# Preface

This Master's thesis concludes the authors' Master of Science degrees in Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU), in the Spring of 2019. The course is a mandatory part for students specializing in Financial Engineering. The thesis is written by Adam Joseph Zinafrazi and Nikolai Nyrud Gobel, under the guidance of Dr. Peter Molnár.

The first part of this thesis attempts to predict the trading activity and usage of Bitcoin, Ether, and Litecoin. We extend the current literature by creating sentiment variables using news articles about cryptocurrencies, in addition to using previously studied control variables. Sentiment analysis is a process of identifying and categorizing opinions articulated in a text. While multiple studies have applied sentiment analysis on social media, few have considered the sentiment in news articles. By performing sentiment analysis using nearly 21,000 news articles about cryptocurrency, we attempt to predict changes in trading activities by looking at four financial variables for each currency: return, volatility, trading volume, and transaction volume. The second part of this thesis investigates the inverse relationship: looking at whether trading activities, usage of cryptocurrencies, and control variables can predict the sentiment of news articles.

We want to extend our gratitude to Dr. Peter Molnár. His enthusiasm and interest in our topic have been inspiring, and his comments and suggestions have been much appreciated.

Trondheim, June 2019

# The Sentiment of News Articles and Trading Activity of Cryptocurrencies

Adam Joseph Zinafrazi and Nikolai Nyrud Gobel

## Abstract

We extract sentiment from news articles about Bitcoin, Ether, and Litecoin and analyze whether it can predict the trading activity and usage of these cryptocurrencies. We analyze nearly 21,000 cryptocurrency-related articles obtained from the ProQuest database, spanning the period 11/2013 – 01/2019. Each article receives a sentiment score, and these scores are aggregated for all articles published within a week. Using three different methods to calculate the sentiment score, we find that all three sentiment measures can contribute to predicting the return, volatility, trading volume, and transaction volume of Bitcoin and Ether, but not for Litecoin. We further conclude that it is not sufficient to only use one dictionary to capture the sentiment in an article. We have several additional findings including 1) number of published articles predicts Bitcoin volatility, 2) number of addresses predicts Bitcoin return, 3) volatility negatively impacts trading volume for all cryptocurrencies, and 4) high return leads to high transaction volume for Bitcoin and Ether. Secondly, we then investigate the inverse relationship: looking at whether the trading activity, usage of cryptocurrencies, and control variables can predict the sentiment of news articles. We find that there are significant differences between the sentiment of the cryptocurrencies. In addition, we find multiple factors which can predict the sentiment of news articles, including return, volatility, Google searches, trading volume, and the number of addresses.

**Keywords and phrases**: Cryptocurrency, Bitcoin, Ether, Litecoin, Sentiment analysis, Text mining, News articles, Prediction

# Sammendrag

Vi måler sentimentet i nyhetsartikler om Bitcoin, Ether og Litecoin og undersøker om det kan forutse handelsaktivitet og bruk av kryptovaluta. Vi analyserer nesten 21,000 artikler relatert til kryptovaluta lastet ned fra ProQuest sin database fra perioden 11/2013-01/2019. Hver artikkel fikk en sentimentverdi, og verdiene ble aggregert til ukentlige verdier. Ved å bruke tre forskjellige metoder for å beregne sentimentet, finner vi at alle tre metoder kan bidra til å modellere fremtidig avkastning, volatilitet, handelsvolum og transaksjonsvolum for Bitcoin og Ether, men ikke for Litecoin. Videre konkluderer vi med at det ikke er nok å bruke kun én ordbok for å måle sentimentet i en artikkel. Vi har flere funn: 1) antall publiserte artikler forutser volatilitet, 2) antall adresser forutser Bitcoin sin avkastning, 3) volatilitet påvirker handelsvolum negativt for alle kryptovalutaer, 4) høy avkastning fører til høyt transaksjonsvolum for Bitcoin og Ether. Videre undersøker vi den motsatte sammenhengen, hvorvidt handelsaktvitet, bruk av kryptovaluta og andre kontrollvariabler kan bli brukt til å forutse sentimentet i nyhetsartikler. For det første observerer vi at det er store forskjeller mellom sentimentet til de ulike kryptovalutaene. Videre finner vi at følgende variabler kan forutse sentimentet: avkastning, volatilitet, Google-søk, handelsvolum og antall adresser.

# Contents

# 1 Introduction

Bitcoin, developed by Nakamoto et al. (2008), is a peer-to-peer electronic cash version that allows payments to be made without involving the trust from a third party, such as a financial institution. Thus, Bitcoin aims to act as an alternative currency to government-issued currencies, such as Euro, US dollar, and Chinese yuan. From June 2016 to 2019, Bitcoin prices have climbed from a minimum price of $525 to a maximum price of $20,089.[1] This makes Bitcoin an attractive investment opportunity for risk-tolerant investors, as discussed by Glaser et al. (2014). They conclude that new users of Bitcoin primarily trade Bitcoin as a speculative investment, rather than using it to pay for goods or services. In addition to being a potentially high return asset, investors are interested in Bitcoin due to its potential as a portfolio diversifier or safe haven (Chuen et al., 2017).

Following Bitcoin, a vast amount of new cryptocurrencies have been developed. These new cryptocurrencies are often referred to as alternative coins, or the shorter version *altcoins*. Two of the most prominent altcoins are Ether and Litecoin, both among the top five in terms of market capitalization.[1] The currency Ether is based on Ethereum: a blockchain-based distributed computing platform and operating system which gives users the ability to create *smart contracts* (Wood et al., 2014). The advantage of smart contracts is the inbuilt trust mechanism which allows users to perform transactions without involving third parties. By incorporating smart contracts, the applications of Ether are quite different from Bitcoin and Litecoin, which both are designed to function as a payment system and store of value. Litecoin is an altcoin that is technically very similar to Bitcoin but aims to decrease the transaction time and improve on some of Bitcoin's limitation as a payment system (Steadman, 2013).

While some attempts have been made to classify cryptocurrencies as commodities or currencies, recent literature agrees that the behavior of cryptocurrencies significantly differs from these asset classes arguing for a separate classification of cryptocurrencies (Chuen et al. 2017; Glaser et al. 2014). Considering these behavior characteristics, and the rise of investor- and media attention in recent years, there is an imminent need to understand the dynamics of cryptocurrencies better. Multiple studies have attempted to explain or predict trading activity-related factors such as return, volatility, and trading volume (Mai et al.

[1] https://coinmarketcap.com

2018; Aalborg et al. 2018; Kristoufek 2013). Others have also researched the usage of cryptocurrencies, proxied by transaction volume (Kim et al. 2016; Koutmos 2018). However, as cryptocurrencies are digital currencies with no backing from any government or central bank, traditional methods of valuation fall short (Mai et al., 2018). For example, as most cryptocurrencies have a fixed supply, the methods used to valuate regular currencies, which have an unlimited supply, do not apply. Furthermore, digital currencies do not generate any cash flow, making the methods used to valuate bonds and stock less applicable (Chuen et al., 2017). Hence, since cryptocurrencies differ from traditional asset classes, researchers have been motivated to consider other methods of valuation. For example, sentiment analysis of media has shown promising results in studying the valuation of cryptocurrencies. (Kaminski 2014; Garcia and Schweitzer 2015; Kim et al. 2016). While several have studied the influence of sentiment in social media, few have considered the sentiment of news articles.

Sentiment analysis, a branch of textual analysis, is a field of study which focuses on analyzing people's opinions, attitudes, and emotions from written language (Liu, 2012). Various methods of sentiment analysis have been used increasingly in later years, as more information has become accessible online. While several studies have performed sentiment analysis on social media about cryptocurrencies, few have considered the sentiment in news articles. This paper uses the sentiment in news articles, along with multiple control variables frequently used in the literature, to predict the trading activity and usage of three major cryptocurrencies: Bitcoin, Ether, and Litecoin. Then we predict the sentiment of news articles related to these cryptocurrencies. Return, volatility, and trading volume are considered to be trading activity-related variables. Transaction volume is considered a proxy for the usage of a cryptocurrency.

Little research has been done on the impact of sentiment in news articles on the trading activity of cryptocurrencies. Among the few examples, Polasik et al. (2015) analyze news articles and discover that the sentiment of Bitcoin-related news reports drives Bitcoin return. Most studies have, however, focused their analysis on how sentiment on social media, such as Twitter and Reddit, influence Bitcoin return and volatility. For example, Garcia and Schweitzer (2015) perform sentiment analysis on Bitcoin-related tweets to develop a profitable trading strategy for Bitcoin. Bukovina et al. (2016) use the social media Reddit as its primary source of sentiment and conclude that the explanatory power of sentiment increases during periods of excessive Bitcoin volatility.

This paper extends the topic of sentiment analysis on news articles, by analyzing the sentiment of news articles about Bitcoin, Ether, and Litecoin, using three different measures of sentiment. Nearly 21,000 news articles published in the period between 01/2011-01/2019 are collected from the ProQuest database and analyzed for sentiment. The news sentiment, along with other control variables frequently used in the literature, is then used in two models of multi-linear regressions. These models are used in an attempt to first predict the trading activity and usage of cryptocurrencies, and then predict the sentiment in news articles.

The rest of this paper is organized as follows: Chapter 2 reviews the literature on sentiment analysis, and discusses current studies on cryptocurrencies; Chapter 3 presents the methodology of sentiment analysis used in this paper; Chapter 4 explains how the data was collected, standardized and transformed; Chapter 5 presents the results; Finally, Chapter 6 concludes and presents ideas for further research.

# 2 Literature Review

This chapter first introduces the concept of sentiment analysis and then provides an overview of the related literature about cryptocurrencies.

## 2.1 Sentiment analysis

Sentiment analysis is defined as the "process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic or product is positive, negative, or neutral"[1]. The idea of using news sentiment to predict returns on financial assets have been an attractive idea for decades, and recent developments in computational power have made it possible to expand the methods of sentiment analysis (Liu, 2012).

Sentiment analysis is a branch of textual analysis. As presented in Figure 2.1, the most common methods of textual analysis are based either on a lexicon-based approach or machine learning. There are four common methods of machine learning: Naive Bayes, Support vector machines, Semantic analysis, and Neural Network. Since these approaches are not used in this paper, the reader is kindly referred to the work by Guo et al. (2016) for a good overview of this field.

The lexicon-based approach can be split into two techniques: readability measure and dictionary-based approach. Readability looks at the degree to which a reader finds a particular text compelling and comprehensive (Mc Laughlin, 1969). This measure has been used in financial literature. For example, De Franco et al. (2015) find that companies that provided more readable analyst reports often had more trade volume after the report date. Similarly, Miller (2010) find that firms with a lower readability score on their annual reports had their shares traded less by investors. Secondly, while readability looks at the ability for a reader to *comprehend* a message, a dictionary-based approach aims to explore the *meaning* of the message (Guo et al., 2016). A dictionary is often referred to as a collection of words where a value is assigned to each word, depending on the particular attribute the dictionary aims to explain. Therefore, each attribute can provide a comparative

[1] https://www.lexico.com/en/definition/sentiment_analysis/

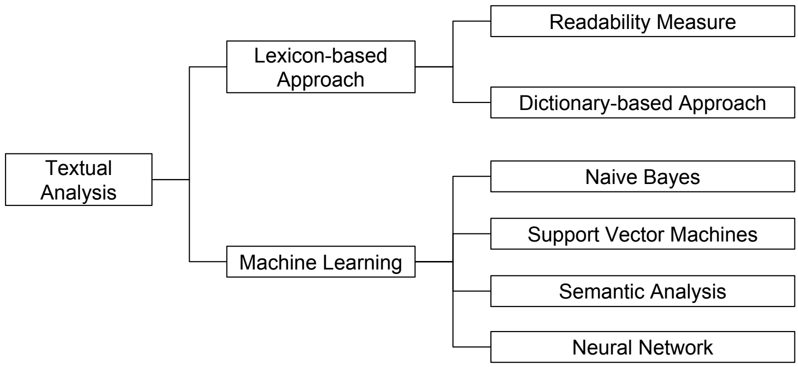



**Figure 2.1:** Popular techniques of textual analysis, diagram modified from Guo et al. (2016).

measure of sentiment. When the order of words and grammar of sentences are ignored in the analysis, the method is referred to as the bag-of-words model (Guo et al., 2016). The assumption that the order of words in a text does not matter may be considered a big assumption. However, Miner et al. (2012) argue that the order of word is less important in terms of classifying text. For example, one can classify the text as either positive or negative, without necessarily considering the order of words. On the other hand, the order of words may be critical in tasks such as information extraction and natural language processing. Some dictionaries follow the bag-of-words model, whereas other more sophisticated dictionaries consider the order of words, by considering negations and degree modifiers for example. These will be discussed further in Chapter 3.

The dictionary-based approach has been applied repeatedly in papers on traditional finance. The Harvard General Inquirer (GI) is an extensive dictionary that can be used to measure the tone sentiment of a message. Tetlock (2007) apply the Harvard GI and find that a pessimistic tone of "Abreast of the Market" in Wall Street Journals can lead to low stock return and high stock volatility. Loughran and McDonald (2011) argue that the Harvard GI does not work well with financial terms, and thus construct a new word list based on analyzing the meaning of words in a business context. Heston and Sinha (2016) perform sentiment analysis on more than 900,000 Thomson-Reuters news and find that the tone in news positively relates to short-term stock return.

## 2.2 Trading activity and usage of cryptocurrencies

Several studies attempt to model the financial dynamics of cryptocurrencies. As introduced earlier, cryptocurrencies are a new speculative asset class which can be used as investment vehicles. As the number of investors in cryptocurrencies have risen, there is a growing need to understand the underlying price mechanics in the market. Furthermore, by modeling and understanding the volatility and trading volume of cryptocurrencies, we can improve our understanding of the trading activity and investor behavior. This information is crucial in the field of risk management and for hedging purposes. Lastly, the transaction volume of a cryptocurrency is a measure of the use of a cryptocurrency, and by modeling this factor, one gets insight into the user behavior of the cryptocurrency.

The following sections discuss the current literature for each financial variable: return, volatility, trading volume, and transaction volume. While many papers attempt to predict return, only a few attempts to predict the latter three financial factors. Each section is discussed first in the lights of sentiment analysis and then using other factors. The key characteristics of the discussed research are summarized in Tables 2.1-2.2.

### 2.2.1 Return

The most common usage of sentiment analysis on text related to cryptocurrencies is to model future returns. One of the first sentiment analysis on cryptocurrency was performed by Kaminski (2014), who analyze the sentiment of Twitter posts containing the keyword "Bitcoin". Using a small 15-word dictionary containing words of both positive and negative attributes, they find a correlation between emotional tweets and the closing price of Bitcoin. However, after a dynamic Granger causality analysis, they conclude that the tweets mirror the market rather than predict it. Garcia and Schweitzer (2015) apply the valence-dictionary developed by Warriner et al. (2013) to study whether the emotional valence in Bitcoin-related tweets can predict Bitcoin return. They reveals that increases in opinion polarization and trading volume precede rising Bitcoin prices and that emotional valence precedes opinion polarization and rising trading volumes. Using their findings, they develop a highly profitable trading strategy for Bitcoin. The previous two studies both look at social media. Polasik et al. (2015), on the other hand, are one of the few studies who perform sentiment analysis on news articles about Bitcoin. They apply the tone-dictionary developed by Henry (2008) to discover that the tone of newspaper articles with the keyword "Bitcoin" act as one of the primary drivers of Bitcoin return, a finding which the author claim to be the first one to report.

As newer dictionaries become developed over time, the vader-dictionary developed by Hutto and Gilbert (2014) has frequently been used in sentiment analysis to predict return, providing inconclusive results. Kim et al. (2016) apply the vader-dictionary to score user-comments and replies in online cryptocurrency communities for Bitcoin, Ether, and Ripple. In an attempt to predict price fluctuations of the three cryptocurrencies, they conclude that positive user comments significantly affect the price of Bitcoin, whereas negative comments significantly influence the prices of Ether and Ripple. Stenqvist and Lönnö (2017) use a dataset consisting of 2.27 million Bitcoin-related tweets and look for sentiment fluctuations that can be used to predict changes in Bitcoin price. Using intraday data, they conclude that their prediction model for Bitcoin returns yield up to 79% accuracy. In more recent years, Ghiasvand (2018) study the relationship between online comments on Reddit with prices of Bitcoin, Ether, and Litecoin. They conclude that the sentiment of Reddit comments was not a successful predictor of price changes. Similarly, Abraham et al. (2018) find that sentiment on Twitter posts does not predict the price of Bitcoin and Ether, but rather the number of posts over time, referred to as *tweet volume*. In contrast to these results, Steinert and Herff (2018) perform sentiment analysis on Twitter posts over a timeframe of 71 days to predict the returns of 181 altcoins. They conclude that the sentiment of tweets can predict fluctuations in altcoins, such as Ether. Finally, Mai et al. (2018) study to what extent social media can impact the value of Bitcoin, with sentiment estimates from the forum BitcoinTalk, Twitter and news articles from the Thomson Reuters News Analytics database for news articles that contained the keyword "Bitcoin". To estimate the sentiment, they apply a finance-based sentiment dictionary containing 2,329 negative and 297 positive words. Using vector error correction models (VECMs) to study the dynamic relationships between the sentiment scores and Bitcoin return, they find positive (negative) forum posts to positively (negatively) predict return, while neither tweets nor news articles provide any significant results. To summarize, the literature is inconclusive on the impact of sentiment analysis on the return of cryptocurrencies.

