UCLouvain NHH LOUVAIN School of Management

Louvain School of Management
and The Norwegian School of Economics

Reddit Sentiment and the Relationship to Subsequent Cryptocurrency Returns across Different Segments

Author:
Thomas Bergerud

Master's thesis with the view of getting the degrees:
Master in Economics and Business Administration +
Master in Business Engineering, Professional Focus

Supervisor:
Roberto Ricco

Academic Year: *2022-2023*

Table of Content

1. Executive Summary

This paper investigates the relationship between sentiment and subsequent cryptocurrency returns for 37 different cryptocurrencies across various sectors. The analysis employs three models: Ordinary Least Squares (OLS), random forest, and cross-section returns. The sentiment is gathered from the CryptoCurrency subreddit using RedditExtractoR and the PushShift API. The posts that contain a certain name or ticker, e.g. Bitcoin or BTC, are classified accordingly. The posts are then tokenized and a score is assigned to each word based on a self-made algorithm. It provides a weighting scale that values words that appear in the title the most, followed by words in text and comments. Also, the number of upvotes and comments on a post is used to provide a more genuine representation of the sentiment. Loughran-McDonald is the sentiment dictionary that finally classifies the sentiment, chosen due to its strength with financial data.

A data frame with price, lagged sentiment and several explanatory variables, such as a global uncertainty index and Twitter popularity for Bitcoin is then created and used in three different models. The OLS model performs poorly in gauging a real connection between sentiment and return as the p-values for most of the cryptocurrencies are above a threshold of 10%. Random forest however is showing signs of larger sentiment importance by looking at the node purity and placement of the sentiment variable. However, the results are ambiguous. At last, a cross-section method provides useful results and

shows that a portfolio consisting of meme coins, smart contracts and DeFi tokens outperforms a base portfolio significantly. The sentiment portfolio that ranks cryptocurrencies by Reddit sentiment, buys top 20% and sells bottom 20%, gathered a cumulative return from 2021 to 2023 at 331.99%. In comparison, a base portfolio for these cryptocurrencies gets a cumulative return of 36.73%.

2. Introduction

Nearly 60% of institutional investors have used Reddit for investment decisions according to a survey by Brunswick Group (*Brunswick's 2023 Digital Investor Survey — Brunswick Group* 2023). The fact that a social media platform such as Reddit is a contributor to financial decisions by institutional investors is an indicator of a different world. Although media historically has had an effect on stocks is it interesting to observe how anonymous social media platforms provide more importance in 2023. Investment information now reaches all age demographics in contrast to previous decades. Social media allows a simple way for millions of users worldwide - from all ages and cultures - to engage in conversations about anything. This goes from jokes on the internet to financial discussions about stocks and cryptocurrencies. Large amounts of data and a diversified user group with different skills and backgrounds are available today only one click away. It raises the question if there is useful information to be gathered if done properly.

One event that brought mainstream media attention to Reddit as a financial tool was the short squeeze of GameStop in 2021. The share price pumped from $17.25 to $325.00 between January 4th 2021 and January 29th 2021. There was a large number of institutional investors, such as Citron, that had leveraged short orders on GameStop. This piece of information inspired the subreddit WallStreetBets to invest strongly in the stock, and the price sky-

rocketed causing hedge funds to lose millions. It again gathered mainstream media attention as more people wanted to take part in this. As a result of the large interest, the retail brokerage firm Robinhood had to restrict investor purchases to protect its own liquidity. Many institutional investors lost millions on their leveraged shorts as they got margin called due to the price appreciation. It became a symbol of "rich versus poor", where the poor had its strength in numbers and with a social media culture and hype to gather more interest as well (D. Hu et al. 2021). Examples such as the 2021 Gamestop shortage demonstrate the causation between social media presence and the value proposition of financial assets.

Now, what if Reddit sentiment possesses predictability for other assets as well? The Gamestop shortage of 2021 has gained lots of media attention and has many experts looking to Reddit for financial insight and predictions into alternative assets. Assets such as cryptocurrencies have begun to see massive adoption in the last decade. In fact, more than 500 cryptocurrency subreddits have launched within the last 10 years (Lilya 2023). Currently, the largest crypto-based subreddit is r/CryptoCurrency with over 6.4 million members (*r/CryptoCurrency subreddit stats* 2023). It contains a large number of posts about multiple different cryptocurrencies, with thousands of predictions that may have benefitted many of the users. There may be predictable patterns discovered within these posts that will increase investor returns, similar to WallStreetBets in 2021.

This paper looks deeper into the relationship between sentiment gathered on this subreddit and the corresponding return for a cryptocurrency the next

day. Also, it investigates how the different categories of cryptocurrencies perform among each other. The models will examine cryptocurrency posts on Reddit and calculate a sentiment score to examine how it relates to the cryptocurrency return the next day. The models will determine:

Is there a relationship between Reddit sentiment and subsequent cryptocurrency returns across the different crypto segments?

The three different models will look into a large number of cryptocurrencies to investigate changes among nine different categories. Is a highly-institutionalized and noisy cryptocurrency such as Bitcoin expected to better possess genuine sentiment through Reddit compared to coins with strong communities and less noise? This paper will therefore observe differences in categories such as decentralized finance (DeFi), centralized exchange tokens and meme coins.

The project is mainly focused on the process around coding, but a qualitative methodology was also implemented. An expert interview was done with Loki - the community manager of ThorFi. The protocol is a decentralized finance protocol on the Avalanche blockchain and is one of many cryptocurrency projects that have experienced sentiment change during the last years entering a bear market. His industry knowledge and sentiment experience within decentralized finance provide interesting insight into the concept of sentiment itself. He put emphasis on the overall importance of sentiment as a predictive variable for return and even ranks it as the most important factor. He goes into detail about the representativity of social media sentiment, and how

Reddit is a useful source due to its voting system and interaction. He also challenges the intrinsic value definition of cryptocurrencies (*Defi Sentiment* 2023). That lays an interesting foundation for the results of the different cryptocurrency categories.

The first part of the project is to gather prices for 37 cryptocurrencies. The CryptowatchR API was utilized for this purpose as it is possible to specify which exchange to gather prices from. This makes it easier to handpick the exchange for each cryptocurrency and maximize the chance of collecting elaborate price data. This was helpful due to the huge differences in data availability across the different market caps. Note that Yahoo Finance was used to complement some of the missing data.

Step two of the process is to collect data from the CryptoCurrency subreddit. Both RedditExtractoR and the PushShift API were utilized in order to collect all the relevant posts. If the ticker or the name, e.g. BTC or Bitcoin, occurs in the post, then that post will be connected to that cryptocurrency. The sentiment algorithm is unique as it weights words differently in regards to the placement (title, text, comment) and the number of upvotes. Then a Loughran-McDonald sentiment dictionary is applied to this score to classify the sentiment. Note that it may not properly capture the cryptocurrency- and social media slang such as "hodl", "moon" or "wagmi", but it does possess strength in classifying financial data.

The last step is to use OLS, Random Forest and Cross-Section return as three different methods to investigate the relationship between cryptocurrency sen-

timent and the subsequent return across different segments. This facilitates a linear and non-linear model as well as a cross-sectional return methodology that appeared successful in other literature (Baker and Wurgler 2006). The OLS provide spurious results - a linear model does not seem to be a good fit for the model. Random forest is performing slightly better for certain cryptocurrencies such as Dogecoin (DOGE), as the variable importance of the lagged sentiment is more important than other explanatory variables.

At last, the cross-section method is providing the most insightful results. A sentiment portfolio buys the cryptocurrencies that rank top 20% in yesterday's sentiment and sells the bottom 20%. It performs well compared to a base portfolio that buys all cryptocurrencies. The difference in categories becomes clear as the best portfolio consists of the following categories:

Exchange token (decentralized)/DeFi - Smart Contracts - Memecoin

This provides a cumulative return of 331.99% and outperforms the base portfolio (36.73%). Memecoins and decentralized assets such as DeFi and smart contracts appear to have a stronger sentiment related to the subsequent return according to this model.

The first section goes through existing literature and how this paper provides value to the field. Then there is a methodology section followed by a data description part that looks deeper into the sentiment variable. The results look into the three models: OLS, Random Forest and Cross-Section, before a final section that concludes this paper.

3. Literature Review

This paper aims to build upon existing literature and provide innovation to it. Overall there are three main categories of existing literature that are looked upon:

Stocks and Sentiment

WallStreetBets

Cryptocurrencies and Sentiment

There is extensive literature on the relationship between sentiment and stock return from an overall point of view. Also, the case study of WallStreetBets and GameStop builds upon this literature as well as it utilizes the same social media platform as this paper. This case study is one of the main motivations to look deeper into the relationship between cryptocurrencies and sentiment. There is some existing literature in this field, although it is not as elaborate as the literature for stocks. This is due to the relatively new implementation of cryptocurrency as a whole.

3.1 Stocks and Sentiment

A vital paper in the literature of stocks and sentiment is Malcolm Baker and Jeffrey Wurgler's "Investor Sentiment and the Cross-Section of Stock Returns" with over 6500 citations on Google Scholar per May 3rd 2023 (Baker and Wurgler 2006). The paper challenges classical finance theory to investigate whether a sentiment proxy affects the cross-section of stock returns.

They looked into stock returns from 1962 to 2001 and make the discovery that stocks that are more volatile, newer, smaller and that possess higher growth potential are more likely to be affected by a shift in the investor sentiment. This is an interesting stepping stone for this paper. The concept of cryptocurrencies is relatively new compared to stocks, and many of them are yet to achieve a large market cap although the volatility and growth potential are there. The results of Baker may be an indicator of predictability within a sentiment variable for subsequent cryptocurrency returns.

Also, this paper builds upon the cross-section methodology used in the paper of Baker. It sets a foundation for how to compare a portfolio with and without sentiment and observe the differences. Where Baker is investigating the sentiment portfolio for different firm characteristics, this paper looks into market caps and cryptocurrency sectors. There are multiple sectors within the industry, and investigating the differences between them is a vital part of this project.

3.2 WallStreetBets

One of the most famous subreddits is "WallStreetBets" - a forum where users can share their opinion about stocks. Reddit describes the subreddit as: "Like 4chan found a Bloomberg Terminal" (Reddit 2023). In other words, it highlights the crossing of an anonymous platform and the exchange of financial analysis. This combination lays the foundation of great speculation and definite potential.

The share price of GameStop increased from $17.25 to $325.00 between January 4th 2021 and January 29th 2021. GameStop was highly shorted before this incident, and it created a powerful short-squeeze that threatened the liquidity of multiple institutions that leveraged GameStop. A strong factor for this price appreciation was the collective cooperation from WallStreetBets (D. Hu et al. 2021).

This event contributed to several papers investigating the importance of Reddit for stock prediction, and these methodologies and results have been important to facilitate this paper. Bradley et al. contribute to this research with their paper "Place your bets? The market consequences of investment research on Reddit's Wallstreetbets" (Bradley et al. 2021).

