# Forecasting of the cryptocurrency market through social media sentiment analysis

Author: Adam Salač
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

**ABSTRACT,**
This paper aims to evaluate the possibility of using data from social media sites to run sentiment-analysis-based predictions on the Bitcoin price developments. In contrast to preexisting literature, it also aims to compare the feasibility of Reddit data in comparison to the current de facto benchmark source of sentiment data, namely Twitter. The data scraped from these social media sites is evaluated using the VADER sentiment analysis toolkit and compared in different time intervals to historical Bitcoin price developments over the course of three months.

**Graduation Committee members:**
**Dr. Fons Wijnhoven**
**Dr. Matthias de Visser**

**Keywords**
Cryptocurrencies, Bitcoin, sentiment analysis, forecasting, social media, Twitter, Reddit

# 1. INTRODUCTION

Artificial intelligence has emerged in the past years as a topic that shifted from works of fiction into first real-world applications as computing power for implementation has become sufficiently available.

One of the aspects of AI is the so-called sentiment analysis, which is commonly defined as the computational treatment, evaluation and coding of opinion, sentiment, and subjectivity in text (Pang, 2008) and represents a tool which can process extraordinarily large text segments, which couldn't feasibly been analyzed manually or using previous tool in a relatively short time periods. One particularly exciting source of data to be used for sentiment analysis are social media posts, particularly on platforms where posts are publicly available. Sentiment analysis of social media posts can be used in different business contexts, such as marketing (Rambocas, 2013) or stock return (Mittal, 2012). It is also to be noted that the use of social media is on the rise and it is currently estimated that 2.77 billion people use social media and according to Statista (2019), this number is expected to rise in the next years. While the cryptocurrency market shrank significantly after the bursting bubble which was precursed by a sudden surge of interest of the general public in late 2017, which was mainly triggered by the vision of quick enrichment, cryptocurrencies have since then matured and challenger banks are now offering cryptocurrency-based services. One example is Revolut, which has implemented in 2018 a cryptocurrency wallet into their banking application, which is currently being used by more than 4 million users across Europe and Australia. This factually means that customers can pay in regular stores by card and the amount can be converted on-the-fly and deducted from their cryptocurrency wallet. Other providers like Coinbase offer a prepaid debit card, which is entirely backed by cryptocurrencies. This indicates a shift in the market, but also that cryptocurrencies are not to be seen as a relic of the recent past. However, cryptocurrencies are still very volatile, and to this date, there is no established model for forecasting. Past research has also assessed the value of Twitter for forecasting altcoins (alternative non-mainstream cryptocurrencies) (Steinert, 2018) and stock market developments (Kirlic, 2018) with success. Furthermore, there are also several commercial providers of cryptocurrency forecasting/prediction services, who claim to evaluate social media sentiment; however, the specific methodology applied by any of these services (such as Coinpredictor) is unclear and undocumented. Twitter is currently ranked by Alexa.com as 11th most visited website and has already been subject to many pieces of research using sentiment analysis, in comparison Reddit ranks slightly lower as 21st most visited website, but the amount on research using Reddit in regards to sentiment analysis is comparably negligible, as Google Scholar provides over 160000 potential results for Twitter and merely 21000 for Reddit as of April 2019. The focal point of the bachelor thesis is therefore to explore a potentially novel source for sentiment analysis concerning a market that has previously been found to be well-forecastable employing sentiment analysis and to provide a comparison. There is a variety of reasons why Reddit should be considered in more details as source for sentiment data and these are both topic-related and general. As a general advantage, Reddit posts are not bound by any restrictions on character count and due to the thread-centered design of Reddit, comments are usually on-topic and even moderated if this requirement isn't fulfilled, while Twitter is basically unmoderated unless there is a direct violation of the Terms and Conditions of Twitter (e.g. racial discrimination or hate speech). This suggests a higher quality of posts. Furthermore, the topic-related advantage should be mentioned. Cryptocurrencies in general have been a very common topic on Reddit and it even became the origin of several high-key cryptocurrencies, such as DogeCoin which was basically developed by frequent "Redditors".

# 2. BACKGROUND

Unlike traditional investment options such as stocks or forex trade, which based on current research seem to be strongly correlated with macroeconomic news and trends (Birz, 2011), cryptocurrencies are not backed by value creation in the real world, and their value is, therefore, more speculative. This suggests that the correlation link between macroeconomic developments and cryptocurrency price developments is weak.


**Figure 1: Comparison between Dow Jones Industrial Average (DJI:NYSE) index and BTC/USD price between January 2016 and April 2019 (Investing.com)**

For this purpose, Figure 1 presents a graph indicating the price in USD of the cryptocurrency with the highest market capitalization, which is Bitcoin (BTC) in comparison to a more traditional indicator of economic development, which in the example of this research is the Dow Jones Industrial Average index. From the graph, it is immediately visible that the Bitcoin price development has been far more volatile and prone to sudden and significant changes throughout the specified timeframe from January 2016 to April 2019. Additionally, a link between the DJI index (representing simplified macroeconomic situation) and the BTC price cannot be assumed. This assumption was further validated in existing research on this topic. Yermack (2013) concluded that daily Bitcoin price developments show no correlation at all with common fiat currency tradings.