Attempts to predict return has not been limited to sentiment analysis. Prevalent examples of other factors used in an attempt to predict return include Google- and Wikipedia searches, transaction volume, number of unique addresses, trading volume, volatility, and macro-financial indicators. Google- and Wikipedia search can be used as a measure of investor interest. Kristoufek (2013) find a strong bidirectional causality between the prices of Bitcoin and the search queries for Bitcoin on Google Trends and Wikipedia. Ciaian et al. (2016) reach a similar conclusion for Wikipedia. Similarly, Matta et al. (2015a) and Mai et al. (2018) show that search queries for Bitcoin on Google trends significantly correlate with the price movements of Bitcoin. Also, studies confirm that when Bitcoin price is either below or above its trendline, the interest in Bitcoin has increased in terms of Google searches (Kristoufek 2013; Polasik et al. 2015). Dickerson (2018) develop a largely profitable Bitcoin-trading algorithm based on Google and Wikipedia searches. On the other hand, Aalborg et al. (2018) do not find any relation between Bitcoin searches on Google and Bitcoin returns. For transaction volume, Ciaian et al. (2016) find that the total number of unique Bitcoin transactions per day had more impact on the Bitcoin return than the number of Bitcoin addresses available in the market. Koutmos (2018) consider the impact a shock in new Bitcoin addresses has on return and find an immediate positive effect on return, which gradually faded over 12 days. On the other hand, Aalborg et al. (2018) conclude that the number of unique Bitcoin address has no predictive power for Bitcoin return. A fourth factor used for predicting return is trading volume. Balcilar et al. (2017) perform a causality-in-quantiles test which reveal that trading volume can predict return. A similar relationship was shown by Garcia and Schweitzer (2015) and Sovbetov (2018). In contrast, Meland and Øyen (2017) and Mai et al. (2018) find that trading volume is negatively correlated to price movements. In terms of volatility, Aalborg et al. (2018) and Mai et al. (2018) do not find any evidence for volatility predicting Bitcoin return, whereas Sovbetov (2018) find volatility to be significant determinant both in long- and short runs for Bitcoin, Ether and Litecoin. Lastly, macro-financial indicators such as the VIX-index, Dow Jones stock market index, oil price, and gold price have all been found insignificant in the price formation of Bitcoin (Ciaian et al. 2016, Aalborg et al. 2018, Kjærland et al. 2018). In short, the attempts to predict return using macro-financial indicators have proven faulty, while the results for the other factors are inconclusive.

### 2.2.2 Volatility, trading- and transaction volume

**Volatility**

Only a few studies have considered the sentiment of social media and news articles, and its impact on volatility. Bukovina et al. (2016) use the tool SentDex on comments from the social media platform Reddit as a proxy for investor sentiment, and conclude that the explanatory power of sentiment for Bitcoin volatility increase during periods of excessive volatility, and that positive sentiment has a greater impact on volatility than negative sentiment. Mai et al. (2018), on the other hand, test the sentiment of Twitter posts, forum posts, and news articles, and find that none of them can predict future volatility. While there is little research considering the relationship between volatility and sentiment, several studies look at other variables impact on future volatility. First of all, using past volatility to predict volatility is widely recognized. For example, Dyhrberg (2016) conclude that Bitcoin shared similarities with gold in terms of its volatility. Ghiasvand (2018) find that Google trends can explain volatility, but not predict it. This is in contrast to the results of Aalborg et al. (2018), who find Google trends, trading volume, and return to predict volatility on daily data, but not on weekly. Secondly, while trading volume has been found to be able to predict return, it is unclear whether it does the same to volatility. Balcilar et al. (2017) find that trading volume can not predict volatility, while Aalborg et al. (2018) and Mai et al. (2018), on the other hand, find that trading volume can improve the volatility prediction model. Figa-Talamanca and Patacca (2018) also find trading volume to affect the conditional variance of Bitcoin returns, and Google trends too, by applying non-linear models. In summary, volatility appears to be impacted by return and Google trends. On the other hand, the impact of sentiment and trading volume is inconclusive.

**Trading volume**

Limited research has been done when it comes to predicting trading volume with sentiment analysis. McAteer (2014) find the number of tweets about Bitcoin to correlate with trading volume strongly. Mai et al. (2018) attempt to predict trading volume using the sentiment of news articles, Twitter, and forum post. However, they report no significant results. Jerdack et al. (2018) use the tool Natural Language Understanding to estimate the sentiment of news articles posted on a Facebook group related to Bitcoin. They find no significant impact of the news' sentiment on the trading volume of Bitcoin. While few studies have considered sentiment's impact on trading volume of cryptocurrencies, some studies explore other predicting factors. Blau (2017) finds no correlation between the trading volume of Bitcoin with its return or volatility. Glaser et al. (2014) find that trading volume was neither predicted by return nor transaction volume. Mai et al. (2018) find high transaction volume to predict high trading volume, while Google trends had no significant predictive ability in their research. This last result contrasts Aalborg et al. (2018), who find that Google trends and transaction volume do predict trading volume on weekly data. In general, most studies find Google trends to be a predictor of trading volume (Matta et al. 2015b; Jerdack et al. 2018; Ghiasvand 2018, Nasir et al. 2019). In conclusion, the sentiment in social media does not appear to influence trading volume. Also, while there are some disagreements, most studies find that Google trends and transaction volume predict trading volume.

**Transaction volume**

In an attempt to predict transaction volume of Bitcoin, Ether, and Ripple, Kim et al. (2016) perform sentiment analysis on user comments in online communities using the vader-dictionary. They find that user comments and replies in cryptocurrency communities can predict the number of transactions to some extent, and the significance is greatest for Bitcoin. They argue that the greater significance is a consequence of the relatively larger size of the Bitcoin communities. Mai et al. (2018) apply sentiment analysis both on forum- and Twitter posts. They find that positive forum posts are a significant predictor of transaction volume. On the other hand, their measure for news sentiment does not predict transaction volume. Moreover, significant predictors of transaction volume in their findings include trading volume and VIX-index, whereas Google trends, volatility, and return do not predict. Glaser et al. (2014) also find that return do not predict transaction volume, whereas Koutmos (2018) observe the opposite. Lastly, Glaser et al. (2014) does not find transaction volume to be predicted by trading volume. To summarize, the sentiment of user comments influences the transaction volume. The ability of return and trading volume to predict transaction volume are inconclusive.

**Table 2.1:** Key characteristics of included quantitative research with sentiment analysis.

| Authors | Name | Dep. Var | Dictionary | Source | Cryptocurrency | Timeperiod | Frequency | Findings |
|---|---|---|---|---|---|---|---|---|
| Kaminski (2014) | Nowcasting the Bitcoin Market with Twitter Signals | Price | Custom model | Twitter | Bitcoin | 23/11/2013-07/03/2014 | Daily | Tweets mirror market rather than predict it. |
| McAteer (2014) | Twitter Sentiment Analysis to Predict Bitcoin Exchange Rate | Price | Custom model | Twitter | Bitcoin | 28/07/2014-10/08/2014 | Daily | Positive correlation between Twitter sentiment and Bitcoin price. Tweet volume strongly correlate trading volume. |
| Matta et al. (2015a) | Bitcoin Spread Prediction Using Social And Web Search Media | Price | Sentistrength | Twitter | Bitcoin | 01/2015-03/2015 | Daily | Significant correlation between Bitcoin price and Google searches. |
| Garcia and Schweitzer (2015) | Social signals and algorithmic trading of Bitcoin | Price | Warriner et al. (2013) | Twitter | Bitcoin | 01/02/2011-31/12/2014 | Daily | Increases in opinion polarization and exchange volume precede rising Bitcoin prices, and valence precedes opinion polarization and rising exchange volumes. |
| Polasik et al. (2015) | Price Fluctuations and the Use of Bitcoin: An Empirical Inquiry | Price | Henry (2008) | Nexis news | Bitcoin | 04/2011-03/2014 | Monthly | Bitcoin returns are driven primarily by Bitcoin's popularity, the sentiment expressed in newspaper reports on cryptocurrency, and total number of transactions. |
| Kim et al. (2016) | Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies | Price,<br>Transaction volume | Hutto and Gilbert (2014) | BitoinTalk,<br>Ethereum.org,<br>XRP Chat | Bitcoin,<br>Ether,<br>Ripple | 01/12/2013-08/02/2016 | Daily | User comments and replies in online communities predict the number of transactions among users. |
| Bukovina et al. (2016) | Sentiment and Bitcoin Volatility | Volatility | SentDex | Reddit | Bitcoin | 12/12/2013-31/12/2015 | Daily | The explanatory power of sentiment significantly increases during the period of excessive volatility. Positive sentiment is more influential for Bitcoin excessive volatility. |
| Stenqvist and Lönnö (2017) | Predicting Bitcoin price fluctuation with Twitter sentiment analysis | Price | Hutto and Gilbert (2014) | Twitter | Bitcoin | 11/05/2017-11/06/2017 | Intraday | Twitter sentiment improves prediction model for Bitcoin price. |
| Abraham et al. (2018) | Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis | Price | Hutto and Gilbert (2014) | Twitter | Bitcoin,<br>Ether | 04/03/2018-03/06/2018 | Daily | Tweet volume, rather than tweet sentiment, is a predictor of price direction. |
| Steinert and Herff (2018) | Predicting altcoin returns using social media | Price | Hutto and Gilbert (2014) | Twitter | 181 currencies | 21/03/2017-04/06/2017 | Intraday | Short-term returns can be predicted from activity and sentiments on Twitter. |
| Ghiasvand (2018) | Investigating cryptocurrencies' return, exchange volume and volatility with investor's attention and investor sentiment: An empirical analysis | Price,<br>Volatility,<br>Trading volume | Hutto and Gilbert (2014) | Reddit | Bitcoin,<br>Ether,<br>Litecoin | 01/09/2017-07/05/2018 | Daily | Consistent long run relationship between search queries and the cryptocurrencies' exchange volume and volatility. The sentiment proxy gave inconsistent results, showing no unilateral direction towards return, exchange volume nor volatility. |
| Mai et al. (2018) | How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis | Price,<br>Volatility,<br>Trading volume,<br>Transaction volume | Loughran and McDonald (2014) | Twitter,<br>BitcoinTalk,<br>Reuters | Bitcoin | 01/01/2012-31/12/2014 | Daily | Social media sentiment is an important predictor in determining Bitcoin's valuation. |
| Jerdack et al. (2018) | Understanding What Drives Bitcoin Trading Activities | Trading volume | IBM Watson | Facebook | Bitcoin | 24/07/2017-19/04/2018 | Daily | The sentiment of news articles published on Facebook-page does not predict trading volume. |

**Table 2.2:** Key characteristics of included quantitative research without sentiment analysis.

| Authors | Name | Dep. Var | Cryptocurrency | Timeperiod | Frequency | Findings |
|---|---|---|---|---|---|---|
| Kristoufek (2013) | BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era | Price | Bitcoin | 01/05/2011-30/06/2013 | Daily | Search queries is related with Bitcoin price with a strong asymmetry between the effect of an increased interest in the currency while being above or below its trend value. |
| Glaser et al. (2014) | Bitcoin - Asset or Currency? Revealing Users' Hidden Intentions | Trading volume, Transaction volume | Bitcoin | 01/01/2011-08/10/2013 | Daily | Trading volume is neither predicted by return nor transaction volume. |
| Ciaian et al. (2016) | The economics of Bitcoin price formation | Price | Bitcoin | 01/01/2009-01/01/2014 | Daily | Wikipedia searches impact price, macro-financial indicators do not drive Bitcoin price. |
| Matta et al. (2015b) | The predictor impact of Web search media on Bitcoin trading volumes | Trading volume | Bitcoin | 06/2014-07/2015 | Daily | Significant correlation between trading volume and Google searches. |
| Dyhrberg (2016) | Bitcoin, gold and the dollar – A GARCH volatility analysis | Volatility | Bitcoin | 19/07/2010-22/05/2015 | Daily | Bitcoin shares several similarities with gold and dollar indicating hedging capabilities and advantages as a medium of exchange. |
| Meland and Øyen (2017) | Explaining Bitcoin's price fluctuations | Price | Bitcoin | 18/09/2011-05/02/2017 | Weekly | Trading volume of Bitcoin has a negative relationship with Bitcoin price. Google trends has a positive relationship with Bitcoin price. |
| Balcilar et al. (2017) | Can volume predict Bitcoin returns and volatility? A quantiles-based approach | Price | Bitcoin | 19/12/2011-25/04/2016 | Daily | Trading volume can predict return. |
| Blau (2017) | Price dynamics and speculative trading in bitcoin | Price, Volatility | Bitcoin | 17/07/2010-01/06/2014 | Daily | During 2013, speculative trading would not contribute to the unprecedented rise and subsequent crash in Bitcoin's value. Nor is speculative trading directly associated with Bitcoin's unusual level of volatility. |
| Sovbetov (2018) | Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero | Price | Bitcoin, Ether, Dash, Litecoin, Monero | 01/2010-01/2018 | Weekly | Trading volume and volatility appear to be significant determinant for all five cryptocurrencies. |
| Dickerson (2018) | Algorithmic Trading of Bitcoin Using Wikipedia and Google Search Volume | Price | Bitcoin | 01/07/2015-03/03/2018 | Daily | A highly profitable trading strategies can be constructed based on Google- and Wikipedia searches. |
| Kjærland et al. (2018) | An Analysis of Bitcoin's Price Dynamics | Price | Bitcoin | 01/01/2013-20/02/2018 | Weekly | Technological factor Hashrate is irrelevant for modeling Bitcoin price dynamics. Bitcoin price is affected by S&P500 and Google trends, while not by VIX, oil, gold, nor transaction volume. |
| Figa-Talamanca and Patacca (2018) | Does market attention affect Bitcoin returns and volatility? | Price, Volatility | Bitcoin | 01/01/2012-31/12/2017 | Daily | Trading volume related measures affect both the mean and the conditional variance of Bitcoin returns while internet searches volume mainly affects the conditional variance of returns. |
| Aalborg et al. (2018) | What can explain the price, volatility and trading volume of Bitcoin? | Price, Volatility, Trading volume | Bitcoin | 01/03/2012-19/03/2017 | Daily, Weekly | Trading volume improves volatility model and can be predicted from Google searches. |
| Nasir et al. (2019) | Forecasting cryptocurrency returns and volume using search engines | Price, Trading volume | Bitcoin | 01/2014-12/2017 | Weekly | Google searches leads to positive returns and a surge in Bitcoin trading volume. |
| Koutmos (2018) | Bitcoin returns and transaction activity | Price, Transaction volume | Bitcoin | 02/01/2013-20/09/2017 | Daily | A shock to transaction activity leads to positive return, followed by a reversal in price behavior. The contribution of return shocks to transaction activity is larger in magnitude. |

# 3 Methodology

While most studies on sentiment analysis for cryptocurrencies are based on social media, few studies consider the sentiment of news articles. This paper extends the current literature by investigating whether sentiment in news articles can be used to predict financial factors of the cryptocurrencies Bitcoin, Ether, and Litecoin. We also investigate whether financial factors, as well as previously studied control variables, can be used to predict sentiment in news articles. The four financial variables investigated for each currency were return, volatility, trading volume, and transaction volume. Three dictionaries are used to evaluate the sentiment of each article: tone-, vader, and valence-dictionary. All three dictionaries had shown promising results in previous papers when it came to predicting aspects of cryptocurrencies.

The relationship between the financial factors and the sentiment is investigated by using multiple-linear regressions, which included several control variables which had been widely used in the literature. These include Google trends data, new addresses on the blockchain and the number of published articles. The number of articles was included as a variable because tweet volume, an analogy from social media, had shown promising results in the literature.

