Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

Håkon C. Bjørgan
Karl G. Nakken
Erik B. Dukefoss

# Predicting Cryptocurrency Returns Using Market, Social Media, Search Volume and Blockchain Data

Master's thesis in Industrial Economics and Technology Management
Supervisor: Peter Molnár
June 2021

NTNU
Norwegian University of
Science and Technology

Håkon C. Bjørgan
Karl G. Nakken
Erik B. Dukefoss

# Predicting Cryptocurrency Returns Using Market, Social Media, Search Volume and Blockchain Data

Master's thesis in Industrial Economics and Technology Management
Supervisor: Peter Molnár
June 2021

Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

NTNU
Norwegian University of
Science and Technology

# Preface

This master's thesis is written as the fulfillment of our Master of Science in Industrial Economics and Technology Management at the Norwegian University of Science and Technology. The purpose of our thesis is to investigate aspects of the cryptoasset market. In particular, we explore the predictability of asset returns and its implications for market efficiency.

During our years at NTNU, cryptoassets and blockchain technology have evolved from a niche interest to a conversation topic among average Joes. Situated at the intersection of technology and finance, it has triggered our curiosity as students in these academic fields. With last year's formidable bull market as a backdrop, we decided to dive into the world of cryptoassets to gain insight into what this alleged financial revolution is all about.

We sincerely wish to thank our supervisor, Peter Molnár. He has provided us with much crucial guidance and help during our work. Also, we would like to express appreciation for the initial assistance in surveying possible research topics offered by Torbjørn Bull Jensen and Vetle Gusgaard Lunde at Arcane Crypto.

# Abstract

This thesis examines if it is possible to predict cryptocurrency returns. To do this, we have constructed a unique dataset consisting of social media, search volume, blockchain and market data for 54 different cryptocurrencies. First, returns are forecast with a linear regression model using only market data. Afterward, features collected from Twitter, Reddit, Google Trends and the underlying blockchains are added to the model. Lastly, we compare this extended linear model to an advanced machine learning model. These models are all backtested on the period from March 2020 to March 2021.

Our analysis finds that the extended linear regression model combined with a realistic trading strategy delivers high risk-adjusted returns. The model beats the market even when we account for transaction costs. This shows that cryptocurrency prices are predictable as of March 2021. Furthermore, we find that market and Twitter data significantly explain price movements. On the other hand, Google Trends, Reddit or blockchain data does not improve the model's forecasts. Nor do we find that machine learning models outperform linear models, contrary to much of the literature on this topic.

# Contents

# List of Figures

# List of Tables

# Glossary and Terminology

The emergence of crypto has introduced new vocabulary. Most of which unknown for all but the people actively participating in the crypto community. In the following thesis we use the word cryptoasset as a collective term for cryptocurrencies and related assets like tokens. Although different practices are endorsed, we have chosen to adhere the following standard with regard to capitalisation: Cryptoassets themselves are written with an initial lowercase letter, while the underlying blockchain is capitalized. E. g. bitcoin refers to the tradeable asset, while Bitcoin refers to the blockchain. No proper community standard has been set with regards to word compounding. For consistency we write cryptoassets and cryptomarkets in their closed compounded form. Following is a glossary of domain specific expressions used:

**Altcoin** Collective term for all other cryptoassets than bitcoin.

**Blockchain** A digital ledger of transactions that is distributed across the entire network of computer systems on the blockchain.

**Stablecoin** A kind of cryptoasset designed to trade in a fixed relationship with another asset (e.g., the U.S. Dollar).

**Wallet** In the world of crypto, a wallet is a way to hold the keys (a sequence of alphabetic and numeric characters) you need to interact with your cryptoassets. Mostly people use dedicated software or hardware, but in theory a wallet could be a piece of paper with the keys written on it.

# Chapter 1

# Introduction

Over the last decade, a new class of financial assets, so-called cryptoassets, has garnered exceptional attention. While bitcoin (BTC) is the most well-known, a plethora of different cryptoassets have been devised over the last years. The growth in market capitalization of these assets and the interest of academic economists for them have increased in tandem.

As history has proven, predicting financial asset returns can be a confounding and laborious exercise. Elementary financial economics tells us that any easily identifiable mispricing will quickly be identified, exploited and thus disappear. With cryptomarkets still in their infancy, studying whether they adhere to the efficient market hypothesis is of great interest for stakeholders and academia alike. Given the lack of consensus on fundamental valuation, significant effort has already been devoted to identifying drivers of cryptoasset prices. With prices arguably being more speculation-driven than for traditional financial assets, social media is often pointed to as a potent data source. In combination with data from other sources, such as search volumes and the underlying blockchains, academics have to some degree been able to establish price-driving factors. Common for most research done on this topic is, however, the limited selection of assets. The vast majority focus solely on bitcoin. The remaining few expand to include popular altcoins such as ether (ETH) and litecoin (LTC). Truly wide-ranging research into the drivers of cryptomarkets do, to the best of our knowledge, until now remain undone.

We have constructed a large and unique dataset that comprises data extracted from Reddit and Twitter, as well as Google Trends, different cryptoasset exchanges and the underlying blockchains. From this data, we have constructed variables, which are used to calibrate a single panel data regression equation to retrodict daily returns for 54 different cryptoassets. Subsequently, we use backtesting to check if we can achieve robust cumulative returns significantly above the market benchmark.

In principle, our thesis attempts to answer a three-part research question. Firstly, are models calibrated using richer datasets able to deliver higher returns than those using smaller datasets? Secondly, do more advanced machine learning methods improve upon the predictions made by linear prediction models? Thirdly and ultimately, can we systematically generate abnormal returns?

The structure of the thesis is as follows: Chapter 2 provides a survey of the scientific literature addressing the use of specific data and methodology in cryptoasset price prediction, as well as overall cryptomarket efficiency. Chapter 3 details the data extraction, cleaning and transformation procedures used to engineer our variables. In Chapter 4 we introduce the methodology used in this thesis. Chapter 5 details and discusses the model performances and the results of our backtesting exercise. The last chapter contains a short conclusion and ideas for further inquiry.

# Chapter 2

# Literature Review

The following chapter surveys literature pertinent to our research inquiry. The two first sections present some of the principal literature on using social media, Google Trends, blockchain data and machine learning in cryptoasset return prediction. Lastly, we present some important papers discussing the efficiency of the cryptomarket in light of the efficient market hypothesis (EMH).

## 2.1 Social Media, Google Trends and Blockchain Data in Cryptoasset Price Prediction

A principal aspect of our thesis is to examine which regressors influence price predictions. In this subsection, we review some of the chief contributions to studying the use of search volumes, social media sentiment and on-chain features as regressors.

The use of search volume data from Google Trends has a long and relatively positive track record in cryptoasset price prediction. An early article by Kristoufek (2013) finds that search volumes from Google Trends and Wikipedia and cryptoasset prices mutually affect each other. Urquhart (2018) asserts that bitcoin price volatility and trading volume influence Google Trends search volume but states that search volumes does not have predictive power on returns. Kim et al. (2017) on the other hand, found that Google Trends and Wikipedia data could inform a bitcoin price prediction model. Another frequently-cited article by Matta et al. (2015) also finds that fluctuations in Google Trends data were associated with significant movements in the future bitcoin price.

Many studies also leverage data extracted from social media in price prediction. However, the overall evidence on the regressors' potency is somewhat mixed. Matta et al. (2015) find that volumes of positive messages on Twitter were able to significantly predict movements in the bitcoin price even three to four days in advance. Similarly, Abraham et al. (2018) asserts that using Twitter post volume and Google Trends data helped predict next-day price movements for ether and bitcoin. Sentiment values were, on the other hand, not found to be informative. In contrast with this, Shen et al. (2019) state that Twitter message volumes significantly explain bitcoin returns. Lamon et al. (2017) claim they can predict extraordinary price changes using sentiment analysis of Twitter data, while Pant et al. (2018) interestingly find that these dependencies can be asymmetric. More specifically, negative sentiments were shown to be a stronger predictor of price movements than positive sentiments. Kaminski (2014) is to the contrary not able to demonstrate that sentiment on Twitter affects prices. They instead argue that price changes induce changes in expressed sentiments.

Data from Reddit is less frequently used in price prediction analysis than Twitter data. However, the work done in Wooley et al. (2019) and Phillips and Gorse (2018) suggests that both sentiment and message board activity can serve as significant regressors in cryptoasset return prediction models.

Some articles also attempt to use features derived from each cryptoasset's blockchain. Jang and Lee (2017) are successful in using blockchain features and macroeconomic variables to predict bitcoin prices. Similarly, Saad et al. (2019) predict cryptoasset prices using variables such as the blockchain hash rate, transactions rates, the number of users and total currency supply. Another recent article by Ji et al. (2019) used blockchain data in deep learning models to make profitable trades using a simple trading strategy.

The literature is replete with examples of isolated use of sentiment data, Google Trends and Blockchain data. However, very few articles use such a wide array of data sources as we do, making this a principal contribution of our thesis. Furthermore, it enables us to assess the impact made by each data type more accurately than what previous studies do.

## 2.2 Machine Learning and Cryptoasset Price Predictions

Despite being relatively new fields, many articles have attempted to couple machine learning with cryptoasset price predictions. In short, researchers have applied a wide variety of algorithms to the problem with varying degrees of success.

A seminal paper by Madan et al. (2015) states that the random forests method provides better binomial predictions than both generalized linear models (GLM) and linear regression. Mallqui and Fernandes (2019) conducted a similar exercise using artificial neural networks (ANNs), support vector machines (SVMs) and ensembles to predict daily maximum, minimum and closing rates, finding that the SVMs performed the best. Jang and Lee (2017) employed a Bayesian neural network (BNN) to predict bitcoin prices and showed that it could outperform SVMs and models based on linear regression.

In the literature, the use of long short-term memory (LSTM) networks is widespread. Lahmiri and Bekiros (2019) show that the time series for bitcoin, digital cash (DASH) and ripple (XRP) exhibit fractal dynamics, long memory and self-similarity. The authors used an LSTM-network to produce return predictions and found it superior to a general regression neural network. Similarly, McNally et al. (2018) were able to show that recurrent neural networks (RNN) and LSTM-networks produced more accurate predictions than simpler ARIMA models. The LSTM-model calibrated with high-dimensional data presented in Chen et al. (2020) outperforms statistical methods like logistic regression on time intervals shorter than a day. However, using a daily horizon like in our thesis, the simpler models outperform the LSTM-network. Additionally, Mudassir et al. (2020) find that their LSTM-model could outperform a regular ANN and SVM for daily as well as longer time horizons. The relative success of LSTM-networks in this field motivates our choice of model.

## 2.3 Cryptoasset Market Efficiency

Researchers have devoted effort to investigating the cryptoasset market in light of the Efficient Market Hypothesis (EMH). One implication of the hypothesis is that consistent abnormal returns are unobtainable from following a simple investment strategy. Accordingly, consistently predicting returns should not be possible in an efficient market. As shown below, many articles argue that the cryptoasset market is inefficient. However, several also differ in their assessment of how market efficiency changes over time.

Urquhart (2016) applied a series of robustness tests to the bitcoin market from 2010 to 2016 and find that the market is inefficient over the entire period. They also provide evidence suggesting that the largest cryptoasset market, namely the bitcoin market, is becoming more efficient. Tiwari et al. (2018) built on this work by introducing a battery of long-range dependence estimators, which indicated progressively increased efficiency in the bitcoin market. This claim is also supported by Bariviera (2017). On the other hand, though, Jiang et al. (2018) found no evidence of the bitcoin market becoming more efficient over time when applying a Hurst exponent analysis.

Other articles have studied the EMH on a longer time horizon using broader sets of cryptoassets. Caporale et al. (2018) examine the movement persistence evident in the cryptoassets bitcoin, ripple, dash and litecoin. They find through the use of long-memory methods that there is evidence of contracting market inefficiency across assets. In their analysis, Aggarwal (2019) found evidence of market inefficiency due to the presence of asymmetric volatility clustering from 2010 to 2018. Khuntia and Pattanayak (2018) argue that the Adaptive Market Hypothesis described in behavioral economics more aptly describes the development in bitcoin prices. They point out that behavior biases and herd mentality make it possible for speculators and arbitrageurs to gain excess returns.

In addition to using a unique combination of features, our thesis contributes to the literature by studying a comparatively broad set of cryptoassets. Economists have so far devoted most of their time to studying popular assets like bitcoin and ether. Our thesis, however, provides a comprehensive perspective on the market by analyzing tens of cryptoassets at once. Analyzing a broad cross-section of the market also has implications for the validity of our analysis of overall market efficiency.

# Chapter 3

# Data

Besides market data (i.e., prices and trading volumes), our model uses data from Google Trends, Twitter, Reddit, as well as the underlying blockchains. Below, we describe the data extraction, feature engineering and variable transformations used to produce our final dataset.

## 3.1 Asset Selection

Our exercise presupposes a broad set of available cryptoassets. Assets were initially selected based on having been part of the cryptoasset index, CCi30. Assets with especially noisy price data, assets that went broke before or were launched during the testing period, were pruned. Subsequently, we removed assets with a median daily trading volume of less than $1,000,000 in 2019. Firstly, such assets generally have poor data quality, making them hard to use. Secondly, if assets are so illiquid that we cannot reasonably act as price-takers, using them in our analysis could produce invalid results. Stablecoins were also excluded. Finally, we excluded all assets that did not have enough associated Twitter or Reddit data. Figure 3.1 summarizes the selection process and the number of assets removed. Which assets are removed in each step is detailed in Appendix C.

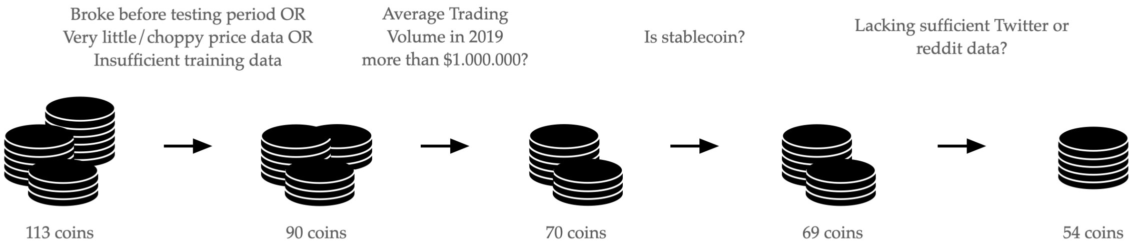
Broke before testing period OR
Very little/choppy price data OR
Insufficient training data
Average Trading
Volume in 2019
more than $1.000.000?
Is stablecoin?
Lacking sufficient Twitter or
reddit data?
113 coins
90 coins
70 coins
69 coins
54 coins

**Figure 3.1:** The asset selection process.

## 3.2 Returns, Trading Volumes and Volatility

Returns are used as both regressors and output in our prediction models. Prices and trading volumes are sourced from the CoinGecko API (Coingecko, 2021). Figure 3.2 shows indexed prices for four popular cryptoassets. One recognizes immediately that the prices are highly correlated and that they have increased sharply since 2015.

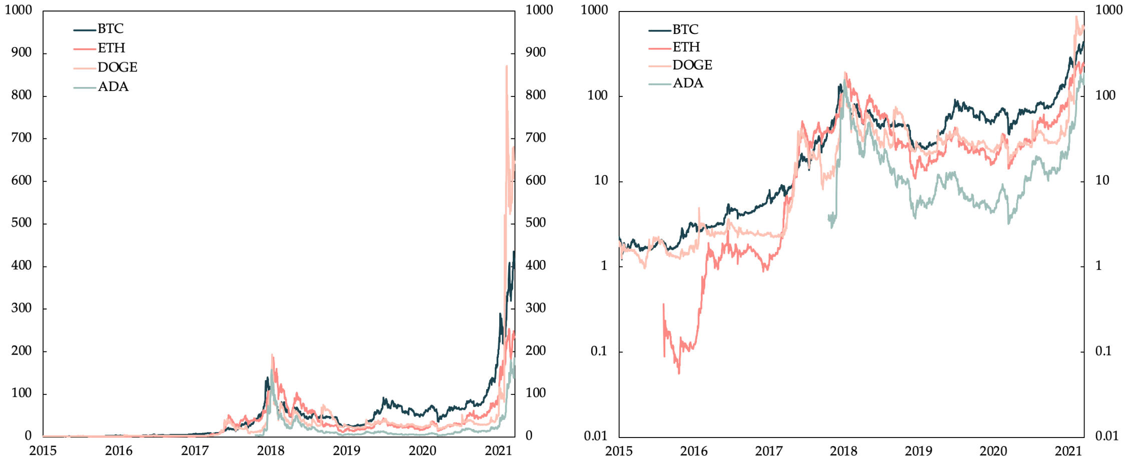

**Figure 3.2:** Prices for bitcoin (BTC), ether (ETH), dogecoin (DOGE) and cardano (ADA). Prices are indexed to be 100 at 1 January 2018. Plotted on a linear (left) and logarithmic scale (right).