Existing literature also suggests that there is a link between composite sentiment scores and intraday cryptocurrency changes based on Twitter data (Zamuda, 2019).

Furthermore, literature evaluating the forecasting ability for stock prices suggests that it's particularly suitable for short-term trading (Kirlić, 2018; Costa, 2018), which is also known as day-trading. This can be applied as well to the cryptocurrency, where fast-paced trading due to low entry barriers and low transaction fees can be assumed.

This leads to the problem statement. Despite the crash in early 2017, the overall capitalization and the total transaction number of transactions with cryptocurrencies is on the rise, and if cryptocurrencies are meant to become more present in daily life, even if marginalized as an investment option within a diversified portfolio, one or more clear forecasting models should be developed. Due to the fast pace of developments, this is and will be subject to constant change. Regardless, this won't necessarily

fit the scope of this research, but this research could provide some basis for the development of such model.

# 3. PROBLEM STATEMENT AND RESEARCH QUESTION

As mentioned previously, there is a clear research gap on utilizing Reddit as forecasting source through sentiment analysis. Furthermore, no adequate research could be found on the particular combination of comparing Reddit posts for sentiment analysis and prediction in regards to cryptocurrencies. Furthermore, as options for cryptocurrency utilization become more available end users, both as means of payment and investment due to the ongoing maturation process of cryptocurrency, any reasonable research on this topic can yield results which are beneficial for different stakeholders. As an example of practical relevance, financial service providers who offer cryptocurrency-based product could potentially create a simple scoring system for different cryptocurrencies as an indicator of forecasted value change. Future research could subsequently aim to develop a model for cryptocurrency forecasting based on the source assessment intended as part of this research (Twitter vs. Reddit). Another potential use scenario would be an automatized trading service as part of an investment scenario. Different disciplines within artificial intelligence such as machine learning in the case of research done by Alessandretti (2019; 2019), has found that cryptocurrency investment portfolio can yield profit even in an overall declining market through the use of machine learning tools. From an academic perspective, this research could provide a starting point for further research on Reddit sentiment analysis and its measured reliability for forecasting measures, as the demographics using Reddit differs from other social media platforms. It also should be noted that Reddit is different by design as a platform, as it more like a topic-based forum with clear distinction, as the off-topic discussion is not permitted on a particular subreddit within Reddit. Twitter, on the other hand, is unrestricted in that sense as any user can write anything anywhere. This distinction makes the use of these two alternate sources viable, because they complement each other, each catering a different target user group. While it can be assumed that Twitter will be more dominant on the quantity of posts, Reddit will prevail in average length of posts and this is not a negligible aspect for sentiment analysis.

However, in order to develop a structured research design, a set of research questions has to be phrased.

*Main research question:*

To which extent can sentiment mining of social media forecast Bitcoin price developments?

*Subquestion 1:*

How does Reddit compete as a source for sentiment mining in comparison to Twitter?

*Subquestion 2:*

Can a prediction model based on sentiment changes yield better than random accuracy?

This set of research questions also leads to the adequate hypotheses.

*H0: Social media-based sentiment analysis can predict Bitcoin prices with a higher than random accuracy.*

*HA: Social media-based sentiment analysis cannot predict Bitcoin prices with a higher than random accuracy.*

# 4. OUTLINE OF SENTIMENT ANALYSIS IN SOCIAL MEDIA

Sentiment analysis has found widespread use in combination with social media, as social media is a good source of valuable and sentimental, however unstructured data itself is of little value for real world applications (IBM, 2017) and social media posts fall into this category. Therefore, sentiment analysis is the ideal tool to transform this unstructured data into tangible and processable information. In terms of sentiment analysis, it is to be differentiated between two main concepts, one is the wordlist-/lexicon-based approach such as VADER and the other one is artificial intelligence based, such as Google NLP engine. Artificial intelligence based sentiment analyzers however, to this point, reach lower correlation with real human sentiment evaluations as described by Hutto (2014), which leads to the assumption that lexicon-based analyzer still represent the more matured and reliable choice for sentiment analysis, even in the context of social with high informality of language.

# 5. LITERATURE REVIEW / THEORETICAL FRAMEWORK

Stenqvist (2017) developed a model for forecasting BTC price developments based on Twitter posts over a time frame of 31 days, gathering 2271815 tweets and subsequently analyzing them using the VADER lexicon method. The compound sentiment scores with a threshold of 0.5 were then compared to the BTC/USD price based on time-series with intervals ranging from 5 minutes to 4 hours. The findings concluded that the data can predict price development directions with an accuracy of up to 83%.

Kaminski (2014) also evaluated Twitter data using the “bitcoin” keyword over a timeframe of 104 days, yielding approximately 160000 tweets and using a simpler word-based sentiment analysis, however concluding that the causality is reversed and Twitter post sentiments merely emotionally reflecting the intraday Bitcoin market developments.

Karalevičius (2018) used sentiment analysis to predict intraday BTC price movements using data from expert news sites and evaluating them using document-based sentiment analysis by using a combination of the Harvard psychosocial dictionary and a financial lexicon developed by Loughran and McDonald (2011), concluding the growing maturity of the Bitcoin trading market and suggesting that expert news can predict semi-short term BTC price developments.