The variables are defined in the Chapter 4, and the regression models for predicting the financial variables are presented in the Chapter 5. This chapter deals with the methodology used to score the sentiment of an article. The following steps were taken to perform sentiment analysis on news articles:

1. Collecting the news articles
2. Choosing dictionary and sentiment method
3. Preprocessing of news articles
4. Performing sentiment analysis

Each step is described in the sections to follow.

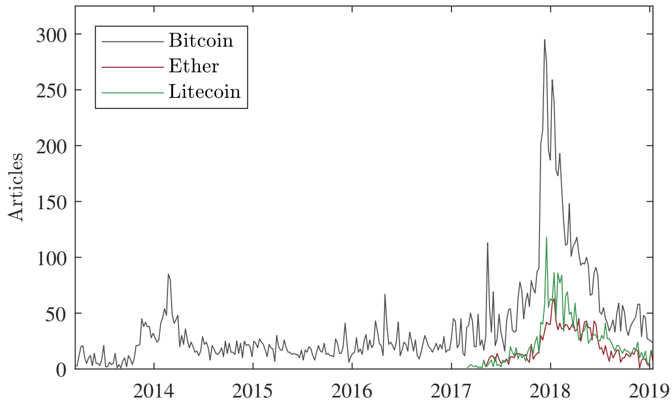



**Figure 3.1:** Number of articles published each week over the period 2013-2019.

## 3.1 Collecting news articles

The news articles used to perform sentiment analysis were downloaded from ProQuest[1]. Their news database consists of publications from more than 3,000 news publishers worldwide including *The Wall Street Journal*, *Financial Times*, and *The Australian Financial Review*. The database was accessed 25/01/2019 and the number of articles for each cryptocurrency (Bitcoin, Ether, and Litecoin) is shown in Figure 3.1. The following search parameters were chosen:

1. Language: English
2. Source category: Newspapers
3. Limit: Full-text articles only
4. Keyword: "Bitcoin", "Ether" or "Litecoin"

Only one keyword was used at a time, creating unique datasets for Bitcoin, Litecoin, and Ether. Frequently, the same article is published by several different publishers. To avoid redundancy in the dataset, the duplicates are removed, creating a final dataset of unique articles. The total number of articles for each keyword, both before and after the removal of duplicates, are given in Table 3.1.

**Table 3.1:** Number of articles downloaded from ProQuest and the number of articles which remained after removing duplicates.

| Keyword | Downloaded | After removing duplicates |
|---|---:|---:|
| Bitcoin | 14,889 | 11,858 |
| Ether | 2,362 | 1,994 |
| Litecoin | 3,437 | 2,945 |

[1]https://www.proquest.com/

Some publishers released articles more frequently than others. A list of the top ten publishers from the Bitcoin dataset is presented in Table 3.2, with the number of unique articles published by each publisher. As shown in the table, a moderate number of large publishers contributed with most of the articles. However, in total, there were more than 240 unique publishers in the dataset.

**Table 3.2:** Top ten publishers of news articles with keyword "Bitcoin".

| Publisher | Published articles |
|---|---|
| Wall Street Journal (Online) | 1,753 |
| NASDAQ OMX's News Release Distribution Channel | 1,219 |
| Financial Times | 1,081 |
| University Wire | 859 |
| Investor's Business Daily | 755 |
| American Banker | 697 |
| Asia News Monitor | 678 |
| The Australian Financial Review | 508 |
| Wall Street Journal, Eastern edition | 454 |
| National Post | 320 |

## 3.2 Dictionary-based approach

The dictionary-based approach consists of three steps: feature extraction, feature scoring and score aggregation (Medhat et al., 2014).

Feature extraction refers to the extraction of the components in a text which is used to measure the sentiment. While most dictionary-based methods extract the words in a text, some methods also extract punctuation, emoticons and emojis, negations, and degree modifiers. Extracting punctuation and emoji/emoticons may, for example, put more accurate emphasis on the text. For example, it is reasonable to believe that *Excellent!* has a stronger positive sentiment than just *Excellent* without the exclamation mark. Similar reasoning applies for emoji/emoticons, but these are more likely to be relevant for sentiment analysis in social media rather than news articles. Additional relevant feature extractions include negation and degree modifiers. Negation refers to the use of negative words such as *no, none, not, never* before other words in a sentence which may change the orientation of the message. A dictionary that extracts negation would thus be able to interpret *not good* as equivalent to *bad*, whereas simpler dictionaries would treat the phrase word by word and potentially giving a less accurate sentiment measure. Extracting degree modifiers can also be relevant in order to give a stronger sentiment score to phrases such as *very risky*, instead of potentially just capturing the word *risky*. Preprocessing of the text, discussed further in Section 3.3, is a critical step in order to extract these features.

The second step of the dictionary-based approach, feature scoring, involves giving each extracted feature a score. Scores for words and emoji/emoticons are defined in the dictionary used. Simpler dictionaries use binary scoring, whereas more sophisticated ones use a scale. Binary scoring could, for example, classify words as either positive (1) or negative (-1). A scale implies giving a score to each feature extracted a value within a range, for example -4 to 4. If the dictionary extracts negation, the score of the word would be given the opposite sign. To account for degree modifiers in front of a word, the original word score can either be up- or down scaled. The sentiment scores for words (and emoji/emoticons depending on the dictionary) can be classified by either the author of the dictionary or by a group of people. When the scorings are classified by the latter, it is common to take the average score. The final step is score aggregation, which refers to how the extracted scores of a text should be combined. Simple dictionaries aggregate the scores feature by feature, whereas other dictionaries aggregate based on more complex relations.

Thus, the complexity of these steps varies depending on the dictionary being used. This paper used three approaches to capture the sentiment of an article: tone, vader, and valence-dictionary. While the tone- and valence-dictionary utilize the bag-of-words model, the vader-dictionary includes more sophisticated textual understanding, which in some instances considers the orders of words in a sentence. In general, there are differences between the dictionaries in the feature extraction, feature scoring, and score aggregation. Table 3.3 provides a summary of the differences between the methods. Sections 3.2.1-3.2.3 discuss each dictionary in detail.

**Table 3.3:** A feature-level comparison of the three dictionaries utilized in this paper.

| Dictionary | Feature extraction | Feature scoring | Score aggregation |
|---|---|---|---|
| Tone | Words | Binary: {-, +} | Word |
| Vader | Words, emoticons/emojis, punctuation, negation, degree modifiers | Scale: -4 to 4 | Sentence |
| Valence | Words | Scale: 1 to 10 | Word |

### 3.2.1 Tone

The tone-dictionary was designed by Henry (2008) to perform sentiment analysis on financial reports. It contains 198 financially related words. The feature scoring of the tone-dictionary is binary, defining words as either positive or negative. Of the included 198 words, 112 are classified as positive, while the remaining 86 words are negative. The full list of the 198 words is included in Appendix A. As introduced in Section 2.1, the tone-dictionary showed promising results for Bitcoin-related news in the study by Polasik et al. (2015).

The tone-dictionary method extracts the words from an article. Score aggregation is based on the words in the article that also match the words in the dictionary. The sentiment score of an article $i$ using the tone-dictionary is defined as $S_i^{Tone}$. It is calculated by finding the ratio between the difference in the number of positive and negative words $(p_i - n_i)$ and the total amount of matched words ($p_i + n_i$), in article $i$. This ratio is shown in Eq. (3.1). Thus, using the definition of $S_i^{Tone}$, the sentiment of an article based on the tone-dictionary will always have a value between -1 and +1, where the end points suggest that the article has an extremely negative or positive tone respectively.

$$S_i^{Tone} = \frac{p_i - n_i}{p_i + n_i} \tag{3.1}$$

### 3.2.2 Vader

The vader-dictionary (VADER, Valence Aware Dictionary and sEntiment Reasoner) was developed by Hutto and Gilbert (2014) to be used as a measure of sentiment. While it was specifically designed to analyze social media, it has shown promising results in other textual environments. The dictionary contains more than 7,000 scored words, emojis, and emoticons. Each word (or emoji/emoticon) in the dictionary was scored by ten individuals, on a scale from -4 to 4. A word would only be included in the dictionary if the standard deviation of the responses was less than 2.5, in which the word was given the average score of the individuals' scoring. In addition to considering the words (and potential emojis/emoticons), the method also extracts punctuation, negation and degree modifiers in the sentence before scoring it, as summarized in Table 3.3. Each sentence is given a sentiment score, computed by summing the sentiment score of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). The dictionary had shown promising results in several papers (Kim et al. 2016; Stenqvist and Lönnö 2017; Steinert and Herff 2018; Mai et al. 2018).

Let $C_{i,j}$ be defined as the normalized compound score of sentence $j$ in article $i$. The sentiment of each article is found by taking the average of the normalized compound scores of all the sentences in an article. However, Hutto and Gilbert (2014) recommend to only include the normalized sentence scores with absolute values of greater than 0.05. Thus, if $x_i$ is the total number of sentences having an absolute compound score above 0.05 in article $i$, the sentiment of an article based on the vader-dictionary, $S_i^{Vader}$, is calculated as shown in Eq. 3.2.

$$S_i^{Vader} = \frac{\sum_{j=1}^{x_i} C_{i,j}}{x_i}, |C_{i,j}| > 0.05 \tag{3.2}$$

As the vader-dictionary aggregates sentiment scores on a sentence-level, it was necessary to find a tool which could split the articles into sentences. The Natural Language Toolkit[2] (NLTK) provides such a tool. It is an established library for building Python-based programs when analyzing human language data. The library includes the function

[2]https://www.nltk.org/index.html/

*Tokenizer*, which was created with the specific purpose to divide longer texts into individual sentences. The library was recommended by Hutto and Gilbert (2014) to be used when analyzing larger text using the vader-dictionary.

### 3.2.3 Valence

The valence-dictionary developed by Warriner et al. (2013) contains nearly 14,000 scored words, making it one of the largest databases of sentiment scores available. The words were scored by individual raters, with the majority of the words being rated by obtaining responses from over 18 individuals (and some cases of up to 70 responses). The words are scored on a scale from 1 to 10. To use the valence-dictionary, one extracts all words from an article and aggregates the dictionary score of every word found in the article that matches the dictionary. Then, the average score of the words are recorded as the sentiment score of the text. This method showed promising results in the research performed by Garcia and Schweitzer (2015) as introduced in Section 2.1.

The sentiment score of an article $i$ using the valence-dictionary is defined as $S_i^{Valence}$. As shown in Eq. 3.3, it is calculated by finding the average score of each word in the article also matched in the dictionary. The notation $V_{i,k}$ is the dictionary score of word $k$ found in article $i$. The total number of matched words in the article is $y_i$.

$$S_i^{Valence} = \frac{\sum_{k=1}^{y_i} V_{i,k}}{y_i} \tag{3.3}$$

## 3.3 Preprocessing of news articles

Preprocessing of text is regarded as a critical step in text classification and may improve classification accuracy significantly (Miner et al. 2012; Angiani et al. 2016; Uysal and Gunal 2014). Some key steps include:

1. *Choose text scope*: Use entire article as text or specific paragraphs etc.
2. *Remove unimportant characters*: Character which do not add meaning are removed.
3. *Tokenize*: Break text into discrete words.
4. *Normalize case*: Covert text to either all upper- or lower case.
5. *Detect sentences*: Apply tools to recognize sentences of a text
6. *Consider negations*: Treat negations like *not good* as *bad*
7. *Consider degree modifiers*: Account for phrases like *very good* and not just *good*

Following step 1, the scope of the text was an entire article. Moreover, a step which is not listed above involves correcting misspellings. However, in the case of this paper, it was a minor issue as the analyzed newspaper articles had few spelling errors. Another step not listed above is changing words with prefixes and suffixes to their original state, without prefixes or suffixes. For example, *pay*, *paying*, and *pays* would all be changed to *pay*. However, this step is only relevant in dictionaries which do not have words with prefixes and suffixes in their lexicon. Since the tone-dictionary already contained words with prefixes and suffixes, the step was not considered.

Some dictionaries have in-built preprocessing techniques, whereas others do not. When it comes to the vader-dictionary, by applying the NLTK-library introduced in Section 3.2.2 to detect sentences, all of the enumerated preprocessing steps are achieved. On the other hand, the tone- and valence-dictionaries have no preprocessing techniques built in, so they had to be developed. For these two dictionaries, the steps given in 1-4 were performed. The MATLAB code is given in Appendix B.

## 3.4 Performing sentiment analysis

The MATLAB code used to perform sentiment analysis with the tone- and valence-dictionary, and the Python code used to run the Vader program, can all be found in Appendix B. While the tone- and valence-dictionaries did not have any code readily available online, the vader-dictionary already had a published Python program[3] which was used.

The distributions of sentiment scores of the articles for each respective dictionary are shown in Figure 3.2. Noticeably, the distributions of the sentiment scores for each article are positively skewed for all dictionaries. In addition, the differences between the distributions of the cryptocurrencies are minor. Moreover, the distribution of scores using the tone-dictionary is different in comparison to using the other two dictionaries. While the vader- and valence-dictionary appear to be more bell-curved, the tone-dictionary is more uniformly distributed. This is likely due to the small amount of words included in the tone-dictionary, and the binary scoring of the words.

**Table 3.4:** Sample words and their dictionary score, based on tone, vader and valence.

| Word | Tone | Vader (lexicon) | Valence |
|---|---|---|---|
| positive | 1 | 2.6 | 7.57 |
| beat | 1 | N/A | 4.38 |
| negative | -1 | -2.7 | 2.52 |
| risk | -1 | -1.1 | 5.17 |
| room | N/A | N/A | 5.55 |

To get an intuition of how the three chosen dictionaries work in practice, some examples of measuring the sentiment of words and sentences are presented in Table 3.4 and Table 3.5 respectively. The first table presents words and their dictionary score. Since the samples are words, the normalized compound score from the vader-dictionary is not used, but rather the lexicon score for each word. For example, the word *positive* has a tone score of 1,

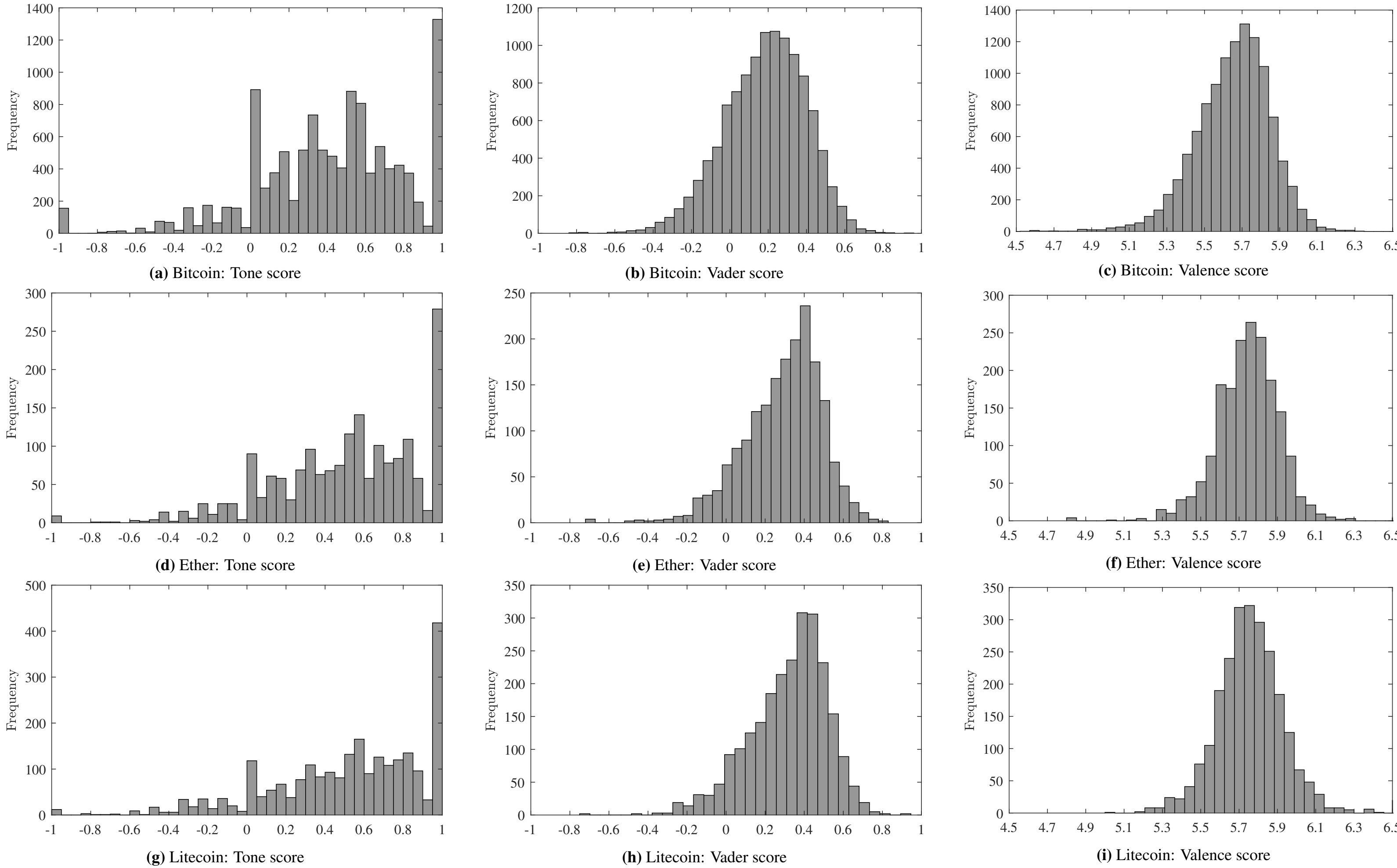
[3]https://github.com/cjhutto/vaderSentiment/

vader score of 2.6, and valence score of 7.57. The tone-dictionary has a binary score, which means that a score of 1 classifies the word as positive. Vader's lexicon values range from -4 to 4, which means that a score of 2.6 is in the upper range. Valence gives a score between 1 to 10, and naturally the word *positive* falls within the upper range. The word *risk* on the other hand, which has more negative connotations, is given negative scores for both tone and vader, whereas the valence score is 5.17. Lastly, since the dictionaries have different amounts of words in their lexicons, with the tone-dictionary only containing 198 words, some words are not found in the dictionaries. For example, the word "room" is missing in both the tone- and vader-dictionary.