They implement due diligence reports (DD) to determine whether a post represents a buy- or sell signal. They find DD recommendations to be a significant predictor for return one month ahead. However, the results before

and after the GameStop incident provide a distinct difference in predictor abilities. They discover comments being less important after this event and argue that the subreddit is diluted from new investors wanting to join the hype. Nevertheless, the fact that the DD reports tilt towards young and volatile stocks sets the foundation for interesting results for cryptocurrencies.

A net DD approach was considered for this paper as well, but the smaller amount of data in the CryptoCurrency subreddit compared to peak WallStreetBets made it insufficient to do this appropriately. The sentiment is therefore represented continuously with a thorough method that will be elaborated on later in this paper.

Another paper that focuses on Reddit and stocks is Hu et al. with the paper: "The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery" (D. Hu et al. 2021). They discover the predictive power of future returns being quite short and limit it to ten days ahead. The paper also investigates different categories and discovers that meme trading is a huge driver for sentiment prediction, more than facts- and experience-based submissions. This fits well in conjunction with cryptocurrencies with less intrinsic value but strong community backing.

Both papers have in common a linear regression model to investigate the coefficient and p-value of a sentiment variable. This paper will use an ordinary least squares model in this approach. Note that the other papers used a panel regression model. The main reason OLS was chosen over a panel regression was due to the fact that this paper looks deeper into the differences among

crypto segments, and how the sentiment correlates to the subsequent return across these sectors. OLS makes it possible to look at one cryptocurrency at a time, and the results can be looked at in conjunction with the different segments. Another reason for it being the chosen method is due to the highly variable amount of sentiment data for the different cryptocurrencies. In short, some cryptocurrencies are naturally mentioned more than others on Reddit. Due to the high differences in available data is therefore preferred to look at the cryptocurrencies individually through an OLS. It is worth noting that a panel regression could be more relevant if the choice of cryptocurrencies were narrower and the segment differences weren't as importantly regarded. However, OLS is chosen due to the goal of the research and the amount of available data.

3.3 Cryptocurrencies and Sentiment

The literature on stock sentiment and WallStreetBets lead up to some existing literature on Reddit and cryptocurrency return. Camou's paper: "Reddit as a prediction tool for crypto-assets" is one example of this (Camou 2022). He looks into BTC, ETH, XRP, LTC and DOGE to forecast volatilities and returns. He finds Reddit sentiment to be useful to reduce the volatility of forecasting errors. Nevertheless, the effect of the return is mixed.

He uses a random forest as the non-linear model in the paper. This paper builds upon Camou's by looking at the variable importance and placement

of a random forest model in conjunction with OLS results and cross-section return from a sentiment portfolio.

There is also a paper from Filippou et al. called "Boosting Cryptocurrency Return Prediction" (Filippou, Rapach, and Thimsen 2021). Reddit is not as much in focus here but rather one of 39 predictors in a decision tree to forecast excess returns for BTC and ETH. These predictors also include Google trend searches and some news-specific articles. Filippou has some similarities in the data extraction of Reddit in the sense that comments are also included and the number of posts plays a role. However, this paper builds upon Filippou's method by implementing a new way to calculate sentiment by doing a weighted sentiment score that takes placement and upvotes into account.

In addition, most of the existing literature does not have a wide choice of cryptocurrencies. Wooley et al. focus on Bitcoin and Ethereum with a focus on Granger causality (Wooley et al. 2019). Prajapati focuses on a residual mean squared error from a forecast perspective of Bitcoin getting the sentiment data from Google News and Reddit (Prajapati 2020). The common factor is that there is a narrower availability of existing literature that choose to focus on a large number of cryptocurrencies. By doing it in this paper it is possible to investigate the predictive power of sentiment across different cryptocurrencies, both market cap and sector. This is one of the main contributions of this paper, in addition to the wide use of models and the weighted sentiment algorithm.

4. Methodology

This section covers the computation and complexity of the data-gathering process as well as investigating the relationship between cryptocurrency sentiment and subsequent returns. The entire process can be split into five parts:

Expert Interview - Gather prices - Scrape Reddit - Calculate Sentiment - Predictive Analytics

First, an expert interview was performed to gather industry knowledge about the experience of sentiment within DeFi. Then, the prices of cryptocurrencies were downloaded. The next step was to scrape Reddit and calculate a sentiment score for different cryptocurrencies. At last, the sentiment score, price data and other metrics were used to investigate the relationship between price and lagged Reddit sentiment across different sectors. This provided the final results.

4.1 Expert Interview

Existing literature on stocks, Reddit and cryptocurrency sentiment provides a baseline for a hypothesis of Reddit sentiment's importance for cryptocurrency

returns. However, in an ever-evolving industry and after a volatile 2022 is it interesting to gather input from people within the industry. That is why an interview was carried out with the community manager of a DeFi project.

ThorFi is a decentralized finance protocol on the Avalanche blockchain and launched on December 8th 2021. It started out as a node protocol like many others, but it is one of the few survivors within the niche after a large industry price depreciation in 2022. It has later pivoted towards being a game finance project with products such as an NFT marketplace, decentralized exchange and subnet in its road map. The protocol is led by Loki. He started out as a moderator in early December 2021 and later emerged as the community manager of the project. As the outer representative of a multi-million market cap DeFi project does he possess industry knowledge and first-hand experience with how sentiment emerges on social media within the cryptocurrency space (*ThorFi price today, THOR to USD live, marketcap and chart* 2023). He was so kind to participate in an interview on the 29th of March 2023 to discuss this.

During the interview, Loki shared his views on the different social media and their reliability as sentiment representations. He mentions among others Twitter, Reddit, medium posts and Discord as channels to capture sentiment, but highlights a potential bias to be aware of. "The problem is when you're looking at individuals that are promoting whatever motive that they want to promote, they're especially on crypto Twitter for example. A lot of them just want to pump their own bags" (*Defi Sentiment* 2023). He points out that all posts do not necessarily represent a genuine feeling for the industry

or a specific project.

Capturing the genuine sentiment is a challenge. On the question of disregarding posts that contain a certain keyword as spam, he highlights that "... it far more matters on the context that word is being used. And I think that that's something that would take a little bit more of a human review process" (*Defi Sentiment* 2023). He believes the best representation of sentiment is found through an aggregate social media platform with an "... AI algorithm that would be able to decide what's an honest piece of feedback versus a dishonest piece of feedback" (*Defi Sentiment* 2023). Although, Reddit sentiment may possess representativity of the entire sentiment. Compared to Twitter he says "I don't think they'd be massively different because it's the same people talking about the same thing on a different platform, so it's a good way to start off"(*Defi Sentiment* 2023).

On a specific question about Reddit as a platform to capture cryptocurrency sentiment, he describes it as a "massive form with an incredible amount of user base that allows you to have a wide birth of opinions and have far more in the way of nitty-gritty conversations between user A and user B. While that exists on Twitter and that exists in Discord and other social media platforms, I don't think there is a better platform than Reddit" (*Defi Sentiment* 2023). He also mentions the voting system of Reddit as an important factor to build upon this.

On a more general note regarding the predictability of cryptocurrency returns, he puts sentiment as the most important factor. "Sentiment is the

spark to the dynamite. So while there may be all these other factors, which is the dynamite in this metaphor, the thing that sparked and initiated the explosion would have been that sentiment spark"(*Defi Sentiment* 2023).

At last, addressing the difference in Bitcoin sentiment and Dogecoin sentiment in regards to intrinsic value he does challenge the concept itself. "The intrinsic value of a Bitcoin versus an ounce of gold bullion are virtually identical. I think until crypto is at the point where it is a realistic payment platform that facilitates payments, with financial transactions that supersede our traditional ones. I think that until that happens, I don't think that there is a huge amount of intrinsic value within crypto, and what we're all participating in is a big speculation game on to where the future is going"(*Defi Sentiment* 2023).

On the whole, with great first-hand experience of cryptocurrency sentiment as a DeFi community manager Loki contribute with some interesting input. The value of Reddit as a representative source for example, although bias and context will be important factors that are difficult to capture through a simple scraping model. This kind of bias will need to be taken into account when evaluating the results. Also, the importance of sentiment and the challenge of the concept of intrinsic value may facilitate an interesting expectation that all cryptocurrencies may contain a high sentiment-return relation regardless of the sector. However, the noise of data due to a simple scraping model and not-aggregated data may not capture this hypothesis.

4.2 Gather prices

4.2.1 Choice of cryptocurrencies

The first part of the coding methodology is collecting price data for different cryptocurrencies. Previous literature was limited to only a few tokens such as Bitcoin(BTC) and Ethereum(ETH). For this purpose, the top 100 market caps during January 2022 will be considered for a wider segment of the market space. It also facilitates comparison between tokens with high institutionalized investment and value proposition to retail investment tokens with strong communities and little-to-none intrinsic value. Although this interpretation of value proposition was challenged by Loki and opens up the overall importance of sentiment. The market cap tokens were pulled from CoinMarketCap - a widely used price-tracking website for cryptocurrencies (CoinMarketCap 2023). Stablecoins were excluded from the analysis due to their goal of being equal to the United States fiat currency. It is also worth noticing that the top 100 tokens per market cap naturally will change over the course of a year. For example, Terra Luna was in the top 10 range but drastically fell after an incident later in 2022.

There were other reasons to exclude certain cryptocurrencies from the list. For example, OMG and ETC have ambiguous names that may bring noise to the model and were therefore excluded. Another factor that excludes cryptocurrencies is the number of occurrences on Reddit. Some of the tokens

examined did not have a sufficient amount of Reddit posts, and will therefore not be included in this project. Note that the occurrence of sentiment is more important in a lagged state. It is of interest to combine today's return with yesterday's sentiment. This means that a model with a large number of missing values risks connecting today's return with the sentiment a long time ago, i.e. 14 days. As sentiment only is expected to possess short-term predictability as per Bouteska et. al, this approach should be avoided (Bouteska, Mefteh-Wali, and Dang 2022). Creating a data frame with all the dates within the range, binding sentiment, lagging and excluding NAs is therefore done to gather a representative sentiment measure. Tokens with too large gaps between the posts were then excluded from the project. After the removal of cryptocurrencies for the aforementioned reasons - the total amount of cryptocurrencies for this project is 37 - displayed in the figure below.