Returns are derived using Equation 3.1.

$$Return_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{3.1}$$

Growth in trading volume has historically developed exponentially, as is decipherable from Figure 3.3. The model input variable, $RelativeLnTradingVolume_t$ is the natural logarithm of the daily trading volume relative to its own weekly average (see Equation 3.2).

$$RelativeLnTradingVolume_t = \frac{LnTradingVolume_t}{\frac{1}{7}\sum_{\tau=t-7}^{t} LnTradingVolume_\tau} - 1 \tag{3.2}$$

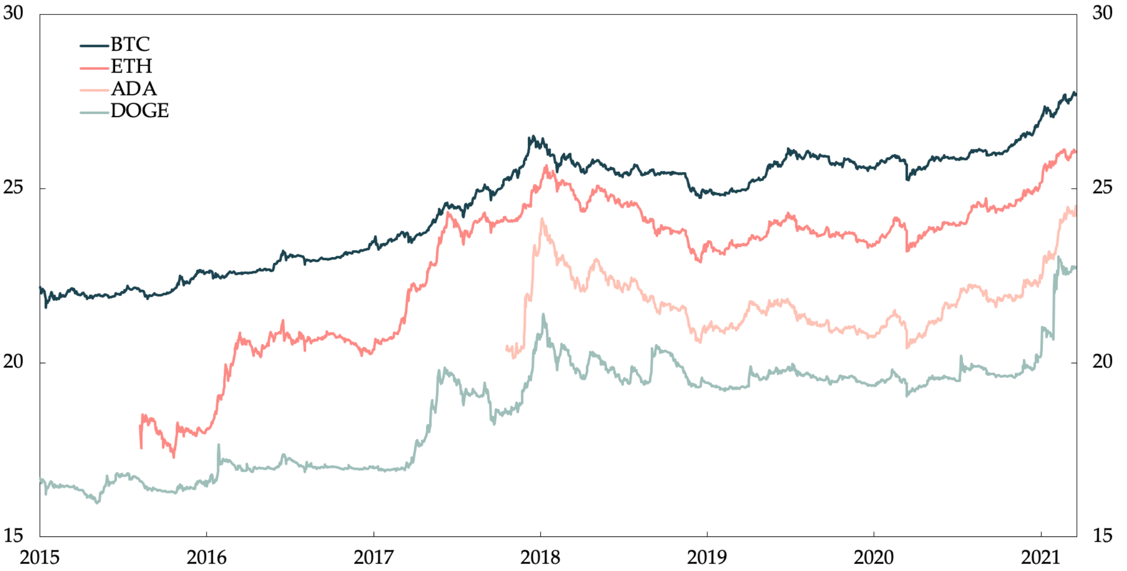

**Figure 3.3:** The natural logarithm of trading volumes for bitcoin (BTC), ether (ETH), dogecoin (DOGE) and cardano (ADA). The exponential growth motivates our choice of transformation.

The finance literature intimately links asset return and volatility. We, therefore, include both a weekly and monthly rolling volatility estimate in our models. We measure volatility using the rolling sample standard deviation where the sample mean is assumed to be zero (Alexander, 2008). Equation 3.3 shows the mathematical expression for the weekly and monthly volatility with $n = 7$ and $n = 30$, respectively.

$$\sigma_t = \sqrt{\frac{\sum_{\tau=t-n}^{t} r_\tau^2}{n}} \tag{3.3}$$

## 3.3 Google Trends

Google is by far the most used online search engine, with more than 3.5 billion queries processed each day (InternetLiveStats, 2021). The internet service giant provides indexed search volume data through its Google Trends service. We composed an array of queries (see Table B1) to accurately estimate the relative search traffic for each asset. Google only provides daily data for query periods shorter than 270 days. Therefore, we acquired and concatenated overlapping time series using the *rtrends* software package (Blinder, 2016). Figure 3.4 shows the estimated search volume related to the asset ether plotted along with the asset's price in USD. The co-movement of the two time series motivates our use of the data.

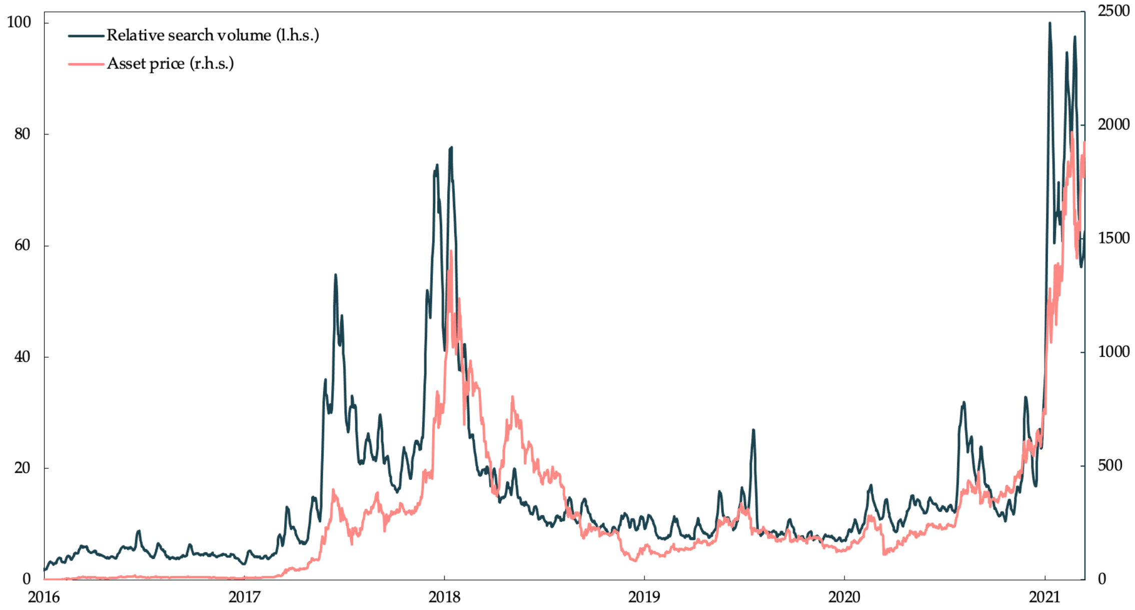

**Figure 3.4:** Relative Google search volume and ether (ETH) asset price in USD. Search index scaled to have a maximum of 100 over the period.

The search volume relative to its weekly average (Equation 3.4) and a rolling weekly average of search volumes (Equation 3.5) enter our model as regressors. While the latter transformation corrects for weekly seasonality, the former is deseasoned prior to the regression using the Python package *statsmodels* (Seabold & Perktold, 2010).

$$RelativeGoogleTrendsVolume_t = \frac{SearchVolume_t}{\sum_{\tau=t-7}^{t} SearchVolume_{t-7}} - 1 \tag{3.4}$$

$$GoogleTrendsWeeklyVolume_t = \sum_{\tau=t-7}^{t} SearchVolume_{t-7} \tag{3.5}$$

## 3.4 Twitter Data

The microblogging application Twitter was launched in 2006. Its more than 352 million active users make it an ideal place for us to gauge marketplace interest. Twitter lets users post public messages or *tweets* of up to 280 characters. So-called *hashtags* (#) are commonly used to identify the topic of a tweet and make them searchable. In the finance and cryptoasset realm of Twitter, users also extensively employ so-called *cashtags* ($, followed by the asset ticker, e.g., $BTC for bitcoin). We use these tags to isolate tweets related to specific cryptoassets.

### 3.4.1 Data Collection

To test our hypothesis, we collected 24.8 million tweets using the Python script *twint* (Twint-Project, 2017) and estimated the sentiment they express. Twint selects tweets based on provided search terms such as cashtags. Where only using cashtags results in sparse data, we included tweets that mentioned the asset name. For example, our sentiment and activity indicators for reddcoin use tweets containing either “reddcoin” or “$rdd.” A full list of terms used to isolate tweets can be found in Appendix B.

On the flip side, some assets generate enormous quantities of Twitter activity. For bitcoin, ether, dogecoin and ripple, only tweets with a minimum number of likes were downloaded to not skimp on valuable computation time. Finally, some assets have ambiguous ticker names (e.g., BTS, the ticker for bitshares, is also a Korean boyband). In these instances, we have only used the asset name to target relevant tweets.

### 3.4.2 Variable Construction

Tweets require processing prior to the sentiment analysis. First, the publishing time was adjusted to align the timezone with the pricing data. We removed duplicate tweets from our dataset. Messages containing words like “free”, “win”, “game”, “bet” and “pic” are filtered out to remove bot-generated content. Tweets generally contain a considerable amount of noise that does not contribute information to the sentiment analysis. Elements such as hyperlinks, hash- and cashtags, HTML-tags, mentions of other users and various signs and numbers are removed prior to estimating sentiments. As our analysis tool handles emojis (Shoeb and de Melo, 2021), these are left in. Messages are also *lemmatized* (i.e., words are transformed into their dictionary form) using the commonly used Wordnet lemmatizer from the NLTK library (Bird et al., 2009). ”Walking” and ”contracts” are for example transformed into ”walk” and ”contract” respectively. Figure 3.5 shows the processing steps for a sample tweet.

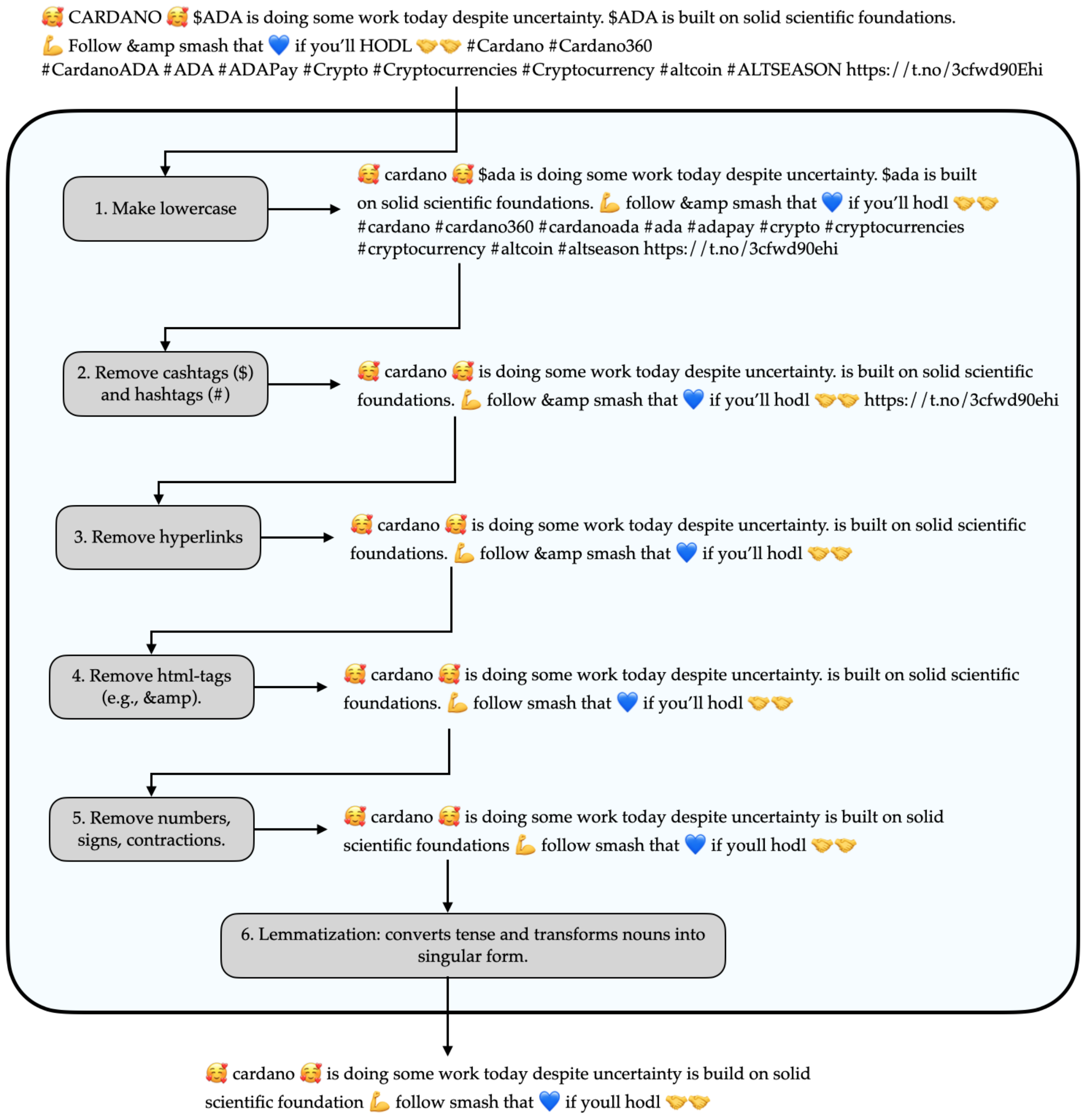
🥰 CARDANO 🥰 $ADA is doing some work today despite uncertainty. $ADA is built on solid scientific foundations. 💪 Follow &amp smash that 💙 if you'll HODL 🤝🤝 #Cardano #Cardano360 #CardanoADA #ADA #ADAPay #Crypto #Cryptocurrencies #Cryptocurrency #altcoin #ALTSEASON https://t.no/3cfwd90Ehi
1. Make lowercase
🥰 cardano 🥰 $ada is doing some work today despite uncertainty. $ada is built on solid scientific foundations. 💪 follow &amp smash that 💙 if you'll hodl 🤝🤝 #cardano #cardano360 #cardanoada #ada #adapay #crypto #cryptocurrencies #cryptocurrency #altcoin #altseason https://t.no/3cfwd90ehi
2. Remove cashtags ($) and hashtags (#)
🥰 cardano 🥰 is doing some work today despite uncertainty. is built on solid scientific foundations. 💪 follow &amp smash that 💙 if you'll hodl 🤝🤝 https://t.no/3cfwd90ehi
3. Remove hyperlinks
🥰 cardano 🥰 is doing some work today despite uncertainty. is built on solid scientific foundations. 💪 follow &amp smash that 💙 if you'll hodl 🤝🤝
4. Remove html-tags (e.g., &amp).
🥰 cardano 🥰 is doing some work today despite uncertainty. is built on solid scientific foundations. 💪 follow smash that 💙 if you'll hodl 🤝🤝
5. Remove numbers, signs, contractions.
🥰 cardano 🥰 is doing some work today despite uncertainty is built on solid scientific foundations 💪 follow smash that 💙 if youll hodl 🤝🤝
6. Lemmatization: converts tense and transforms nouns into singular form.
🥰 cardano 🥰 is doing some work today despite uncertainty is build on solid scientific foundation 💪 follow smash that 💙 if youll hodl 🤝🤝

**Figure 3.5:** Stages of pre-processing of a sample tweet prior to the sentiment analysis.

VADER (Valence Aware Dictionary and sEntiment Reasoner), described further in section 4.1, is used for the sentiment analysis. The goal of the analysis is to determine whether a text generally has a positive or negative disposition. By averaging over the sentiment for tweets published on a given day, we obtain a single sentiment time series for every asset. The number of tweets per day provides an estimate of the overall activity level.

While the daily sentiment value is included in the model as is, $RelativeTwitterVolume_t$ is the number of tweets per day relative to its weekly average (see Equation 3.6). A rolling weekly average of the number of posts is also included as an independent variable (see Equation 3.7).

$$RelativeTwitterVolume_t = \frac{TwitterVolume_t}{\frac{1}{7}\sum_{\tau=t-7}^{t} TwitterVolume_\tau} - 1 \tag{3.6}$$

$$TwitterWeeklyVolume_t = \frac{1}{7} \sum_{\tau=t-7}^{t} TwitterVolume_\tau \tag{3.7}$$

Figure 3.6 displays tweet volume related to ether plotted against that asset's price in USD. The evident co-variation between the series, as well as the consistent use of such data in the literature, motivate our use of the volume variables.

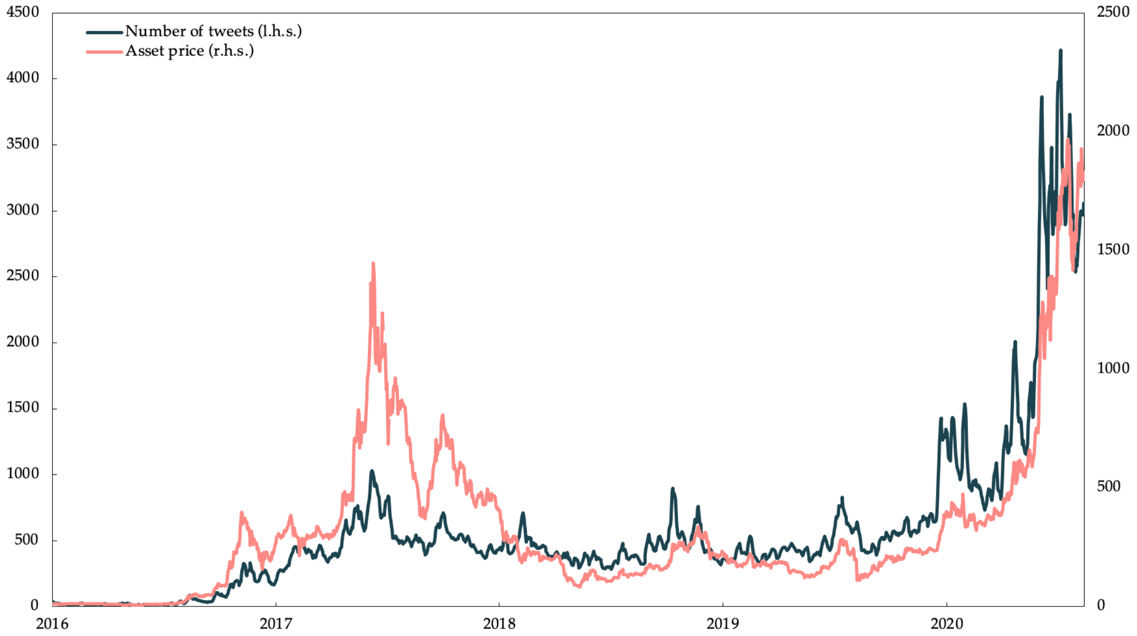

**Figure 3.6:** Seven-day average Twitter message volume related to ether and ether (ETH) asset price in USD.

## 3.5 Reddit Data

Reddit is most aptly described as a collection of forums called subreddits. Subreddits are devoted to particular topics like pictures of space (r/spaceporn), lifehacks (r/lifehacks) or cryptoassets like bitcoin (r/Bitcoin) and dogecoin (r/dogecoin). Within a subreddit, users can post anything that conforms to the subreddits' rules and guidelines. In crypto-related subreddits, this could include anything from so-called memes to detailed assessments of the state of the currency. Users also regularly discuss future technical developments and possible improvements. Once published, other users can up- or downvote a post. The net number of up-votes partly determines a post's visibility.

Although less known than Facebook or Twitter, Reddit jolted the mainstream in early 2021. Extraordinary price fluctuations in the GameStop (GME) and AMC Theatres (AMC) stocks have been attributed to activity in the subreddit r/wallstreetbets. While the impact might be more slight, we hypothesize that sentiment and activity in crypto-associated subreddits might correlate with future asset returns.