Table 3.5 presents three example sentences designed to be positive, negative, and neutral. It can be observed that all dictionaries appear to give meaningful sentiment scores, with all dictionaries reaching its upper region for the positive sentence. Interestingly, not all the dictionaries classify the second sentence as the most negative, as the vader-dictionary classifies the third sentence as the most negative instead. This example illustrates that the dictionaries can give different results about the sentiment of a sentence, which highlights why it is interesting to consider using different dictionaries for comparison.

**Table 3.5:** Sample sentences and their sentiment score, based on tone, vader and valence.

| Sentence | Tone | Vader (compound) | Valence |
|---|---|---|---|
| Bitcoin is performing great with large price increases in recent days | 1.00 | 0.62 | 5.91 |
| Decline in investor faith in Bitcoin after China banned cryptocurrencies | -1.00 | -0.05 | 5.02 |
| Risky investments in cryptocurrencies may lead to high return | 0.00 | -0.20 | 5.34 |




**(a)** Bitcoin: Tone score
**(b)** Bitcoin: Vader score
**(c)** Bitcoin: Valence score
**(d)** Ether: Tone score
**(e)** Ether: Vader score
**(f)** Ether: Valence score
**(g)** Litecoin: Tone score
**(h)** Litecoin: Vader score
**(i)** Litecoin: Valence score

**Figure 3.2:** Distributions of sentiment scores of individual articles based on keywords "Bitcoin", "Ether", and "Litecoin". Time periods: 17/11/2013 - 13/01/2019 for Bitcoin, and 30/04/2017 - 13/01/2019 for Ether and Litecoin.

# 4 Data

The data used in this paper was collected from BitInfoCharts[1], Blockchain.com[2], CoinMarketCap[3], Etherscan[4], Google Trends[5], ProQuest[6], and Quandl[7]. The data was collected on the 25/01/2019, and span the time period from 17/11/2013 to 19/01/2019 for Bitcoin, and 30/04/2017 to 19/01/2019 for Ether and Litecoin. As Bitcoin has been in existence for a longer time, there are more data available. To make the data comparable between cryptocurrencies, the dataset for Bitcoin was split into three time periods of equal length, where the last period matched the time period covering Ether and Litecoin. The time span of each period are presented in Table 4.1. The following sections describe how each variable was constructed.

**Table 4.1:** Time span of each dataset.

| Dataset | Time period |
|---|---|
| Bitcoin period 1 | 17/11/2013 - 08/08/2015 |
| Bitcoin period 2 | 09/08/2015 - 29/04/2017 |
| Bitcoin period 3 | 30/04/2017 - 13/01/2019 |
| Ether | 30/04/2017 - 13/01/2019 |
| Litecoin | 30/04/2017 - 13/01/2019 |

## 4.1 Articles

The articles were downloaded from ProQuest and the total number of published articles in week $t$ is denoted as $n_t$. $Articles_t$ is the standardized value of published articles in week $t$. Using the average value over the past 8 weeks, and the standard deviation $\sigma_t$ over the same period, $Articles_t$ was calculated as shown in Eq. (4.1).

[1] https://bitinfocharts.com/
[2] https://www.blockchain.com/
[3] https://coinmarketcap.com/
[4] https://etherscan.io/
[5] https://trends.google.com/
[6] https://www.proquest.com/
[7] https://www.quandl.com/

$$Articles_t = \frac{n_t - \frac{1}{8}\sum_{i=1}^{8} n_{t-i}}{\sigma_t} \tag{4.1}$$

## 4.2 Tone

$Tone_t$, as presented in Eq. (4.2), is the average tone score in week $t$, where $n_t$ is the total number of published articles in week $t$, and $S_i^{Tone}$ is the sentiment score, using the tone-dictionary, of article $i$.

$$Tone_t = \frac{\sum_{i=1}^{n_t} S_i^{Tone}}{n_t} \tag{4.2}$$

## 4.3 Vader

$Vader_t$, as presented in Eq. (4.3), is the average vader score in week $t$, where $n_t$ is the total number of published articles in week $t$, and $S_i^{Vader}$ is the sentiment score, using the vader-dictionary, of article $i$.

$$Vader_t = \frac{\sum_{i=1}^{n_t} S_i^{Vader}}{n_t} \tag{4.3}$$

## 4.4 Valence

$Valence_t$, as presented in Eq. (4.4), is the average valence score in week $t$, where $n_t$ is the total number of published articles in week $t$, and $S_i^{Valence}$ is the sentiment score, using the valence-dictionary, of article $i$.

$$Valence_t = \frac{\sum_{i=1}^{n_t} S_i^{Valence}}{n_t} \tag{4.4}$$

## 4.5 Addresses

The total number of unique addresses registered on the blockchain in week $t$ is defined as $A_t$. Notice that $A_t$ will be strictly increasing, as used addresses cannot be removed from the blockchain. To consider the growth rate of new addresses, $Addresses_t$ is defined as the logarithmic weekly change of unique addresses recorded on the blockchain, as shown in Eq. (4.5).

$$Addresses_t = \log(A_t) - \log(A_{t-1}) \tag{4.5}$$

The data was downloaded from Quandl and Etherscan for Bitcoin and Ether respectively. Unfortunately, as there were no available sources for the number of addresses of Litecoin, $Addresses_t$ was not included for Litecoin.

## 4.6 Google trends

Google Trends provides a normalized index value for the number of Google searches of a specific search phrase, within a specified period of time. It uses a standardized, integer scale from 0 - 100, where 100 represents the highest query volume within a chosen time period. To standardize the index value $G_t$, the method used in Bijl et al. (2016) was utilized, as shown in Eq. (4.6), where $\sigma_t$ is the standard deviation over the past 8 weeks. Subscript $i$ is used to calculate the average value over the past 8 weeks. The standardized value is referred to as $GoogleTrends_t$.

$$GoogleTrends_t = \frac{G_t - \frac{1}{8}\sum_{i=1}^{8} G_{t-i}}{\sigma_t} \tag{4.6}$$

The keywords used to download the data were "Bitcoin", "Ether" and " Litecoin". The data was downloaded in segments of one and a half years at a time, with half a year overlap. For example, the first interval for Bitcoin span from 17/03/2013 - 20/09/2014, the second from 17/03/2014 - 20/09/2015, and so on. This way, it is possible to increase the resolution of the data, which would be lacking if a longer time period was chosen. By using a process known as stitching, which involves having an overlap between the downloaded time periods to scale all time periods to the same scale, it was possible to keep the greater resolution of the results given by downloading several, shorter time periods, while maintaining the longer time span covered.

## 4.7 Return

Daily prices were downloaded from CoinMarketCap. Subscript $t$ is the weekly time-unit. The weekly return was calculated as shown in Eq. (4.7), where $P_{t-1}$ is the closing price in week $t-1$, and $P_t$ is the closing price in week $t$.

$$Return_t = log(P_t) - log(P_{t-1}) \tag{4.7}$$

## 4.8 Volatility

Volatility was calculated using the volatility estimator by Garman and Klass (1980) as discussed in Molnár (2012). The data for prices was downloaded from CoinMarketCap. Subscript $d$ refers to the day in the week, where daily closing price ($close_d$), opening price ($open_d$), high ($high_d$) and low ($low_d$) values are provided. Daily variance, $\sigma_d^2$ , was first calculated as shown in Eq. (4.8).

$$\sigma_d^2 = \frac{1}{2}(h_d - l_d)^2 - c_d^2(2log(2) - 1) \tag{4.8}$$

where

$$c_d = log(close_d) - log(open_d)$$
$$l_d = log(low_d) - log(open_d)$$
$$h_d = log(high_d) - log(open_d)$$

Finally, weekly volatility was calculated as the square root of the average daily variance in week $t$, as shown in Eq. (4.9):

$$Volatility_t = \sqrt{\frac{1}{7}\sum_{d=1}^{7}\sigma_d^2} \tag{4.9}$$

## 4.9 Trading volume

Trading volume $V_t$ is the total USD value of all trades of a specific cryptocurrency on major cryptocurrency exchanges within a given week $t$. The data was downloaded as daily data from CoinMarketCap, and weekly data was created as the sum of the daily values within a week. The data was standardized as shown in Eq. (4.10), using the average value and standard deviation over the past 8 weeks, and the standardized value is referred to as $TradingVolume_t$. The standard deviation of $V_t$ is denoted as $\sigma_t$, and it is calculated over the past 8 weeks.

$$TradingVolume_t = \frac{V_t - \frac{1}{8}\sum_{i=1}^{8} V_{t-i}}{\sigma_t} \tag{4.10}$$

## 4.10 Transaction volume

Transaction volume $T_t$ is the USD value of all transactions registered on the blockchain from one address to another of a specific cryptocurrency in week $t$. The data for Bitcoin was downloaded from Blockchain.info, and from BitInfoCharts for Ether and Litecoin. Transaction volume was standardized using the average value and standard deviation over the past 8 weeks, and the standardized value is referred to as $TransactionVolume_t$. $\sigma_t$ is the standard deviation of the $T_t$, calculated from the past 8 weeks.

$$TransactionVolume_t = \frac{T_t - \frac{1}{8}\sum_{i=1}^{8} T_{t-i}}{\sigma_t} \tag{4.11}$$

## 4.11 Descriptive statistics

The descriptive statistics and correlation matrices are presented for all five datasets. Following these tables, the time evolution of the variables are presented in Figure 4.1.

**Table 4.2:** Descriptive statistics of the Bitcoin variables during period 1, 17/11/2013 - 08/08/2015.

| **Bitcoin Period 1** | Mean | Median | Maximum | Minimum | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Tone | 0.393 | 0.392 | 0.634 | 0.056 | 0.110 | -0.300 | 3.193 |
| Vader | 0.159 | 0.160 | 0.318 | 0.029 | 0.059 | -0.112 | 3.057 |
| Valence | 5.658 | 5.668 | 5.792 | 5.508 | 0.062 | -0.306 | 2.873 |
| Articles | 0.023 | -0.194 | 4.485 | -2.994 | 1.455 | 0.597 | 3.446 |
| Addresses | 0.009 | 0.007 | 0.050 | 0.004 | 0.009 | 3.335 | 14.517 |
| GoogleTrends | 0.148 | -0.446 | 8.437 | -2.138 | 1.810 | 2.162 | 8.586 |
| Return | -0.003 | -0.005 | 0.258 | -0.209 | 0.058 | 0.471 | 8.428 |
| Volatility | 0.029 | 0.019 | 0.252 | 0.005 | 0.039 | 4.126 | 22.129 |
| TradingVolume | -0.013 | -0.074 | 0.889 | -0.441 | 0.260 | 1.084 | 4.306 |
| TransactionVolume | 0.007 | -0.015 | 0.810 | -0.289 | 0.187 | 1.845 | 8.519 |

**Table 4.3:** Descriptive statistics of the Bitcoin variables during period 2, 09/08/2015 - 29/04/2017.

| **Bitcoin Period 2** | Mean | Median | Maximum | Minimum | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Tone | 0.448 | 0.458 | 0.708 | 0.134 | 0.103 | -0.268 | 3.550 |
| Vader | 0.188 | 0.185 | 0.399 | 0.014 | 0.066 | 0.150 | 3.949 |
| Valence | 5.650 | 5.651 | 5.858 | 5.467 | 0.066 | 0.286 | 3.973 |
| Articles | 0.198 | -0.289 | 4.878 | -2.011 | 1.470 | 1.359 | 4.781 |
| Addresses | 0.006 | 0.006 | 0.013 | 0.004 | 0.001 | 1.386 | 7.137 |
| GoogleTrends | 0.416 | -0.213 | 15.841 | -2.660 | 2.664 | 3.964 | 21.448 |
| Return | 0.008 | 0.007 | 0.096 | -0.082 | 0.031 | -0.119 | 3.979 |
| Volatility | 0.016 | 0.012 | 0.054 | 0.003 | 0.011 | 1.439 | 4.575 |
| TradingVolume | 0.066 | 0.028 | 0.968 | -0.362 | 0.259 | 1.086 | 4.536 |
| TransactionVolume | 0.037 | 0.033 | 0.482 | -0.241 | 0.112 | 0.617 | 4.909 |

**Table 4.4:** Descriptive statistics of the Bitcoin variables during period 3, 30/04/2017 - 13/01/2019.

| **Bitcoin Period 3** | Mean | Median | Maximum | Minimum | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Tone | 0.311 | 0.392 | 0.648 | -0.173 | 0.199 | -0.661 | 2.059 |
| Vader | 0.195 | 0.201 | 0.303 | -0.091 | 0.068 | -1.406 | 6.326 |
| Valence | 5.658 | 5.670 | 5.745 | 5.436 | 0.059 | -1.332 | 5.315 |
| Articles | -0.026 | -0.331 | 11.424 | -3.894 | 2.013 | 2.669 | 14.969 |
| Addresses | 0.004 | 0.004 | 0.017 | 0.002 | 0.002 | 2.424 | 11.363 |
| GoogleTrends | 0.151 | -0.094 | 5.022 | -4.616 | 1.836 | 0.751 | 3.770 |
| Return | 0.005 | 0.006 | 0.148 | -0.156 | 0.063 | 0.047 | 2.665 |
| Volatility | 0.026 | 0.024 | 0.062 | 0.003 | 0.013 | 0.839 | 3.424 |
| TradingVolume | 0.038 | 0.001 | 0.545 | -0.300 | 0.205 | 0.472 | 2.328 |
| TransactionVolume | 0.012 | 0.007 | 0.356 | -0.380 | 0.164 | 0.161 | 2.311 |

**Table 4.5:** Descriptive statistics of the Ether variables, 30/04/2017 - 13/01/2019.

| **Ether** | Mean | Median | Maximum | Minimum | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Tone | 0.515 | 0.527 | 0.757 | 0.140 | 0.132 | -0.277 | 2.649 |
| Vader | 0.286 | 0.292 | 0.447 | -0.140 | 0.098 | -1.491 | 6.574 |
| Valence | 5.746 | 5.755 | 6.028 | 5.384 | 0.083 | -1.009 | 9.174 |
| Articles | 0.087 | 0.020 | 3.781 | -5.056 | 1.647 | -0.228 | 3.364 |
| Addresses | 0.017 | 0.009 | 0.054 | 0.003 | 0.014 | 0.935 | 2.789 |
| GoogleTrends | -0.048 | -0.456 | 4.264 | -3.327 | 1.491 | 0.946 | 3.628 |
| Return | 0.003 | 0.000 | 0.178 | -0.186 | 0.079 | 0.163 | 2.891 |
| Volatility | 0.028 | 0.025 | 0.076 | 0.005 | 0.014 | 1.174 | 4.388 |
| TradingVolume | 0.069 | 0.027 | 0.791 | -0.468 | 0.248 | 0.766 | 3.566 |
| TransactionVolume | -0.007 | -0.029 | 0.611 | -0.751 | 0.276 | -0.021 | 3.352 |