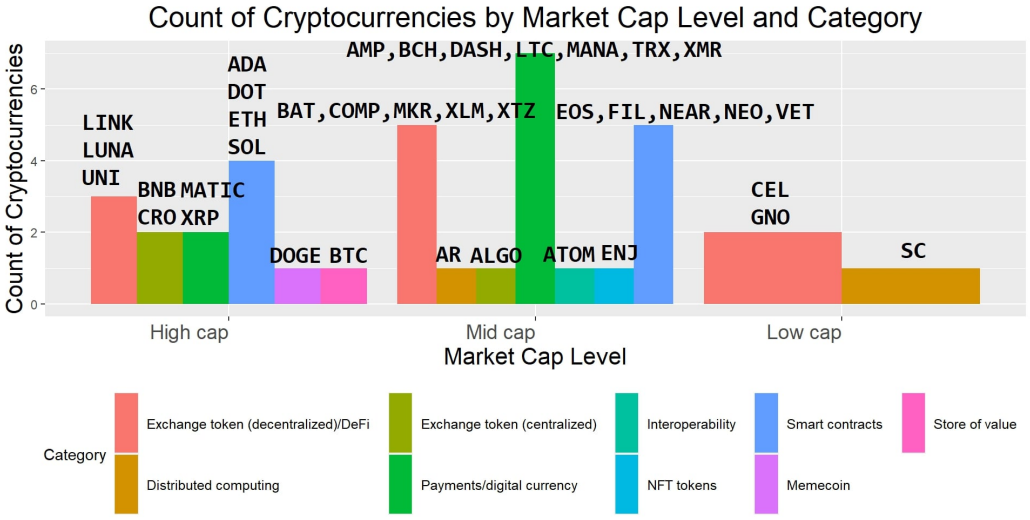

Figure 4.1: Overview of cryptocurrencies in this paper

It displays 9 different categories across 3 different market caps. Most of the cryptocurrencies appear in the mid-cap. The market cap and categorization are gathered from Statista (*Crypto market cap ranking 2022* 2023). It highlights the biggest cryptocurrencies in the world on January 10, 2022. The high market cap is defined as above 10 billion U.S. dollars, the low cap is below one billion whereas the mid-cap is in between these ranges. The figure shows a diversification among categories and market cap that is useful to investigate the differences among sentiment importance for different sectors later in this paper.

4.2.2 CryptowatchR

The chosen API to gather the prices of each cryptocurrency is Cryptowatch - cryptowatchR in R. It lets the user gather the historic crypto pair (in this case paired with USD). The specification of exchange makes it flexible and easier to facilitate finding prices of all types of market cap. For example, Kraken may have more data on some cryptocurrencies, whereas Coinbase has it on others. This makes it a bit more complicated. The script that gathers prices, therefore, creates a list of exchanges in descending quality and tries to gather the price of a pair and utilize another exchange if the data is insufficient. Note that Yahoo Finance was used to complement price data where it was needed and available. It had its strength in newer data but did not contain sufficient data for all 37 cryptocurrencies to be the only source of data.

4.3 Scrape Reddit

This part is by far the most complicated and time-consuming programming part of the project. The goal is to gather all Reddit posts from the CryptoCurrency subreddit where a specific cryptocurrency is mentioned, scrape it, tokenize the words and output a data frame with these results. However, there are quite a few decisions that need to be made in this step. First, RedditExtractoR was used as a simplistic way to gather data, but it appeared to be insufficient due to a low amount of data. Therefore, a rather manual process of investigating the PushShift servers was initiated to increase the number of posts within the analysis. This includes extracting the URLs in Python, assigning a weight-based sentiment algorithm to calculate scores for each word in a post, and at last using the Loughran-McDonald sentiment dictionary to assign it to a positive or negative sentiment.

4.3.1 RedditExtractoR

First, how to gather the relevant Reddit URLs. The Reddit API facilitates this, but there are other APIs that build on Reddit's and make it more user-friendly to work with. RedditExtractoR is a R-package that builds on the Reddit API in a simple user-friendly way in R. This package contains a function that inputs a keyword, subreddit and start date to output URLs of posts containing said keyword from the given subreddit in the given time frame. The downside here is that the available time period backwards is

limited. It is possible to get posts one year back. However, due to the early beginning of working on this project, the package was able to approximately gather Reddit posts for 2 years: 2021-2023.

The URLs from RedditExtractoR were used further on in the process, but it was problematic. The results from the predictive analytics appeared to be spurious and random. This might be due to the low amount of data available for the next steps. Therefore, this step was revisited with the goal of gathering more data that hopefully will provide less spurious results later in the analysis.

4.3.2 PushShift API

A natural proceedor to using the RedditExtractoR is diving into the PushShift API. This API has been used in earlier literature, and it is a well-known API that is widely used due to its higher capacity than the Reddit API itself. There were however some difficulties with this API. It turned out that the API was not able to return results before November 2022. Deeper research into the PushShift API's subreddit confirmed this error. Naturally, this does not facilitate sufficient amounts of data for this project, and a stalemate was reached.

Looking deeper into the PushShift API, it turned out that all the historic posts are publically available to access through their servers without the API itself. Therefore, it is possible to avoid the API's bug and access earlier

data. Nevertheless, this process is significantly more complicated than just utilizing the API. It involves downloading a large torrent with all data for the CryptoCurrency subreddit, extracting it to a .zst file, and working with it as a .ndjson file. This means that each line is a new JSON object.

4.3.3 Extracting URLs in Python

The complex process of downloading the torrent and reading the .zst file was thoroughly elaborated in PushShift's Reddit (Watchful1 2023). However, the Python script needed to be tweaked for this specific purpose. That is extracting the following measurements:

date - time - title - text - url

The title and text would then be used to search for any occurrence of a cryptocurrency. For example, each cryptocurrency can be referred to through its name (Bitcoin) or ticker (BTC), both being case insensitive. If either the name or the ticker occurs in the title or text of a post, then that entire post is regarded as a Bitcoin-related post. So if a Reddit post in the CryptoCurrency subreddit has the following title: "Bitcoin is on a streak!", then that post will be connected to the Bitcoin topic.

This is a simplistic solution, and therefore also includes some challenges. First, the cryptocurrency can be mentioned in conjunction with another one,

but not be the most important one. For example: "This brilliant new innovation from Litecoin makes transactions significantly faster than Bitcoin". Looking at this fake title for itself, it represents a positive sentiment for Litecoin and is not really relevant to Bitcoin. If this post were to have sentiment for Bitcoin, it should be negative as a competitor is taking steps to cannibalize its market cap. However, by the code in this project, this will be regarded as positive for both Bitcoin and Litecoin.

The fact that tickers may appear in longer names is taken care of with the code. For example "adamant" is not the same as "ada" (cardano). Ideally, the process of classifying posts to cryptocurrencies could have been more thorough, which fits Loki highlighting the complexity of capturing the sentiment online. More exceptions could have been looked into, and even a machine learning approach with topic modelling could be a solution, but this has not been a priority and the scope excludes this from the project. Nevertheless, the output of this step is an enormous data frame with the following values:

date - time - crypto - url

4.3.4 Insufficiency of PushShift data

The next step is to scrape Reddit through the different URLs. This step may seem superfluous as the data from the PushShift servers already contain the content from each post. There are two reason why this is needed, and it is

all about the structure of the PushShift servers. It contains millions of posts from thousands of different subreddits. It is almost computationally impossible to make this data live, in the sense that is often refreshed. Therefore it only contains a snapshot of the subreddit at a given time. This means that:

1) The newest comments are not included

2) The upvotes are not up to date

Both of these measurements are vital for the sentiment algorithm of this paper, so the extra step of scraping again must therefore be done.

4.3.5 Establishing scraping in R

The vital part of the scraping in this project is the different valuation of word placement. For example, a word placed in the title is more likely to be seen by a user than a word in a comment. Therefore, these words should also be weighted differently. That is why three parameters have been implemented:

$$\textbf{title\_value} = \textbf{50}$$

$$\textbf{text\_value} = \textbf{20}$$

$$\textbf{comment\_value} = \textbf{10}$$

These numbers are based on the logical foundation that words placed in the title are by far the highest weighted, then followed by words in the text with the lowest weighting for words in the comments. One of the disadvantages of long computational time is the lack of tuning opportunities for these parameters, so the parameters are based on a non-tested rough estimation. Nevertheless, the parameters are used to calculate a score for different words. It is based on the number of times a word occurs (n), upvotes and a weighted placement. Note that words that appear in the title or text use the upvote score of the post, whereas words in comments utilize the upvote for the comment itself. All in all, the score of each word is calculated in the following way:

$$score = n \times weighted\_value \times upvote\_score \tag{4.1}$$

Be aware that this does not cover positive or negative sentiment yet. It is simply a procedure done before sentiment is introduced later on in the process. This facilitates that occurrence (n), placement (weighted_value) and support/agreement (upvote_score) are taken care of.

4.3.6 Computational Scraping

At this stage, the push_url_df consists of 1 348 608 observations/URLs. This is why this step needs some extra tweaking. First of all, parallelization is introduced. In addition, a background job in RStudio makes it possible to

run multiple scripts simultaneously. Although, the limit as to how much data can be run without exceeding the memory requires trial and error. Ideally, the code would be uploaded to a cloud service that would run it all. This was looked at, but it was difficult to find solutions that can handle such large amounts of data for free. Therefore, the code was executed in small bunks over the course of several days, where the code often failed due to memory shortage. To give a sense of the computational difficulties: Running 1 background process for 100 000 observations takes approximately 10.5 hours, and takes up more than 100GB in temporary memory space. This process is probably the biggest bottleneck within the project, as a long computational process makes it difficult to tune, trial and error.

4.4 Calculate Sentiment

With occurrence, placement and support taken care of for each word, it also needs to be some sort of sentiment dictionary applied. Establishing a sentiment dictionary towards Reddit sentiment for cryptocurrency could be a master thesis on its own. Instead, it is necessary to make a choice that although not perfect, does fit some of the needs. Some literature such as Hu et. al (D. Hu et al. 2021) or Bradley et. al (Bradley et al. 2021) chooses to implement their own sentiment dictionary to capture the specificity of financial terms and social slang. There is also some literature that chooses to implement VADER as it is facilitated for social media slang (Wooley et al. 2019).

Elaborate work on a sentiment dictionary would in all likelihood improve the sentiment gauge, but it goes beyond the scope of this project. VADER would have been a viable option and is popular among the existing literature. However, this paper wishes to provide value by interpreting sentiment differently. Instead of calculating sentiment in paragraphs, it will rather be interpreted as single tokens. This provides more flexibility within the model with the ability to remove outliers on a token level. As VADER is more appropriate for context-based sentiment within social media, then Loughran-McDonald was chosen instead. It is a well-known sentiment dictionary that performs well for financial data (Loughran and Mcdonald 2011). Nevertheless, it is not specified towards cryptocurrencies or social media, so some misclassification is bound to happen to this imperfect sentiment dictionary. The idea is that although not perfect, the token approach and the financial strength of the dictionary will provide more advantages than disadvantages. Nevertheless, a VADER approach would be interesting with the weighting scheme of this paper for future research.

4.5 Predictive Analytics

The final part of the project contains predictive analytics where different models are utilized to investigate the sentiment relationship. The results will be covered in the next chapter, but the methodology behind the different models will be elaborated on here. This includes the choice of explanatory variables, the OLS approach, the variable importance of random forest and

the methodology of the cross-section.

4.5.1 OLS

Ordinary least squares is utilized as a simple linear model and a base methodology for the project. By running 37 OLS models and getting the p-value for the linear, square and cubic term LagSent_1 does it output a table with multiple p-values. Note that robust standard errors are implemented to improve the reliability of the model. It is likely to worsen the results, but as the model does not fit all requirements for a best linear unbiased estimate anyway, then it is implemented to reduce the heteroskedasticity of the residuals.