### 3.5.1 Data Collection

The premier step in collecting the Reddit data involves finding the main subreddit for each cryptoasset. In instances where an asset has several associated subreddits, we decided to probe the ostensibly most popular one. A list of subreddits used in our analysis is on display in Appendix B. The second step calls for scraping all posts published between July 2014 and March 2021. However, Reddit prevents such mass collection of data through its API. Glenski et al. (2019) and Burnie and Yilmaz (2019) circumvent these restrictions through the use of *Pushshift* (Baumgartner et al., 2020). We adopt the same practice. Pushshift is a free to use, independent third-party project that maintains a clone of Reddit's post history. While posts themselves are available through this service, some pertinent information is lost when compared to fetching Reddit's data directly. Ideally, we would prefer to acquire snapshots of subreddits on any arbitrary day. This luxury would enable weighing sentiment estimates by a post's popularity, the number of up- and downvotes, or other metadata. Pushshift collects data frequently but does not retroactively update its database with changes to previously seen posts. This practice implies that only a tiny fraction of the existing metadata is available through the service. Nevertheless, we were able to build a raw dataset of 2,262,761 posts extracted from 54 subreddits.

### 3.5.2 Variable Construction

The number of posts published on a given day is used to indicate the activity level in a given subreddit. Each individual post is processed prior to estimating its expressed sentiment. As in our Twitter analysis, posts are pre-processed by removing formatting characters like '\n', hyperlinks and other noise. We extract sentiment estimates from each post's title and body using VADER.

Many Reddit-posts only consist of a title and a graphical element like a gif or a picture. Since we do not want to discard such posts, we let the sentiment measure for a post consist of the average sentiment value of the title and the body. By averaging over this value for each post published within the same day, we obtain a daily sentiment estimate.

Textboxes 3.1 and 3.2 show two posts made to the subreddit *r/cardano*. In 3.1 we see an example of a title VADER gives a score of 0, meaning that its sentiment is estimated to be completely neutral. The body of 3.1 is a story of how crypto can provide access to capital in low-income countries. To the human eye the body of 3.1 seems bullish on cryptoassets in general and cardano (ADA) specifically. The sentiment score for the body is 0.9939 and thus in accordance with our human judgement. Textbox 3.2 shows a post on the opposite side of the sentiment spectrum. The post expresses concerns that the user might not see profits on their investment in cardano, and both title and body appear strongly negative. The accompanying sentiment scores of -0.6096 and -0.9653 for title and body respectively seem to accurately reflect the post's sentiment. The complete posts can be found in Appendix A.

The aforementioned processing results in three separate variables: $RedditWeeklyVolume_t$, $RelativeRedditVolume_t$ and $RedditDailySentimentValue_t$. $RedditWeeklyVolume_t$ and $RelativeRedditVolume_t$ enter into the model as described in Equation 3.8 and Equation 3.9. Like

*Title:* 'The Power of Capital in Africa: A Personal Example'
*Body:* 'I served in the Peace Corps in an African country. The village was small, 300 people at most. Everyone was a farmer, growing peanuts mostly. (...) Thank you to this community and lets push forward!Tldr: Gave a loan to a African farmer and he used it to benefit the whole village. Cardano could help millions become food secure and rise from poverty.'

**Textbox 3.1:** Excerpt of a post to *r/cardano* analyzed by VADER. The accompanying scores were 0.0 and 0.9939 for the title and body respectively.

*Title:* Why is there so much hate on ADA?
*Body:* 'I am below newbie in all the crypto stuff. I started buying by the end of January (and Im also not wealthy at all). (...) Ofc im not planning to become millionaire with such a low amount of crypto but damn it feels like I wont be able to make any money at all with my investment.'

**Textbox 3.2:** Excerpt of a post to *r/cardano* analyzed by VADER. The accompanying scores were -0.6096 and -0.9653 for the title and body respectively.

the Twitter sentiment, the Reddit sentiment is not further transformed.

$$RedditWeeklyVolume_t = \frac{1}{7}\sum_{\tau=t-7}^{t} RedditVolume_\tau \tag{3.8}$$

$$RelativeRedditVolume_t = \frac{RedditVolume_t}{\frac{1}{7}\sum_{\tau=t-7}^{t} RedditVolume_\tau} - 1 \tag{3.9}$$

Figure 3.7 shows the daily number of posts in the subreddit *r/dogecoin* against the asset price of dogecoin. The two measures seemingly co-vary, which motivates our use of the data in our return prediction.

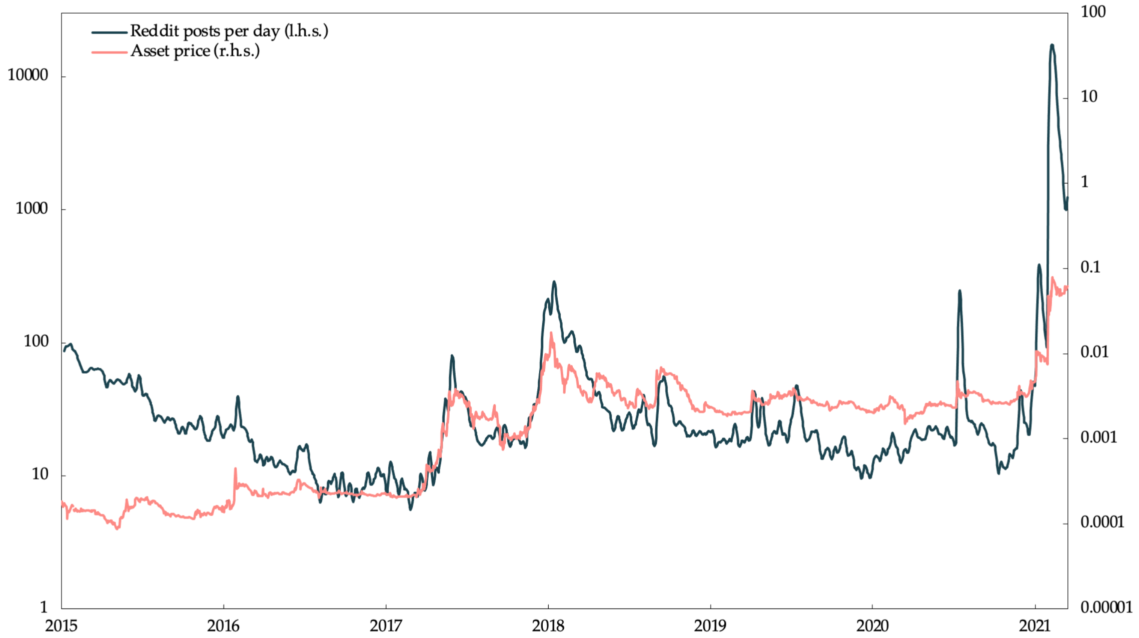

**Figure 3.7:** Seven-day moving average of number of posts made to *r/dogecoin* and dogecoin (DOGE) asset price in USD. Logarithmic scale.

## 3.6 Blockchain Data

In general, all decentralized cryptoassets have an associated blockchain. Some have their independent blockchain, while others are issued on top of existing ones. Without delving into the technical details, one can think of blockchains as public ledgers keeping track of a set of accounts or *wallets*. Blockchains are generally public and anyone with some technical know-how can survey all transactions made between the cryptoasset wallets. Having this transaction history allows for analyzing changes in transaction patterns, changes in which wallets are interacting and a plethora of other insights. While not exactly social media data, blockchains do contain information about human intentions and actions. For example, a surge in new wallets could indicate an uptick in adaptation rate, while increases in transaction size might suggest that institutional investors are entering the market.

### 3.6.1 Data Collection

Extracting data from all the blockchains related to assets is a monumental task. At the time of writing (i.e., May 2021), both Bitcoin and Ethereum are above 300GB in size. While the other blockchains are mostly smaller, they are in sum too large to handle without specially dedicated hardware. Consequently, we rely on third-party actors who have performed blockchain analyses and exposed their results publicly through APIs. We have used IntoTheBlock's analyses, elicited through the free version of CryptoCompare's API (CryptoCompare, 2021). Roughly a third of the selected assets have readily available blockchain data. For some of the remaining assets the data is unavailable either because the analysis remains unpublished or simply has not been performed. For others, the data is unavailable due to the protocol followed by the blockchain. An example of this is Monero (XMR), which is designed to obfuscate the transaction history, making any attempt at useful analysis essentially impossible.

### 3.6.2 Variable Construction

Appendix D contains a complete list of data available for the blockchains with accompanying variable descriptions. We deemed *transaction count*, *large transaction count* and *new addresses* to be promising independent variables. The idea being that the change in these on-chain features could correlate with future price movements in the same way as changes in social media activity seem to. Having the variables already extracted from the blockchains makes any large-scale feature engineering redundant. The variables can be used as is, with the addition of data cleaning and fitting variable transformations.

Like many of the other variables, the blockchain features enter into the model as the daily value relative to its rolling seven-day average given by Equation 3.10, Equation 3.11 and Equation 3.12. The number of transactions and new addresses are also included as weekly averages, given by Equation 3.13 and Equation 3.14

$$RelativeTransactionCount_t = \frac{TransactionCount_t}{\frac{1}{7}\sum_{\tau=t-7}^{t} TransactionCount_\tau} - 1 \tag{3.10}$$

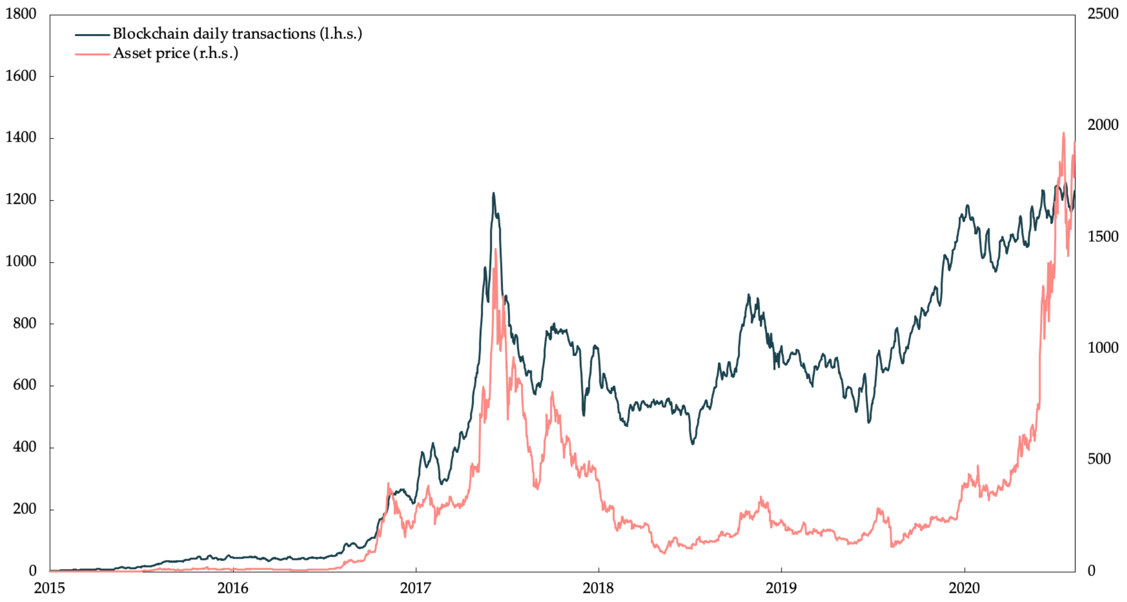

**Figure 3.8:** Seven-day moving average of transaction count on Ethereum in thousands and ether (ETH) asset price in USD.

$$RelativeLargeTransactionCount_t = \frac{LargeTransactionCount_t}{\frac{1}{7}\sum_{\tau=t-7}^{t} LargeTransactionCount_\tau} - 1 \tag{3.11}$$

$$RelativeNewAddresses_t = \frac{NewAddresses_t}{\frac{1}{7}\sum_{\tau=t-7}^{t} NewAddresses_\tau} - 1 \tag{3.12}$$

$$WeeklyTransactionCount_t = \sum_{\tau=t-7}^{t} TransactionCount_\tau \tag{3.13}$$

$$WeeklyNewAddresses_t = \sum_{\tau=t-7}^{t} NewAddresses_\tau \tag{3.14}$$

Upon inspection these variables seem to co-vary with asset price developments. Figure 3.8 shows the rolling average of daily transactions on Ethereum plotted against the ether asset price. During the uptick in prices in 2017 the connection is especially evident.

## 3.7 Data Treatment and Variable Scaling

From the processes described in this chapter we end up with dataset reaching as far back as 2015 for some assets. In forecasting, having larger datasets is mostly associated with generating more robust results. However, if the underlying relationships in the data change throughout the dataset, its comprehensiveness might be an impediment. Cryptomarkets have likely undergone multiple structural shifts in the period leading back to 2015. Rudimentary linear regression analysis suggested that calibrating our model on data from January 2019 to February 2020 could be appropriate.

During processing, particular attention was paid to ensure that no information from the future leaked into the past. For example, while we linearly interpolate missing values in the training set, only forward-filled values are used in the testing period. Calculated averages are always backward-looking, and measures like volatility are only calculated based on information available when returns are predicted. Deseasoning is always performed based on patterns in the training data, never in the full dataset.

We scale all variables to have a minimum of 0 and maximum of 1 in the training period. This serves a dual purpose. Firstly, data from different assets is normalized and can form a uniform joint dataset. Secondly, the LSTM-model used is sensitive to magnitudes of the variables. We note that this scaling is sensitive to outliers, but observe that the results seem unaffected.

Finally, the datasets for the individual assets fused to produce a single panel dataset. The combined dataset is used to calibrate a single model that estimates one relationship between the regressors and output variables for all assets.

# Chapter 4

# Methodology

The following chapter provides a short introduction to our sentiment analysis tool, VADER, as well as linear regression and LSTM-networks. Finally, we present the trading strategies in addition to the evaluation metrics and benchmarking methods used to validate our prediction results.

## 4.1 Sentiment Analysis

Much of online data is in the form of unstructured text. The millions of published news articles, social media posts and emails convey a myriad of beliefs and opinions. Recognizing the potential insight that analyzing such content could yield have in part lead to the development of the intersectional academic field of *natural language processing* (NLP). NLP encompasses a set of methods used for computational analysis of textual data (Cambria & White, 2014). The natural language processing tool leveraged in this thesis is commonly referred to as *sentiment analysis*. Sentiment analysis is the act of extracting and measuring the subjective emotions or opinions expressed in text.

We utilize the software package VADER from the NLTK library (Bird et al., 2009). The method performs lookup in a reference lexicon to label words and phrases with their associated sentiments (Taboada et al., 2011). VADER has proven to be a reliable estimator of Twitter sentiment (Park & Seo, 2018).

**Table 4.1:** Examples of Twitter posts and associated VADER-scores.

| **Tweet** | **Score** |
|---|---|
| ”i ignored the chat and kept watching the progress update future is bright keep it growing this year will be cardano year” | 0.62 |
| ”when youve finally broke even on that shitty altcoin you bought into at the top” | -0.68 |
| ”growing strong 💪🏼” | 0.61 |
| ”right place right time 😍” | 0 |
| ”crypto nerd be like this is the digital currency of the future then lose on the trade” | -0.34 |

VADER assigns words and phrases a decimal number in the range $[-1, 1]$, where higher scores are associated with posts being more positive. VADER can correctly categorize complex syntactic constructs like "not good" as negative and that exclamation marks intensify the expressed sentiment. The method is also capable of processing various slang words and emojis. Table 4.1 shows some examples of tweets and their associated sentiment value.

## 4.2 Linear Regression

Our simplest prediction model is a pooled linear regression model. Since we are mainly looking at the impact of adding new variables we eschew using more than a single lag in our model formulation. The response variable is the next-day predicted asset return, while the independent variables are incorporated as described in Equation 4.1. The coefficients are estimated in R using OLS.

$$\begin{aligned}
Return_{i,t+1} = {} & \beta_{Const} + \beta_R \cdot Return_{i,t} + \beta_{RLTV} \cdot RelativeLnTradingVolume_{i,t} + \\
& \beta_{WV} \cdot WeeklyVolatility_{i,t} + \beta_{MV} \cdot MonthlyVolatility_{i,t} + \\
& \beta_{GTWV} \cdot GoogleTrendsWeeklyVolume_{i,t} + \beta_{RGTV} \cdot RelativeGoogleTrendsVolume_{i,t} + \\
& \beta_{TWV} \cdot TwitterWeeklyVolume_{i,t} + \beta_{TDSV} \cdot TwitterDailySentimentValue_{i,t} + \\
& \beta_{RTV} \cdot RelativeTwitterVolume_{i,t} + \beta_{RWV} \cdot RedditWeeklyVolume_{i,t} + \\
& \beta_{RDSV} \cdot RedditDailySentimentValue_{i,t} + \beta_{RRV} \cdot RelativeRedditVolume_{i,t} + \epsilon_{i,t}
\end{aligned} \tag{4.1}$$

## 4.3 Recurrent Neural Network

Artificial neural networks have in recent years become one of the most popular model types for academic research. With the flexibility of adding different layer variants and activation functions, models can range from relatively simple to deeply complex. Exemplified, an ANN with a single layer and a linear activation function is equivalent to a linear regression model, while an ANN with a combination of different layers and activation functions is theoretically capable of capturing complex non-linear relationships between the input and output variables.

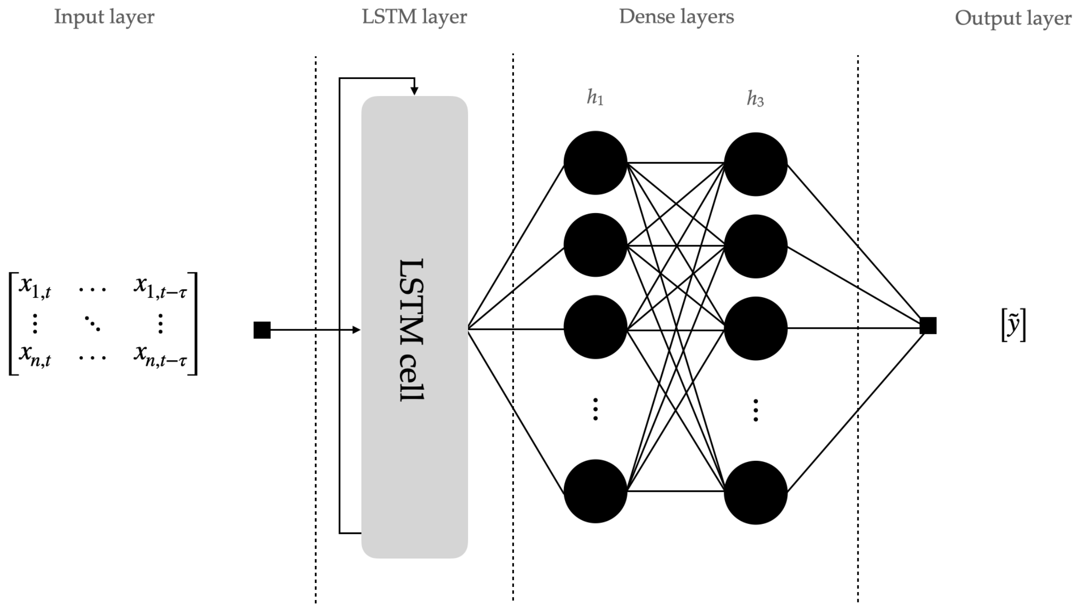
Input layer
LSTM layer
Dense layers
Output layer
$h_1$
$h_3$
$\begin{bmatrix} x_{1,t} & \cdots & x_{1,t-\tau} \\ \vdots & \ddots & \vdots \\ x_{n,t} & \cdots & x_{n,t-\tau} \end{bmatrix}$
LSTM cell
$[\tilde{y}]$

**Figure 4.1:** Schematic design of an LSTM-network with $n$ inputs, an LSTM-layer, two dense layers and one output. The self arrow in the LSTM-layer illustrates where the output is used at time $t - \tau$. At time $t$ the output is propagated through the dense layers.