**Table 4.6:** Descriptive statistics of the Litecoin variables, 30/04/2017 - 13/01/2019.

| **Litecoin** | Mean | Median | Maximum | Minimum | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Tone | 0.558 | 0.576 | 0.875 | 0.230 | 0.128 | -0.407 | 2.989 |
| Vader | 0.334 | 0.337 | 0.472 | 0.109 | 0.068 | -0.335 | 3.291 |
| Valence | 5.780 | 5.786 | 5.916 | 5.564 | 0.064 | -0.313 | 3.368 |
| Articles | 0.197 | -0.185 | 6.128 | -2.505 | 1.815 | 1.428 | 5.292 |
| GoogleTrends | 0.190 | -0.233 | 9.950 | -2.646 | 2.014 | 2.249 | 9.600 |
| Return | 0.003 | -0.005 | 0.284 | -0.158 | 0.084 | 0.860 | 4.235 |
| Volatility | 0.030 | 0.027 | 0.087 | 0.005 | 0.015 | 1.276 | 4.635 |
| TradingVolume | 0.048 | -0.003 | 1.032 | -0.617 | 0.287 | 0.672 | 3.951 |
| TransactionVolume | -0.052 | -0.349 | 2.901 | -2.047 | 1.317 | 0.467 | 2.126 |

**Table 4.7:** Correlation matrix between the Bitcoin variables during period 1, 17/11/2013 - 08/08/2015. The symbols *, and ** denote significance at the 5%, and 1% levels, respectively.

| **Bitcoin Period 1** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) Tone | 1.00 | | | | | | | | | |
| (2) Vader | 0.57** | 1.00 | | | | | | | | |
| (3) Valence | 0.47** | 0.74** | 1.00 | | | | | | | |
| (4) Articles | -0.08 | -0.18 | -0.05 | 1.00 | | | | | | |
| (5) Addresses | -0.17 | -0.09 | 0.08 | 0.33** | 1.00 | | | | | |
| (6) GoogleTrends | -0.20 | -0.11 | -0.07 | 0.52** | 0.33** | 1.00 | | | | |
| (7) Return | 0.28** | 0.10 | 0.07 | 0.09 | 0.16 | 0.08 | 1.00 | | | |
| (8) Volatility | -0.32** | -0.27* | -0.13 | 0.44** | 0.43** | 0.49** | -0.08 | 1.00 | | |
| (9) TradingVolume | -0.25* | -0.13 | -0.09 | 0.38** | 0.52** | 0.49** | 0.06 | 0.42** | 1.00 | |
| (10) TransactionVolume | -0.04 | 0.06 | 0.06 | 0.32** | 0.55** | 0.46** | 0.31** | 0.17 | 0.67** | 1.00 |

**Table 4.8:** Correlation matrix between the Bitcoin variables during period 2, 09/08/2015 - 29/04/2017. The symbols *, and ** denote significance at the 5%, and 1% levels, respectively.

| **Bitcoin Period 2** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) Tone | 1.00 | | | | | | | | | |
| (2) Vader | 0.42** | 1.00 | | | | | | | | |
| (3) Valence | 0.30** | 0.75** | 1.00 | | | | | | | |
| (4) Articles | -0.12 | -0.36** | -0.29** | 1.00 | | | | | | |
| (5) Addresses | 0.20 | 0.31** | 0.12 | 0.16 | 1.00 | | | | | |
| (6) GoogleTrends | -0.01 | -0.03 | 0.03 | 0.33** | 0.12 | 1.00 | | | | |
| (7) Return | 0.07 | 0.11 | 0.17 | -0.10 | -0.07 | 0.02 | 1.00 | | | |
| (8) Volatility | -0.01 | -0.10 | -0.15 | 0.27** | 0.06 | 0.35** | -0.17 | 1.00 | | |
| (9) TradingVolume | 0.10 | 0.03 | -0.09 | 0.16 | 0.19 | 0.28** | 0.01 | 0.56** | 1.00 | |
| (10) TransactionVolume | 0.13 | -0.10 | -0.09 | 0.23* | 0.12 | 0.32** | 0.15 | 0.52** | 0.62** | 1.00 |

**Table 4.9:** Correlation matrix between the Bitcoin variables during period 3, 30/04/2017 - 13/01/2019. The symbols *, and ** denote significance at the 5%, and 1% levels, respectively.

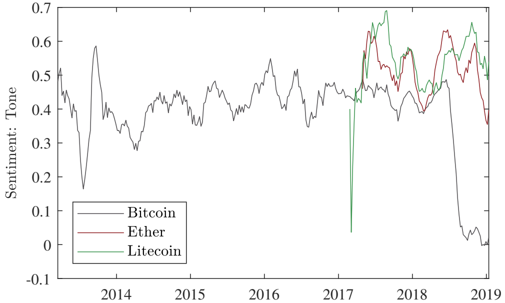
| **Bitcoin Period 3** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) Tone | 1.00 | | | | | | | | | |
| (2) Vader | 0.03 | 1.00 | | | | | | | | |
| (3) Valence | 0.20 | 0.87** | 1.00 | | | | | | | |
| (4) Articles | 0.05 | -0.38** | -0.25* | 1.00 | | | | | | |
| (5) Addresses | 0.33** | -0.21* | -0.14 | 0.41** | 1.00 | | | | | |
| (6) GoogleTrends | 0.09 | -0.27* | -0.18 | 0.54** | 0.35** | 1.00 | | | | |
| (7) Return | 0.17 | -0.09 | -0.01 | 0.20 | 0.17 | 0.04 | 1.00 | | | |
| (8) Volatility | 0.32** | -0.25* | -0.23* | 0.26* | 0.16 | 0.47** | -0.05 | 1.00 | | |
| (9) TradingVolume | 0.09 | -0.34** | -0.29** | 0.51** | 0.39** | 0.75** | 0.15 | 0.56** | 1.00 | |
| (10) TransactionVolume | 0.16 | -0.15 | -0.04 | 0.54** | 0.46** | 0.65** | 0.23* | 0.17 | 0.71** | 1.00 |

**Table 4.10:** Correlation matrix between the Ether variables, 30/04/2017 - 13/01/2019. The symbols *, and ** denote significance at the 5%, and 1% levels, respectively.

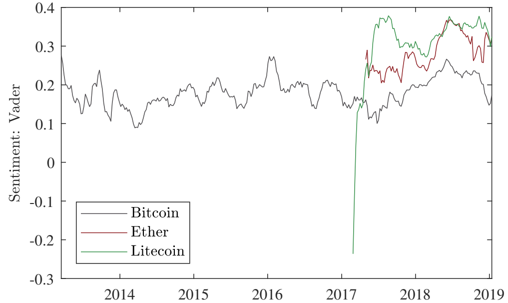
| **Ether** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) Tone | 1.00 | | | | | | | | | |
| (2) Vader | 0.46** | 1.00 | | | | | | | | |
| (3) Valence | 0.50** | 0.72** | 1.00 | | | | | | | |
| (4) Articles | 0.13 | -0.01 | 0.06 | 1.00 | | | | | | |
| (5) Addresses | 0.10 | -0.34** | -0.07 | 0.47** | 1.00 | | | | | |
| (6) GoogleTrends | -0.14 | -0.06 | -0.04 | 0.11 | 0.15 | 1.00 | | | | |
| (7) Return | 0.18 | -0.09 | -0.02 | 0.10 | 0.36** | -0.07 | 1.00 | | | |
| (8) Volatility | -0.19 | -0.35** | -0.27* | 0.22* | 0.49** | 0.32** | 0.11 | 1.00 | | |
| (9) TradingVolume | 0.11 | -0.14 | -0.02 | 0.39** | 0.66** | 0.51** | 0.42** | 0.58** | 1.00 | |
| (10) TransactionVolume | 0.23* | -0.16 | 0.05 | 0.39** | 0.62** | 0.35** | 0.37** | 0.36** | 0.81** | 1.00 |

**Table 4.11:** Correlation matrix between the Litecoin variables, 30/04/2017 - 13/01/2019. The symbols *, and ** denote significance at the 5%, and 1% levels, respectively.

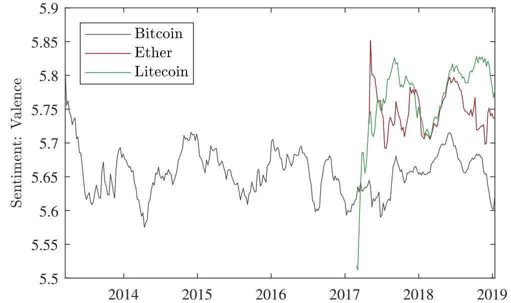
| **Litecoin** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| (1) Tone | 1.00 | | | | | | | | |
| (2) Vader | 0.62** | 1.00 | | | | | | | |
| (3) Valence | 0.30** | 0.59** | 1.00 | | | | | | |
| (4) Articles | -0.02 | -0.06 | -0.02 | 1.00 | | | | | |
| (5) GoogleTrends | 0.18 | -0.15 | -0.09 | 0.36** | 1.00 | | | | |
| (6) Return | 0.36** | 0.16 | 0.15 | 0.22* | 0.48** | 1.00 | | | |
| (7) Volatility | -0.21* | -0.35** | -0.38** | 0.35** | 0.54** | 0.20 | 1.00 | | |
| (8) TradingVolume | 0.10 | -0.04 | -0.14 | 0.36** | 0.65** | 0.39** | 0.62** | 1.00 | |
| (9) TransactionVolume | 0.19 | -0.07 | 0.04 | 0.34** | 0.48** | 0.37** | 0.33** | 0.60** | 1.00 |



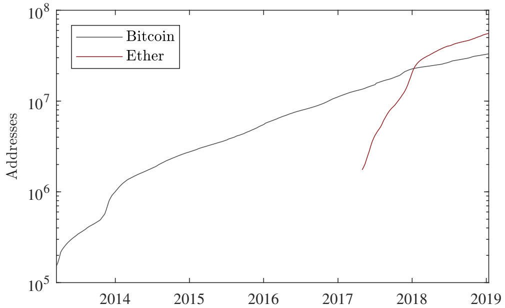

**(a)** Tone 8-week moving average



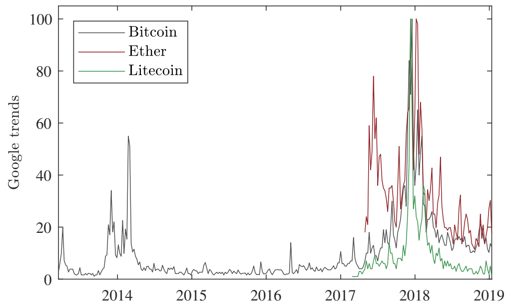

**(b)** Vader 8-week moving average



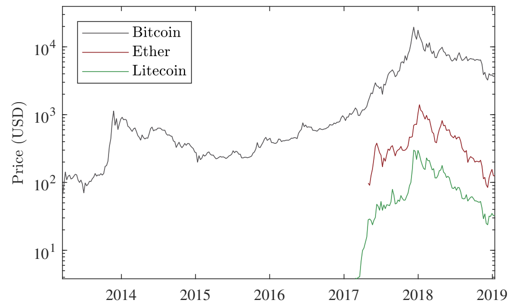

**(c)** Valence 8-week moving average



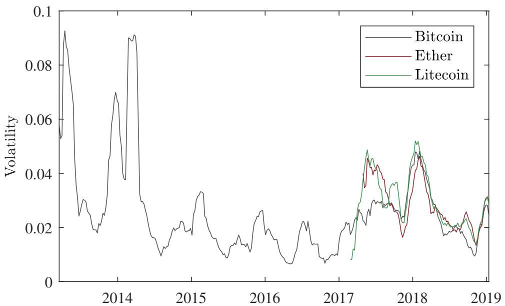

**(d)** Addresses



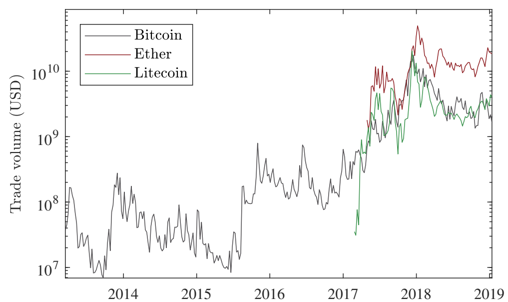

**(e)** Google trends



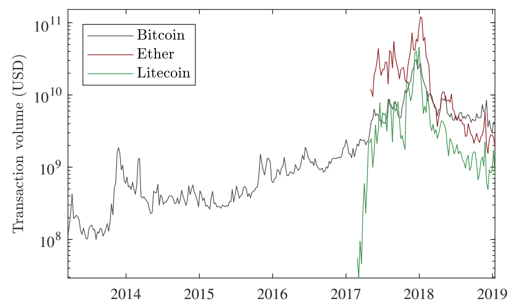

**(f)** Price




**(g)** Volatility 8-week moving average




**(h)** Trade volume




**(i)** Transaction volume

**Figure 4.1:** Time evolution of original variables.

# 5 Results

Regression models were used to investigate the relationships between the sentiment of news articles and various cryptocurrency-related variables. This chapter is organized into two sections: the estimated predictive models for financial- and sentiment variables are presented in Section 5.1 and Section 5.2, respectively.

For each dependent variable, a set of full models and a restricted model are estimated. The regressions were performed on weekly data. Statistical interference was based on robust standard errors. $R^2$ and $R^2_{adj}$ are reported for each regression. We denote the independent variable as $X_{i,t}$, and the dependent variable as $Y_{t+1}$. Here $t$ is indexing weeks, and $i$ is indexing independent variables. Each regression includes a constant $\alpha$ and the one-week lagged dependent variable, $Y_t$. The error term is denoted as $\epsilon_t$. Finally, $\beta_i$ are regression coefficients. The full model is defined in Eq. (5.1).

$$Y_{t+1} = \alpha + \beta_0 Y_t + \sum_{i=1}^{n} \beta_i X_{i,t} + \epsilon_{t+1} \tag{5.1}$$

In restricted models, only one explanatory variable is included at a time. The restricted model also includes the AR(1) term. The restricted model is defined in Eq. (5.2).

$$Y_{t+1} = \alpha + \beta_0 Y_t + \beta_i X_{i,t} + \epsilon_{t+1} \tag{5.2}$$

For each dependent variable, a summarized table with the regression results from the restricted models is presented first, while the complete results are included in Appendix C. The tables for the full models are presented at the end of each section.

## 5.1 Predictive models for financial variables

This section presents the results for the financial variables: *Return*, *Volatility*, *TradingVolume*, and *TransactionVolume*. In the full model, only one sentiment variable was included at a time, as *Tone*, *Vader* and *Valence* are different measures of the same factor - sentiment. Thus, the results from these regressions are presented side-by-side in the tables, such that the different methods of sentiment analysis can be easily compared.

Each subsection discusses the most important findings in the order of Bitcoin, Ether, and Litecoin respectively. In the discussion, more emphasis is put on the results from the full model.

### 5.1.1 Return

**Table 5.1:** Summary of the regression results for the restricted models of *Return*, which can be found in Appendix C.1. A significant result is reported with the symbol + for a positive coefficient, and - for negative. Specifically, one, two, and three symbols represent 10%, 5%, and 1% significance respectively.

| **Return**$_{t+1}$ | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| Return$_t$ | | | | | |
| Tone$_t$ | | | | + | |
| Vader$_t$ | | | - - | | |
| Valence$_t$ | | | | | |
| Articles$_t$ | | | + + | | |
| Addresses$_t$ | | | + + + | + | |
| GoogleTrends$_t$ | | - - | | | |
| Volatility$_t$ | | | | + | |
| TradingVolume$_t$ | | | | | |
| TransactionVolume$_t$ | | | | | |

The results for the restricted models for predicting *Return* are summarized in Table 5.1, and Table 5.5 presents the results for the full models. During the first period of Bitcoin, no significant variables were found in the full models. On the other hand, during the second period, *GoogleTrends* predicted *Return* with a negative coefficient. In other words, weeks with a higher amount of Google searches for Bitcoin precede weeks of lower returns. While this observation might seem counter-intuitive, it could be that investor interest, proxied by Google searches, was a result of negative events in the Bitcoin space during this period. For instance, an example of a negative event occurred in August 2016 during the major hacking of the cryptocurrency exchange Bitfinex[1], and more noticeably the hacking of the digital decentralized autonomous organization The DAO in June 2016.[2]

[1] https://www.bitfinex.com/

[2] https://www.wired.com/2016/06/50-million-hack-just-showed-dao-human/

Furthermore, in the full model (see Table 5.5), a week of larger trading volumes also had a negative impact on the following week's return, whereas transaction volume had a positive impact.