Bukovina (2016) is also to be considered, as it is one of the few papers reviewing the option of evaluating Reddit sentiment to explain BTC value volatility. Herefore, external sentiment analysis data from sentdex.com was used in a modified model by Saxa (2015) (originally used for mortgage forecasting). The results correlated with those by Georgoula (2015), claiming that it only explains for part of the volatility, but that there is statistically significant difference between negative and positive sentiments, whereas positive sentiments carry a stronger explanation to positive changes in BTC value.

Another research paper that can be used to get a better overview of the Bitcoin market and the sentiments of social media users who are active in this market, is the paper by Hernandez (2014). Here it is suggested that Twitter users interested in Bitcoin are less expressive about their emotions and sentiments on social media, which is an aspect that is of high importance and consideration for sentiment analysis.

Using this short summarization of existing research, a basic theoretical framework can be drafted. It can be concluded, that findings on the explanatory value of social media sentiment is

varied, however most papers conclude that it can indeed explain some of the volatility and possibly even predict changes to an extent. It also signifies the value of natural language processing and sentiment analysis for this topic. Most papers use a time-series based approach for evaluation of BTC value. This should also be applied to this research.

# 6. METHODOLOGY

In order to develop a model which would create a suitable methodological approach.

The goal of the methodological approach is to evaluate the link between social media sentiments and forecasted value of cryptocurrencies. A set of variables is to be defined so that regression analysis can be done. In this case, the dependent variable will be defined as the price of BTC in USD, while the independent variable will be represented by a sentiment analysis score.

While Tabbari (2019) suggests an approach of only scraping Twitter data from verified accounts to ensure the public interest of data, Bitcoin is vastly different from stock trading due to the rather informal nature and therefore this requirement should be dropped within this study.

VADER has been found to be a suitable tool for sentiment mining, and it was also used in previous related research done by Kirlić (2018) or Steinert (2018). VADER itself is originally a Python library that is used for dictionary-based sentiment analysis, but has been ported to different languages and platforms throughout the time.

## 6.1 Sample selection

In accordance to existing relevant research, data for sentiment analysis was collected over prolonged timeframe, in the case of this research, the timeframe was set to three months, ranging from November 1st 2018 to January 31st 2019. A longer time frame was considered and attempted; however this attempt was met with grave complications as it made calculations exponentially more difficult as some programs used in the workflow (e.g. Excel) had a limit of ~1 million rows. Furthermore, the limited scope of this research as part of a Bachelor thesis should be considered, however despite these limitations this research is as of June 2019 larger in scale than existing related research which has been conducted on this topic.

## 6.2 Data collection

Data needed to be mainly collected from two different sources, namely Twitter and Reddit. For this purpose, a Linux-based virtual machine has been deployed and as for collection software, twint was used to gather data from Twitter. The aforementioned tag words were utilized and the data output was set to JSON for easy data manipulation. Twint offers an integrated option to select timeframes and languages, which in case of this research was split up into monthly chunks and language was specified to English to make subsequent data cleaning steps easier.

| ***Key*** | ***Description*** |
|---|---|
| id | Unique ID of captured Tweet |
| conversation_id | Unique ID of conversation in which the Tweet was posted (also known as "thread") |
| created_at | Timestamp of Tweet in Epoch UNIX UTC millisecond format |
| date | Date of Tweet (YYYY-MM-DD) |
| time | Time of Tweet (HH:mm:ss) |
| timezone | Timezone of Tweet (hardcoded to UTC) |
| user_id | Unique user ID who posted this Tweet |
| username | Username of user who posted this Tweet |
| name | Canonical name / nickname of user |
| place | Geolocation of Tweet |
| tweet | Content of Tweet |
| mentions | Other users mentioned in Tweet ("@username") |
| urls | URLs included in Tweet |
| photos | URL to photo included in Tweet |
| replies_count | Number of replies to this Tweet |
| retweet_counts | Number of "retweets"/shares of this Tweet |
| likes_count | Number of likes of this Tweet |
| location | Tweet geolocation (obsolete) |
| hashtags | Hashtags mentioned in Tweet |
| link | URL permalink to Tweet |
| retweet | URL to retweets (obsolete) |
| quote_url | URL to quote |
| video | URL to video included |

**Table 1. Description of Twitter JSON output keys**

It should also be noted that not all variables contain information in all tweets. This is particularly noticeable with “place”, which is only present in ~0.8% of captured tweets and also often contains bogus data such as URLs. It can therefore be assumed that this data, while on the first sight potentially useful, is of no use for this study.

The second data source is Reddit, with the main focus on the Bitcoin-relevant subreddits. However, considering the scope of this research, only the most popular Bitcoin-related subreddits was chosen as data sources, namely ”/r/Bitcoin”. Data collection was initiated using the Reddit API and a Python-based scraper; however this approach wasn’t satisfactory for this purpose as the API can only provide about 1000 entries per hour to generic users with a normal API key.

| ***Key*** | ***Description*** |
|---|---|
| author | Unique ID of captured Tweet |
| author_* | Specific settings for creating user (visual) |
| can_gild | Ability of user to gild a post |
| can_mod_post | Moderator rights |
| collapsed | Whether comment is collapsed by default or not |
| collapsed_reason | Reason in case comment is collapsed by default |
| controversiality | Extent to which the comment is being met with mixed reactions |
| distinguished | Special post |
| edited | Whether comment was edited after posting |
| permalink | Permanent URL to comment |
| stickied | Whether comment is stickied in thread |
| subreddit | Subreddit in which the comment was mentioned |
| created_at | Epoch timestamp of comment |

**Table 2. Description of JSON Reddit output keys**

This speed was insufficient to gather a sufficient dataset within the timeframe for this thesis. Therefore, a preexisting scrape of Reddit comments maintained by Jason Baumgartner at pushshift.io was used. This dataset contains all Reddit comments from all subreddits and is formatted in JSON format, containing the actual content, as well as 37 different metadata attributes including a timestamp.