Explanatory variables

It would be possible to perform this project with no further variables, and only utilize:

date - time - price - LagSent_1

However, this is not recommended due to the bias-variance trade-off. There would be a bias in the LagSent_1 coefficient that might affect the statistical significance and even the sign of the coefficient. Therefore, several explanatory variables are included to reduce this bias and then be able to investigate the relation between lagged Reddit sentiment and price.

Finding proper explanatory variables to include in the model is a challenge. The use of explanatory variables in OLS and random forest for cryptocurrency cases has been less studied compared to stocks, likely due to the relatively short tenure of cryptocurrency in the financial industry. In addition, the interview with Loki emphasizes the relative importance of sentiment variables compared to more traditional fundamentals. Also, there is a debate about the predictability of cryptocurrencies in the first place. Nevertheless, Bitcoin following a random walk has been disproved in some research (Palamalai, Kumar, and Maity 2021).

The choice of explanatory variables is directed towards the blockchain environment in general, and not for specific cryptocurrencies (Bergerud and Hu 2023). This is due to the inclusion of smaller cryptocurrencies as well as not having sufficient data for fundamentals to be included as explanatory variables. However, the idea is that due to the interconnectedness of the market, there may be explanatory power for all cryptocurrencies by looking at overall blockchain-, global- and Bitcoin metrics. Note that searching for daily data instead of monthly data drastically reduces the available content online. A lagged return variable is also added to the model in addition to the following explanatory variables.

Halving

There is a concept within Bitcoin that is vital for its existence and deflationary nature: halving. Being decentralized, there is no government that

can control the supply of Bitcoin. Therefore, there needs to be some sort of mechanism to ensure that it does not become inflationary and loses its value over time. To understand this deflationary mechanism better, it is vital to investigate what happens when transactions take place on the blockchain. A "miner" verifies a transaction by solving a computationally difficult math problem. The first miner to solve this problem is rewarded with Bitcoins. This reward is halved every 210 000 blocks, which roughly makes up to be every fourth year. This is regarded as halving - a process which increases the value of Bitcoin(Meynkhard 2019). The idea is that decreasing the rate of new supply while maintaining the same demand will increase the price.

Halving reward as a factor variable did not contribute to the model and was excluded. Therefore there is only one halving variable included:

countdown_halving

The number of days estimated to the next halving. The past halving dates are known, but the next one is based on an average estimate of the daily number of blocks and assumes a future halving date to be the 7th of April 2024(*Next Bitcoin Halving 2024 Date & Countdown [BTC Clock]* 2023)

Twitter Sentiment

This paper focuses on Reddit sentiment for individual cryptocurrencies. The idea is that it will capture a strong community of retail investors. However, Twitter is another social media platform that may possess predictability

for future returns. Also, it was widely mentioned in the interview with Loki. Multicollinearity tests were performed to make sure that it does not correlate too much with Reddit. As it passed the test, it will remain in the model.

twitterHits

Tweets per day with #Bitcoin(*Bitcoin, Ethereum, Dogecoin, Litecoin stats* 2023)

Blockchain

This category covers different characteristics of the overall blockchain of Bitcoin. The idea is that blockchain metrics represent the overall usage of the largest cryptocurrency in the world and may be a relevant inclusion. There are three different blockchain measurements included. The average transaction value covers the average Bitcoin amount that people trade for. Next, each transaction has a fee. This fee changes over time, and it may be a useful inclusion as a higher fee is a result of higher demand at a current time. At last, the use case of Bitcoin and any cryptocurrency is based on how well it is spread. Therefore, the number of unique Bitcoin addresses is included as an explanatory variable.

avgTrans

Average transaction value of Bitcoin transactions(*Bitcoin, Ethereum, Dogecoin, Litecoin stats* 2023)

transFee

Average transaction fee on the Bitcoin network(*Bitcoin, Ethereum, Dogecoin, Litecoin stats* 2023)

uniqueAdd

Number of unique Bitcoin addresses(*Bitcoin, Ethereum, Dogecoin, Litecoin stats* 2023)

Mining

Although slightly covered in other categories, mining is a vital part of the proof-of-work cryptocurrency Bitcoin. It allows users to verify transactions and ensures the safety of the network. This process is criticized for its environmental footprint as it takes a large amount of power to carry out this mining. However, the power needed to mine Bitcoin (also known as difficulty) is ever-changing to achieve a block time of 10 minutes. The power may therefore be an indication of mining popularity and overall mining sentiment.

power

Power demand to mine Bitcoin (*Cambridge Bitcoin Electricity Consumption Index (CBECI)* 2023)

Uncertainty Index

The relationship between cryptocurrencies and the global economy is conflicted. Bitcoin is meant to be a hedge against stocks. The stock market does not affect Bitcoin in times with low and medium uncertainty, although it significantly impacts its return in high uncertainty periods (Nguyen 2022). Therefore, there are three variables included in this category.

uncert

Equity Market-related Economic Uncertainty Index. Used as a benchmark for global uncertainty(Baker, Scott R., Bloom, Nick, and Davis, Stephen J. 1985)

worldIndex

MSCI World Index covers large- and mid-range stocks in 23 developed countries. Used as a benchmark for the performance of global equity(investing.com 2023)

worldIndexInt

World index in times of uncertainty.

If uncertainty is bigger than its mean: worldIndex

else: 0

4.5.2 Random Forest

The non-linear model to investigate the relationship between sentiment and return will be a random forest. The same explanatory variables will be used in this model as well. The variable importance from the random forest model will be a useful indicator of the relative importance of sentiment compared to the other explanatory variables. It will capture both node purity and relative placement. The node purity is a measurement of the similarity of sentiment within the node and provides information about the distinguishability of sentiment in regards to regressing return. Placement on the other hand provides information on the relative importance of sentiment compared to other explanatory variables. Is it the most important variable to predict return (placement = 1), or the worst (placement = 11)?

Tuning of random forest

The model will first need to be tuned to optimize the result for each specific cryptocurrency. The following parameters within the randomForest() model will be tuned:

mtry

ntree

maxnodes

nodesize

Random forest in short is a decision tree model with many trees found by a bootstrap data selection and a random feature selection. *mtry* is the number of variables for each split of the trees. *ntree* is the number of trees in the random forest, whereas *maxnodes* and *nodesize* indicate a range of terminal nodes. The amount of data for the different cryptocurrencies are rather spread, so it is imprecise to tune the model on a one-fit-for-all approach. Therefore, the tuning ranges for the parameters vary on the number of rows in the data frame, and the model will be tuned specifically for one cryptocurrency data frame. This will be more computational but is regarded as a necessity to gather reliable results.

Variable Importance

There are 11 explanatory variables in the model, one of which is LagSent_1. They are however not equally important to predict the return of a cryptocurrency. Therefore, there are two measurements that will be looked further into:

placement

purity

Placement is a measurement of the variable importance of a random forest

model in decreasing order. It ranges from 1 (most important variable) to 11 (least important variable). Purity however is a measurement of the homogeneity within a node and will range between 0 (no homogeneity - bad) and 1 (total homogeneity - good). These two variables need to be looked at in conjunction with each other as they can be misleading on their own.

4.5.3 Cross-Section

This method differentiates from the others as it is disregarding the explanatory variables and is only looking at the return and lagged sentiment. First, there needs to be a base portfolio to have something to compare it with. It consists of buying all the cryptocurrencies that are used on 1/1/21 and selling them on 31/12/22. For example, if a high-market cap approach is regarded, then the base portfolio will buy all cryptocurrencies within this genre and sell them after two years. Note that this is value-weighted by the market cap values in 2021 (*Historical Snapshot - 01 January 2021* 2023). In addition, there will be a Bitcoin portfolio that buys Bitcoin on 1/1/21 and sells it on 31/12/22. This is useful due to the overall importance of Bitcoin and will provide another portfolio to compare with.

The main portfolio however is the sentiment portfolio. The first step here is to create a data frame with multiple cryptocurrencies and sort it by date and LagSent_1. Then assign a threshold of how many cryptocurrencies to buy or sell. This is set to 20% - found by trial and error. The sentiment portfolio goes long on the top 20% cryptocurrencies and goes short on the

bottom 20%.

However, a transaction fee is included in the model to make it more realistic. The sentiment portfolio performs a lot more transactions than the base portfolio and the Bitcoin portfolio and would in real life have its return diluted by the number of transactions. The transaction fee is set to 0.03% as it is the fee to swap tokens on Uniswap - one of the largest decentralized exchanges (*Fees — Uniswap* 2023). The fee is taken when a position opens and when it closes. That means it occurs twice for the Bitcoin portfolio but for each cryptocurrency in the base portfolio on 1/1/21 and 31/12/22. For the sentiment portfolio, however, it occurs significantly more often. Every day that has Reddit data for 5 or more cryptocurrencies, the top 20% will go long and the bottom 20% will go short for the sentiment portfolio. This requires a fee for each transaction. The next day these need to be closed with a fee, and the new ones need to be opened with a fee. This will decrease the overall return of the sentiment portfolio, but it will improve the reliability of the results.

The cumulative return of these three portfolios will give a benchmark of the importance of sentiment. If the sentiment portfolio outperforms the base model even with transaction fees, then that may be an indication that it is indeed a valuable variable that possesses predictive power for return.

5. Data Description

5.1 Differences among cryptocurrencies

The sentiment for the different cryptocurrencies is a result of how often words occur, the Loughran-McDonald sentiment of the words, a weighted placement of its occurrence and the number of upvotes. This will facilitate large differences in the sentiment among the different cryptocurrencies. Larger cryptocurrencies that are spoken about more, such as Bitcoin, are expected to have stronger absolute values than cryptocurrencies that are rarely mentioned and have little support in the form of upvotes.

The figure below visualizes the different raw sentiments among a few of the cryptocurrencies. It displays Bitcoin, Ethereum and Doge. All of them have a peak towards the end of 2017, something that matches well with the price evolvement for the cryptocurrency market as well. There are stronger absolute values within this period before a quieter period occurs followed by a slight up-rise around 2022. Note that BTC and ETH are more similar in regards to sentiment compared to DOGE, something that is interesting from an institutional versus retail investing point of view. Also, the large standard deviations of the sentiment indicate that a removal of outliers will be necessary.