Neural networks in the traditional sense do not explicitly model a time dimension. Consequently, such models do not always perform well with time series or panel data. Recurrent neural networks were conceived to transcend this limitation. In RNNs output values may depend upon relationships that are apparent only along the time dimension of the data. In practice, however, capturing such relationships has proven to be a difficult task mainly due to an issue called the Vanishing Gradient Problem (Bengio et al., 1994). Long Short-Term Memory networks were developed to address this problem (Hochreiter & Schmidhuber, 1997). With the use of more complex gated cells than vanilla RNNs, learning can more effectively take place across longer sequences of data. This feature has made LSTM-networks a preferred model for handling sequential data.

Figure 4.1 shows a neural network containing a combination of recurrent and dense layers, as our model does. Compared to the traditional neural networks containing only dense layers it differs by taking a sequence of data, represented by a matrix, as input, rather than a single vector. Each matrix row represents a single variable at different points in time. The LSTM-layer iteratively ingests the data for each timestep, combining it with encoded data from previous timesteps. The final encoding of all timesteps is propagated through the succeeding dense layers.

### 4.3.1 Model Specifications

We use *PyTorch* (Paszke et al., 2019), a Python machine learning library, to construct the neural network. The specification of the network is shown in Table 4.2. The hidden layers are constituted of a single LSTM-layer and two dense layers. The hidden state in the LSTM-layer is the internal encoding of data from previous time steps. The sequence length is the number of time steps used as input for the model. As we have daily data, a sequence length of seven means that the past week's data is used to predict the next-day returns. We apply no activation function between the LSTM-layer and the first dense layer. Between the two dense layers we apply the ReLU activation function (see Equation 4.2). No activation function is applied after the second and final dense layer. Note that the LSTM-layer does employ both the sigmoid and hyperbolic tangent activation function in its internal processes by design. We use the Adam (Adaptive Moment Estimation) optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.001 to tune the weights of the network. The number of epochs is the amount of times the training data is passed through the network.

$$f(x) = max(0, x) \tag{4.2}$$

We landed on these model specifications by performing a grid search in the hyper-parameter space using the last three months of training data as a validation set. The grid search revealed that more complex models (i.e. additional or wider layers) were less accurate measured by the metrics in subsection 4.4. Increases in the number of epochs led to overfitting of the model.

**Table 4.2:** LSTM-model parameter specification.

| **Model parameter** | **Value** |
|---|---|
| Number of layers | 3 |
| Size of hidden state in LSTM-layer | 20 |
| Number of neurons in dense layer | (20, 32) |
| Sequence length | 7 |
| Activation function | ReLU |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Number of epochs | 50 |

## 4.4 Model Evaluation

Several metrics are used to compare and evaluate the prediction models. *Root mean square error* (RMSE) captures the goodness of a fit and shows the average error in predicted returns. Exemplified, an RMSE-value of 0.02 implies that the model's predictions on average are off by 2%. RMSE is defined as,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(PredictedReturn_i - ActualReturn_i)^2}, \tag{4.3}$$

where $N$ denotes the size of the test set.

We are also interested in whether the models can correctly guess the sign of the future return. Based on the relationship between the predicted and real return each prediction can be classified as either a true or false positive or a true or false negative (see Table 4.3).

**Table 4.3:** Confusion matrix for classifying predictions.

| | | **Predicted return** | |
|---|---|---|---|
| | | **Positive:** | **Negative:** |
| **Actual return** | **Positive:** | *True Positive (TP)* | *False Negative (FN)* |
| | **Negative** | *False Positive (FP)* | *True Negative (TN)* |

A high *accuracy*, defined in Equation 4.4, is associated with being able to correctly identify the sign of the next-day return.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

*Recall*, defined in Equation 4.5, is a measure of how good the model is at identifying positive predictions. The *true negative rate*, defined in Equation 4.6, is the proportion of negative returns correctly predicted by the model.

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

$$TrueNegativeRate = \frac{TN}{TN + FP} \tag{4.6}$$

Model *precision*, defined in Equation 4.7, tells us how likely it is for a positive prediction to be true, while the *negative predictive value*, defined in Equation 4.8, is the ratio between the total number of correctly predicted negative returns and the total number of times a negative prediction was made.

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

$$NegativePredictiveValue = \frac{TN}{TN + FN} \tag{4.8}$$

When presenting these metrics in the results section, they are always calculated without any transformation to the dependent variable, to preserve the interpretability of the metrics.

## 4.5 Trading Strategies

We test our prediction models by instituting a simple trading strategy outlined in Table 4.4. The trader obtains a set of return estimates for all the coins in our selection. The $n$ coins with the highest return estimates are selected by the trader. Coins are sold when they are no longer among the top estimates for the following day. We rule out shorting, and all coins are weighted equally in the portfolio. We assume that our investments are sufficiently small so that they alone do not influence market prices. When comparing models, transaction costs are disregarded as they would be roughly equal across models.

**Table 4.4:** Simple trading strategy used for model comparison.

| | **Simple Trading Strategy** |
|---|---|
| Buy | Buy the $n$-assets with the largest predicted returns from the model. |
| Sell | All assets not included in the next day's picks are sold at the end of the day. |
| Shorting | No shorting. |
| Weighting | All assets are equally weighted. |

To realistically test our models we incorporate trading costs and adjust the simple trading strategy described above. We assume a trading fee of 0.1% based on the exchange listings such as Binance (Binance, 2021). This is on the high end of what one could expect as an institutional trader, but is chosen to ensure the robustness of our results. Additionally, we assume an average bid-ask spread of 0.25%. While being subject to variation throughout the test period, the spread for liquid pairs such as ETH/BTC and BTC/USDT is normally close to 0%. When trading in more illiquid assets, the spreads range between 0.3-0.5%. To bring down transaction costs the new strategy only allows for investing in assets with a predicted return of more than 1%. If no assets meet this requirement, a cash-equivalent stablecoin is held until the model again finds a worthwhile investment. The model is additionally barred from holding more than ten assets overall.

We note that a bid-ask spread of 0.25% is not based on historical day by day bid-ask spreads for the test period. Such data has proven very difficult to come by for most trading pairs. Therefore, the 0.25% estimate is based on current spreads, plus an added premium to account for the low market liquidity in the early weeks of the test period.

**Table 4.5:** Realistic trading strategy used to gauge the achievable returns in real world markets.

| | **Realistic Trading Strategy** |
|---|---|
| Buy | Buy the 10 or fewer assets with the largest predicted returns above 1%. |
| Sell | All assets not included in the next day's picks are sold at the end of the day. |
| Shorting | No shorting. |
| Weighting | All assets are equally weighted. |
| Transaction costs | 0.35% per trade. |

## 4.6 Benchmarking

Validating our prediction analysis results constitutes a final challenge. In our particular case, it is interesting to see if our models perform statistically better than performing the same trading strategy but picking coins at random.

Our benchmark is constructed using Monte Carlo simulations. By simulating a daily pick of $n$ random assets 10,000 times, we institute a benchmark for how well our prediction model must do before we confidently can call its excess return non-random.

Figure 4.2 shows the distribution of the portfolio value for a random portfolio of 27 assets. We see that the cumulative returns vary considerably. In particular, portfolios within the 95%-band have between a 350% and 740% overall return at the end of the test period. However, establishing this benchmark help us validate our findings even though the market has been very volatile during the test period.

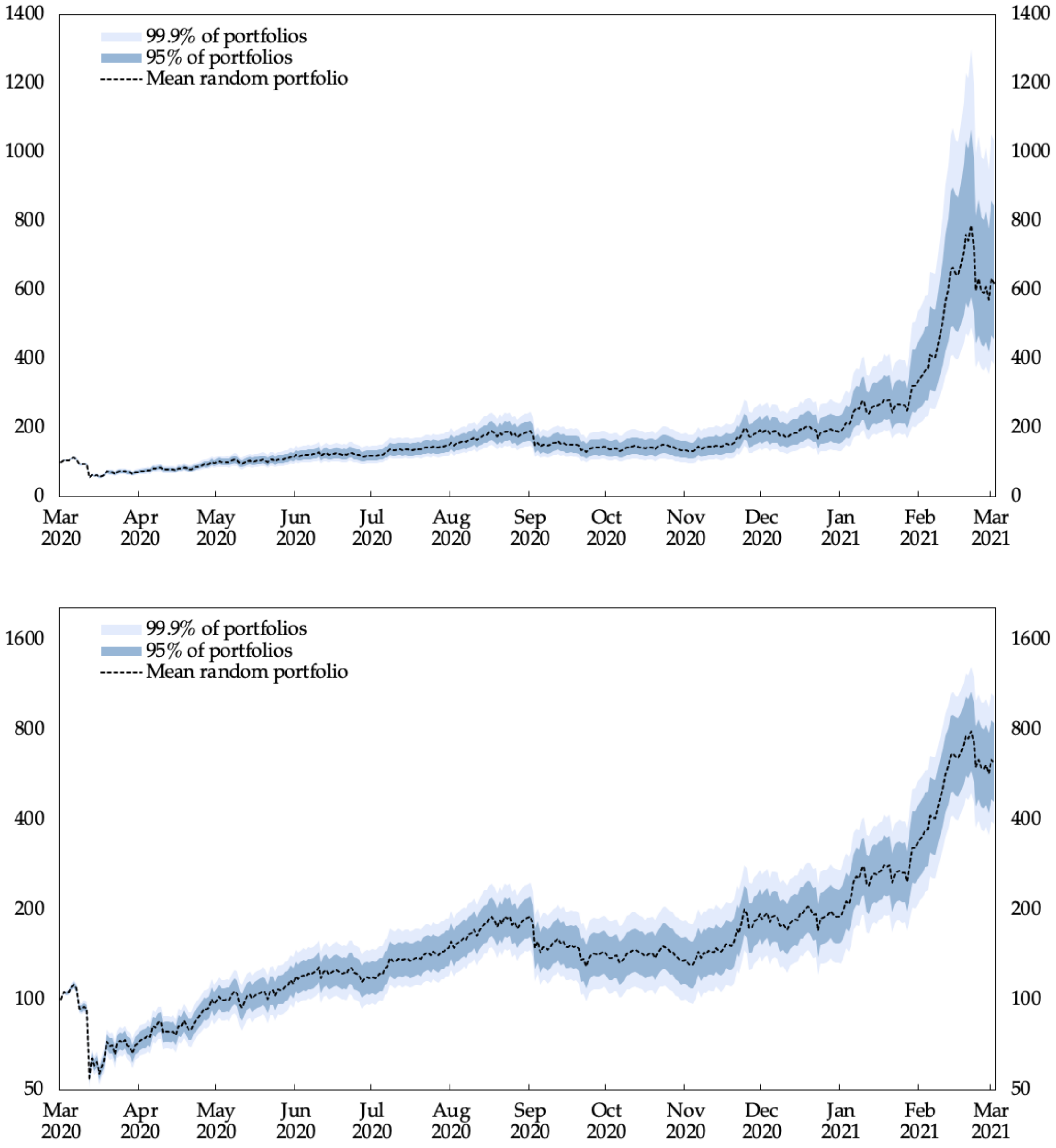

**Figure 4.2:** Indexed portfolio value for portfolios generated at random. Index = 100 on 1 March 2020. 10,000 simulations. Top: Linear scale. Bottom: Logarithmic scale.

## 4.7 Risk-Adjusting Returns

A final metric to shed light on the portfolio performance is the risk-adjusted return. Introductory portfolio theory tells us that increased returns normally come at the cost of increases in risk. Adjusting for risk is therefore crucial when comparing the trajectory of two portfolios. For this we use the Sharpe and Sortino ratios as given in Equation 4.9 and Equation 4.10. The Sortino ratio is a variation of the well-known Sharpe ratio, in which downside risk is isolated. In both equations $N$ is the number of trading days, $\bar{r}_p$ is the mean daily return of the portfolio, $r_f$ is the risk-free rate. In Equation 4.9 $\sigma_p$ is the daily volatility of the portfolio while $\sigma_{d,p}$ in Equation 4.10 is the daily downside volatility. The risk-free rate is assumed to 0.1% based on the US 3-month Treasury Bill yields during the test period. The daily effect is thus close to 0% and accordingly neglected in our calculations. Having 366 trading days in our test set, the presented $Sharpe_p$ and $Sortino_p$ ratios are annualized.

$$Sharpe_p = \sqrt{N}\Big(\frac{\bar{r}_p - r_f}{\sigma_p}\Big) \tag{4.9}$$

$$Sortino_p = \sqrt{N}\Big(\frac{\bar{r}_p - r_f}{\sigma_{d,p}}\Big) \tag{4.10}$$

# Chapter 5

# Results

Below, we present and discuss the results of our retrodiction of returns and backtesting exercise. Firstly, we introduce a simple linear regression model and use it to perform some elementary technical analysis. Secondly, we assess how model performance depends on the portfolio size and how it is influenced by extending the set of input variables. Figure 5.1 serves as a visual aid by showing in what order each subset of features is included in the model. The model numbers are used for reference throughout this chapter. Afterward, the performance of our linear model is compared to that of an LSTM neural network model. Lastly, we analyze the impact of trading costs on our model performance and show that a common-sense trading strategy can be used to curb the cumulative effect of such fees.

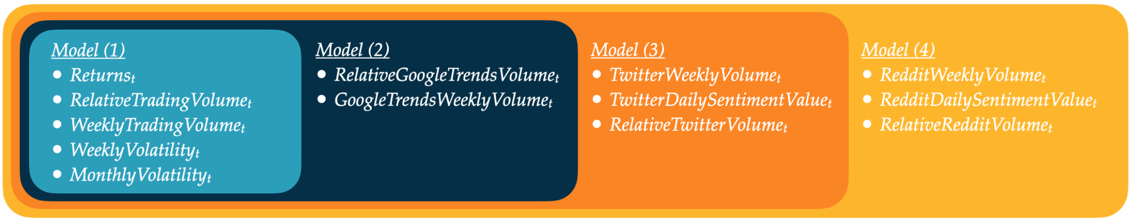
Model (1)
• Returns$_t$
• RelativeTradingVolume$_t$
• WeeklyTradingVolume$_t$
• WeeklyVolatility$_t$
• MonthlyVolatility$_t$
Model (2)
• RelativeGoogleTrendsVolume$_t$
• GoogleTrendsWeeklyVolume$_t$
Model (3)
• TwitterWeeklyVolume$_t$
• TwitterDailySentimentValue$_t$
• RelativeTwitterVolume$_t$
Model (4)
• RedditWeeklyVolume$_t$
• RedditDailySentimentValue$_t$
• RelativeRedditVolume$_t$

**Figure 5.1:** Stages of inclusion of variable subsets in the models.

## 5.1 Regression and Initial Technical Analysis

We specify a pooled linear regression equation according to Equation 5.1 using a combined dataset for all assets. All variables are standardized directly prior to the regression in order to increase interpretability. Table 5.1 contains the results of the regression. We use White standard errors as a Breusch-Pagan test establishes that the data significantly exhibits heteroskedasticity.

$$\begin{aligned} Return_{i,t+1} = \beta_{Const} + \beta_R \cdot Return_{i,t} + \beta_{RLTV} \cdot RelativeLnTradingVolume_{i,t} \\ +\beta_{WV} \cdot WeeklyVolatility_{i,t} + \beta_{MV} \cdot MonthlyVolatility_{i,t} + \epsilon_{i,t} \end{aligned} \tag{5.1}$$

One has to be careful in interpreting the signs and values of the coefficients, considering the variable

**Table 5.1:** Linear regression coefficients for model (1).

| | *Dependent variable: $Return_{i,t+1}$* |
|---|---|
| | Model (1) |
| $Return_{i,t}$ | $-0.050^{***}$ |
| | (0.008) |
| $RelativeLnTradingVolume_{i,t}$ | $0.021^{***}$ |
| | (0.007) |
| $WeeklyVolatility_{i,t}$ | $0.029^{***}$ |
| | (0.009) |
| $MonthlyVolatility_{i,t}$ | $-0.031^{***}$ |
| | (0.008) |
| Constant | 0.000 |
| | (0.007) |
| Observations | 22,947 |
| $R^2$ | 0.003 |
| Adjusted $R^2$ | 0.003 |
| Residual Std. Error | 0.052 (df = 22,942) |
| F Statistic | $18.576^{***}$ (df = 4; 22,942) |

$^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

transformations and scaling. Nevertheless, some general assertions can be made. All independent variables are significant at the 1% level. Especially concerning the trading volume, there was uncertainty beforehand as to how valid the data was. Wash trading is a known problem on many cryptoasset exchanges (Cong et al., 2020), and the Coingecko API is unclear on how, or even if, this is addressed. With the coefficient on $RelativeLnTradingVolume_{i,t}$ being significant, it seems that the data is of sufficient quality.