Subsequently, supplementary data on number of Bitcoin transactions as well as number of Blockchain wallets were scraped from Coinbase. Finally, a dataset containing the BTC/USD price in hourly intervals was downloaded from coinmetrics.io.

The resulting datasets overall 999879 Tweets (JSON, 474481056 byte / ~474.48 MB), 124681323 Reddit comments (JSON, 139680032464 byte / ~139.7 GB), 92 time points of number of wallets (JSON), 92 time points of daily transaction numbers (JSON) and 364 BTC/USD price values. The discrepancy between number of Reddit comments and tweets can be explained by the difference in data collection methods, as in the case of Twitter the data is being fetched with a pre-existing filter on keywords, while the raw dataset from Reddit contains all comments on all topics in all subreddits in the given timeframe.

## 6.3 Data cleaning

The acquired datasets however need to be altered to remove undesirable data which could negatively influence the credibility of this study. Based on data from table A, it is observable, that various data is either redundant (e.g. “created_at” and “time”, “date”, timezone”) or of no value for the sake of this sentiment analysis. To speed up the sentiment analysis, the redundant and invaluable data has to be removed.

The most important data for this analysis is the timestamp and the content of the tweet or comment. For timestamp, first, the UNIX Epoch timestamp format has to be briefly outlined. As per the current definition by IEEE Open Group, Epoch time specifies the time as a number, which outlines how many milliseconds have passed since 00:00:00 UTC, January 1st 1970. Given the number 1546300798000 from the first row in the Twint output and the aforementioned conditions, it can be calculated that the human-readable timestamp of this Tweet is 23:59:58 UTC, December 31st 2018. This format was chosen as it allows easy manipulation using only one parameter for time and date. Cutoffs to narrow down data to a certain day or hour can also easily be used by reducing data to entries with timestamps between these two Epoch values (e.g. data for December 31st 2018 have a timestamp between 1546214400000 and 1546300799000). In result, the Twitter data was stored in JSON format merely containing two keys, namely “created_at” and “tweet”. The reduction of keys and objects was entirely achieved using jq JSON command line editor.

| ***created_at*** | ***tweet*** |
|---|---|
| 1546300798000 | Bitcoin Could Revolutionize Governance, Says Cypherpunk Jameson Lopp: According to Jameson… https://goo.gl/fb/55LxMg |
| 1546300791000 | Looking at my slightly over a year-old Bitcoin price prediction for the end of 2018, wow, I was way off https://twitter.com/bascule/status/937014918359932928 … |
| 1546300780000 | Wishing you a bright 2019 from your friends at Electrumdark. It's been an extremely great year, but we aim to achieve even more in the coming year. We thank you for your continuous support and look forward to impressing you in the future.#electrumdark #bitcoin #NewYearsEve pic.twitter.com/RlygtIWAo2 |

| | |
|---|---|
| 1546300772000 | #Happy2019 #Bitcoin Ends the Year #2018 in the #Red @ $3742 #GMT Fingers crossed for 2019 #btc pic.twitter.com/jcQXp2Jt9i |
| 1546300763000 | Bitcoin (BTC) Holds Steady Above $3,400, But Analysts Still Believe Further Losses Could Be in Store http://bit.ly/2GbiwlE |

**Table 3. Sample of formatted Twitter data**

The Reddit dataset was first narrowed down for each month to the most relevant subreddit for the scope of this study, which is "/r/Bitcoin", then these segments were combined into a large file again. This yielded a dataset with 260348 comments from "/r/Bitcoin". Language is not a concern for this dataset, as English is the only permitted language on both subreddits and posts or comments violating this rule are actively monitored and deleted. The keys of the JSON file were truncated to the equivalent of the aforementioned keys in the Twitter dataset, which represent the time and the content of the comment. In this case, it should be noted that the order of the data is reversed as the keys are sorted alphabetically and in this dataset the content of the comment, which is the unit of analysis in this research, is coded as "body". This leads to a problem, as for the process of running VADER sentiment analysis using the Python script that was created for this purpose, unified datasets are required. Therefore, the Twitter was adjusted to match the format of the Reddit dataset, which means that the keys were set as "body" and "created_by". This was achieved with a bash script which replaced "tweet" with "body" (using the Linux program sed) and then the order was adjusted using the sorting function of the Linux-based JSON data editor jq. The precise steps for this are in the scripts in appendix A and B of this paper.