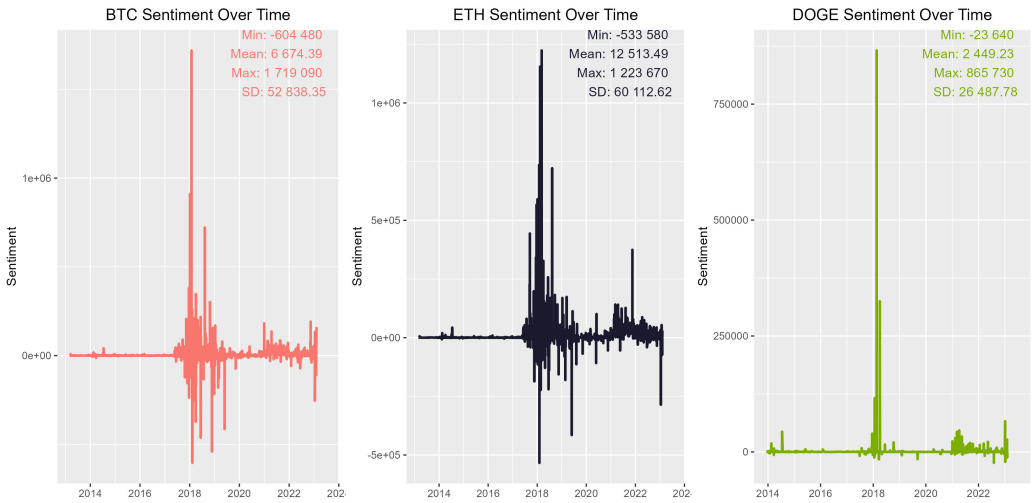

Figure 5.1: Raw Sentiment of BTC, ETH and DOGE

5.2 Remove Outliers

The previous plots indicate volatility within the sentiments. However, there appear to be some outliers that have an effect on the totality of the dataset. This is better visualized through density plots. Here we observe the steepness of the curve and that there are some positive and negative outliers although most values are ranged in the middle.

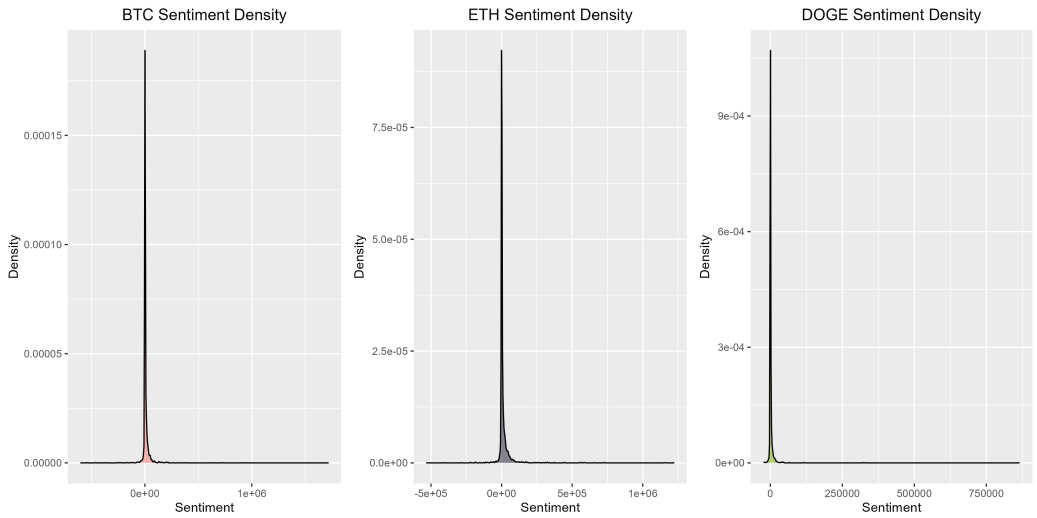

Figure 5.2: Raw density plot for sentiment of BTC, ETH and DOGE

There is a choice whether or not to keep these outliers in the data. There may be some error with the model, where the model misinterprets the context of the content and therefore represent it poorly. However, it might also represent genuine strong sentiment on a particularly bullish/bearish day. It is also worth noticing the amount of data is not too large (less than 2 000 observations), and that outliers may have a large effect on all the data. Also, the plots indicate that even though sentiment indeed changes over time, the outliers appear sporadically. All in all, the outliers will be removed from the variable. Multiple methods were tried such as Winsor and a quantile exclusion. By trial and error, a median method was implemented. The algorithm excludes values that have an absolute difference between the value and the median larger than a threshold multiplied by a median absolute deviation (MAD). The density plots after removing these outliers are showing a clear difference:

Figure 5.3: Density plot after removal of outliers for BTC, ETH and DOGE

5.3 Further adjustment of sentiment

The outliers of sentiment have been removed, but there are still some difficulties with the characteristics of the variable. First, it is on rather different inter- and intrascales. In other words, the sentiment between the cryptocurrencies is different, and so is the scale between sentiment and the other variables as well. This makes sense because high-interaction cryptocurrencies as more talked about than the lower tiers, and will therefore get higher absolute value scores. Therefore, scale() was used in R to centre the data.

Next, there are three possible adjustments that were used in a trial-and-error manner to improve the characteristics of the sentiment variable:

Log-transformation

First-difference

Quantization

The best results came from first scaling the sentiment before a log-diff operation. This facilitates the sentiment to first be transformed to a standardized scale, something which is not the case in the beginning. The sentiment algorithm returned numbers of high absolute values, but after scaling it can be comparable with other variables.

Next, a log difference is performed after increasing the sentiment by the absolute value of the smallest value. In other words, make sure the sentiment is not negative before performing a log difference to make the operation possible. The log was performed due to the skewness of the data observed earlier and to try to make it more normally distributed. The first difference was added by trial and error and provided better results. This may be a contributor to removing trends within the data, although the interpretability of the coefficient will be more complicated. Quantization did not improve the results are were therefore not utilized.

5.4 Conjunction with return

After exploring and adjusting the characteristics of the sentiment variable would it be interesting to observe it in conjunction with its return before diving deeper into the regression models. The following graph shows the

return variable for the chosen cryptocurrencies and the lagged sentiment variable after the aforementioned adjustments.

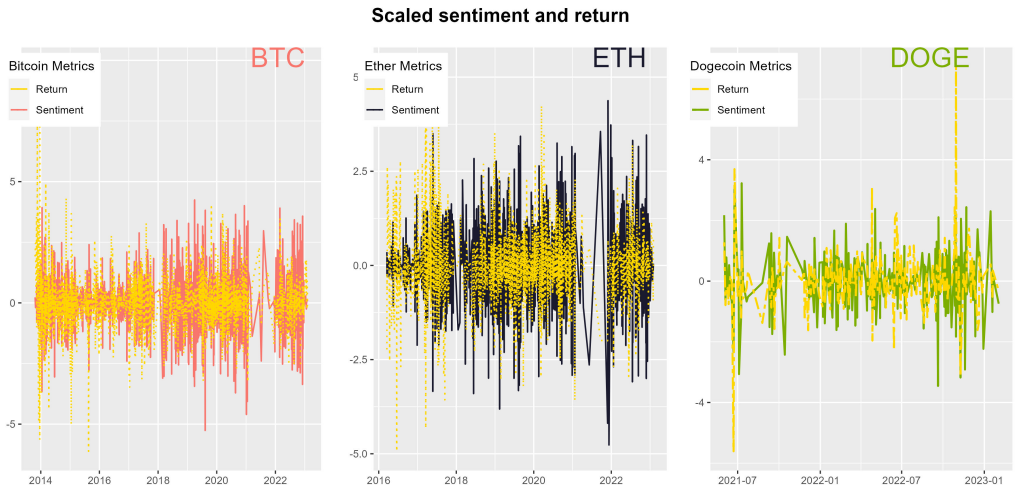

Figure 5.4: Scaled sentiment and return for BTC, ETH and DOGE

The sentiment and return for BTC, ETH and DOGE are scaled to plot them all together. Sentiment is also log differenced and outliers are removed. First, BTC and ETH have more observations, and the plot is affected by this. The volatility in return and sentiment appear for all cryptocurrencies, but it is difficult to make any conclusion regarding the sentiment-return relationship.

6. Results

This section covers the last part of the project after prices have been gathered, Reddit has been scraped and sentiment has been calculated. There are three different methods that are utilized to investigate the relationship between sentiment and cryptocurrency returns.

First, an Ordinary Least Squares (OLS) is used to create a table of p-values for the polynomial OLS. It will visualize the success rate of using OLS as a method to establish a relation between sentiment and return.

Random Forest is also used to better capture the potential non-linearity of a sentiment variable. After tuning the model then the variable importance of sentiment will be investigated. This covers both relative placement and node purity. A single decision tree will also be investigated.

At last, a cross-section method is utilized to compare a base portfolio with a portfolio where lagged sentiment is the driver of which cryptocurrencies to buy or sell. Together these three methods cover a wide theoretical background that is useful when looking at the importance of Reddit sentiment.

6.1 Ordinary Least Squares

A quick glance at the table below shows a poor relation between sentiment and return through the OLS model. A majority of the p-values are insignificant. In fact, most of the cryptocurrencies have insignificant p-values for the linear, quadratic and cubic sentiment.

There are some exceptions to the poor results. CEL, COMP, ETH and TRX all have one statistically significant p-value for one of the sentiment coefficients. LUNA is performing well, which is interesting given the volatile 2022 with the ANKR crash. NEAR however is performing the best. It is difficult to interpret why it is like that. It may be a strong community backing the scalable blockchain platform of NEAR protocol. On the other hand, it may just be noise from Reddit posts that use "near" as a synonym for close.

All in all, OLS appear to not be the appropriate methodology to establish a relation between sentiment and cryptocurrency return. A model that facilitates non-linearity in a better fashion may perform better for this problem.

Table 6.1 - P-Values from OLS

This table displays p-values for the linear-, square- and cubic variable LagSent_1 after performing an OLS. *, ** and *** denote the p-value being below respectively 0.10, 0.05 and 0.01

Crypto	**LagSent_1**	**LagSent_1^2**	**LagSent_1^3**
ada	0.298	0.981	0.639
algo	0.119	0.227	0.434
amp	0.367	0.227	0.772
ar	0.280	0.200	0.703
atom	0.996	0.819	0.748
bat	0.758	0.224	0.506
bch	0.888	0.939	0.830
bnb	0.598	0.413	0.839
btc	0.523	0.506	0.475
cel	**0.009*****	0.153	0.141
comp	**0.033****	0.336	0.169
cro	**0.074***	0.433	**0.037****
dash	0.911	0.286	0.917
doge	0.970	0.253	0.178
dot	0.771	0.139	0.668
enj	0.986	0.956	0.990
eos	0.914	0.631	0.818
eth	**0.078***	0.253	0.143
fil	0.222	0.656	0.411
gno	0.603	0.325	0.834
link	0.333	0.720	0.269
ltc	0.614	0.157	0.888
luna	**0.037****	0.317	**0.098***
mana	0.714	0.776	0.751
matic	0.335	0.112	0.907
mkr	0.296	0.854	0.431
near	**0.037****	**0.002*****	**0.000*****
neo	0.479	0.971	0.250
sc	0.344	0.542	0.512
sol	0.802	0.823	0.767
trx	0.410	**0.032****	0.156
uni	0.977	0.406	0.771
vet	0.758	0.367	0.433
xlm	0.444	0.248	0.326
xmr	0.472	0.790	0.757
xrp	0.390	0.209	0.992
xtz	0.331	0.492	0.585

6.2 Random Forest

The methodology in this section will go further into two aspects: a concrete decision tree for DOGE and a variable importance table from the random forest model. This will facilitate the visualization of a typical decision tree for DOGE but also look at it in conjunction with the variable importance table of DOGE and the rest of the cryptocurrencies.