In the third period, *Addresses* showed strong predictive power at a 1% significance level. In other words, weeks with a high amount of new addresses preceded weeks with high returns. A possible interpretation is that new users of Bitcoin must first generate an address before they can use the cryptocurrency. As a consequence, new addresses can be considered a demand-related factor, and weeks with a high number of new addresses is thus a sign of increased demand for Bitcoin. Interestingly, as the same relationship was not observed in previous periods, the owners of new addresses in period 3 might have different intentions compared to earlier periods. One possible explanation could be that many new addresses in the past were generated for mining, instead of trading. If that were the case, new addresses would not precede rises in returns, as the addresses were not created for trading purposes. Also, during the third period, the restricted model found 5% significance levels for both *Articles* and *Vader*, with a negative coefficient for the latter. As a high sentiment predicted a lower return the following week, we may have observed a case of "Buy on bad news, sell on good news, a trading strategy where the trading activity is opposite of the sentiment of news. The idea behind this strategy is that a strong market reaction to news will be followed by a mean reversion in price. Using this strategy, buying on bad news and selling on good news could grant a positive return. In terms of *Articles*, the number of articles published in a week can be considered a measure of market interest, as journalists can be assumed to write about topics the public find interesting. Hence, an increase in the number of articles leading to an increase in return implies that information supply, in the form of articles, created demand for Bitcoin.

As for the remaining two cryptocurrencies, similar results were not observed. For Ether, *Tone* was significant in the full model. Interestingly, the positive coefficient suggests that investors reacted differently to the sentiment of news for different cryptocurrencies, as the coefficient of *Vader* for Bitcoin was negative. For Litecoin, there were no significant coefficients. Overall, the results for the different cryptocurrencies suggest that distinct trading strategies are needed for each currency since they reacted differently to changes in market conditions. In general, the predicting power $R^2$ of the significant variables were quite low, with a surprising exception in the case of *Addresses* for Bitcoin. While this finding was surprising, the generally low $R^2$ was expected, since predictability in the return of an asset should quickly be nullified in an efficient market.

### 5.1.2 Volatility

The results for the restricted models for predicting *Volatility* are summarized in Table 5.2, and Table 5.6 presents the results from the full model. All periods of Bitcoin showed autocorrelation, which was expected and is similar to other financial asset classes. In the second period, several other variables were significant: *Articles* showed a strong significance level in both regression models, and this pattern remained in period 3. Interestingly,

**Table 5.2:** Summary of the regression results for the restricted models of *Volatility*, which can be found in Appendix C.2. A significant result is reported with the symbol + for a positive coefficient, and - for negative. Specifically, one, two, and three symbols represent 10%, 5%, and 1% significance respectively.

| **Volatility$_{t+1}$** | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| Volatility$_t$ | + + + | + + + | + + + | + + + | + + + |
| Tone$_t$ | | | + + | | |
| Vader$_t$ | | | | | |
| Valence$_t$ | | | | | |
| Articles$_t$ | | - - - | + + + | + + | |
| Addresses$_t$ | | | + + | + + + | |
| GoogleTrends$_t$ | - | | | | + + |
| Return$_t$ | | + + + | | | |
| TradingVolume$_t$ | | + | | + | |
| TransactionVolume$_t$ | | + + | | + | |

the coefficient changed from negative in period 2 to positive in period 3. In other words, if many articles were published in period 2, the volatility dropped the next week, while in period 3 many published articles preceded an increase in volatility. Moreover, *Return* also predicted *Volatility* in Bitcoin's second period, where a week of greater returns was followed by an increase in volatility. In the stock market, the opposite is usually observed where increases in return are followed by a decrease in volatility.[3] One interpretation of this result is that the Bitcoin market in period 2 was quite immature, and the observed behavior could be that investors faith in the stability of the price was low, leading to increases in volatility after weeks of high returns. In the third period, the pattern reversed, as high return lead to lower volatility the week after. This may have been a sign that the market had matured in period 3, to more closely mirror the pattern from the stock market. Lastly, a positive *Tone* preceded an increase in *Volatility*.

Similar to Bitcoin, a large degree of autocorrelation was present in the volatility of both Ether and Litecoin. For Ether, a week with a high number of new *Addresses* was followed by an increase in *Volatility*, a relationship which also was observed in Bitcoin's third period in the restricted model. As new users joined the market, they influenced the stability of the price equilibrium. Interestingly, even though the volatility of Ether exhibits significant autocorrelation (see Table 5.2), past volatility becomes insignificant in the full model (see Table 5.6). The reason is that the variable *Addresses*, which is correlated with *Volatility* (see Table 4.10) is a better predictor of volatility than the past volatility. In traditional financial markets, the best predictor of volatility is almost always past volatility (see e.g., Kim et al. (2019)), and other variables improve the model only marginally. We consider this as a strong indication that the Ether market is very immature. We therefore expect that

[3]https://www.investopedia.com/articles/financial-theory/08/volatility.asp

in the future, the market will evolve in such a way that in a volatility model, past volatility will become more important than past addresses. Lastly, for Litecoin, it was observed that a rise in *GoogleTrend* preceded an increase in *Volatility*.

### 5.1.3 Trading volume

**Table 5.3:** Summary of the regression results for the restricted models of *TradingVolume*, which can be found in Appendix C.3. A significant result is reported with the symbol + for a positive coefficient, and - for negative. Specifically, one, two, and three symbols represent 10%, 5%, and 1% significance respectively.

| **$\text{TradingVolume}_{t+1}$** | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| $\text{TradingVolume}_t$ | + + + | + + + | + + + | + + + | + + + |
| $\text{Tone}_t$ | + + + | | | + + | |
| $\text{Vader}_t$ | + | | | | |
| $\text{Valence}_t$ | + | | | + + + | |
| $\text{Articles}_t$ | | - - - | + + + | | |
| $\text{Addresses}_t$ | | | + + + | | |
| $\text{GoogleTrends}_t$ | | | + + + | | + |
| $\text{Return}_t$ | + | + + + | + | | + + |
| $\text{Volatility}_t$ | | - - | - - - | - - - | - |
| $\text{TransactionVolume}_t$ | | | + + + | | |

Estimated restricted models for predicting *TradingVolume* are summarized in Table 5.3, and Table 5.7 presents the results from the full model. First of all, *TradingVolume* for all three cryptocurrencies exhibited strong autocorrelation in both models: weeks with a high trading volume often preceded weeks with higher trading volume. In the first period of Bitcoin, a positive *Tone* and *Vader*, as well as *Valence* in the restricted model, predicted a high trading volume. This result means that the sentiment in news had an impact on trading volume in the early years of Bitcoin, but the relationship did not hold in the later periods. In the second period, an increase in *Articles* lead to a lower *TradingVolume* the week after. Interestingly, while the significance remained strong in period 3, the coefficient changed from negative to positive. Recall from Section 5.1.2 that the relationship between *Articles* and *Volatility* behaved in a similar fashion. The result suggests that the market reaction to news about Bitcoin in period 2 was fundamentally different from period 3.

In the second period, a week of high *Volatility* was usually followed by a week of lower *TradingVolume* in the restricted model. This behavior became even more evident in period 3, where weeks of high volatility were followed by weeks of lower trading volume for all three cryptocurrencies. In other words, investors appetite for trading diminished after weeks of high volatility. Starting in the second period, *GoogleTrends* predicted *TradingVolume* of Bitcoin, with an even stronger relationship in the third period. A similar relationship was seen for Litecoin, although the significance was weaker. Thus, as the public interest increased, the trading activity followed. Also in Bitcoin period 3, *Addresses* lead to an increase in *TradingVolume*.

Ether's trading volume was affected by the sentiment of news articles, where an increase in *Valence* was followed by an increase in *TradingVolume*, and similar results were found for *Tone* in the restricted regression. As this was not observed for Bitcoin in the same period, it appears that investors of Ether were more sensitive to the sentiment of news than investors of Bitcoin. Litecoin's trading volume, on the other hand, did not appear to be influenced by the sentiment of news. However, both *GoogleTrends* and *Return* influenced the *TradingVolume* of Litecoin, where an increase in either factor would lead to an increase the following week in trading volume.

### 5.1.4 Transaction volume

**Table 5.4:** Summary of the regression results for the restricted models of *TransactionVolume*, which can be found in Appendix C.4. A significant result is reported with the symbol + for a positive coefficient, and - for negative. Specifically, one, two, and three symbols represent 10%, 5%, and 1% significance respectively.

| **TransactionVolume$_{t+1}$** | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| TransactionVolume$_t$ | + + + | + + + | + + + | + + + | + + + |
| Tone$_t$ | + + | | | | |
| Vader$_t$ | | | | + | |
| Valence$_t$ | | | | + + | |
| Articles$_t$ | | | | | |
| Addresses$_t$ | | | | | |
| GoogleTrends$_t$ | | | | | + |
| Return$_t$ | + + + | + + + | + + + | + + + | |
| Volatility$_t$ | | - - | - - | | |
| TradingVolume$_t$ | | | | + + | |

Estimated restricted models for predicting *Volatility* are summarized in Table 5.2, and Table 5.6 presents the results from the full model. First, *TransactionVolume* for all three cryptocurrencies exhibited a strong autocorrelation. Furthermore, in all three periods of Bitcoin and for Ether, a week with high *Return* was followed by a week of higher *TransactionVolume*. This suggests that investors, after a week of high returns, increased their transaction activities. In Bitcoin's first period, a week of high *Volatility* was followed by increased *TransactionVolume* during the following week. This relationship reversed in the following periods, where a week of high volatility preceded a week of lower transaction volume, and the relationship was strongest in period 3.

The same relationship was observed for Ether: a week with high volatility usually preceded a week of lower transaction volume. The relationship between volatility and transaction volume in period 3 is similar to the relationship between volatility and trading volume in the same period, discussed in Section 5.1.3. In Bitcoin's third period, an increase in *Addresses* was followed by an increase in *TransactionVolume*. With more addresses comes more possible destinations for transactions, increasing the network's value. As for Ether, both a high *Vader* and *Valence* predicted a high *TransactionVolume*. Thus, if the sentiment in news was high, the usage of Ether increased the following week. Lastly, *TradingVolume* had a positive impact on *TransactionVolume*. For Litecoin, no other variables were significant.

**Table 5.5:** Regression results from the full model with *Return* as dependent variable. Explanatory variables sampled one week prior to the dependent variable. Every column reports the result from one regression. *Sentiment* is a placeholder for either *Tone*, *Vader*, or *Valence*. Coefficients are presented first, with robust standard errors reported in parentheses. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: Return$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence |
| Intercept | 0.014 | -0.013 | -0.169 | 0.022 | 0.012 | 0.039 | -0.050** | -0.036 | -0.062 | -0.110** | -0.094** | -0.991* | -0.045 | 0.003 | 0.783 |
| | (0.028) | (0.021) | (0.405) | (0.023) | (0.017) | (0.375) | (0.022) | (0.032) | (0.623) | (0.046) | (0.043) | (0.593) | (0.052) | (0.057) | (0.889) |
| Return$_t$ | -0.129 | -0.154 | -0.153 | -0.060 | -0.053 | -0.063 | -0.075 | -0.086 | -0.084 | 0.065 | 0.086 | 0.087 | -0.062 | -0.029 | -0.008 |
| | (0.179) | (0.183) | (0.183) | (0.115) | (0.117) | (0.120) | (0.108) | (0.107) | (0.107) | (0.138) | (0.136) | (0.134) | (0.117) | (0.111) | (0.112) |
| Sentiment$_t$ | -0.053 | 0.027 | 0.028 | -0.041 | -0.048 | -0.006 | -0.016 | -0.067 | 0.002 | 0.123** | 0.145 | 0.164 | 0.078 | 0.009 | -0.133 |
| | (0.056) | (0.089) | (0.071) | (0.037) | (0.064) | (0.067) | (0.031) | (0.107) | (0.110) | (0.062) | (0.092) | (0.102) | (0.074) | (0.146) | (0.152) |
| Articles$_t$ | -0.004 | -0.004 | -0.004 | -0.002 | -0.003 | -0.002 | 0.003 | 0.003 | 0.004 | -0.000 | -0.001 | -0.000 | 0.003 | 0.003 | 0.003 |
| | (0.004) | (0.004) | (0.004) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.006) | (0.006) | (0.006) | (0.005) | (0.005) | (0.005) |
| Addresses$_t$ | -0.838 | -0.761 | -0.802 | 1.937 | 2.121 | 1.315 | 12.681*** | 12.128*** | 12.286*** | 1.096 | 1.491 | 1.134 | - | - | - |
| | (1.431) | (1.456) | (1.433) | (1.966) | (2.209) | (2.150) | (2.166) | (2.095) | (2.049) | (1.000) | (1.028) | (0.995) | - | - | - |
| GoogleTrends$_t$ | 0.001 | 0.001 | 0.001 | -0.003* | -0.003* | -0.003* | 0.001 | 0.001 | 0.001 | -0.002 | -0.003 | -0.003 | 0.010 | 0.011 | 0.012 |
| | (0.003) | (0.003) | (0.003) | (0.002) | (0.002) | (0.002) | (0.007) | (0.007) | (0.007) | (0.006) | (0.006) | (0.006) | (0.009) | (0.009) | (0.009) |
| Volatility$_t$ | 0.277 | 0.315 | 0.314 | -0.454 | -0.429 | -0.434 | 0.157 | 0.016 | 0.014 | 1.085 | 1.091 | 1.073 | 0.027 | -0.238 | -0.515 |
| | (0.242) | (0.264) | (0.264) | (0.364) | (0.370) | (0.367) | (0.923) | (0.811) | (0.816) | (0.864) | (0.837) | (0.819) | (0.846) | (0.846) | (0.960) |
| TradingVolume$_t$ | -0.021 | -0.015 | -0.015 | -0.018* | -0.017* | -0.018* | 0.024 | 0.024 | 0.033 | -0.026 | -0.044 | -0.026 | -0.036 | -0.033 | -0.034 |
| | (0.038) | (0.035) | (0.035) | (0.010) | (0.010) | (0.010) | (0.070) | (0.067) | (0.069) | (0.080) | (0.083) | (0.080) | (0.045) | (0.045) | (0.046) |
| TransactionVolume$_t$ | 0.059 | 0.052 | 0.053 | 0.082* | 0.072 | 0.076* | -0.091 | -0.089 | -0.099 | 0.005 | 0.029 | 0.016 | -0.007 | -0.006 | -0.006 |
| | (0.055) | (0.050) | (0.052) | (0.042) | (0.044) | (0.044) | (0.068) | (0.065) | (0.069) | (0.055) | (0.055) | (0.054) | (0.008) | (0.008) | (0.008) |
| $R^2$ | 0.09 | 0.08 | 0.08 | 0.15 | 0.14 | 0.13 | 0.24 | 0.25 | 0.24 | 0.13 | 0.12 | 0.13 | 0.06 | 0.05 | 0.06 |
| $R^2_{adj}$ | -0.00 | -0.01 | -0.01 | 0.06 | 0.05 | 0.04 | 0.17 | 0.17 | 0.17 | 0.05 | 0.04 | 0.04 | -0.02 | -0.03 | -0.02 |