| ***created_at*** | ***body*** |
|---|---|
| 1546300737000 | The best thing to do and I'm sure others will agree is to take a break from the charts for a little while "missed out" on the drop from 6k because I was having a break. Taking that month break was the best thing I have ever done directly because of having some time off. However, expecting an increase after the ETH mess up. Happy new year :) |
| 1546300377000 | So you're targeting a rally above 4? 4.5k? from around here first in January, then another drop to test new lows right after? |
| 1546300254000 | What was the drop on the last difficulty adjustment? |
| 1546299516000 | I think with crypto such a shitshow and eth having dropped so much, any real development could result in a bump. Im not expecting much, and some of its already happened the last ciuple weeks, but it should be something and hopefully it extends to btc a bit. |

**Table 4. Sample of formatted Reddit comment data**

The supplementary data of number of transactions and wallets from Coinbase also had to be processed to match the aforementioned format. Timestamps were converted from the original CSV file to UNIX Epoch millisecond timestamps.

| ***created_at*** | ***wallets*** |
|---|---|
| 1546293272000 | 31914414 |
| 1546275272000 | 31906385 |
| 1546257272000 | 31894081 |
| 1546239272000 | 31894081 |

**Table 5. Sample format of cleaned wallet number data**

| ***created_at*** | ***transactions*** |
|---|---|
| 1546214400000 | 259684 |
| 1546128000000 | 267639 |
| 1546041600000 | 264980 |
| 1545955200000 | 308267 |

**Table 6. Sample format of cleaned transaction number data**

The number of wallets and transaction number as well as transaction volume and market capitalization are intended as control variables or potential information to better perceive the scale of the research and datasets in relation to the entire Bitcoin market.

## 6.4 Sentiment analysis

For the actual sentiment analysis of gathered data, VADER is being used due to its proven track record in previous research. The default lexicon of VADER is used for sentiment analysis, as it also contains colloquial and informal expressions, emotional abbreviated expressions and finally emojis as well as emoticons, which are common on social media posts due to the character count limitation, as research by Sari (2014) suggests. The data excerpts in the previous section also suggest that the gathered posts may contain profanity, which is something that is also better handled by the VADER lexicon than by for example the Harvard psychosocial lexicon used by Karalevičius (2018), which has been checked (http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm) and observed not to contain the profane language as seen in the Reddit data excerpt.

This procedure was fulfilled by running a self-developed Python script which called the VADER Python library and piped each text through the sentiment analyzer and subsequently appended the compound score as new key in a new output JSON file. In accordance with existing research, posts with a compound sentiment score of zero were truncated from all resulting datasets to minimize bias to the center as these text samples were of no value for forecasting and instead dampened the meaningfulness and explanatory power of the data. This reduced the size of the datasets from 999879 samples from Twitter to 610290 samples and in the case of Reddit, the dataset was reduced from 260348 samples to 186026 samples. This descriptive data should not be neglected, as we can see that Reddit has a higher ratio of useable data (71.45% vs. 61.04%). Additional descriptive statistics of the data can be found in the outputs in the appendix.

| ***body*** | ***created_at*** | ***compound*** |
|---|---|---|
| According to @DowJones Market Data, January will mark the 6th consecutive month of losses for #Bitcoin, causing more FUD to be spread. | 1546293272000 | -0,7028 |
| Bitcoin for me represents hope | 1546293272000 | 0,8478 |

| | | |
|---|---|---|
| for a better future for all humans. Society is decaying and at the root of it all is soft money (fiat) and centralized controls led by humans. People talk about leaving things better for future generations. #bitcoin inspires conviction! | | |

**Table 7. Resulting sample table with compound sentiment scores**

It should also be noted that it was observed that most of the aforementioned samples with zero sentiment were caused due to very specific slang, empty contents or typographical or grammatical errors which hinder the accurate sentiment analysis.

Subsequently, the data was imported into Excel into three different spreadsheets – one for Reddit, one for Twitter and one for the combined data - and a pivot table was used to generate average compound sentiment scores as well as the count of scores in each table for each timeframe (24 hours, 12 hours and 6 hours). The resulting data was imported into IBM SPSS as preparation for the statistical tests described below. Hereby, three datasets were created in SPSS, which each contained all necessary data for the respective timeframe.

## 6.5 Correlation and Regression

The regression model for this study aims to model the relationship between the dependent variable, which is Bitcoin price in USD and the independent variable, which is the compound sentiment score provided by VADER for each chosen time interval. This approach was adapted from the research by Greaves (2015). This linear regression model is individually created for each time interval, each for using Twitter and separately for using Reddit as data source and finally using the compound data from both sources to check whether a mixed source approach yields more desirable outcomes than the individual sources.

An alternative approach for consideration can be adapted from research by Stenqvist (2017), which uses binary prediction vectors and multiple price change thresholds, however, this approach was not deemed suitable for the type of this research.

The prepared data is processed using SPSS for each time frame. First, the Pearson correlation between variables is being checked. It's noticeable that there is a surprisingly low correlation between the sentiments of Twitter and Reddit posts. In terms of the daily data, it can be observed that both data sources correlate statistically significantly with the Bitcoin price difference to the following day, more so than representing the status quo Bitcoin price for that given day.