6.2.1 Decision tree for Dogecoin

A model like a random forest should in theory be better suited to capture a relation between sentiment and return if the relation were to be non-linear. The random forest is built upon creating many decision trees. However, it is interesting to look deeper into how one decision tree could be visualized for Dogecoin. A decision tree with a max depth of five nodes is displayed in the figure below. Interestingly, LagSent_1 is the first variable in the decision tree. If it is below the threshold of -0.0603 then the decision tree moves to the left. As expected, low sentiment and high uncertainty are associated with a worse return than low sentiment and low uncertainty.

On the other side, if the uncertainty is higher than -0.603, then there is a new LagSent_1 split. If it is higher or equal to -0.04 it goes to the right. Counter-intuitively, there is a larger sentiment when this is not the case. In other words, when sentiment is in the range (-0.063, -0.04) then the predicted

return is at its highest with 0.047. This accounts for 7% of the observations. It would be expected that although the relationship might not be linear that the highest predicted return would occur for higher sentiment than within this range.

This single decision tree does show LagSent_1 to be an important variable to predict return, however, the logic is not always ideal. Low sentiment and high uncertainty provide the lowest return which were to be logically expected. On the other hand, the highest return is not found at the highest level of lagged sentiment, but rather in the range (-0.063, -0.04). This may be an indicator that the model itself is flawed. However, it may also be an indicator that a slightly pessimistic sentiment on Reddit is associated with the best returns. For a strong-sentiment coin such as DOGE there may be excessive usage of bullish sentiment with investors "hoping for the moon". This bullish sentiment may be a false representation of the genuine sentiment of the community as a whole and may be subject to noise. This does challenge the expectancy of a simple high-sentiment equals a high-return mindset.

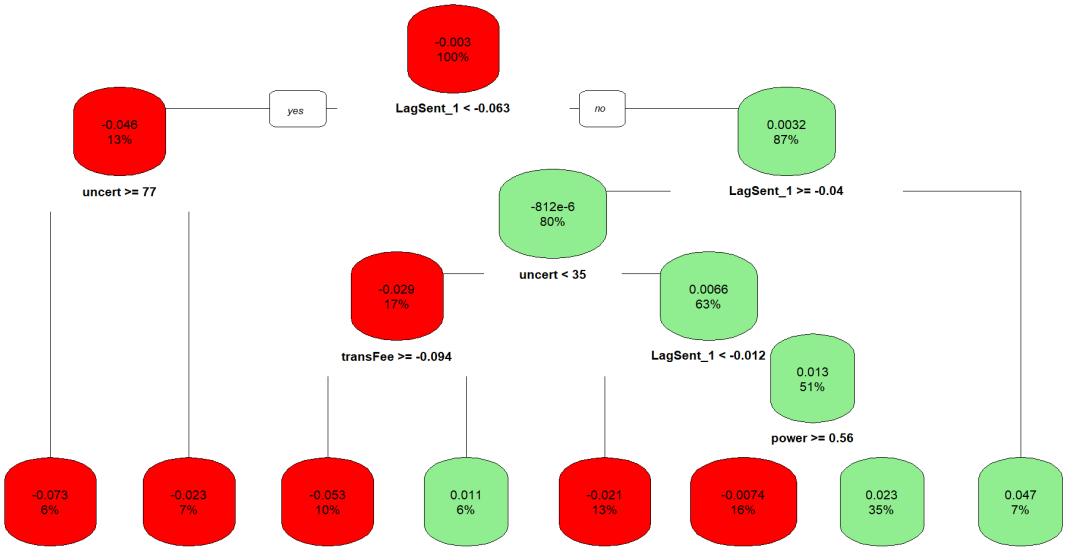

Figure 6.1: Decision Tree for Dogecoin

6.2.2 Variable Importance

The variable importance results show a large variation in sentiment importance for the different cryptocurrencies. The lagged sentiment is ranked low on placement for many of the large cryptocurrencies such as BTC, ETH, LTC and XRP. However, DOGE, MATIC, NEAR and UNI are examples of cryptocurrencies where the lagged sentiment variable possesses more importance for the random forest model. Looking at this in conjunction with node purity, there are some contradicting results. For example, although a high placement, the node purity for DOGE is 0.08 whereas for LTC it is 0.236. This may be an indication that LTC is easier to predict with a random forest model, but the other explanatory variables are better for this use case.

Table 6.2 - Variable Importance Random Forest

This table displays the variable importance of all 37 cryptocurrencies through a random forest model. Placement ranges between 1 and 11, with 1 being the best. Purity ranges between 0 and 1, with 1 being the best.

Crypto	Placement	Purity
ada	6.635	0.097
algo	6.575	0.163
amp	2.405	0.099
ar	6.670	0.054
atom	6.915	0.130
bat	3.985	0.288
bch	9.740	0.119
bnb	9.905	0.033
btc	8.325	0.190
cel	1.925	0.232
comp	4.970	0.163
cro	6.835	0.031
dash	9.215	0.109
doge	3.180	0.080
dot	4.165	0.097
enj	8.575	0.290
eos	9.975	0.174
eth	8.935	0.291
fil	6.215	0.118
gno	7.735	0.180
link	6.265	0.144
ltc	9.775	0.236
luna	4.715	0.494
matic	2.605	0.092
mkr	5.175	0.169
near	2.030	0.183
neo	9.645	0.224
sc	2.940	0.207
sol	5.445	0.156
trx	5.580	0.178
uni	2.410	0.180
vet	9.740	0.236
xlm	9.905	0.156
xmr	4.410	0.194
xrp	7.110	0.210
xtz	6.110	0.134

6.3 Cross-Section Return

This section covers a cross-section methodology by looking at the full portfolio first. Then, the different market caps will be evaluated, and differentiation into categories of cryptocurrencies will be looked further into. This will be done to establish an optimal portfolio consisting of cryptocurrencies that improve the portfolio when included and reduce the noise as well. Ideally, all combinations of cryptocurrencies in the portfolio would have been tested out, but this has not been looked into due to computational resources.

6.3.1 Full Portfolio

The first cross-section model is looking at all cryptocurrencies in the range from 2021-01-01 to 2022-12-31. As mentioned earlier in the methodology, there will not be data for every cryptocurrency for every day. This model only looks at the dates where there are 5 different cryptocurrencies mentioned on Reddit. This exclusion is necessary to not dilute the outcomes. For example, if there is only one cryptocurrency mentioned on Reddit on a certain day, then the sentiment portfolio would go long and go short on it. Also, even if there were two cryptocurrencies, the sentiment might be really good or bad for both of them, and a split would be inappropriate for the model.

The Bitcoin portfolio gives a cumulative return of -42.8%. The price followed large fluctuations before ending up at a lower price at the end of the two years,

but the highs at the end of 2021 ensure that the overall return is not too bad. Note that the Bitcoin portfolio will be the same for all different scenarios. The base model on the other hand is performing roughly the same, which is natural as Bitcoin makes up a large portion of the market cap value-weighted base portfolio. It ends up with a cumulative return of -37.86%.

Sentiment however is not as drastically affected by the downturn towards the end of 2021 but performs well at the end of 2022. This makes it the cumulative return of 74.63% for the sentiment portfolio containing all 37 cryptocurrencies. It is therefore outperforming both the Bitcoin- and the base portfolio. Although it does not reach the highs of the base portfolio in 2021 - it does appear less volatile and is good at picking up the overall sentiment and which cryptocurrency to go long or to go short. The fact that the sentiment portfolio contains a lot more transaction fees as well is another positive argument towards its realistic performance.

Looking at all the cryptocurrencies for the cross-section calculation gave positive backing to the hypothesis of a relation between sentiment and return. However, it is unknown what size of cryptocurrencies provides the better results. Therefore, looking at the portfolio for different market caps can be useful to further investigate this.

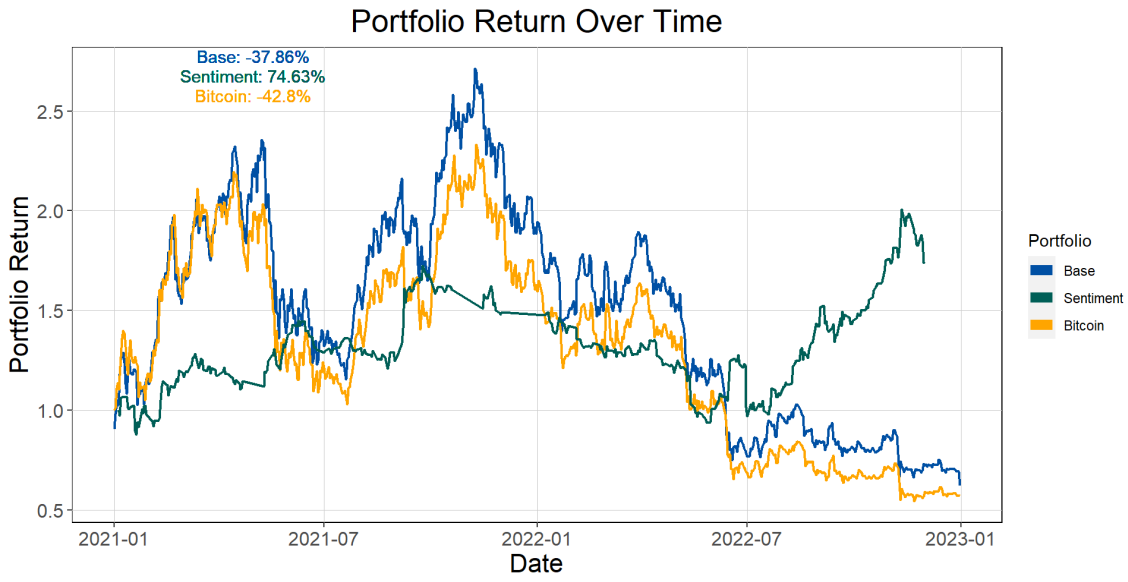

Figure 6.2: Sentiment portfolio return from all 37 cryptocurrencies between 2021-2023

6.3.2 Market Cap

High Market Cap

Choosing to look at just the high market cap cryptocurrencies affect the portfolio massively. First, the base portfolio is improved, and the cumulative return for the two years is -29.29%. The fact that it performs better than the base portfolio with all cryptocurrencies may be due to the fact that it simply consists of assets with less relative volatility. Also, the fact that it is a high market cap is also an indication of the long-term price action in comparison with the newest low market cap asset on the market.

The sentiment portfolio is still not reaching the highs of the Bitcoin- or base portfolio. However here it does not have the late 2022 spike. Therefore, the cumulative returns end up being -7.06%. It is still the best among the three portfolios, but it is performing worse than the sentiment portfolio of all cryptocurrencies. One reason for this may be an enlarged level of noise in the high market-cap cryptocurrencies.

The most popular cryptocurrencies are in the high market cap category. They are talked about the most on social media and reach out to the largest range of investors. This does however not necessarily mean that the social media sentiment is more genuine. For many, it is a first step into a new industry of blockchain technology and an unrealistic expectation of future return may create an impatient social media presence that doesn't capture the genuine sentiment behind the cryptocurrency itself. This fits Loki's comment about the bull run creating unrealistic expectations for the investors (*Defi Sentiment* 2023).