Returns today are negatively associated with returns tomorrow. This implies that the price on average exhibits a reversion pattern, given that the returns have an approximate zero mean. Increases in trading volume relative to that of the past week pull up the next-day return estimate. The volatility measures should be interpreted together. With similar magnitudes but opposite signs, their relationship is interesting. When weekly volatility is higher than monthly volatility, the overall contribution is positive, and vice versa. A theory to explain this phenomenon is that high short-term volatility compared to long-term volatility acts as a proxy for "hype" around an asset. On the other hand, higher long-term volatility than short-term volatility indicates decreasing "hype" and has a negative price effect. Notably, the adjusted $R^2$ value for the model is only 0.3%. The model as a whole thus explains very little of the total variation in next-day returns.

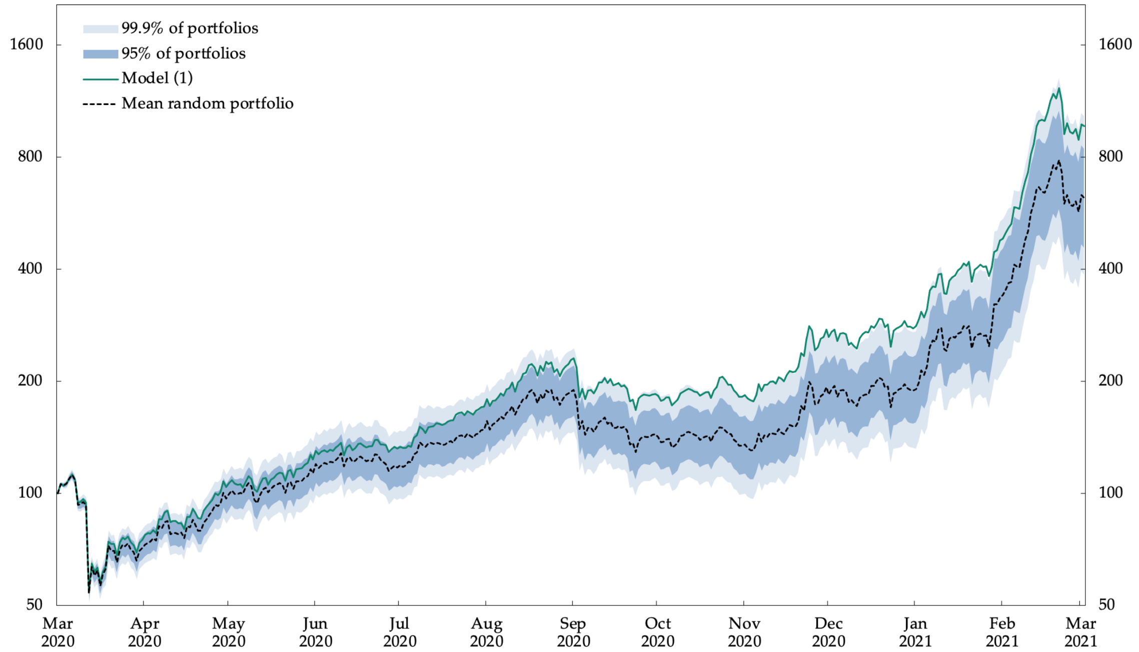

**Figure 5.2:** Indexed portfolio value when retrodicting returns using simple linear regression. Index = 100 on 1 March 2020. 27 assets. Logarithmic scale.

Figure 5.2 shows the performance of model (1) when combined with the simple trading strategy described in section 4.5 while holding a portfolio of 27 assets. The benchmark portfolio consists of 27 equally weighted cryptoassets selected at random. The model outperforms 95% of the randomly generated portfolios for the majority of the period. In fact, our model almost performs as well as the top 0.1% of the random portfolios from September 2020 and onward. That such a simple model using information that is readily available for people with some technical know-how can deliver excess returns hints at market inefficiency. However, transaction costs remain unaccounted for. As we will elaborate on in section 5.6, the cumulative effect of such costs can quickly cancel out excess returns when using a daily trading strategy.

We use the model from this section as a point of departure for the results presented in the following two sections. There we investigate the impact of decreasing the number of assets held in the portfolio and extending our model with more input features.

## 5.2 Returns and Portfolio size

In the preceding section, our trading strategy entailed investing in half of all available assets each day. The results we obtained seemed to indicate that the model could outperform the market. However, going long on half the market is probably not the profit-maximizing strategy. In fact, on a number of days, several of the top 27 predictions are negative. This fact motivates experimenting with the portfolio size to see how cumulative returns are impacted. If the model can correctly identify positive future returns, capital should be dedicated to the most promising investment opportunities. If limiting the portfolio size consistently increases returns, it is a tell-tale sign that our model indeed is capable of picking winning assets. Figure 5.3 shows the portfolio value achieved by the model presented in Table 5.1, when selecting 5, 10, 15, 20 and 27 assets.

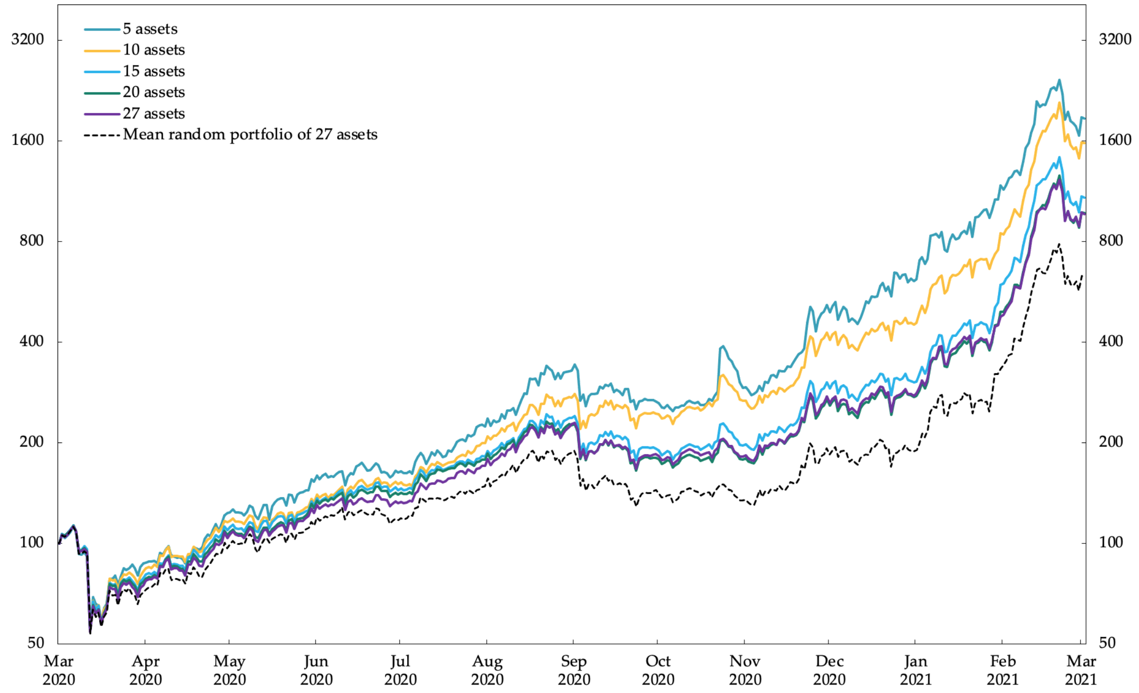

**Figure 5.3:** Indexed portfolio value when selecting 5, 10, 15, 20 and 27 assets using model (1). Index = 100 on 1 March 2020. Logarithmic scale.

Inspecting the chart closely, portfolios of sizes 20 and 27 perform almost identically, while the 15-asset portfolio barely eclipses them. It is likely that on any given day, only a small subset of assets have predictions that strongly indicate positive returns. Evidently, the overall return is a decreasing function in the portfolio size. This supports the theory that our model is capable of finding the most profitable assets in the selection.

Investing in fewer assets also increases portfolio volatility. Accordingly, the width of the confidence bands generally increases when the size of the portfolio decreases. However, all of our portfolios attain excess returns above the range where they could plausibly be random. Several of the smaller portfolios make huge gains towards the end of the period. As cumulative returns are not memory-less, single lucky picks can significantly impact the overall returns. However, this does not invalidate our findings as such since the jumps happen towards the end of the period, while the excess returns are convincingly non-random throughout.

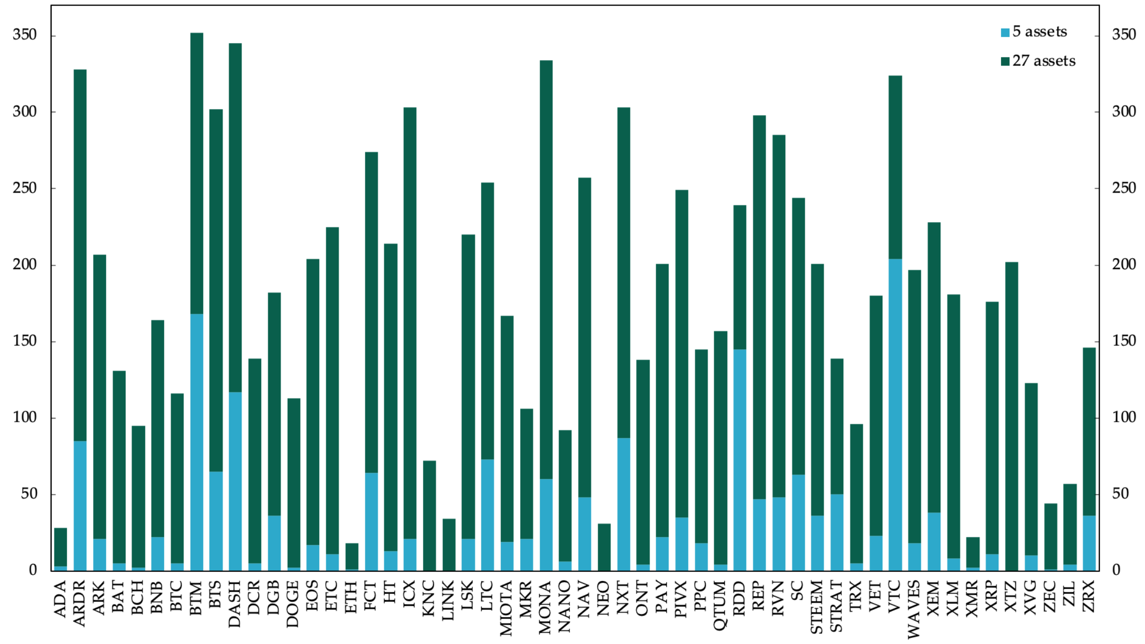

**Figure 5.4:** Frequency distribution of selected assets when model (1) selects portfolios of 5 and 27 assets.

Figure 5.4 shows the number of days each asset was selected into portfolios of sizes 5 and 27. It would be problematic if the models predominantly chose from a small subset of the available assets, which just happened to perform well. However, the chart shows that the model picks from the entire set. Interestingly, some of the larger assets, measured by market capitalization (e.g., ADA and ETH), are infrequently selected. Assets that are larger in terms of market capitalization probably experience more modest relative changes in variables like trading volume. In turn, this generates less extreme return predictions, which rarely make these assets one of the most promising investments in the view of our model. While we do not explore this peculiarity any further, it could potentially be a weakness of our model.

## 5.3 Exploring the Impact of Search Volume and Social Media Data

This thesis partly asks if social media and search volume data improves return predictability. To answer this question, we iteratively add features elicited from Google Trends, Twitter and Reddit to the model. As Table 5.2 shows, many of these variables seemingly have significant explanatory power for next-day returns. Note, however, that $R^2$ remains low.

Pivoting from model (1) by including the most readily available social data, namely the data from Google Trends, yields model (2). Introducing these variables has mixed effects. The weekly search volume is strongly significant across all models, albeit somewhat less so in model (4). $RelativeGoogleTrendsVolume_{i,t}$ is statistically insignificant across all models configurations. As to why the relative volume is insignificant, there are several possible explanations. It could be that relative trading volume captures much of the same information as search volumes. Another possibility is that the daily Google Trends data itself is too inaccurate to act as a regressor. When

**Table 5.2:** Regression coefficients for iteratively more elaborate linear regression models. The first column corresponds to Table 5.1.

| | Dependent variable: $Return_{i,t+1}$ | | | |
|---|---|---|---|---|
| | Model (1) | Model (2) | Model (3) | Model (4) |
| $Return_{i,t}$ | $-0.050^{***}$<br>(0.008) | $-0.052^{***}$<br>(0.008) | $-0.058^{***}$<br>(0.008) | $-0.059^{***}$<br>(0.008) |
| $RelativeLnTradingVolume_{i,t}$ | $0.021^{***}$<br>(0.007) | $0.021^{***}$<br>(0.007) | $0.021^{***}$<br>(0.007) | $0.022^{***}$<br>(0.007) |
| $WeeklyVolatility_{i,t}$ | $0.029^{***}$<br>(0.009) | $0.032^{***}$<br>(0.009) | $0.036^{***}$<br>(0.009) | $0.035^{***}$<br>(0.009) |
| $MonthlyVolatility_{i,t}$ | $-0.031^{***}$<br>(0.008) | $-0.030^{***}$<br>(0.008) | $-0.026^{***}$<br>(0.008) | $-0.024^{***}$<br>(0.008) |
| $RelativeGoogleTrendsVolume_{i,t}$ | | 0.011<br>(0.007) | 0.009<br>(0.007) | 0.009<br>(0.007) |
| $GoogleTrendsWeeklyVolume_{i,t}$ | | $-0.021^{***}$<br>(0.007) | $-0.021^{***}$<br>(0.007) | $-0.018^{**}$<br>(0.007) |
| $TwitterWeeklyVolume_{i,t}$ | | | $0.019^{***}$<br>(0.008) | $0.017^{**}$<br>(0.009) |
| $TwitterDailySentimentValue_{i,t}$ | | | $0.029^{***}$<br>(0.007) | $0.030^{***}$<br>(0.007) |
| $RelativeTwitterVolume_{i,t}$ | | | $-0.019^{**}$<br>(0.008) | $-0.017^{**}$<br>(0.008) |
| $RedditWeeklyVolume_{i,t}$ | | | | $0.012^{*}$<br>(0.007) |
| $RedditDailySentimentValue_{i,t}$ | | | | 0.008<br>(0.007) |
| $RelativeRedditVolume_{i,t}$ | | | | $-0.013^{*}$<br>(0.007) |
| Constant | 0.000<br>(0.007) | 0.000<br>(0.007) | 0.000<br>(0.007) | 0.000<br>(0.007) |
| Observations | 22,947 | 22,947 | 22,947 | 22,947 |
| $R^2$ | 0.003 | 0.004 | 0.005 | 0.005 |
| Adjusted $R^2$ | 0.003 | 0.003 | 0.005 | 0.005 |
| Residual Std. Error | 0.052 (df = 22,942) | 0.052 (df = 22,940) | 0.052 (df = 22,937) | 0.052 (df = 22,934) |
| F Statistic | $18.576^{***}$ (df = 4; 22,942) | $14.200^{***}$ (df = 6; 22,940) | $12.986^{***}$ (df = 9; 22,937) | $10.403^{***}$ (df = 12; 22,934) |

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

fetching the search traffic data, there were several incidents where performing identical queries to their API returned different results. These inconsistencies seemed to be most prevalent on a day-to-day basis and would average out for a weekly value. Consequently, this could explain the differing significance between the two Google Trends variables. A final possibility is that the effect, if any, is reflected in the price with little delay. If so, the effect on next-day returns will of course be non-existent.

Inspecting Figure 5.5, we see that model (2) does not improve upon model (1) when picking the five assets with the highest predictions each day. In the first half of the test period, model (2) performs poorer than model (1). However, the model makes a comeback in the second half, yielding a similar overall result. This result suggests that the Google Trends variables in isolation add little value to the model. While we do find that the weekly average significantly explains returns, it would not have increased cumulative returns alone in combination with our trading strategy in the test period.

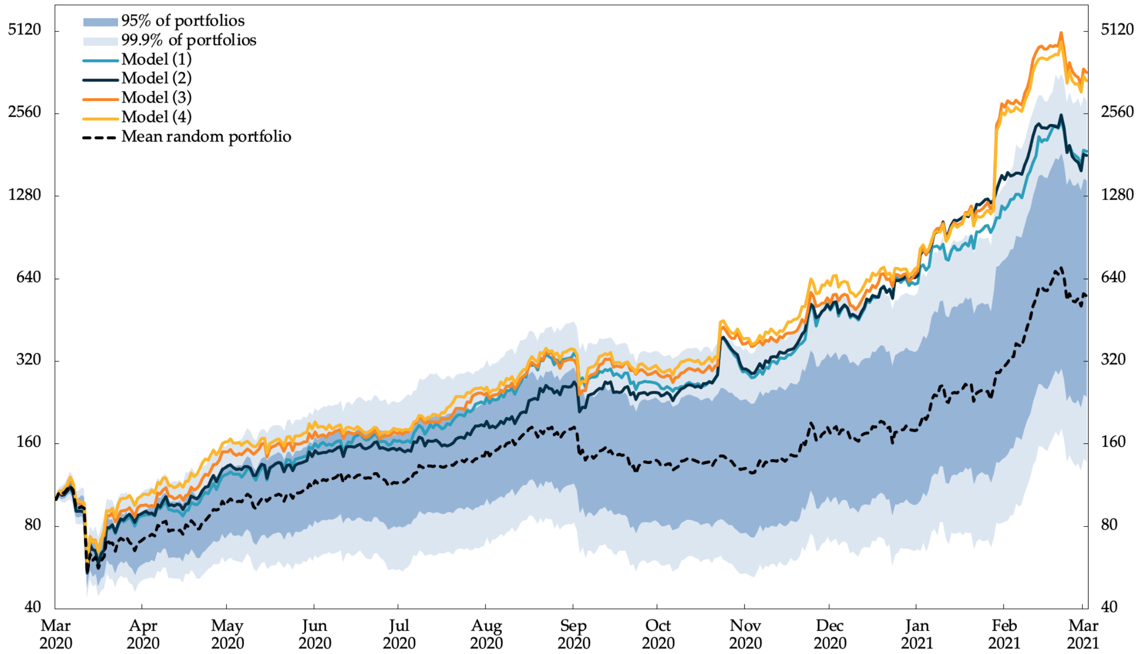

**Figure 5.5:** Indexed portfolio value when retrodicting returns using model (1)-(4). Index = 100 on 1 March 2020. Five assets. Logarithmic scale.