Subsequently, a linear regression is being calculated. It is interesting to observe that in all three timeframes, the mixed model has a higher R-squared value, which signifies that a mixed model using data from both sources can better explain the variance of the independent variable.

| *R square* | Reddit | Twitter | Mixed | N |
|---|---|---|---|---|
| **24 hours** | 0.095 | 0.312 | 0.337 | 92 |
| **12 hours** | 0.128 | 0.317 | 0.442 | 184 |
| **6 hours** | 0.111 | 0.372 | 0.415 | 368 |

**Table 8. R-square table for Bitcoin price difference**

Also, it is noteworthy that the observation from the correlation pointing out a potential connection between the price difference and compound sentiment score, rather than the Bitcoin price itself, is being confirmed by the linear regression. The R square value again is significantly higher in all cases in comparison to the linear regression with Bitcoin price as independent variable.

| *R square* | Reddit | Twitter | Mixed | N |
|---|---|---|---|---|
| **24 hours** | 0.067 | 0.234 | 0.255 | 92 |
| **12 hours** | 0.043 | 0.192 | 0.211 | 184 |
| **6 hours** | 0.088 | 0.185 | 0.201 | 368 |

**Table 9. R-square table for Bitcoin value**

We observe that the highest R-squared value (0.442) is in the mixed model for Bitcoin price difference as dependent variable and the mix of Reddit and Twitter sentiments as independent variables for a timeframe of 6 hours.

# 7. RESULTS

The results basically can be seen as confirming existing research regarding the explicability of Bitcoin variance with sentiments from social media. We therefore reject the null hypothesis as there is not a statistically significant explanation for the dependent variable in any of the models. Additionally, is should be noted that the transaction volume as well as transaction number and number of Bitcoin wallets by far outweigh the number of samples in this research, which leads to the conclusion that while there is some correlation between social media sentiment and Bitcoin value, we cannot assume a causation as the mass of users is too insignificant in relation to the Bitcoin userbase.

# 8. DISCUSSION

This research is as of the time of writing unprecedented in scale with almost 800 thousand valid samples of analysis encompassing an approximate of 180 million characters and 20 million words. These numbers also highlight the potential of sentiment analysis, as processing these large sets of data would have been impossible a decade ago. Therefore, research remains unoriented and novel due to the early stage of maturity and applicability of this technology. This is also envisioned in the results, which show seeming confirmation of previous research, namely that the real-world usage of big datasets in combination with sentiment analysis is still a tool of limited and debatable power.

## 8.1 Limitations

The most significant limitation of this research can be appointed to the used sentiment analysis framework, in this case VADER. This is due to the fact that it is a dictionary-based sentiment analyzer and especially social media posts are, as proven by this research, informal by nature and often tainted by scene-specific slang, typographical and grammatical errors and common use of irony and cynicism, which is at this stage very difficult to evaluate by sentiment analyzers. Furthermore, in line with observations from past research and the indications in the Github project page, VADER does seem to have difficulties with evaluating negations in phrases. Additional limitations were

given by the scale of this research and imperfections can be observed in the dataset. As an example, despite multiple attempts through multiple approaches, it was not possible to remove hashtags, links and unusual characters ("\n" – line feed) without corrupting the dataset and therefore inhibiting further analysis. This data overhead increased the overall wordcount of comments or tweets and the wordcount classified as neutral by VADER, which resulted in many cases in a compound sentiment score biased towards the center (0) and therefore decreasing the strength of sentiment. These occur due to an imperfect data cleaning process and it can be assumed that better data cleaning by a more skilled scientist would result in more meaningful research results. However, as this is research conducted at the University of Twente under the open access principle highlighted by the licensing (CC-BY-NC), this research can be reproduced and improved and the datasets, scripts, as well as the precise steps taken will be published in a Github repository (https://github.com/adamsalac/bsc_thesis/). Furthermore, the time series intervals were longer than in existing research, this was due to the lower amount of Reddit comments which upon cleaning would yield irrationally low number of samples for certain timeframes and would not yield any meaningful results. This can already be partially observed in the results in this research in the 6 hour timeframe as the R squared value deteriorates.

## 8.2 Business-related implications

In terms of business implications, we observe that at this stage sentiment analysis of social media, regardless whether Reddit or Twitter, is not yet suitable as exclusive predictor and forecasting model basis for the price development of Bitcoins. Further research would be required to see whether data collected in a similar manner could be useful to aforementioned stakeholders in the cryptocurrency market, be it private users, investors, fintech startups or enthusiasts. Bitcoin remains a rather speculative investment and its movements are often unpredictable, not only by magnitude, but even by direction.

# 9. CONCLUSION

The research has been successful to the extent that the given research gap was well targeted and exploited, however we fail to observe groundbreaking observations. However, it is to be noted that a combination of data sources such as in this case Twitter and Reddit can add a dimension and a different perspective to the model and potentially yield better results than a model based on a single data source. In either case, further researched on this topic should be considered desirable and could with new technological advancements yield new and potentially more promising results.

# 10. ACKNOWLEDGEMENTS

At this point I would like to express my thankfulness to my supportive thesis supervisors, Dr. Fons Wijnhoven and Dr. Matthias de Visser, for continuous support and feedback as well as the creative freedom to devote my time for research that not only correlated with my preexisting interests, but also significantly helped me acquiring new skills and improving existing skills. Furthermore, I would like to thank my family for continuously supporting and embracing my studies despite all hardship endured along the way. Apart from that, I also want to thank my study advisors for not giving up on me, supporting and fighting for me, which enabled me to come to this point. Finally, I also want to thank my friends for their support and the help with specific problems encountered during this research.