That is one potential explanation for the noise of high market cap sentiment. Another element to take into account is the imperfect methodology in capturing sentiment. As long as a cryptocurrency (or its ticker) is mentioned in a post, then it is categorized as that cryptocurrency. However, the post may in reality be about something else. The inclusion of large market cap cryptocurrencies can be used as a remedy for the Reddit user to make some form of comparison, and will therefore provide noise for the large cryptocurrency sentiment itself.

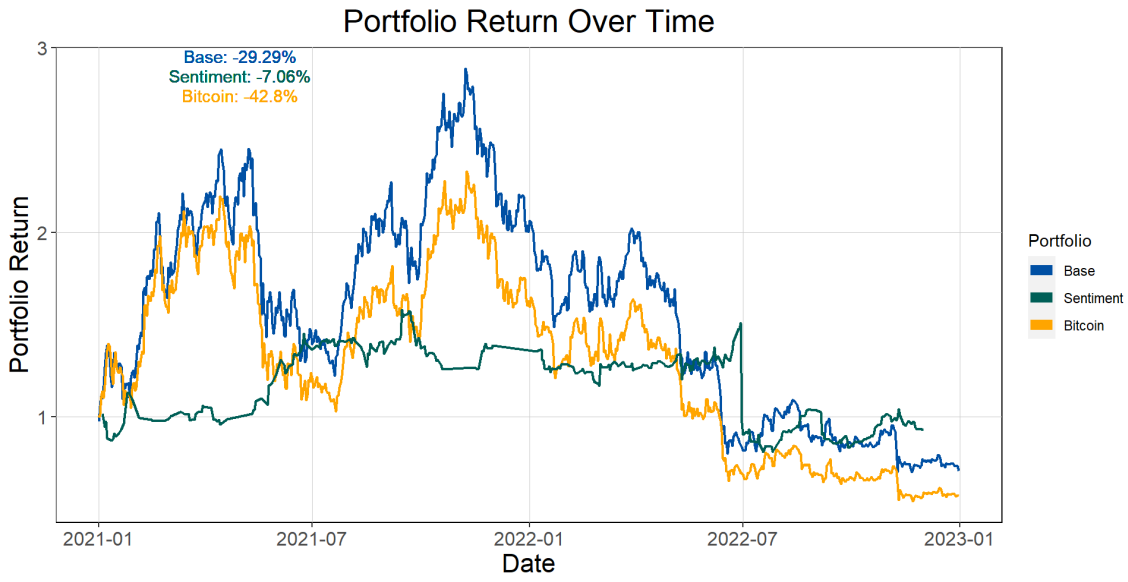

Figure 6.3: High market cap portfolio 2021-2023

Therefore, observing the cross-section return for the mid-market cap model is interesting. The base model performs roughly the same and has a cumulative return of -28.28%. The sentiment portfolio on the other hand performs worse with a cumulative return of -20.6%. Here as well the sentiment portfolio fails to reach the highs of the base portfolio in 2021 but remains rather stable throughout the two years.

The fact that high-market cap cryptocurrencies perform better than mid-market caps is interesting to observe. This may be due to the total amount of Reddit posts regarding the most popular cryptocurrencies. There are more mid-market cap cryptocurrencies in the model, but by popularity, there are more posts that contain the high-market cap ones. The increase in posts provides more choices for the sentiment portfolio, and a better possibility to

separate the positive and negative sentiment for specific days. This may be why the high-market cap portfolio performs slightly better. At last, although it would be interesting to compare these results with the low market cap portfolio, this will not be possible. This is due to the fact that there is not enough data for the low market cap to implement a cross-section return model.

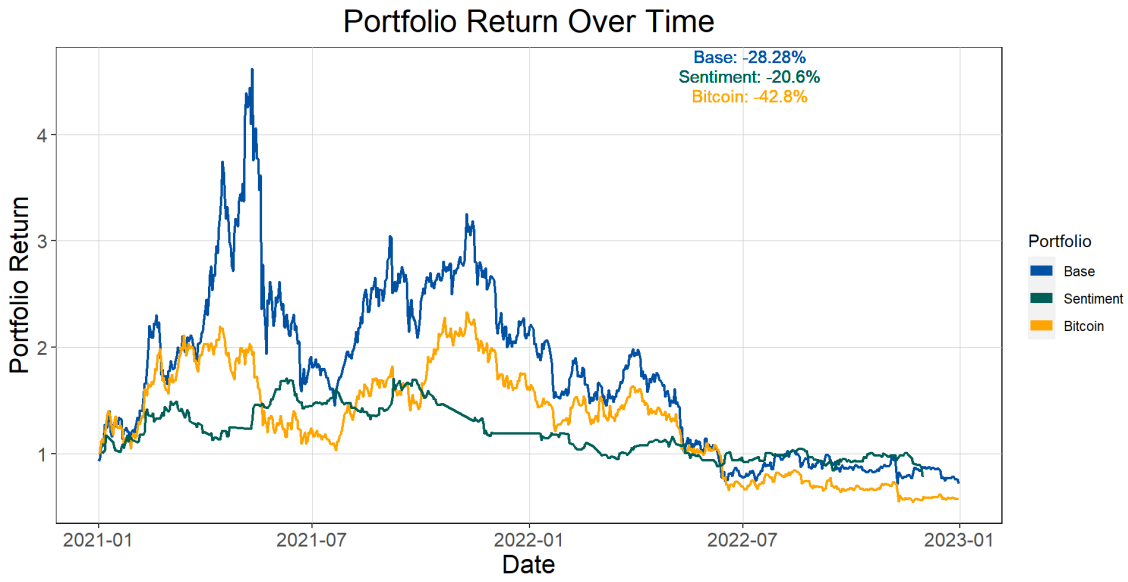

Figure 6.4: Mid market cap portfolio 2021-2023

6.3.3 Sector

The results from the market cap differentiation indicate a preference for cryptocurrencies of a high-level market cap to select a portfolio with Reddit sentiment as the sole decision variable. However, this market cap is rather wide and covers many different segments. Therefore, it is interesting to

categorize the cryptocurrencies by segment instead of market cap to observe which segments to include in a sentiment portfolio.

The category selection method is similar to a stepwise backward regression starting from nine available categories. Although all combinations were not able to be tested due to insufficient data. The worst-performing sentiment portfolio consists of :

Exchange token (centralized)

Payments / digital currency

Distributed computing

NFT tokens

Store of value

From the respective categories, these are some examples of cryptocurrencies: BNB, LTC, HNT, AXS, BTC. A sentiment portfolio of these categories performs poorly and gets a cumulative return of -61.05%. Although the base portfolio also is performing poorly with -43.54%, it is better than the sentiment one. It is worth noticing that Bitcoin is included in all three portfolios, but the least bad of these three portfolios is the one that just contains Bitcoin.

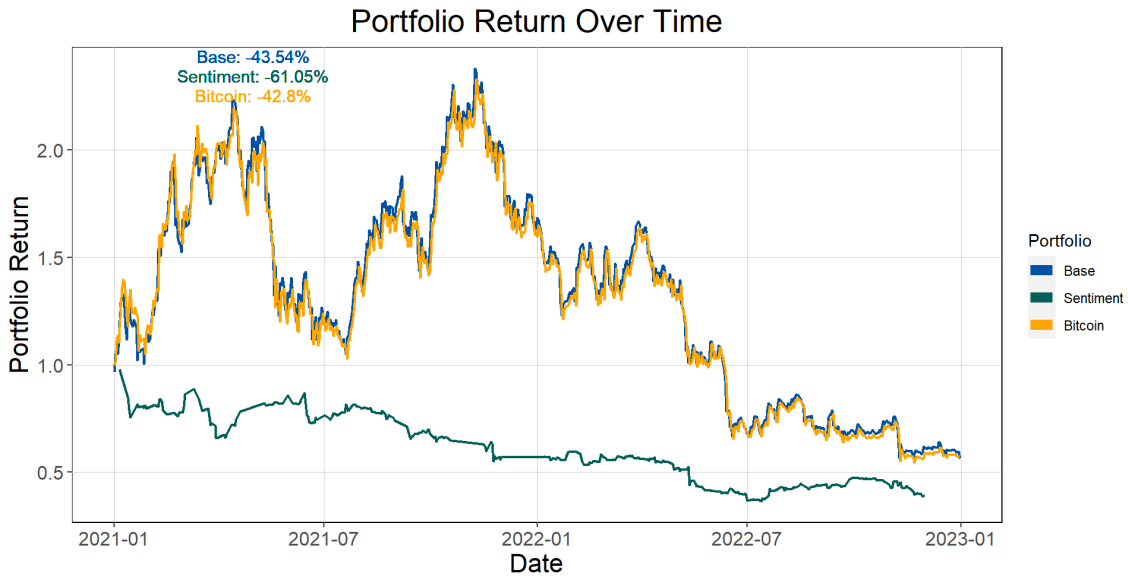

Figure 6.5: Worst category portfolio 2021-2023

On the other hand, the best portfolio consists of the following categories:

Exchange token (decentralized)/DeFi

Smart contracts

Memecoin

Note that although Luna goes within DeFi, it decreases the return of the sentiment portfolio and it has therefore been excluded. It experienced a large drop in value in 2022, but this negative sentiment and price depreciation turned out to not be properly picked up by the sentiment portfolio as a genuine decision variable. The categories in this portfolio are more related to

DeFi and less about traditional finance, with smart contract cryptocurrencies being closely connected to DeFi as well. Some examples of all respective categories could be: AVAX, ETH, DOGE. The fact that memecoins is included in this list as well fits well with the previously mentioned hypothesis of stronger sentiment relation for these type of assets.

This portfolio of the best categories is 331.99% and is outperforming the other sentiment models significantly. In comparison, the base portfolio has a cumulative return of 36.73%. The sentiment portfolio manages to maintain an overall positive trend, with its biggest increase in the second half of 2022 (which also happened for the full sentiment portfolio).