Expanding model (2) with the Twitter variables yields model (3). From Table 5.2 we see that the weekly volume of posts, as well as the sentiment value, are significant at a 1% level. The relative message volume, however only has a $p$-value of less than 5%. These variables are all also significant in model (4). The coefficients indicate that trailing weeks with abnormally high activity and days with positive sentiment are associated with subsequent positive returns. High relative volume, on the other hand, has a negative impact on estimated returns. Since the variables now roughly have the standard normal distribution, the coefficients now approximately measure the variables' feature importance. The coefficients thus imply that the sentiment value has a particularly high impact on the estimated return.

The curves in Figure 5.5 indicate that adding Twitter data contributes positively to cumulative returns. Bar some weeks at the end of 2020, model (3) generally outperforms model (1) and (2). While the final result seems convincing, part of the gains is attributable to spectacular returns in January 2021. Such leaps obfuscate the models' relative performance. All in all, however, the evidence seems to point to the fact that the Twitter features help the model make better predictions. In contrast with some of the other works cited, we find that both volume and sentiment significantly help predict future returns.

Including the variables from Reddit leads us to model (4). Table 5.2 shows that $RedditWeeklyVolume_{i,t}$ and $RelativeRedditVolume_{i,t}$ only have an associated $p$-value of less than 0.1, and that the sentiment indicator is far from being significant. This result is not entirely surprising as these variables likely capture much of the same information as the Twitter features. As Twitter is the richer data source, hoping that Reddit features would substantially improve the results is, in retrospect, perhaps a longshot. Nevertheless, to test whether the Reddit and Twitter features captured the same information would be an interesting exercise. Being vastly different types of social media, it can be that they capture information from different parts of the public, with a dissimilar impact on cryptomarkets.

Even though the coefficients on the Reddit variables are relatively insignificant, their inclusion does not stymie model performance. Figure 5.5 suggests that their inclusion gives model (4) a slight edge over model (3) throughout the test period. Only towards the end does model (3) surpass model (4) before narrowly clinching the win.

**Table 5.3:** Model evaluation metrics for model (1)-(4).

| Model | **Model (1)**<br>$n = 19{,}764$ | **Model (2)**<br>$n = 19{,}764$ | **Model (3)**<br>$n = 19{,}764$ | **Model (4)**<br>$n = 19{,}764$ |
|---|---|---|---|---|
| *RMSE* | 0.0802 | 0.0806 | 0.0820 | 0.0936 |
| *Accuracy* | 0.5399 | 0.5315 | 0.5312 | 0.5334 |
| *Recall* | 0.6437 | 0.5837 | 0.5794 | 0.6059 |
| *True Negative Rate* | 0.4257 | 0.4740 | 0.4781 | 0.4535 |
| *Precision* | 0.5524 | 0.5499 | 0.5500 | 0.5498 |
| *Negative Predicted Value* | 0.5203 | 0.5083 | 0.5079 | 0.5110 |

The results in Table 5.3 complicate the story told by Figure 5.5. The RMSE increases markedly when the Reddit variables are added. Correspondingly, the overall accuracy is slightly worse in models (2), (3) and (4) compared to model (1). With the significance of the Google Trends and Twitter variables demonstrated in Table 5.1 it is somewhat surprising that the RMSE increases in model (2) and (3). One possible explanation is that these variables have more extreme values in the test set, pulling the predictions far to either side. If these extreme values are correlated across regressors, it might go a long way in consolidating these seemingly incongruent results. If this issue only affects single assets, a portfolio of them might be able to sustain its value even though single estimates are very far off the mark. One could, for instance, imagine this being the case with the weekly Twitter volume during the volatile period in early 2021. Nevertheless, it is quite possible for the return to increase alongside an increase in RMSE. In essence, the goal of the model is to identify which cryptoassets seem promising *relative* to others. Under such a scheme, the absolute error is of lesser importance.

Figure 5.6 displays the frequency distribution of the selected cryptoassets for model (1) and (4). Similarly to the distribution shown in Figure 5.4, the models still pick assets from the entire selection.

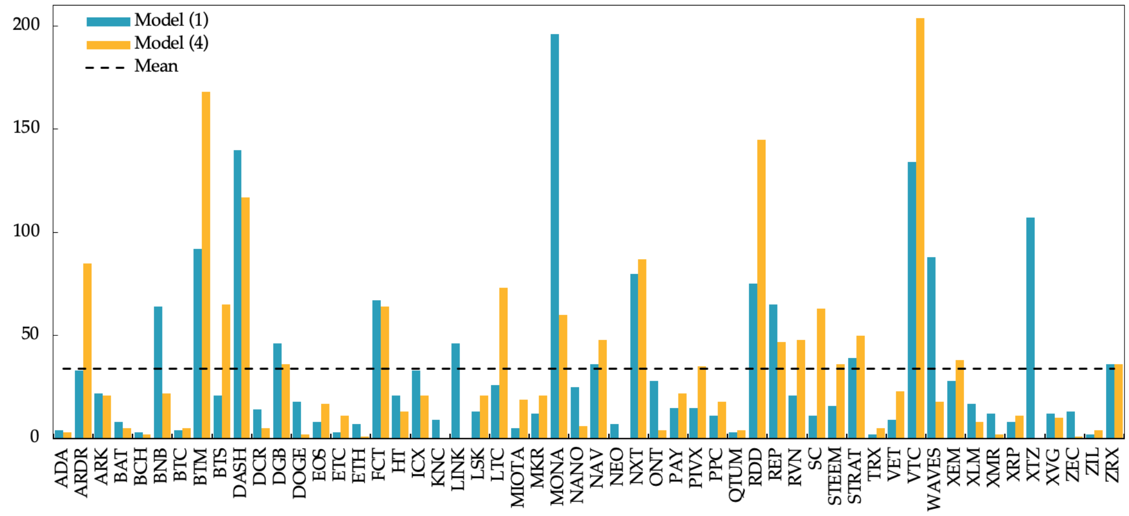

**Figure 5.6:** Distribution of assets selected when returns are retrodicted using the linear regression models (1) and (4). Portfolio size of five cryptoassets.

## 5.4 Exploring the Impact of On-chain Variables

Only 16 of our 54 cryptoassets have easily available data sourced from the underlying blockchain. To explore if this data produces any added value in price prediction, we have devised the following two models. The first uses the variables described in Table 5.2 as model (4). The second model uses these variables with the addition of the five blockchain variables described in subsection 3.6.2. The results, displayed in Figure 5.7, are obtained by letting the model invest in the 8 most promising assets each day.

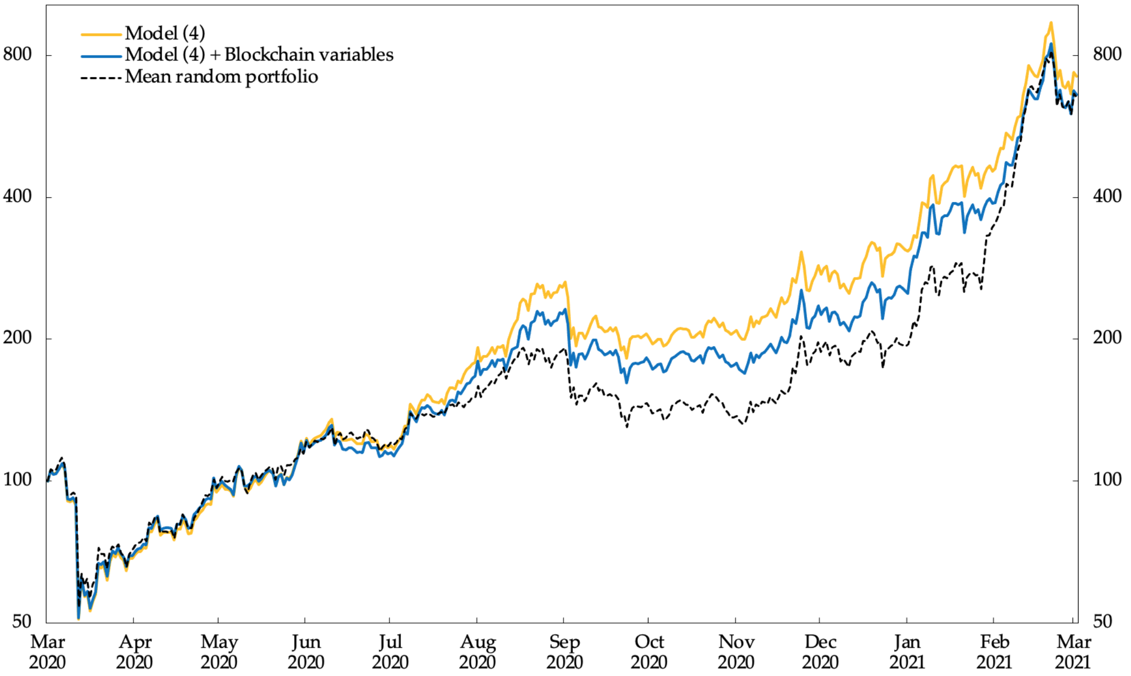

**Figure 5.7:** Indexed portfolio value for model with and without on-chain features. The model invests in eight out of 16 available assets. The sixteen assets are ADA, BAT, BCH, BTC, BTM, DASH, DOGE, ETH, HT, KNC, LINK, LTC, MKR, PAY, REP and ZRX. Index = 100 on 1 March 2020. Logarithmic scale.

As stated in our literature review, several other papers have been able to find that features derived from the underlying blockchain have predictive power. Figure 5.7 implies that we cannot find any similar result in our prediction exercise. Rather, performance declines when the blockchain variables are introduced, suggesting that the variables provide more noise than information. These results do not disprove that the blockchain variables contain valuable information. We merely state that we are unable to find any such value using our models and data. It is plausible that other variables capture the information from the blockchains. Consider the case where increases in the number of new addresses and increases in social media activity are indicators of market adaption. If the latter is a more potent predictor of returns, adding the number of new addresses might increase the level of noise in the model, thus impeding its performance. As shown in Table 5.4 the RMSE does decrease when the on-chain features are included. While this is a positive signal for the model as a whole, the picture is complicated by the fact that the recall and true negative rate move in opposite directions. With these metrics and the portfolio return taken into account, it is not clear that these features provide any added value to the analysis.

**Table 5.4:** Model evaluation metrics for a linear regression model with and without on-chain variables.

| Model | **Model excl. on-chain**<br>$n = 6,800$ | **Model incl. on-chain**<br>$n = 6,800$ |
|---|---|---|
| *RMSE* | 0.0958 | 0.0856 |
| *Accuracy* | 0.5166 | 0.5137 |
| *Recall* | 0.4921 | 0.4472 |
| *True Negative Rate* | 0.5440 | 0.5882 |
| *Precision* | 0.5474 | 0.5490 |
| *Negative Predictive Value* | 0.4886 | 0.4870 |

## 5.5 Linear Model vs. LSTM-network

In section 4.3 we describe the capabilities of RNNs and LSTM-networks specifically. As part of determining the predictability of asset returns and cryptomarket efficiency, we wish to examine whether there are any non-linear relationships between the variables or along the time dimension. If such relationships are evident, a machine learning model like an LSTM-network should deliver superior results to the regression model used so far.

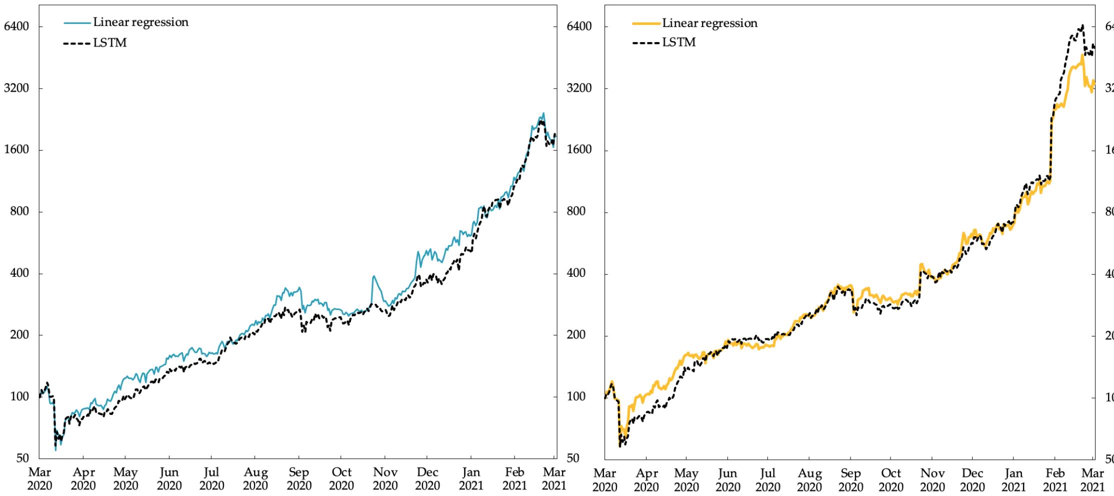

**Figure 5.8:** Indexed portfolio value when retrodicting returns using an LSTM-network and a linear regression model. Using the sets of variables in model (1) (left) and (4) (right) from Table 5.2. Index = 100 on 1 March 2020. Five assets. Logarithmic scale.

**Table 5.5:** Model evaluation metrics for the LSTM-network and linear regression model.

| Model | **Lin. reg. (1)**<br>$n = 19,764$ | **Lin. regr. (4)**<br>$n = 19,764$ | **LSTM (1)**<br>$n = 19,764$ | **LSTM (4)**<br>$n = 19,764$ |
|---|---|---|---|---|
| *RMSE* | 0.0802 | 0.0802 | 0.0805 | 0.0803 |
| *Accuracy* | 0.5399 | 0.5240 | 0.5354 | 0.5323 |
| *Recall* | 0.6437 | 0.5128 | 0.8944 | 0.8049 |
| *True Negative Rate* | 0.4257 | 0.5364 | 0.1400 | 0.2321 |
| *Precision* | 0.5524 | 0.5492 | 0.5339 | 0.5358 |
| *Negative Predictive Value* | 0.5203 | 0.4999 | 0.5462 | 0.5193 |

The results shown in Figure 5.8 and Table 5.5 suggests that no such relationships have been captured. Other than in the last month, the LSTM-network performs on par with or worse than the linear model. This is in line with what was described in Chen et al. (2020), but stands in contrast with several studies cited in section 2.2. If complex variable relationships actually were captured, one would expect to see consistently increased performance throughout the test period. The first quarter of 2021 was an especially volatile period in cryptomarkets. The excess return delivered by the LSTM-model in that period could simply be attributed to a few lucky picks rather than superior predictive prowess. Table 5.5 tells the same story, in that both model types have very similar error scores. Notably, the RMSE the LSTM-network achieved here is similar to that found in McNally et al. (2018). The recall and true negative rates also indicate that the LSTM-model tends to predict positive returns on most days. The linear model, on the other hand, is more balanced in its predictions. Producing a distribution of prediction outcomes that shares characteristics with the true returns seems like an attractive property if, for example, the model is used for individual instead of relative valuation.

Figure 5.8 and Table 5.5 are representative for all combinations of variables presented in this thesis. All prediction exercises have been performed using both an LSTM-model as well as a linear model. None of the results provided clear evidence that the LSTM-network could outperform the linear regression model. By similar reasoning as with the effect of adding blockchain variables, these results do not imply that non-linear relationships between the model variables do not exist. We can only state that our model was unable to find any. It might very well be the case that a larger dataset, different variable transformations, model architectures and hyper-parameters would make such relationships evident.

## 5.6 Transaction Costs and Trading Strategies

So far, our analysis has not accounted for fees and transaction costs. These have a bearing on whether or not markets can be considered efficient and accordingly if the predictability is priced in. The simple trading strategy described in section 4.5 would incur substantial cumulative costs in a real-world market. To realistically gauge the returns achievable by our model, we test the strategy described in Table 4.5. The results are shown in Table 5.6 and Figure 5.9. Contrary to the previous benchmarks used, the benchmark portfolio here is not re-balanced daily to avoid transaction costs.

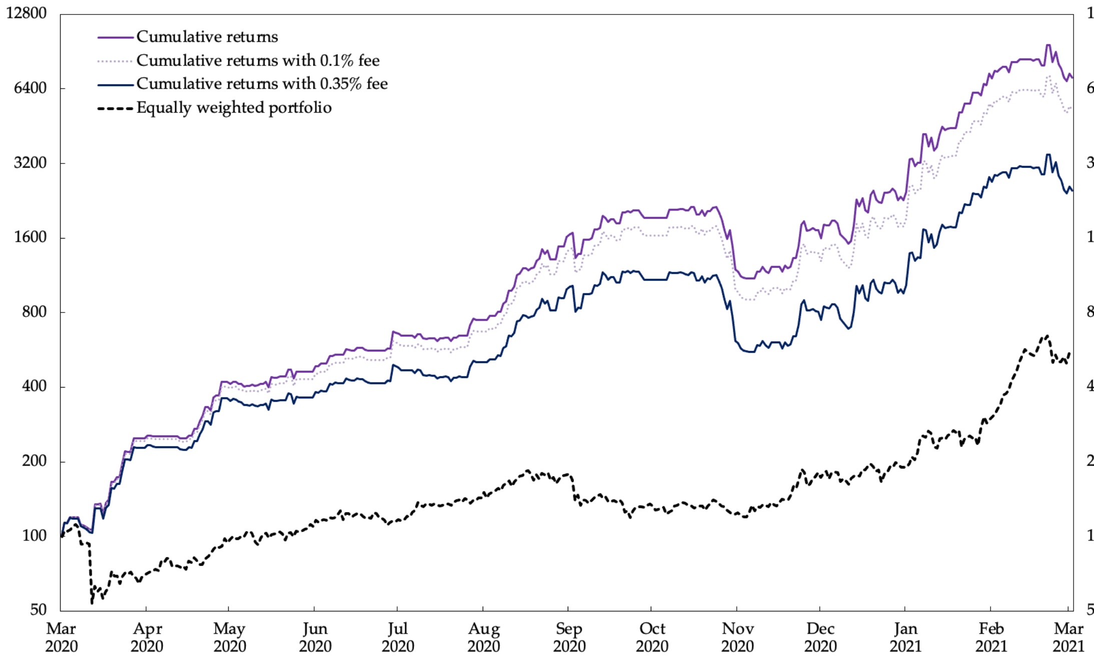

**Figure 5.9:** Indexed portfolio value when incorporating trading fees and using the realistic trading strategy. Index = 100 on 1 March 2020. Benchmarked against a passive, equally-weighted portfolio of 54 cryptoassets. Logarithmic scale.