# 11. REFERENCES

1. Alessandretti, L., ElBahrawy, A., Aiello, L. M., & Baronchelli, A. (2018). Machine learning the cryptocurrency market. arXiv preprint arXiv:1805.08550. Rambocas, Meena,and João Gama. Marketingresearch:Theroleofsentimentanalysis. No.489. Universidade do Porto, Faculdade de Economia do Porto, 2013.
2. Alessandretti, L., ElBahrawy, A., Aiello, L. M., & Baronchelli, A. (2018). Anticipating cryptocurrency prices using machine learning. Complexity, 2018.
3. Birz, G., & Lott Jr, J. R. (2011). The effect of macroeconomic news on stock returns: New evidence from newspaper coverage. Journal of Banking & Finance, 35( 11), 2791-2800.
4. Bukovina, J., & Marticek, M. (2016). Sentiment and bitcoin volatility (No. 2016-58). Mendel University in Brno, Faculty of Business and Economics.
5. Dwyer, G. P. (2015). The economics of Bitcoin and similar private digital currencies. Journal of Financial Stability, 17, 81-91.
6. Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D., & Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices. Available at SSRN 2607167.
7. Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. No Data.
8. Hernandez, I., Bashir, M., Jeon, G., & Bohr, J. (2014, June). Are Bitcoin Users Less Sociable? An analysis of users' language and social connections on twitter. In International Conference on Human-Computer Interaction (pp. 26-31). Springer, Cham.
9. 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.
10. Karalevicius, V., Degrande, N., & De Weerdt, J. (2018). Using sentiment analysis to predict interday Bitcoin price movements. The Journal of Risk Finance, 19(1), 56-75.
11. Kirlić, A., Orhan, Z., Hasovic, A., & Kevser-Gokgol, M. (2018). Stock market prediction using Twitter sentiment analysis. Invention Journal of Research Technology in Engineering & Management (IJRTEM) , 2(1), 01-04.
12. Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf) 15 (2012).
13. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2( 1–2), 1-135.
14. Steinert, Lars, and Christian Herff. "Predicting altcoin returns using social media." PloS one 13.12 (2018): e0208119.
15. Sari, Y. A., Ratnasari, E. K., Mutrofin, S., & Arifin, A. Z. (2014). User emotion identification in twitter using specific features: Hashtag, emoji, emoticon, and adjective term. Jurnal Ilmu Komputer dan Informasi, 7(1), 18-23.
16. Saxa, B. (2015). Forecasting mortgages: internet search data as a proxy for mortgage credit demand. NATIONAL BANK OF THE REPUBLIC OF MACEDONIA, 107.

17. Tabari, N., Seyeditabari, A., Peddi, T., Hadzikadic, M., & Zadrozny, W. (2018, September). A Comparison of Neural Network Methods for Accurate Sentiment Analysis of Stock Market Tweets. In ECML PKDD 2018 Workshops (pp. 51-65). Springer, Cham.
18. Costa, M., Marschall, L., Mirsadeghi, S. H., & Sanctis, A. D. (2018). Investigation of sentiment importance on intraday stock returns.
19. Yermack, D. (2015). Is Bitcoin a real currency? An economic appraisal. In Handbook of digital currency (pp. 31-43). Academic Press.
20. Zamuda, A., Crescimanna, V., Burguillo, J. C., Dias, J. M., Wegrzyn-Wolska, K., Rached, I., ... & Salomie, I. (2019). Forecasting Cryptocurrency Value by Sentiment Analysis: An HPC-Oriented Survey of the State-of-the-Art in the Cloud Era. In High-Performance Modelling and Simulation for Big Data Applications (pp. 325-349). Springer, Cham

# 12. APPENDIX

## 12.1 Sample script for data collection and editing – Reddit DS

```
#!/bin/bash

#Step 1 - Downloading the Pushshift Reddit comments dataset (replace YYYY-MM accordingly
- check https://files.pushshift.io/reddit/comments/ for list of available months)

wget https://files.pushshift.io/reddit/comments/RC_YYYY-MM.zst

#Step 2 - Uncompressing the dataset (replace YYYY-MM accordingly)

unzstd RC_YYYY-MM.zst

#Step 3 - Delete unnecessary keys from dataset (replace YYYY-MM accordingly)

cat                        RC_YYYY-MM                        |                        jq
'del(.author,.author_cakeday,.author_created_utc,.author_flair_background_color,.autho
r_flair_css_class,.author_flair_richtext,.author_flair_template_id,.author_flair_text,
.author_flair_text_color,.author_flair_type,.author_fullname,.author_patreon_flair,.ca
n_gild,.can_mod_post,.collapsed,.collapsed_reason,.controversiality,.distinguished,.ed
ited,.gilded,.gildings,.id,.is_submitter,.link_id,.no_follow,.parent_id,.permalink,.qu
arantined,.removal_reason,.retrieved_on,.score,.send_replies,.stickied,.subreddit_id,.
subreddit_name_prefixed,.subreddit_type)' > R_YYYY-MM_intermediate.json.json

#Step 4 - Truncate dataset to only include comments from the desired subreddit (replace
YYYY-MM accordingly)

cat R_YYYY-MM_intermediate.json | jq 'select(.subreddit == "Bitcoin")' > R_YYYY-MM.json

#Step 5 - Cleanup (replace YYYY-MM accordingly)

rm RC_YYYY-MM.zst R_YYYY-MM_intermediate.json
```