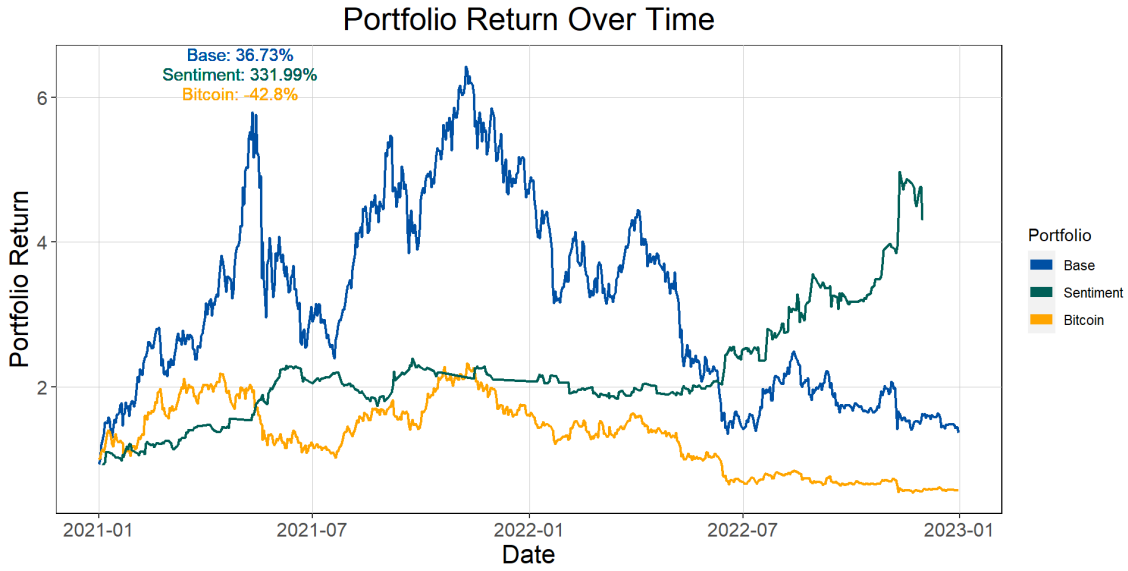

Figure 6.6: Best category portfolio 2021-2023

So, a portfolio with all the cryptocurrencies outperformed the base portfolio and had a cumulative return of 74.63%. Looking deeper into the market

cap, there were some indications that the high-market cap cryptocurrencies were a better decision-maker for a sentiment portfolio than a portfolio built on mid-market caps. The data for the lower market cap was insufficient to regard for itself.

The categorization of market cap appeared to be too wide, so a categorization by sector was then implemented. This led to strong differences in the results. The worst-performing portfolio contains sectors such as store of value and payment systems. BTC and XRP are included in this portfolio. The portfolio had a cumulative return of -61.05%. The best-performing portfolio on the other hand contains the sectors DeFi, smart contracts and meme coins with cryptocurrencies such as AVAX, ETH and DOGE. This gave a cumulative return of 331.99% for the sentiment portfolio which significantly outperformed the Bitcoin portfolio which was -42.8% but also did better than the base portfolio at 36.73%.

To conclude, this may be an indication that there is indeed a relationship between sentiment and cryptocurrency return. However, the cryptocurrency market as a whole is diversified into different market caps and sectors that possess different predictability for tomorrow's return. Failing to identify the cryptocurrencies with a realistic sentiment representation through Reddit may provide a poor decision variable to a sentiment-based portfolio. However, if this is done successfully, as shown here, then there it has the potential as a useful predictor for the return of the right type of cryptocurrencies.

7. Conclusion

Social media is becoming more important for investors as decision variable (*Brunswick's 2023 Digital Investor Survey — Brunswick Group* 2023). The incident with WallStreetBets and GameStop is just one example of events that have highlighted the importance of social media and increased the user base of Reddit through mainstream media coverage. In addition, 15 years after the birth of cryptocurrencies it is commonly referred to in media as highly volatile and sentiment-based. This paper combines these pieces. By looking at the relationship between Reddit sentiment for cryptocurrencies and subsequent returns across different segments - it investigates the importance of sentiment. Loki from ThorFi mentions it as the most important variable for all cryptocurrencies, but are there some differences between the different categories (*Defi Sentiment* 2023)?

This paper uses three methodologies to investigate this sentiment-return relationship. The ordinary least squares method does not perform well in this paper with most of the cryptocurrencies having poor p-values. Random forest on the other hand provides more value as the variable importance differ among the cryptocurrencies - sentiment is more important for some than others. Bitcoin sentiment has a relatively large node purity (0.19) but a low placement of the variable (8.325). Dogecoin however have the sentiment as more important with an average placement of 3.18 but a lower node purity at 0.08. These two models together are not sufficient to fully establish a

sentiment-return relationship.

Cross-section return however is better at displaying the category differences and the importance of Reddit sentiment as a decision variable. A sentiment portfolio that buys and sells the top/bottom 20% of the cryptocurrencies based on yesterday's sentiment outperforms a base model that buys all cryptocurrencies that appear on Reddit between 2021 and 2023. The best model contains DeFi, smart contracts and meme coins, with a cumulative portfolio return of 331.99%. This indicates that there is predictive power in Reddit sentiment, but only for certain segments within the crypto space. Identifying which sectors to use as a sentiment variable may be the difference between a good and bad variable to predict tomorrow's cryptocurrency return. For further research, it would be interesting to utilize another sentiment dictionary that is specific to social media and cryptocurrency terminology. Expanding the social media to create an aggregate with Twitter as well, as per recommendation from Loki(*Defi Sentiment* 2023), is another process that would be interesting to build upon the work of this paper.

Bibliography

Baker, Malcolm and Jeffrey Wurgler (2006). "Investor sentiment and the cross-section of stock returns". In: *The journal of Finance* 61.4. Publisher: Wiley Online Library, pp. 1645–1680.

Baker, Scott R., Bloom, Nick, and Davis, Stephen J. (Jan. 1, 1985). *Equity Market-related Economic Uncertainty Index*. FRED, Federal Reserve Bank of St. Louis. Publisher: FRED, Federal Reserve Bank of St. Louis. URL: https://fred.stlouisfed.org/series/WLEMUINDXD (visited on 05/08/2023).

Bergerud and Hu (Jan. 2023). *Forecasting on Bitcoin*.

Bitcoin, Ethereum, Dogecoin, Litecoin stats (May 9, 2023). BitInfoCharts. URL: https://bitinfocharts.com/ (visited on 05/08/2023).

Bouteska, Ahmed, Salma Mefteh-Wali, and Trung Dang (Nov. 1, 2022). "Predictive power of investor sentiment for Bitcoin returns: Evidence from COVID-19 pandemic". In: *Technological Forecasting and Social Change* 184, p. 121999. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2022.121999. URL: https://www.sciencedirect.com/science/article/pii/S0040162522005200 (visited on 03/01/2023).

Bradley, Daniel et al. (Mar. 15, 2021). "Place Your Bets? The Market Consequences of Investment Research on Reddit's Wallstreetbets". In: DOI: 10.2139/ssrn.3806065. URL: https://papers.ssrn.com/abstract=3806065 (visited on 02/26/2023).

Brunswick's 2023 Digital Investor Survey — Brunswick Group (2023). URL: https://www.brunswickgroup.com/perspectives/digital-investor-survey/2023/ (visited on 05/10/2023).

Cambridge Bitcoin Electricity Consumption Index (CBECI) (2023). URL: https://ccaf.io/cbnsi/cbeci (visited on 05/08/2023).

Camou, Luis Antonio Loredo (May 7, 2022). "Reddit as a prediction tool for crypto-assets". In: *Brazilian Review of Finance* 20.1. Number: 1, pp. 1–39. ISSN: 1984-5146. DOI: 10.12660/rbfin.v20n1.2022.83888. URL: https://bibliotecadigital.fgv.br/ojs/index.php/rbfin/article/view/83888 (visited on 04/07/2023).

CoinMarketCap (2023). *CoinMarketCap*. CoinMarketCap. URL: https://coinmarketcap.com/ (visited on 02/28/2023).

Crypto market cap ranking 2022 (2023). Statista. URL: https://www.statista.com/statistics/1269013/biggest-crypto-per-category-worldwide/ (visited on 05/12/2023).

Defi Sentiment (Mar. 29, 2023). In collab. with Loki.

Fees — Uniswap (2023). URL: https://docs.uniswap.org//contracts/v2/concepts/advanced-topics/fees (visited on 05/24/2023).

Filippou, Ilias, David Rapach, and Christoffer Thimsen (Jan. 1, 2021). "Boosting Cryptocurrency Return Prediction". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3914414.

Historical Snapshot - 01 January 2021 (2023). CoinMarketCap. URL: https://coinmarketcap.com/historical/20210101/ (visited on 05/24/2023).

Hu, Danqi et al. (Mar. 14, 2021). "The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery". In: DOI:

10.2139/ssrn.3807655. URL: https://papers.ssrn.com/abstract=3807655 (visited on 02/26/2023).

investing.com (May 5, 2023). *MSCI World Index (MIWO00000PUS)*. Investing.com. URL: https://www.investing.com/indices/msci-world (visited on 05/08/2023).

Lilya, Kristen (2023). *Which is the most useful Cryptocurrency Subreddits?* Native News Online. URL: https://nativenewsonline.net/advertise/branded-voices/which-is-the-most-useful-cryptocurrency-subreddits (visited on 05/29/2023).

Loughran, Tim and Bill Mcdonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *The Journal of Finance* 66.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01625.x, pp. 35–65. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2010.01625.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x (visited on 03/30/2023).

Meynkhard, Artur (Nov. 28, 2019). "Fair market value of bitcoin: halving effect". In: *Investment Management and Financial Innovations* 16.4, pp. 72–85. ISSN: 18104967, 18129358. DOI: 10.21511/imfi.16(4).2019.07. URL: https://businessperspectives.org/journals/investment-management-and-financial-innovations/issue-334/fair-market-value-of-bitcoin-halving-effect (visited on 03/30/2023).

Next Bitcoin Halving 2024 Date & Countdown [BTC Clock] (2023). Bitcoin Halving. URL: https://buybitcoinworldwide.com/halving/ (visited on 05/08/2023).

Nguyen, Khanh Quoc (May 1, 2022). "The correlation between the stock market and Bitcoin during COVID-19 and other uncertainty periods". In: *Finance Research Letters* 46, p. 102284. ISSN: 1544-6123. DOI: 10.1016/j.frl.2021.102284. URL: https://www.sciencedirect.com/science/article/pii/S1544612321003238 (visited on 03/30/2023).

Palamalai, Srinivasan, K. Krishna Kumar, and Bipasha Maity (Sept. 1, 2021). "Testing the random walk hypothesis for leading cryptocurrencies". In: *Borsa Istanbul Review* 21.3, pp. 256–268. ISSN: 2214-8450. DOI: 10.1016/j.bir.2020.10.006. URL: https://www.sciencedirect.com/science/article/pii/S2214845020300673 (visited on 03/30/2023).

Prajapati, Pratikkumar (Jan. 16, 2020). *Predictive analysis of Bitcoin price considering social sentiments*. DOI: 10.48550/arXiv.2001.10343. arXiv: 2001.10343[cs]. URL: http://arxiv.org/abs/2001.10343 (visited on 03/01/2023).

r/CryptoCurrency subreddit stats (2023). URL: https://subredditstats.com/r/CryptoCurrency (visited on 05/29/2023).

Reddit (2023). *WallStreetBets Description*. URL: https://www.reddit.com/subreddits/search?q=wallstreetbets (visited on 02/27/2023).

ThorFi price today, THOR to USD live, marketcap and chart (2023). CoinMarketCap. URL: https://coinmarketcap.com/currencies/thor/ (visited on 05/11/2023).

Watchful1 (Feb. 28, 2023). *Separate dump files for the top 20k subreddits*. r/pushshift. URL: www.reddit.com/r/pushshift/comments/11ef9if/separate_dump_files_for_the_top_20k_subreddits/ (visited on 03/30/2023).

Wooley, Stephen et al. (Dec. 2019). "Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment". In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 500–505. DOI: 10.1109/ICMLA.2019.00093.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/lsm**