With the imposed restriction of only investing in assets with a predicted return of more than 1%, the model eschews trading on more than 30% of available trading days. Inspecting Figure 5.9 closely, one can see the aforementioned decreased trading activity. Especially in April and October 2020, the flat line indicates consecutive days where the model has no predictions north of 1% and consequently remains outside the market. This has the intended consequence of reducing compounded trading costs considerably. Figure 5.9 shows the results of this less trade-heavy strategy. In the figure, we compare the returns with a passive portfolio, initially composed of all 54 assets equally weighted. Even with steep transaction costs of 0.35% per trade, the model delivers returns far above the passive position. Included in the figure is also the result from the same trading strategy assuming trading costs at 0.1%. The difference between the curves shows how sensitive the results of a daily trading strategy are to an accumulation of transaction costs. Given this sensitivity, it is evident that the assumption made about bid-ask spreads in section 4.5 potentially is very crucial. While we believe to be on the safe side with regards to the transaction costs, this increases the uncertainty of our results.

**Table 5.6:** Risk-adjusted returns for the actively traded and passive portfolio.

| **Portfolio** | $N$ | $\bar{r}_p$ | $\sigma_p$ | $Sharpe_p$ | $\sigma_{d,p}$ | $Sortino_p$ |
|---|---|---|---|---|---|---|
| Model | 366 | 1.06% | 6.13% | 3.31 | 3.66% | 5.54 |
| Benchmark | 366 | 0.59% | 4.96% | 2.29 | 4.50% | 2.52 |

In Table 5.6 we present average daily returns, daily volatility measures and annualized Sharpe and Sortino ratios for the actively traded and passive portfolio. When calculating the Sharpe ratio, we find that the model portfolio is riskier than the benchmark but that the returns compensate handsomely. With a Sharpe ratio 45% higher for the model portfolio, it is seemingly the better investment option. In calculating the Sortino ratio, where only downside risk is included, we find that the model portfolio is less risky than the benchmark. Combined with superior returns, the model portfolio is again the preferred option with a Sortino ratio 119% higher than that of the benchmark. Having said that, there are two points of caution worth stressing. Firstly, both ratios are calculated under the assumption that the returns are normally distributed, which is not the case in our data. The distribution of returns in our dataset is leptokurtic, thus possibly decreasing the relevance of the ratios. Secondly, the volatility of the model portfolio is affected by the model abstaining from the market on a large portion of trading days. Consequently, the true risk of the model portfolio might be higher than the volatility measures express.

We have deliberately kept our trading strategies simple and realistic. More complex strategies allowing for shorting, margin trading and the use of financial derivatives would likely yield higher returns. The danger in devising complex trading rules is that they might generalize poorly. However, following the simple rule of only investing in positions that the model is truly bullish on has the dual benefit of limiting costs while presumably maintaining generalizability. That such a simple strategy so roundly beats the passive position is our most unambiguous evidence of marketplace inefficiency.

# Chapter 6

# Conclusion

This thesis examines if and to what extent one can predict changes in cryptoasset prices. For our analysis, we have built a large and unique dataset for 54 cryptoassets. The data is collected from Twitter, Reddit, Google Trends, the underlying blockchains and cryptoasset exchanges. This vast dataset allows for a broader market analysis compared to those one sees in the literature.

Our first model uses simple linear regression to forecast next-day prices using market data. Next, we add data sources to this model one at a time to estimate each component's contribution to the forecast. The best-performing linear model is then compared to an advanced machine learning model to see if the latter can improve upon the linear model in any meaningful way. Lastly, we show that a realistic trading strategy using the linear model predictions can outperform the market. In particular, the model delivered excess risk-adjusted returns compared to a representative market portfolio from March 2020 to March 2021.

Our study shows that using data from multiple sources improves how well the model performs. While market data produced the most impactful regressors, our model achieved better results by also using Twitter data. One of the regressors built using Google Trends data is also significant, using the Trends data does not translate into higher returns in the test period. We also do not find that using Reddit and blockchain data improves the quality of the predictions.

Our analysis can not verify the claim made in the literature that advanced machine learning models like LSTM-networks can outperform linear regression models. The machine learning model fails to beat our linear model for all combinations of input data, portfolio size and trading strategy. Therefore, it is probably either the case that there are no non-linear relationships in the data or that other types of data or model calibrations would have yielded better results.

In summary, we find clear evidence that cryptoasset prices can be predicted as of March 2021, as all of our models can outperform appropriate benchmarks. Furthermore, since this result also holds when we account for large trading costs, we conclude that it is possible to systematically achieve abnormal returns in the cryptoasset market.

An obvious extension of the analysis conducted in this paper is to look at other time periods. Investigating the performance of similar models over time can give new insights into the predictability of the market. Showing that more complex variable transformations, different model variants and other trading strategies deliver comparable results would support the findings in this thesis.

The high significance of our Twitter variables also suggests that the cryptomarket partially might be "hype-driven." Another avenue of investigation would thus be to investigate this hypothesis directly.

# Bibliography

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

Aggarwal, D. (2019). Do bitcoins follow a random walk model? *Research in Economics*, *73*(1), 15–22.

Alexander, C. (2008). *Market risk analysis, practical financial econometrics* (Vol. 2). John Wiley & Sons.

Bariviera, A. F. (2017). The inefficiency of bitcoin revisited: A dynamic approach. *Economics Letters*, *161*, 1–4.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830–839.

Bengio, Y., Simard, P. & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157–166.

Binance. (2021). *Binance fee schedule.* https://www.binance.com/en/fee/schedule (accessed: 10.04.2021)

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Blinder, A. (2016). *Rtrends.* https://cran.r-project.org/web/packages/rtrends/index.html?fbclid=IwAR03e_zKC1lXcf426rnXilP0w9LMvjlHa1YGVZ6vpg-gO80ZSkmJkVr931k (accessed: 10.03.2021)

Burnie, A. & Yilmaz, E. (2019). An analysis of the change in discussions on social media with bitcoin price. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 889–892.

Cambria, E. & White, B. (2014). Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48–57.

Caporale, G. M., Gil-Alana, L. & Plastun, A. (2018). Persistence in the cryptocurrency market. *Research in International Business and Finance*, *46*, 141–148.

Chen, Z., Li, C. & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395.

Coingecko. (2021). *Coingecko api.* https://www.coingecko.com/en/api (accessed: 20.03.2021)

Cong, L. W., Li, X., Tang, K. & Yang, Y. (2020). Crypto wash trading. *Available at SSRN 3530220.*

CryptoCompare. (2021). *Cryptocompare api.* https://min-api.cryptocompare.com/ (accessed: 19.03.2021)

Glenski, M., Saldanha, E. & Volkova, S. (2019). Characterizing speed and scale of cryptocurrency discussion spread on reddit. *The World Wide Web Conference*, 560–570.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

InternetLiveStats. (2021). *Google search statistics.* https://www.internetlivestats.com/google-search-statistics/ (accessed: 19.05.2021)

Jang, H. & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *Ieee Access*, *6*, 5427–5437.

Ji, S., Kim, J. & Im, H. (2019). A comparative study of bitcoin price prediction using deep learning. *Mathematics*, *7*(10), 898.

Jiang, Y., Nie, H. & Ruan, W. (2018). Time-varying long-term memory in bitcoin market. *Finance Research Letters*, *25*, 280–284.

Kaminski, J. (2014). Nowcasting the bitcoin market with twitter signals. *arXiv preprint arXiv:1406.7577.*

Khuntia, S. & Pattanayak, J. (2018). Adaptive market hypothesis and evolving predictability of bitcoin. *Economics Letters*, *167*, 26–28.

Kim, Y. B., Lee, J., Park, N., Choo, J., Kim, J.-H. & Kim, C. H. (2017). When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PloS one*, *12*(5), e0177630.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Kristoufek, L. (2013). Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, *3*(1), 1–7.

Lahmiri, S. & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, *118*, 35–40.

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

Madan, I., Saluja, S. & Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms. *URL: http://cs229.stanford.edu/proj2014/Isaac%20Madan*, *20.*

Mallqui, D. C. & Fernandes, R. A. (2019). Predicting the direction, maximum, minimum and closing prices of daily bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, *75*, 596–606.

Matta, M., Lunesu, I. & Marchesi, M. (2015). Bitcoin spread prediction using social and web search media. *UMAP workshops*, 1–10.

McNally, S., Roche, J. & Caton, S. (2018). Predicting the price of bitcoin using machine learning. *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*, 339–343.

Mudassir, M., Bennbaia, S., Unal, D. & Hammoudeh, M. (2020). Time-series forecasting of bitcoin prices using high-dimensional features: A machine learning approach. *Neural Computing and Applications*, 1–15.

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K. & Lama, B. K. (2018). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, 128–132.

Park, C. W. & Seo, D. R. (2018). Sentiment analysis of twitter corpus related to artificial intelligence assistants. *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*, 495–498.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703.*

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

Saad, M., Choi, J., Nyang, D., Kim, J. & Mohaisen, A. (2019). Toward characterizing blockchain-based cryptocurrencies for highly accurate predictions. *IEEE Systems Journal*, *14*(1), 321–332.

Seabold, S. & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

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

Shoeb, A. A. M. & de Melo, G. (2021). Assessing emoji use in modern text processing tools. *arXiv preprint arXiv:2101.00430*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267–307.

Tiwari, A. K., Jana, R. K., Das, D. & Roubaud, D. (2018). Informational efficiency of bitcoin—an extension. *Economics Letters*, *163*, 106–109.

TwintProject. (2017). *Twintproject*. https://github.com/twintproject (accessed: 01.04.2021)

Urquhart, A. (2016). The inefficiency of bitcoin. *Economics Letters*, *148*, 80–82.

Urquhart, A. (2018). What causes the attention of bitcoin? *Economics Letters*, *166*, 40–44.

Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S. (2019). Extracting cryptocurrency price movements from the reddit network sentiment. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 500–505.

# Appendix

## A Complete Reddit Post Examples

*Title:* 'The Power of Capital in Africa: A Personal Example'
*Body:* 'I served in the Peace Corps in an African country. The village was small, 300 people at most. Everyone was a farmer, growing peanuts mostly. The harvest was once a year, one paycheck to last until the rains came again. You couldn't trust the banks, I had numerous "fees" that could not be explained to me taken from my account. People held cash and if they could afford bought animals. I ruminated on goats being the African IRA. Some young adults from the village made their way to europe to work on farms. They sent money back, but at a horrible exchange rate and fees on both ends. Still it was enough for the village to get by. I often wondered how access to capital would benefit the village. If people could get loans(even $200-300) how would that change their lives. I raised some money through friends and family to find out. I chose one ambitious farmer and gave him access to the capital. I told him to borrow as much as he thought he would be able to back back(I wasn't really going to have him pay it back). He used $200 to buy a solar panel and a cheap water pump. He pumped water to concrete basins and started a tree nursery. Together we planted fruit trees in his field and sold young trees to neighbors. It was amazing how access to such a small amount of money improved not just his life but the life of the whole village. It shows the immense wealth locked up in Africa. I hope cardano can be the key that unlocks some of this wealth. Access to banking will have cascading benefits that can lift millions out of poverty, and it fills me with optimism. Thank you to this community and lets push forward!Tldr: Gave a loan to a African farmer and he used it to benefit the whole village. Cardano could help millions become food secure and rise from poverty.'

**Textbox 1:** Complete Reddit post used in Textbox 1 in subsection 3.5.2

*Title:* 'Why is there so much hate on ADA? '
*Body:*'I am below newbie in all the crypto stuff. I started buying by the end of January (and Im also not wealthy at all). I got around 900 ADA today, but i keep reading so much discouragement from ppl around forums and YouTube, all of them saying that Cardano is too hyped and probably wont even reach $10. There is a lot of market cap talk and circulating coins and Im telling you, I dont know nothing about any of this. Im just here trying to Hold because im sick tired of my shitty job and dont want to live like this forever. Ofc im not planning to become millionaire with such a low amount of crypto but damn it feels like I wont be able to make any money at all with my investment.'

**Textbox 2:** Complete Reddit post used in Textbox 3.2 in subsection 3.5.2

## B Data Sources Specifications

**Table B1:** Keywords used for Google Trends data, cashtag and additional search terms for used for Twitter data and subreddit used for Reddit data.

| Cryptoasset (Ticker) | Google Trends | Twitter | Reddit |
|---|---|---|---|
| **Cardano (ADA)** | cardano + ADA + cardano crypto + cardano coin + ADA crypto + ADA coin + buy ADA + buy Cardano | $ADA | r/cardano |
| **Aeternity (AE)** | aeternity + AE + aeternity crypto + aeternity coin + AE crypto + AE coin + buy AE + buy Aeternity | $AE | r/Aeternity |
| **Agoras Tokens (AGRS)** | AGRS + agoras crypto + agoras coin + AGRS crypto + AGRS coin + buy AGRS + buy Agoras Tokens | $AGRS + agorastoken + agorastoken | N/A |
| **Ardor (ARDR)** | ARDR + ardor crypto + ardor coin + ARDR crypto + ARDR coin + buy ARDR + buy Ardor | $ARDR | r/Ardor |
| **Ark (ARK)** | ark crypto + ark coin + ARK crypto + ARK coin + buy ARK + buy Ark | $ARK | r/ArkEcosystem |
| **Cosmos (ATOM)** | cosmos + cosmos crypto + cosmos coin + ATOM crypto + ATOM coin + buy ATOM + buy Atomic Coin | $ATOM | r/cosmosnetwork |
| **Basic Attention Token (BAT)** | basic attention token+ basic attention token crypto + basic attention coin + BAT crypto + BAT coin + buy BAT + buy Basic Attention Token | $BAT | r/BATproject |
| **BitBay (BAY)** | bitbay + bitbay crypto + bitbay coin + BAY crypto + BAY coin + buy BAY + buy BitBay | bitbay | r/BitBay |
| **BitconneeectX Genesis (BCCX)** | bitconnect + BCCX + bitconnect crypto + bitconnect coin + BCCX crypto + BCCX coin + buy BCCX + buy BitconnectX Genesis | $BCCX + bitconnect | r/ThebCCapp |
| **Bitcoin Cash (BCH)** | bitcoin cash + BCH + bitcoin cash crypto + trade bitcoin cashBCH crypto + BCH coin + buy BCH + buy Bitcoin Cash | $BCH | r/Bitcoincash |
| **Bytecoin (BCN)** | bytecoin + trade bytecoin + bytecoin crypto + bytecoin token + BCN crypto + BCN coin + buy BCN + buy Bytecoin | $BCN | r/BytecoinBCN |
| **BlackCoin (BLK)** | blackcoin + blackcoin crypto + BLK crypto + BLK coin + buy BLK + buy BlackCoin | blackcoin | r/blackcoin |
| **Blocknet (BLOCK)** | blocknet + blocknet crypto + blocknet coin + BLOCK crypto + BLOCK coin + buy BLOCK + buy Blocknet | $BLOCK + blocknet | r/theblocknet |
| **Binance Coin (BNB)** | binance + trade binance + binance crypto + binance coin + BNB crypto + BNB coin + buy BNB + buy Binance Coin | $BNB | r/binance |
| **Bitcoin SV (BSV)** | bitcoin sv + BSV + bitcoin sv cryptoBSV crypto + BSV coin + buy BSV + buy Bitcoin SV | $BSV | r/bitcoinsv |
| **Bitcoin (BTC)** | bitcoin + BTC + bitcoin crypto + bitcoin token + BTC crypto + BTC coin + buy BTC + buy Bitcoin | $BTC | r/Bitcoin |

| | | | |
|---|---|---|---|
| **BitcoinDark (BTCD)** | bitcoindark + BTCD + bitcoindark crypto + bitcoindark coin + BTCD crypto + BTCD coin + buy BTCD + buy BitcoinDark | bitcoindark | r/bitcoindark |
| **Bitcoin Gold (BTG)** | bitcoin gold + BTG + bitcoin gold cryptoBTG crypto + BTG coin + buy BTG + buy Bitcoin Gold | $BTG | r/btg |
| **Bytom (BTM)** | bytom + bytom crypto + bytom coin + BTM crypto + BTM coin + buy BTM + buy Bytom | $BTM | r/BytomBlockchain |
| **BitShares (BTS)** | bitshares + trade bitshares + bitshares crypto + bitshares coin + BTS crypto + BTS coin + buy BTS + buy BitShares | bitshares | r/BitShares |
| **Celsius (CEL)** | CEL + celsius crypto + celsius coin + CEL crypto + CEL coin + buy CEL + buy Celsium | $CEL | r/CelsiusNetwork |
| **CloakCoin (CLOAK)** | cloakcoin + cloakcoin + cryptoCLOAK crypto + CLOAK coin + buy CLOAK + buy CloakCoin | cloakcoin | r/Cloack_Coin |
| **Compound (COMP)** | compound crypto + compound coin + COMP crypto + COMP coin + buy COMP + buy Compound | compcoin | r/Compound |
| **Crypto.com Coin (CRO)** | crypto com + CRO + crypto com crypto + crypto com coin + CRO crypto + CRO coin + buy CRO + buy Crypto.com Coin | $CRO | r/cro |
| **Dash (DASH)** | dash token + trade dash + dash crypto + dash coin + DASH crypto + DASH coin + buy DASH + buy Dash | $DASH | r/dashpay |
| **Decred (DCR)** | decred + DCR + decred crypto + decred coin + DCR crypto + DCR coin + buy DCR + buy Decred | $DCR | r/decred |
| **DigiByte (DGB)** | digibyte + DGB + digibyte crypto + digibyte coin + DGB crypto + DGB coin + buy DGB + buy DigiByte | $DGB + digibyte | r/Digibyte |
| **DigixDAO (DGD)** | digixdao + DGD + digixdao crypto + digixdao coin + DGD crypto + DGD coin + buy DGD + buy DigixDAO | digixdao | r/digixdao |
| **Dogecoin (DOGE)** | dogecoin + DOGE + dogecoin crypto + doge token + DOGE crypto + DOGE coin + buy DOGE + buy Dogecoin | $DOGE | r/dogecoin |
| **Polkadot (DOT)** | polkadot + DOT + polkadot crypto + polkadot coin + DOT crypto + DOT coin + buy DOT + buy Polkadot | polkadot | r/dot |
| **Emercoin (EMC)** | emercoin + EMC + emercoin cryptoEMC crypto + EMC coin + buy EMC + buy Emercoin | $EMC + emercoin | r/EmerCoin |
| **EOS (EOS)** | eos token + trade eos + eos crypto + eos coin + EOS crypto + EOS coin + buy EOS + buy EOS | $EOS | r/eos |
| **Ethereum Classic (ETC)** | ethereum classic + ETC token + ethereum classic crypto + ethereum classic coin + ETC crypto + ETC coin + buy ETC + buy Ethereum Classic | $ETC | r/EthereumClassic |
| **Ethereum (ETH)** | ethereum + ETH + ethereum crypto + ethereum coin + ETH crypto + ETH coin + buy ETH + buy Ethereum | $ ETH | r/ethereum |