**Table 5.6:** Regression results from the full model with *Volatility* as dependent variable. Explanatory variables sampled one week prior to the dependent variable. Every column reports the result from one regression. *Sentiment* is a placeholder for either *Tone*, *Vader*, or *Valence*. Coefficients are presented first, with robust standard errors reported in parentheses. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: Volatility$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence |
| Intercept | 0.012 | 0.004 | -0.096 | 0.010 | 0.013* | 0.065 | 0.009** | 0.010** | 0.011 | 0.021*** | 0.015* | -0.052 | 0.021*** | 0.020* | 0.304* |
| | (0.016) | (0.011) | (0.254) | (0.009) | (0.007) | (0.120) | (0.004) | (0.005) | (0.110) | (0.007) | (0.009) | (0.088) | (0.007) | (0.012) | (0.166) |
| Volatility$_t$ | 0.572** | 0.584*** | 0.584*** | 0.278** | 0.277** | 0.273** | 0.419*** | 0.534*** | 0.534*** | 0.221 | 0.250 | 0.276* | 0.421*** | 0.429*** | 0.356*** |
| | (0.235) | (0.226) | (0.223) | (0.112) | (0.115) | (0.114) | (0.117) | (0.113) | (0.114) | (0.154) | (0.161) | (0.155) | (0.117) | (0.124) | (0.123) |
| Sentiment$_t$ | -0.014 | 0.012 | 0.018 | 0.004 | -0.013 | -0.009 | 0.013** | -0.004 | -0.000 | -0.010 | -0.000 | 0.011 | -0.008 | -0.011 | -0.049* |
| | (0.032) | (0.050) | (0.045) | (0.015) | (0.023) | (0.021) | (0.005) | (0.017) | (0.020) | (0.010) | (0.020) | (0.015) | (0.011) | (0.030) | (0.028) |
| Articles$_t$ | 0.004 | 0.004 | 0.004 | -0.002*** | -0.002*** | -0.002*** | 0.003*** | 0.002*** | 0.002*** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | (0.003) | (0.003) | (0.003) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Addresses$_t$ | 0.676 | 0.698 | 0.674 | -0.322 | -0.031 | -0.181 | 0.548 | 0.860 | 0.869 | 0.343** | 0.346** | 0.353** | - | - | - |
| | (0.480) | (0.474) | (0.472) | (0.844) | (0.874) | (0.865) | (0.627) | (0.548) | (0.548) | (0.151) | (0.155) | (0.148) | - | - | - |
| GoogleTrends$_t$ | -0.005 | -0.005 | -0.005 | 0.000 | 0.001 | 0.001 | -0.000 | -0.000 | -0.000 | -0.000 | 0.000 | 0.000 | 0.002** | 0.002** | 0.002** |
| | (0.003) | (0.003) | (0.003) | (0.000) | (0.000) | (0.000) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Return$_t$ | 0.059 | 0.052 | 0.052 | 0.072** | 0.076** | 0.076** | -0.036** | -0.030* | -0.029* | -0.013 | -0.014 | -0.013 | 0.013 | 0.011 | 0.016 |
| | (0.058) | (0.054) | (0.054) | (0.032) | (0.033) | (0.033) | (0.017) | (0.017) | (0.017) | (0.026) | (0.026) | (0.026) | (0.015) | (0.015) | (0.016) |
| TradingVolume$_t$ | 0.023 | 0.025 | 0.026 | 0.005 | 0.005 | 0.004 | -0.006 | -0.014 | -0.013 | 0.007 | 0.006 | 0.005 | 0.001 | 0.002 | 0.001 |
| | (0.023) | (0.024) | (0.024) | (0.004) | (0.004) | (0.005) | (0.010) | (0.010) | (0.010) | (0.012) | (0.013) | (0.012) | (0.008) | (0.008) | (0.008) |
| TransactionVolume$_t$ | -0.041 | -0.043 | -0.043 | 0.018 | 0.017 | 0.018 | -0.002 | 0.004 | 0.004 | -0.003 | -0.004 | -0.004 | -0.001 | -0.001 | -0.001 |
| | (0.032) | (0.032) | (0.033) | (0.013) | (0.014) | (0.013) | (0.011) | (0.010) | (0.011) | (0.007) | (0.007) | (0.007) | (0.001) | (0.001) | (0.001) |
| $R^2$ | 0.50 | 0.50 | 0.50 | 0.32 | 0.32 | 0.32 | 0.50 | 0.47 | 0.47 | 0.32 | 0.31 | 0.31 | 0.42 | 0.41 | 0.44 |
| $R^2{}_{adj}$ | 0.45 | 0.45 | 0.45 | 0.25 | 0.25 | 0.25 | 0.45 | 0.42 | 0.42 | 0.25 | 0.24 | 0.24 | 0.36 | 0.36 | 0.40 |

**Table 5.7:** Regression results from the full model with *TradingVolume* as dependent variable. Explanatory variables sampled one week prior to the dependent variable. Every column reports the result from one regression. *Sentiment* is a placeholder for either *Tone*, *Vader*, or *Valence*. Coefficients are presented first, with robust standard errors reported in parentheses. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: TradingVolume$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence |
| Intercept | -0.227** | -0.133* | -2.893 | 0.089 | 0.061 | -0.043 | 0.043 | 0.087 | 0.377 | 0.064 | 0.010 | -2.037** | 0.167 | 0.208 | 1.700 |
| | (0.099) | (0.070) | (2.300) | (0.120) | (0.105) | (1.960) | (0.044) | (0.070) | (1.580) | (0.085) | (0.078) | (0.919) | (0.122) | (0.159) | (2.147) |
| TradingVolume$_t$ | 0.496*** | 0.476*** | 0.479*** | 0.644*** | 0.647*** | 0.643*** | 0.460*** | 0.385*** | 0.403*** | 0.849*** | 0.804*** | 0.830*** | 0.771*** | 0.785*** | 0.766*** |
| | (0.121) | (0.120) | (0.120) | (0.087) | (0.085) | (0.084) | (0.110) | (0.111) | (0.112) | (0.163) | (0.157) | (0.155) | (0.112) | (0.113) | (0.113) |
| Sentiment$_t$ | 0.465** | 0.602* | 0.506 | -0.127 | -0.192 | 0.016 | 0.092 | -0.222 | -0.059 | 0.083 | 0.261 | 0.370** | -0.068 | -0.219 | -0.270 |
| | (0.220) | (0.361) | (0.407) | (0.169) | (0.247) | (0.352) | (0.087) | (0.247) | (0.279) | (0.128) | (0.197) | (0.160) | (0.170) | (0.387) | (0.366) |
| Articles$_t$ | 0.001 | 0.008 | 0.005 | -0.043*** | -0.045*** | -0.041*** | 0.015* | 0.011 | 0.013* | 0.006 | 0.004 | 0.005 | 0.007 | 0.008 | 0.008 |
| | (0.019) | (0.018) | (0.018) | (0.011) | (0.011) | (0.011) | (0.009) | (0.007) | (0.008) | (0.010) | (0.010) | (0.010) | (0.008) | (0.008) | (0.008) |
| Addresses$_t$ | 2.881 | 2.512 | 1.726 | 4.734 | 6.048 | 2.555 | 12.501* | 14.322** | 14.714** | 2.260 | 3.024* | 2.420 | - | - | - |
| | (3.165) | (3.184) | (3.219) | (14.674) | (15.137) | (14.800) | (7.428) | (6.609) | (6.773) | (1.654) | (1.654) | (1.572) | - | - | - |
| GoogleTrends$_t$ | 0.002 | -0.001 | 0.000 | 0.009* | 0.009** | 0.009* | 0.033*** | 0.032*** | 0.032*** | 0.001 | 0.001 | 0.000 | 0.025* | 0.023* | 0.024* |
| | (0.013) | (0.013) | (0.013) | (0.005) | (0.005) | (0.005) | (0.011) | (0.011) | (0.010) | (0.013) | (0.012) | (0.011) | (0.013) | (0.013) | (0.013) |
| Return$_t$ | 0.393 | 0.593 | 0.610 | 1.024 | 1.056* | 1.005 | 0.226 | 0.269 | 0.278 | 0.236 | 0.265 | 0.273 | 0.431* | 0.444* | 0.441* |
| | (0.408) | (0.385) | (0.373) | (0.624) | (0.635) | (0.652) | (0.221) | (0.216) | (0.216) | (0.297) | (0.293) | (0.292) | (0.236) | (0.232) | (0.244) |
| Volatility$_t$ | -0.014 | -0.106 | -0.171 | -3.494* | -3.414* | -3.424* | -4.392*** | -3.556*** | -3.551*** | -4.938*** | -4.512*** | -4.382*** | -4.368** | -4.578** | -4.647** |
| | (0.670) | (0.735) | (0.737) | (1.857) | (1.878) | (1.853) | (1.364) | (1.182) | (1.190) | (1.696) | (1.521) | (1.480) | (2.028) | (2.098) | (2.128) |
| TransactionVolume$_t$ | 0.018 | 0.005 | 0.028 | 0.434 | 0.399 | 0.417 | 0.022 | 0.098 | 0.076 | -0.033 | -0.006 | -0.030 | -0.020 | -0.022 | -0.019 |
| | (0.175) | (0.179) | (0.180) | (0.278) | (0.283) | (0.280) | (0.140) | (0.129) | (0.132) | (0.099) | (0.092) | (0.093) | (0.020) | (0.019) | (0.020) |
| $R^2$ | 0.38 | 0.37 | 0.36 | 0.56 | 0.55 | 0.55 | 0.65 | 0.65 | 0.65 | 0.72 | 0.73 | 0.73 | 0.65 | 0.65 | 0.65 |
| $R^2_{adj}$ | 0.32 | 0.30 | 0.30 | 0.51 | 0.51 | 0.51 | 0.62 | 0.62 | 0.61 | 0.69 | 0.70 | 0.70 | 0.62 | 0.62 | 0.62 |

**Table 5.8:** Regression results from the full model with *TransactionVolume* as dependent variable. Explanatory variables sampled one week prior to the dependent variable. Every column reports the result from one regression. *Sentiment* is a placeholder for either *Tone*, *Vader*, or *Valence*. Coefficients are presented first, with robust standard errors reported in parentheses. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: TransactionVolume$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence |
| Intercept | -0.123* | -0.039 | -2.308 | 0.055 | 0.043 | 0.431 | 0.038 | 0.055 | 0.550 | 0.035 | -0.059 | -2.376*** | 0.827 | -0.083 | -11.108 |
| | (0.068) | (0.054) | (1.514) | (0.057) | (0.047) | (1.081) | (0.026) | (0.040) | (0.941) | (0.103) | (0.074) | (0.905) | (0.894) | (0.932) | (12.494) |
| TransactionVolume$_t$ | 0.554*** | 0.560*** | 0.554*** | 0.721*** | 0.705*** | 0.711*** | 0.572*** | 0.608*** | 0.611*** | 0.508*** | 0.530*** | 0.503*** | 0.335*** | 0.322** | 0.312** |
| | (0.178) | (0.204) | (0.180) | (0.114) | (0.121) | (0.118) | (0.122) | (0.122) | (0.125) | (0.120) | (0.104) | (0.104) | (0.126) | (0.127) | (0.127) |
| Sentiment$_t$ | 0.280* | 0.199 | 0.407 | -0.054 | -0.086 | -0.070 | 0.051 | -0.085 | -0.090 | 0.031 | 0.287* | 0.418*** | -1.361 | 0.018 | 1.890 |
| | (0.151) | (0.292) | (0.268) | (0.088) | (0.191) | (0.191) | (0.064) | (0.138) | (0.165) | (0.162) | (0.170) | (0.156) | (1.244) | (2.183) | (2.138) |
| Articles$_t$ | -0.001 | 0.003 | 0.002 | -0.007 | -0.008 | -0.007 | 0.004 | 0.003 | 0.003 | 0.010 | 0.007 | 0.008 | 0.072 | 0.079 | 0.074 |
| | (0.009) | (0.009) | (0.009) | (0.007) | (0.006) | (0.006) | (0.006) | (0.005) | (0.005) | (0.013) | (0.011) | (0.011) | (0.082) | (0.086) | (0.085) |
| Addresses$_t$ | -1.750 | -2.031 | -2.560* | -1.519 | -0.879 | -1.812 | 7.247** | 8.352** | 8.370** | -0.161 | 0.701 | 0.043 | - | - | - |
| | (1.558) | (1.571) | (1.476) | (5.650) | (5.504) | (5.543) | (3.486) | (3.975) | (4.042) | (1.837) | (1.873) | (1.743) | - | - | - |
| GoogleTrends$_t$ | 0.014 | 0.012 | 0.012 | 0.005 | 0.005 | 0.006 | 0.010 | 0.010 | 0.010 | -0.011 | -0.010 | -0.011 | 0.108 | 0.091 | 0.089 |
| | (0.009) | (0.009) | (0.009) | (0.004) | (0.004) | (0.004) | (0.008) | (0.008) | (0.008) | (0.014) | (0.014) | (0.014) | (0.109) | (0.105) | (0.104) |
| Return$_t$ | 0.547*** | 0.671*** | 0.679*** | 0.857*** | 0.872*** | 0.874*** | 0.549*** | 0.574*** | 0.580*** | 0.654** | 0.683** | 0.693*** | 1.816 | 1.217 | 0.945 |
| | (0.176) | (0.187) | (0.176) | (0.293) | (0.305) | (0.306) | (0.153) | (0.156) | (0.154) | (0.263) | (0.274) | (0.266) | (1.753) | (1.845) | (1.810) |
| Volatility$_t$ | 0.771** | 0.669* | 0.698* | -1.227 | -1.194 | -1.217 | -3.433*** | -2.968*** | -2.953*** | -3.139** | -2.490* | -2.323 | -6.303 | -1.332 | 2.303 |
| | (0.376) | (0.360) | (0.390) | (0.928) | (0.904) | (0.917) | (1.055) | (0.792) | (0.791) | (1.363) | (1.371) | (1.434) | (10.193) | (10.193) | (9.897) |
| TradingVolume$_t$ | 0.011 | -0.007 | 0.007 | -0.003 | -0.001 | -0.004 | 0.080 | 0.043 | 0.040 | 0.371*** | 0.316** | 0.343*** | 0.417 | 0.359 | 0.375 |
| | (0.115) | (0.121) | (0.113) | (0.042) | (0.043) | (0.040) | (0.079) | (0.082) | (0.082) | (0.137) | (0.136) | (0.131) | (0.731) | (0.730) | (0.691) |
| $R^2$ | 0.65 | 0.63 | 0.64 | 0.59 | 0.59 | 0.59 | 0.77 | 0.76 | 0.76 | 0.75 | 0.75 | 0.76 | 0.32 | 0.31 | 0.32 |
| $R^2_{adj}$ | 0.61 | 0.59 | 0.61 | 0.55 | 0.55 | 0.55 | 0.74 | 0.74 | 0.74 | 0.72 | 0.73 | 0.74 | 0.27 | 0.25 | 0.26 |

## 5.2 Predictive models for sentiment variables

This section presents and discusses predictive models for the sentiment variables: *Tone*, *Vader*, and *Valence*. The results for the restricted model for each respective sentiment variable are summarized in Table 5.9, whereas the complete results can be found in Appendix C. The results from the full model regressions are provided in Table 5.10.

Starting with Bitcoin, all sentiment variables were strongly autocorrelated in period 3: the sentiment in news about Bitcoin is strongly related to the sentiment from the previous week. This is in contrast to its earlier periods and the other cryptocurrencies. There are several possible explanations for this observation. One possible explanation is that the news sentiment becomes more persistent over time as a cryptocurrency matures. If this is the case, the persistence could be a consequence of the total amount of available information increasing over time, which lower the relative impact of news. However, it could also be that the measures of sentiment contain too much noise, which makes them unable to capture autocorrelation in earlier periods, when there were less articles published per week.

For Ether, a high *Return* positively impacted all three measures of sentiment in the restricted model. While it is not surprising that high return is considered good news, it is interesting to observe that only Ether behaved in this manner. It would be natural to assume that the sentiment of Bitcoin and Litecoin should have this relationship too. Recall that the two cryptocurrencies are both designed to function as a payment system and store of value, while Ether is not. Apparently, news about Ether is more influenced by the price movement of the cryptocurrency, compared to Bitcoin and Litecoin. Secondly, in the full model, it was found that *Addresses* negatively impacted the sentiment: a week with a high number of new addresses was followed by a decrease in sentiment. Higher *TradingVolume* positively impacted sentiment of Ether news, suggesting that high trading volume is considered good news for Ether.

For Litecoin, the significant result of *Articles* implies that a week with a high number of articles predicted an increase in the sentiment of Litecoin. As such, an increased amount of media attention was considered positive for Litecoin, independent of the sentiment in the articles. Furthermore, *Volatility* showed strong significance: a week of higher volatility would negatively impact the following week's sentiment. This finding is quite intuitive, as volatility is generally considered to be negative. Thus, the sentiment in Litecoin news articles was more sensitive to volatility than articles about Bitcoin and Ether.

The three different measures of sentiment used in this paper have differences in their methodologies. The differences can be found in the number of words and their scores, feature extraction, and preprocessing. While all three methods attempted to capture the sentiment of news articles, their differences lead to surprisingly different regression results. Thus, different measures of sentiment capture distinct elements of the full sentiment in an article. In the following paragraphs, the overall trends for the regression models are discussed in the following order: *Tone*, *Vader*, and *Valence*.

