index. Thus, the following section will attempt to simulate and evaluate the economic value of a simple trading strategy with the sentiment index as the main indicator. The methodology of this section is inspired largely by Hilpisch (2014) and Chen et al. (2019).

Based on the sentiment index, one could try to predict (point estimate) the market return using the regression equation of the VAR model from section 2.2.2, given formally as:

$$\widehat{RM}_t = \widehat{\propto} + \sum_{i=1}^{5} \hat{\beta}_i \Delta SENT_{t-i} \; + \sum_{i=1}^{5} \hat{\delta}_i RM_{t-i} \tag{2.6}$$

where the constant $\widehat{\propto} = 0.002516$ and other lagged coefficients of $\Delta SENT$ and RM ($\hat{\beta}_i$ and $\hat{\delta}_i$, respectively)

could be found in table 5.7. The rule to generate trading signals is simple, in which we go long ($BUY = 1$) when the forecasted return is greater than 0, go short (BUY = -1) when the forecasted return is less than 0, and wait (do nothing) otherwise. In mathematical terms, this could be expressed as:

$$BUY_t = \begin{cases} 1, \ if \ \widehat{RM}_t > 0 \\ -1, \ if \ \widehat{RM}_t \ < 0 \\ 0, \ otherwise \end{cases} \tag{2.7}$$

Thus, the cumulative return of the sentiment strategy at day t is given by:

$$R_t^{SENT} \ = \ \prod_{i=1}^{t} (BUY_i * RM_i \ \ + \ 1) \ - \ 1 \tag{2.8}$$

By fitting the sample data of RM and $\Delta SENT$ from 29/11/2014 to 20/07/2020 (a total of 2035 days) using equation 5.6, we receive the point estimates of RM (or $\widehat{RM}$) from 05/12/2014 to 25/07/2020. Next, trading signals and cumulative returns of the strategy will be derived accordingly, based on equation 5.7 and 5.8. The cumulative returns of sentiment-based strategy are then plotted against the ones of the buy-and-hold strategy in Figure 2.4. It could be seen that although the sentiment strategy could not capture the whole upside during the crypto market boom in late 2017 and early 2018, the strategy outperforms the market quite significantly over the whole sampling period. It appears that our strategy performs relatively well (in comparison with the market index) during the downturn after January 2018. Over the entire examining period, the sentiment-based strategy achieves a daily return of ~31bps[16], around 1.73 times of the buy-and-hold strategy return (~18 bps).

[16] 1 basis point (bp) is equal to 0.01%.

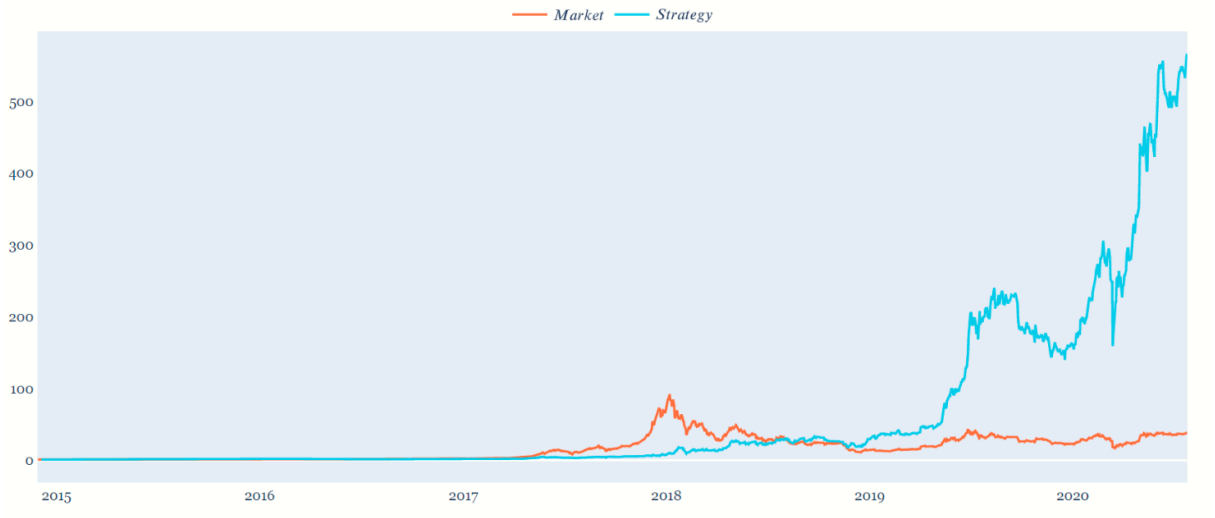

Figure 2.4: Cumulative Returns of the CRIX buy-and-hold Strategy & Sentiment-based Strategy (05/12/2014 to 25/07/2020). Source: Own elaboration.

The results indicate that the composite sentiment index can potentially gain value to the investors who use them as a trading indicator. In the meantime, it further confirms our argument that sentiment plays an important role in predicting future cryptocurrency market returns. However, an important note to be mentioned here is that there is a long way to go from a simulated strategy to a trading system applicable in the real-world. Speaking in an ambitious manner, the strategy in practice could either perform better or worse than the one simulated here. In a way, it could be worse, since in order to backtest the simulated strategy, operational issues (e.g. trade execution) and relevant market microstructure elements (e.g., transaction costs) are completely neglected. Generally speaking, these real-world factors often increase risks and diminish returns in our trading. In another way, the strategy could also be better since the one simulated here is very simple and there are a lot of rooms for

improvement. For example, an investor may attempt to impose several more sophisticated fund management techniques than the one employed in our simulated strategy (i.e, long and short the same amount every time). An interesting question that might arise is what would happen if our trading volume varies according to the confidence interval of the predicted returns. The answer is difficult to say and we save this question for further researches or practical implementations in the future.