| | | | |
|---|---|---|---|
| **FairCoin (FAIR)** | faircoin + FAIR + faircoin cryptoFAIR crypto + FAIR coin + buy FAIR + buy FairCoin | $FAIR + faircoin | r/faircoin |
| **Factom (FCT)** | factom + FCT + factom crypto + factom coin + FCT crypto + FCT coin + buy FCT + buy Factom | $FCT | r/factom |
| **Filecoin (FIL)** | filecoin + filecoin crypto + FIL crypto + FIL coin + buy FIL + buy Filecoin | filecoin | r/Filecoin |
| **Feathercoin (FTC)** | feathercoin + FTC + feathercoin cryptoFTC crypto + FTC coin + buy FTC + buy Feathercoin | $FTC + feathercoin | r/featherCoin |
| **GameCredits (GAME)** | gamecredits + gamecredits crypto + gamecredits coin + GAME crypto + GAME coin + buy GAME + buy GameCredits | gamecredits | r/GameCreditsCrypto |
| **Obyte (GBYTE)** | byteball + GBYTE + byteball crypto + byteball coin + GBYTE crypto + GBYTE coin + buy GBYTE + buy Obyte | $GBYTE | r/byteball |
| **Golem (GLM)** | golem network + GLM + golem network crypto + golem network coin + GLM crypto + GLM coin + buy GLM + buy Golem | $GLM + golemcoin | r/GolemProject |
| **Gnosis (GNO)** | gnosis + GNO + gnosis crypto + gnosis coin + GNO crypto + GNO coin + buy GNO + buy Gnosis | $GNO | r/gnosisPM |
| **Gridcoin (GRC)** | gridcoin + GRC + gridcoin crypto + GRC crypto + GRC coin + buy GRC + buy Gridcoin | $GRC + gridcoin | r/gridcoin |
| **HedgeTrade (HEDG)** | hedgetrade + HEDG + hedgetrade crypto + hedgetrade coin + HEDG crypto + HEDG coin + buy HEDG + buy HedgeTrade | $HEDG + hedgetrade | r/HedgeTrade |
| **Hashshare (HSS)** | hshare + HSS + hshare crypto + hshare coin + HSS crypto + HSS coin + buy HSS + buy Hashshare | $HSS + hashshare | N/A |
| **Huobi Token (HT)** | huobi + HT + huobi crypto + huobi coin + HT crypto + HT coin + buy HT + buy Huobi Token | $HT + huobitoken | r/huobi |
| **Iconomi (ICN)** | iconomi + ICN + iconomi crypto + iconomi coin + ICN crypto + ICN coin + buy ICN + buy Iconomi | $ICN + iconomi | r/ICONOMI |
| **ICON (ICX)** | ICX + icon crypto + icon coin + ICX crypto + ICX coin + buy ICX + buy ICON | $ICX | r/helloicon |
| **I/O Coin (IOC)** | iocoin + IOC + iocoin crypto + IOC crypto + IOC coin + buy IOC + buy I/O Coin | iocoin + ioccoin | r/IODigitalCurrency |
| **Kyber Network (KNC)** | kyber network + KNC + kyber network crypto + kyber network coin + KNC crypto + KNC coin + buy KNC + buy Kyber Network | kyber-network | r/kybernetwork |
| **Aave (LEND)** | aaveaave crypto + aave coin + LEND crypto + LEND coin + buy LEND + buy Aave | $LEND + aave | r/Aave_Official |
| **UNUS SED LEO (LEO)** | unus-sed-leo crypto + unus-sed-leo coin + LEO crypto + LEO coin + buy LEO + buy unus-sed-leo | unus-sed-leo + leo-token | N/A |

|  |  |  |  |
|---|---|---|---|
| **Chainlink (LINK)** | chainlink + chainlink crypto + chainlink coin + LINK crypto + LINK coin + buy LINK + buy Chainlink | $LINK | r/Chainlink |
| **Lisk (LSK)** | lisk + LSK + lisk crypto + lisk coin + LSK crypto + LSK coin + buy LSK + buy Lisk | $LSK | r/Lisk |
| **Litecoin (LTC)** | litecoin + LTC + litecoin crypto + litecoin token + LTC crypto + LTC coin + buy LTC + buy Litecoin | $LTC | r/litecoin |
| **MaidSafeCoin (MAID)** | maidsafe + trade maid + maidsafe crypto + maidsafe coin + MAID crypto + MAID coin + buy MAID + buy MaidSafeCoin | $MAID | r/safeNetwork |
| **IOTA (MIOTA)** | iota token + MIOTA + iota crypto + iota coin + MIOTA crypto + MIOTA coin + buy MIOTA + buy IOTA | $MIOTA | r/IOTA |
| **Maker (MKR)** | MKR + maker crypto + maker coin + MKR crypto + MKR coin + buy MKR + buy Maker | $MKR | r/MakerDAO |
| **Melon (MLN)** | MLN + melon crypto + melon coin + MLN crypto + MLN coin + buy MLN + buy Melon | $MLN | r/melonproject |
| **MonaCoin (MONA)** | monacoin + monacoin crypto + monacoin coin + MONA crypto + MONA coin + buy MONA + buy MonaCoin | $MONA | r/monacoin |
| **Maximine Coin (MXM)** | maximine + MXM + maximine crypto + maximine coin + MXM crypto + MXM coin + buy MXM + buy Maximine Coin | $MXM | r/MaxiMineCoin |
| **Nano (NANO)** | NANO + nano crypto + nano coin + NANO crypto + NANO coin + buy NANO + buy Nano | $NANO | r/nanocurrency |
| **Navcoin (NAV)** | nav-coin + NAV + nav-coin crypto + nav-coin coin + NAV crypto + NAV coin + buy NAV + buy Navcoin | navcoin | r/NavCoin |
| **Neo (NEO)** | trade neo + neo token + neo crypto + neo coin + NEO crypto + NEO coin + buy NEO + buy Neo | $NEO | r/NEO |
| **Gulden (NLG)** | gulden + NLG + gulden crypto + gulden coin + NLG crypto + NLG coin + buy NLG + buy Gulden | $NLG + guldencoin | r/GuldenCommunity |
| **Namecoin (NMC)** | namecoin + NMC + namecoin cryptoNMC crypto + NMC coin + buy NMC + buy Namecoin | namecoin | r/Namecoin |
| **nxt (NXT)** | nxt + NXT + nxt crypto + nxt coin + NXT crypto + NXT coin + buy NXT + buy Nxt | $NXT | r/NXT |
| **OKB (OKB)** | okb + OKB + okb crypto + okb coin + OKB crypto + OKB coin + buy OKB + buy OKB | $OKB | N/A |
| **OMG Network (OMG)** | omisego + omisego crypto + omisego coin + OMG crypto + OMG coin + buy OMG + buy OmiseGO | $OMG | r/omise_go |
| **Ontology (ONT)** | ontology + ontology crypto + ontology coin + ONT crypto + ONT coin + buy ONT + buy Ontology | $ONT | r/OntologyNetwork |
| **TenX (PAY)** | tenx + tenx crypto + tenx coin + PAY crypto + PAY coin + buy PAY + buy TenX | $PAY | r/TenX |

| | | | |
|---|---|---|---|
| **PIVX (PIVX)** | pivx + PIVX + pivx crypto + pivx coin + PIVX crypto + PIVX coin + buy PIVX + buy PIVX | $PIVX | r/PIVX |
| **PotCoin (POT)** | potcoin + potcoin crypto + POT crypto + POT coin + buy POT + buy PotCoin | $POT + potcoin | r/PotCoin |
| **Peercoin (PPC)** | peercoin + PPC + peercoin cryptoPPC crypto + PPC coin + buy PPC + buy Peercoin | $PPC | r/Peercoin |
| **Populous (PPT)** | PPT + populous crypto + populous coin + PPT crypto + PPT coin + buy PPT + buy Populous | $PPT | r/populous_platform |
| **Qtum (QTUM)** | qtum + QTUM + qtum crypto + qtum coin + QTUM crypto + QTUM coin + buy QTUM + buy Qtum | $QTUM | r/Qtum |
| **Revain (R)** | revain + R + revain crypto + revain coin + R crypto + R coin + buy R + buy Revain | revain-coin | r/revain_org |
| **Rubycoin (RBY)** | rubycoin + RBY + rubycoin cryptoRBY crypto + RBY coin + buy RBY + buy Rubycoin | $RBY + rubycoin | N/A |
| **ReddCoin (RDD)** | reddcoin + RDD + reddcoin cryptoRDD crypto + RDD coin + buy RDD + buy ReddCoin | $RDD + reddcoin | r/reddcoin |
| **Augur (REP)** | REP + augur crypto + augur coin + REP crypto + REP coin + buy REP + buy Augur | $REP | r/Augur |
| **Ravencoin (RVN)** | ravencoin + RVN + ravencoin cryptoRVN crypto + RVN coin + buy RVN + buy Ravencoin | ravencoin | r/Ravencoin |
| **Safex Token (SAFEX)** | safe exchange + SAFEX + safe exchange crypto + safe exchange coin + SAFEX crypto + SAFEX coin + buy SAFEX + buy Safex Token | safextoken + safeexchangecoin | r/safex |
| **Siacoin (SC)** | siacoin + SC + siacoin crypto + siacoin coin + SC crypto + SC coin + buy SC + buy Siacoin | $SC | r/siacoin |
| **SolarCoin (SLR)** | solarcoin + SLR + solarcoin crypto + solarcoin coin + SLR crypto + SLR coin + buy SLR + buy SolarCoin | $SLR + solarcoin | r/SolarCoin |
| **SingularDTV (SNGLS)** | singulardtv + SNGLS + singulardtv crypto + singulardtv coin + SNGLS crypto + SNGLS coin + buy SNGLS + buy SingularDTV | $SNGLS + singulardtv | r/SingularDTV |
| **Steem (STEEM)** | steem + STEEM + steem crypto + steem coin + STEEM crypto + STEEM coin + buy STEEM + buy Steem | $STEEM | r/steem |
| **Stratis (STRAT)** | stratis + STRAT + stratis crypto + stratis coin + STRAT crypto + STRAT coin + buy STRAT + buy Stratis | $STRAT + $STRAX | r/stratisplatform |
| **Syscoin (SYS)** | syscoin + syscoin crypto + SYS crypto + SYS coin + buy SYS + buy Syscoin | $SYS + syscoin | r/SysCoin |
| **Theta (THETA)** | theta crypto + theta coin + THETA crypto + THETA coin + buy THETA + buy THETA | $THETA + thetacoin + thetatoken | r/theta_network |
| **TRON (TRX)** | trade tron + TRX + tron crypto + tron coin + TRX crypto + TRX coin + buy TRX + buy TRON | $TRX | r/Tronix |

| | | | |
|---|---|---|---|
| **Uniswap (UNI)** | uniswap + uniswap crypto + uniswap coin + UNI crypto + UNI coin + buy UNI + buy Uniswap | uniswap | r/uniSwap |
| **SuperNET (UNITY)** | supernet unity + supernet unity crypto + supernet unity coin + UNITY crypto + UNITY coin + buy UNITY + buy SuperNET | $UNITY | r/supernet |
| **Tether (USDT)** | trade tether + USDT + tether crypto + tether coin + USDT crypto + USDT coin + buy USDT + buy Tether | $USDT | r/Tether |
| **VeChain (VET)** | vechain + vechain crypto + vechain coin + VET crypto + VET coin + buy VET + buy VeChain | $VET | r/VeChain |
| **Viacoin (VIA)** | viacoin + viacoin crypto + viacoin coin + VIA crypto + VIA coin + buy VIA + buy Viacoin | viacoin | r/viacoin |
| **VeriCoin (VRC)** | vericoin + VRC + vericoin crypto + vericoin coin + VRC crypto + VRC coin + buy VRC + buy VeriCoin | $VRC + vericoin | r/veriCoin |
| **Vertcoin (VTC)** | vertcoin + VTC + vertcoin cryptoVTC crypto + VTC coin + buy VTC + buy Vertcoin | $VTC + vertcoin | r/vertcoin |
| **Waves (WAVES)** | waves crypto + waves coin + WAVES crypto + WAVES coin + buy WAVES + buy Waves | $WAVES | r/Wavesplatform |
| **Wrapped Bitcoin (WBTC)** | wrapped bitcoin + WBTC + wrapped bitcoin crypto + wrapped bitcoin coin + WBTC crypto + WBTC coin + buy WBTC + buy Wrapped Bitcoin | $WBTC + wrappedbitcoin | r/WrappedBitcoin |
| **Xaurum (XAUR)** | xaurum + XAUR + xaurum crypto + xaurum coin + XAUR crypto + XAUR coin + buy XAUR + buy Xaurum | $XAUR | r/xaurum |
| **Counterparty (XCP)** | counterparty + XCP + counterparty crypto + counterparty coin + XCP crypto + XCP coin + buy XCP + buy Counterparty | $XCP | r/counterparty_xcp |
| **DigitalNote (XDN)** | digitalnote + XDN + digitalnote crypto + digitalnote coin + XDN crypto + XDN coin + buy XDN + buy DigitalNote | $XDN | r/digitalNote |
| **NEM (XEM)** | new economy movement + XEM + nem crypto + nem coin + XEM crypto + XEM coin + buy XEM + buy NEM | $XEM | r/nem |
| **Stellar (XLM)** | Stellar lumen + XLM + stellar crypto + stellar coin + XLM crypto + XLM coin + buy XLM + buy Stellar | $XLM | r/stellar |
| **Monero (XMR)** | monero + XMR + monero crypto + monero coin + XMR crypto + XMR coin + buy XMR + buy Monero | $XMR | r/monero |
| **Ripple (XRP)** | ripple token + XRP + ripple crypto + ripple coin + XRP crypto + XRP coin + buy XRP + buy XRP | $XRP | r/ripple |
| **Tezos (XTZ)** | tezos + XTZ + tezos crypto + tezos coin + XTZ crypto + XTZ coin + buy XTZ + buy Tezos | $XTZ | r/tezos |
| **Verge (XVG)** | XVG + verge crypto + verge coin + XVG crypto + XVG coin + buy XVG + buy Verge | $XVG + vergecoin + vergecurrency | r/vergecurrency |

| | | | |
|---|---|---|---|
| **yearn.finance (YFI)** | yearn finance + YFI + yearn finance crypto + yearn finance coin + YFI crypto + YFI coin + buy YFI + buy yearn.finance | yearn-finance | r/yearn_finance |
| **Zcash (ZEC)** | zcash + ZEC + zcash crypto + zcash coin + ZEC crypto + ZEC coin + buy ZEC + buy Zcash | $ZEC | r/zec |
| **Zilliqa (ZIL)** | zilliqa + ZIL + zilliqa crypto + zilliqa coin + ZIL crypto + ZIL coin + buy ZIL + buy Zilliqa | $ZIL | r/zilliqa |
| **0x (ZRX)** | 0x + ZRX + 0x crypto + 0x coin + ZRX crypto + ZRX coin + buy ZRX + buy 0x | $ZRX | r/0xProject |
| **FirstBlood (1ST)** | firstblood + 1ST + firstblood crypto + firstblood coin + 1ST crypto + 1ST coin + buy 1ST + buy FirstBlood | 1stcoin + firstblood-coin | r/FirstBloodio |

# C Asset Selection Process

**Table C1:** Assets removed in the selection process. Tickers are only listed once, even though several exclusion criteria might apply.

| **Criteria** | **Assets Removed** |
|---|---|
| Average daily trading volume  1m USD in 2019 | AGRS, BCCX, BCN, BLK, BLOCK, CEL, CLOAK, EMC, FAIR, FIL, FTC, GAME, GRC, LEND, MLN, NLG, NMC, POT, RBY, SLR, VRC, XAUR, XCP, XDN |
| Not tradeable in 2020/2021 | BTCD, ICN, MXM, SAFEX, UNI, UNITY, YFI |
| Stablecoin | USDT, WBTC |
| Insufficient price, Twitter or Reddit data | AE, ATOM, BAY, BSV, COMP, CRO, DGD, DOT, GBYTE, GLM, HEDG, HSS, LEO, R, VIA, 1ST |

# D Complete Blockchain Variables

**Table D1:** Complete list of data available for the blockchains with accompanying variable descriptions.

| **Variable** | **Description** |
|---|---|
| Block height | Block height represents the max block number for the given day |
| Transaction count | Count of valid transactions for a given day, after filtering out failed transactons |
| Transaction count all time | Count of transactions since inception |
| Large transaction count | Count of large (>100,000 USD) transactions per day |
| Average transaction value | Average transaction value denominated in the native units of the digital asset per day |
| Zero balance addresses all time | Sum of zero balance addresses since inception |
| Unique addresses all time | The sum of addresses that executed at least one transaction since inception. |
| New addresses | The sum of addresses that were created that day |
| Active addresses | The sum of addresses that executed at least one transaction during the last day |
| Hashrate | The hash rate for a day is the average difficulty / the average time between blocks for the day / $10^{12}$. It is expressed in TH/s (1,000,000,000,000 (one trillion) hashes per second) |
| Difficulty | The mean difficulty of finding a hash that meets the protocol-designated requirement (e.g. for Bitcoin it is the the difficulty of finding a new block) that day. |
| Current supply | the sum of all native units issued on the ledger |
| Block time | Average time in seconds it took for each block to be created that day. |
| Block size | The average size in bytes of all blocks created that day. |

NTNU

Norwegian University of
Science and Technology