## 12.2 Sample VADER Python script – Reddit DS

```
__author__ = "Adam Salač"

__credits__ = ["Adam Salač"]

__license__ = "MIT"

__version__ = "1.0"


from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer


import json


analyzer = SentimentIntensityAnalyzer()


#replace MM-YYYY with according filename

def   load_transform_data(input_file_name   =   'T_MM_YYYY.json',   output_file_name   =
"T_MM_YYYY_sentiment.json", encoding ="utf8" ):
    #loading data into a list
    with open(input_file_name, encoding=encoding) as f:
        data = f.readlines()
        data = [str(line) for line in data]
        data = [data[4*i:4*i+4] for i in range(0,int(len(data)/4))]


    # opening the file
    output_file = open(output_file_name, "w")
    for line in data:
        #getting the "text" of the tweet
```

```
        text = line[1].split('"body":' )[1][:-3]
        #analysing it
        score = analyzer.polarity_scores(text[2:])
        compound_score = score['compound']
        #inserting the "compound"
        line.insert(3, f'  "compound": {compound_score} \n')
        #adding a coma in utc
        line[2] = line[2][:-1] + "," + "\n"
        for element in line:
            output_file.write(element)
    output_file.close()
load_transform_data()
```

## 12.3 SPSS outputs - correlations

**Correlations**

| | | Twitter | btcusd |
|---|---|---|---|
| Twitter | Pearson Correlation | 1 | -.193 |
| | Sig. (2-tailed) | | .067 |
| | N | 91 | 91 |
| btcusd | Pearson Correlation | -.193 | 1 |
| | Sig. (2-tailed) | .067 | |
| | N | 91 | 92 |

**Correlations**

| | | Twitter | diff |
|---|---|---|---|
| Twitter | Pearson Correlation | 1 | .354** |
| | Sig. (2-tailed) | | .001 |
| | N | 91 | 91 |
| diff | Pearson Correlation | .354** | 1 |
| | Sig. (2-tailed) | .001 | |
| | N | 91 | 92 |

**. Correlation is significant at the 0.01 level (2-tailed).

**Correlations**

| | | Twitter | Reddit | btcusd | diff |
|---|---|---|---|---|---|
| Twitter | Pearson Correlation | 1 | .242 | -.483** | .609** |
| | Sig. (2-tailed) | | .105 | .001 | .000 |

| | | | | | |
|---|---|---|---|---|---|
| | N | 46 | 46 | 46 | 44 |
| Reddit | Pearson Correlation | .242 | 1 | -.259 | .308* |
| | Sig. (2-tailed) | .105 | | .082 | .042 |
| | N | 46 | 46 | 46 | 44 |
| btcusd | Pearson Correlation | -.483** | -.259 | 1 | -.129 |
| | Sig. (2-tailed) | .001 | .082 | | .405 |
| | N | 46 | 46 | 46 | 44 |
| diff | Pearson Correlation | .609** | .308* | -.129 | 1 |
| | Sig. (2-tailed) | .000 | .042 | .405 | |
| | N | 44 | 44 | 44 | 44 |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

### 12.4 SPSS outputs - regression

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .193[a] | .037 | .026 | 1001.814105398 |

a. Predictors: (Constant), Twitter

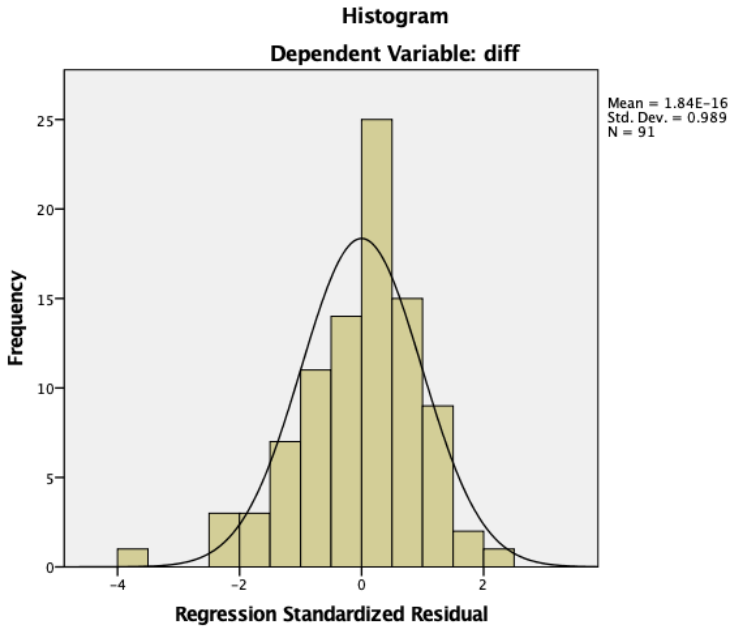
b. Dependent Variable: btcusd

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .427[a] | .183 | .164 | 128.886598639459920 |

a. Predictors: (Constant), Twitter, Reddit

b. Dependent Variable: diff




**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .196[a] | .038 | .017 | 1006.842274526 |

a. Predictors: (Constant), Twitter, Reddit

b. Dependent Variable: btcusd