*Tone* was created using the tone-dictionary, which was designed to evaluate the sentiment in financial reports. The dictionary contains finance-related words (see Appendix A), which affect the way it captures the sentiment of a text. In period 3, *Tone* was autocorrelated for all three cryptocurrencies in the restricted model. Furthermore, when comparing *Tone* to the other two measures of sentiment, the lagged *Tone* had a much higher $R^2$ (see Appendix Tables C.5-C.7). While all three sentiment measures attempt to estimate the sentiment of an article, it becomes clear that they capture different elements of the full sentiment. In terms of autocorrelation, *Tone* is more persistent than the other variables. This could be a consequence of the distribution of tone-scored articles being more uniform than the vader- and valence-scored articles (recall Figure 3.2). As *Tone* measures the sentiment of financial words in an article, the implication is that the financial sentiment is more persistent than the general sentiment, as measured by *Vader* and *Valence.* Lastly, *Articles* in period 3 positively impacted *Tone* of both Bitcoin- and Litecoin.

*Vader* involves more complex preprocessing compared to the other dictionaries and should give a more dynamic understanding of the text based on its rules for understanding negation and degree modifiers. The autocorrelation of *Vader* increased every period for Bitcoin, with both the significance and $\mathrm{R}^2$ increasing over time. This relationship strengthens the hypothesis that the sentiment of a cryptocurrency becomes more persistent as the cryptocurrency matures. Furthermore, out of the three sentiment variables, *Vader* reacted most strongly to *Volatility*. A week of high *Volatility* was usually followed by a week of lower *Vader*, and this result was observed for all three cryptocurrencies. In the case of Bitcoin, the significance of *Volatility* diminished over time, suggesting that as a cryptocurrency matures, the sentiment becomes less sensitive to volatility. In addition, the restricted model indicated that *GoogleTrends* had a negative impact on *Vader* for both Bitcoin and Litecoin. Furthermore, *Addresses* had a negative impact on *Vader* for both Ether and Bitcoin in the full model.

*Valence* has the largest dictionary of words of the three methods used in this paper, but it lacks the more advanced textual understanding *Vader* has. In the restricted model, *TradingVolume* had a negative impact on the *Valence* of Bitcoin and Litecoin. This relationship was not observed for the other sentiment measures, which further emphasizes the differences between the sentiment measures. Lastly, *Volatility* and *GoogleTrends* both negatively impacted the *Valence* of Bitcoin and Litecoin.

**Table 5.9:** Summary of the regression results for the restricted models of *Tone*, *Vader*, and *Valence*, which can be found in Appendix C.5. A significant result is reported with the symbol + for a positive coefficient, and - for negative. Specifically, one, two, and three symbols represent 10%, 5%, and 1% significance respectively.

| **Tone**$_{t+1}$ | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| Tone$_t$ | | | + + + | + + | + |
| Articles$_t$ | - | | + + | | + + |
| Addresses$_t$ | - - | | | | |
| GoogleTrends$_t$ | | + | | | |
| Return$_t$ | | | | + + | |
| Volatility$_t$ | - - - | | | | |
| TradingVolume$_t$ | | | | | |
| TransactionVolume$_t$ | | | | | |

| **Vader**$_{t+1}$ | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| Vader$_t$ | + | + + | + + + | + + | |
| Articles$_t$ | | | | - - | |
| Addresses$_t$ | | | - | - - - | |
| GoogleTrends$_t$ | | - | - - - | | - - - |
| Return$_t$ | | | | + | |
| Volatility$_t$ | - - - | - | | - - - | - - - |
| TradingVolume$_t$ | | | - - - | | |
| TransactionVolume$_t$ | | | - - | - - | |

| **Valence**$_{t+1}$ | Bitcoin | | | Ether | Litecoin |
|---|---|---|---|---|---|
| | Period 1 | Period 2 | Period 3 | Period 3 | Period 3 |
| Valence$_t$ | | | + + + | | |
| Articles$_t$ | | | | | |
| Addresses$_t$ | | | | | |
| GoogleTrends$_t$ | | | - - | | - - |
| Return$_t$ | | | | + + | - - - |
| Volatility$_t$ | | - - | | | - - - |
| TradingVolume$_t$ | | - - - | - - - | | - - - |
| TransactionVolume$_t$ | | | | | - - |

**Table 5.10:** Regression results from the full model with Sentiment (*Tone*, *Vader*, or *Valence*) as dependent variable. Explanatory variables sampled one week prior to the dependent variable. Every column reports the result from one regression. *Sentiment* is a placeholder for either *Tone*, *Vader*, or *Valence*. Coefficients are presented first, with robust standard errors reported in parentheses. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: Sentiment$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence | Tone | Vader | Valence |
| Intercept | 0.520*** | 0.158*** | 4.872*** | 0.381*** | 0.126*** | 5.152*** | 0.029 | 0.105*** | 3.671*** | 0.493*** | 0.364*** | 5.659*** | 0.526*** | 0.426*** | 5.785*** |
| | (0.051) | (0.025) | (0.674) | (0.074) | (0.035) | (0.636) | (0.034) | (0.037) | (0.711) | (0.083) | (0.032) | (0.346) | (0.086) | (0.053) | (0.668) |
| Sentiment$_t$ | -0.222** | 0.098 | 0.139 | -0.026 | 0.120 | 0.085 | 0.810*** | 0.460*** | 0.351*** | 0.119 | 0.041 | 0.023 | 0.146 | -0.098 | 0.010 |
| | (0.101) | (0.121) | (0.119) | (0.138) | (0.123) | (0.112) | (0.074) | (0.139) | (0.126) | (0.111) | (0.071) | (0.060) | (0.135) | (0.131) | (0.114) |
| Articles$_t$ | -0.014 | -0.005 | -0.001 | -0.001 | -0.008 | -0.004 | 0.019** | 0.007 | 0.005 | -0.001 | -0.001 | 0.002 | 0.023** | 0.007* | 0.006** |
| | (0.010) | (0.005) | (0.006) | (0.006) | (0.006) | (0.006) | (0.009) | (0.004) | (0.003) | (0.009) | (0.005) | (0.005) | (0.009) | (0.004) | (0.003) |
| Addresses$_t$ | -3.346** | -0.482 | 0.751 | 10.773 | 9.520* | 5.409 | -0.847 | -2.120 | -3.204 | -1.532 | -3.461*** | -1.748* | - | - | - |
| | (1.674) | (0.759) | (0.739) | (8.128) | (5.447) | (4.881) | (4.142) | (2.332) | (2.643) | (1.677) | (1.049) | (0.971) | - | - | - |
| GoogleTrends$_t$ | 0.015** | 0.006* | 0.002 | 0.004 | -0.003 | -0.003 | -0.001 | -0.006 | -0.004 | -0.025** | -0.011 | -0.012 | -0.011 | -0.010** | 0.004 |
| | (0.007) | (0.003) | (0.003) | (0.003) | (0.002) | (0.003) | (0.009) | (0.005) | (0.005) | (0.011) | (0.009) | (0.007) | (0.012) | (0.005) | (0.003) |
| Return$_t$ | 0.305 | 0.107 | 0.088 | 0.357 | -0.044 | -0.104 | 0.242 | 0.104 | 0.145 | 0.312 | 0.300** | 0.302* | 0.097 | 0.009 | -0.119 |
| | (0.210) | (0.102) | (0.111) | (0.377) | (0.230) | (0.250) | (0.160) | (0.104) | (0.091) | (0.236) | (0.136) | (0.156) | (0.225) | (0.114) | (0.082) |
| Volatility$_t$ | -0.401 | -0.361** | -0.289 | 0.788 | -0.841 | -0.606 | 1.246 | 0.508 | 0.637 | -1.140 | -1.698*** | -0.799 | -1.790* | -1.998*** | -2.145*** |
| | (0.280) | (0.156) | (0.183) | (1.239) | (0.885) | (1.064) | (1.262) | (0.487) | (0.478) | (1.530) | (0.584) | (0.803) | (0.934) | (0.500) | (0.547) |
| TradingVolume$_t$ | -0.070 | -0.008 | -0.014 | -0.025 | -0.025 | -0.081** | -0.081 | -0.067 | -0.093* | 0.254* | 0.234*** | 0.114 | 0.039 | 0.048 | 0.006 |
| | (0.083) | (0.040) | (0.042) | (0.058) | (0.034) | (0.035) | (0.095) | (0.055) | (0.052) | (0.143) | (0.083) | (0.075) | (0.074) | (0.043) | (0.039) |
| TransactionVolume$_t$ | 0.064 | 0.012 | 0.005 | -0.008 | 0.036 | 0.089 | -0.109 | -0.020 | 0.035 | -0.057 | -0.112* | -0.028 | 0.003 | 0.003 | -0.007 |
| | (0.106) | (0.053) | (0.057) | (0.156) | (0.087) | (0.068) | (0.101) | (0.074) | (0.055) | (0.083) | (0.059) | (0.052) | (0.011) | (0.007) | (0.007) |
| $R^2$ | 0.20 | 0.13 | 0.07 | 0.04 | 0.18 | 0.16 | 0.74 | 0.33 | 0.27 | 0.19 | 0.32 | 0.18 | 0.16 | 0.22 | 0.27 |
| $R^2_{adj}$ | 0.12 | 0.04 | -0.02 | -0.05 | 0.09 | 0.07 | 0.72 | 0.26 | 0.20 | 0.10 | 0.25 | 0.10 | 0.08 | 0.15 | 0.21 |

# 6 Conclusions

Cryptocurrencies have shown remarkably high returns and volatilities since Bitcoin, the first cryptocurrency, was introduced in 2008. Unlike other asset classes, cryptocurrencies are fully digital. Hence, their trading activity and usage are likely to behave differently compared to traditional asset classes. Sentiment analysis has been used increasingly in later years to analyze and predict price movements of assets. This paper attempted to predict the trading activity and usage of Bitcoin, Ether, and Litecoin, using the sentiment in news articles with previously studied control variables. Also, we studied which factors impacted the sentiment in news articles. To perform the sentiment analysis, three different dictionary-based approaches were used and compared, namely: tone-, vader-, and valence-dictionary. Surprisingly, the sentiment measures often gave different results, despite being significantly correlated with each other. It appeared that each method captured different elements of the sentiment in an article and that each thus provided unique and valuable information. All studies in the reviewed literature had only applied one sentiment measure. We conclude that it is not sufficient to only use one dictionary to capture the sentiment in an article.

Different cryptocurrencies react differently to the sentiment in news articles. For Bitcoin, the sentiment has significant predicting ability for return, volatility, trading volume, and transaction volume. However, the result varied depending on the period and model. For Ether, the sentiment has significant predicting ability for return, trading volume, and transaction volume. The sentiment has no impact on the financial variables of Litecoin. In the literature, only two studies have applied the dictionary-based approach to study the impact of news sentiment on financial factors of Bitcoin. Polasik et al. (2015) find that the sentiment of news positively impacts Bitcoin price, while our results suggest that a similar conclusion can only be drawn for Ether. Mai et al. (2018) do not find any significant result for news sentiment. However, our findings suggest that there is valuable information in the news sentiment, and that different cryptocurrencies react differently to sentiment.

In the model predicting return, an increase in addresses has a strong positive impact on the return of Bitcoin from 30/04/2017-13/01/2019. This confirms the finding by Koutmos (2018), and leads to the conclusion that Bitcoin's network value affects its return. Several studies also find Google searches and trading volume to predict the return of Bitcoin. We find both variables to be significant for Bitcoin only during 09/08/2015-29/04/2017.

When it comes to predicting volatility, this paper confirms the consensus that the volatilities of cryptocurrencies exhibit autocorrelation. Furthermore, several studies find that trading volume and Google searches predict volatility. We find no convincing predicting ability for trading volume, whereas Google searches predict Litecoin volatility. Interestingly, we find that the number of articles strongly predicts the volatility of Bitcoin, a relationship previously unreported in the literature. The direction of the impact of the number of articles changed depending on the period. In addition, return predicts the volatility of Bitcoin, with changing direction over the periods. The volatility of Ether was positively impacted by the number of new addresses.

We find several significant results for predicting trading volume. Most studies report that Google searches predict trading volume. We find this to be true for Bitcoin and Litecoin, where an increase in Google searches lead to an increase in trading volume. In addition, we discover that a week of high volatility is usually followed by a week of lower trading volume across all three cryptocurrencies, another relationship previously unreported in the literature. Lastly, it appears that return has a positive impact on trading volume for Bitcoin and Litecoin.

Transaction volume, a measure of the usage of a cryptocurrency, has not frequently been studied in the literature. We find strong evidence that a week of high return predicts a higher transaction volume for Bitcoin and Ether. Also, volatility appears to have a negative impact on the transaction volume of Bitcoin and Ether, especially in the later periods of Bitcoin. Furthermore, the trading volume of Ether positively impacts transaction volume. Lastly, the transaction volume of Litecoin is mostly unaffected by the variables included in this research.

The last part of this paper investigated which factors impacted the sentiment in news articles. In general, there are surprisingly large differences related to which factors predict which sentiment variable. However, some common trends are observed. The sentiment in news articles about Litecoin is predicted by both volatility and the number of published articles. While volatility has a negative impact on all measures of sentiment, the number of articles has a positive impact. Moreover, all three measures of sentiment exhibit strong autocorrelation for Bitcoin from 30/04/2017-13/01/2019. This leads to the conclusion that the persistence of the sentiment variables either increase over time, or that data from earlier periods were too noisy to show autocorrelation. Return positively impacts the sentiment of Ether, a relationship not observed for neither Bitcoin nor Litecoin. In addition, new addresses appear to have a negative impact, while trading volume appears to have a positive impact on Ether's sentiment. Lastly, there seems to be a general trend that volatility negatively impacts news sentiment for all three cryptocurrencies.

### Further work

First of all, the research was performed on weekly data, as the number of available news articles did not allow for daily data. If the research could be replicated using daily data, one could test smaller increments of lagged sentiment to improve the understanding of both how sentiment is created, and how it affects other factors. Secondly, this paper measured the sentiment on English language articles. Since cryptocurrencies are traded globally, there may be more information in the sentiment of foreign language articles. Thirdly, all articles were weighted equally in this study, but the relevance, importance, and reach of the articles differed. Thus, the accuracy of the sentiment may be increased if the articles were to be categorized and weighted. Moreover, by using more advanced methods of textual understanding, it may be possible to improve the accuracy in the sentiment analysis. One example is to use dictionaries specifically designed to analyze the sentiment of text about cryptocurrencies. Another example is to use more advanced textual processing algorithms, for example, by applying machine learning. Lastly, given the different behaviors of the cryptocurrencies, one may attempt to study the reason behind these differences and compare them with additional cryptocurrencies.

# Appendices

# A Tone-dictionary Developed by Henry (2008)

Positive and negative words developed by Henry (2008) to measure the tone within earnings press releases. Tone is defined as the count of positive words subtracted by the count of negative words, divided by the sum of positive and negative words.

**POSITIVITY** word list:
positive positives success successes successful succeed succeeds succeeding succeeded accomplish accomplishes accomplishing accomplished accomplishment accomplishments strong strength strengths certain certainty definite solid excellent good leading achieve achieves achieved achieving achievement achievements progress progressing deliver delivers delivered delivering leader leading pleased reward rewards rewarding rewarded opportunity opportunities enjoy enjoys enjoying enjoyed encouraged encouraging up increase increases increasing increased rise rises rising rose risen improve improves improving improved improvement improvements strengthen strengthens strengthening strengthened stronger strongest better best more most above record high higher highest greater greatest larger largest grow grows growing grew grown growth expand expands expanding expanded expansion exceed exceeds exceeded exceeding beat beats beating

**NEGATIVITY** word list:
negative negatives fail fails failing failure weak weakness weaknesses difficult difficulty hurdle hurdles obstacle obstacles slump slumps slumping slumped uncertain uncertainty unsettled unfavorable downturn depressed disappoint disappoints disappointing disappointed disappointment risk risks risky threat threats penalty penalties down decrease decreases decreasing decreased decline declines declining declined fall falls falling fell fallen drop drops dropping dropped deteriorate deteriorates deteriorating deteriorated worsen worsens worsening weaken weakens weakening weakened worse worst low lower lowest less least smaller smallest shrink shrinks shrinking shrunk below under challenge challenges challenging challenged

# B MATLAB and Python Code

## Preprocessing (MATLAB)

```
%**********************************************************************
% Description: Preprocess input text by:
%       1)Remove unimportant characters such as # $ % & @ ( ) etc.
%       2)Convert text to lower case
%
% Returns: Preprocessed text based on the criteria given above
%**********************************************************************

function processedText = preProcess(text)
%load list of unimportant characters to remove from input text
characterList = loadCharacterList('characterlist.txt');

processedText = '';
for c = 1:length(text)
    character = text(c);
    isCharacterInList = false; %character is not in characterlist
    for i = 1:length(characterList)
        if strcmp(character, char(characterList(i)))
            isCharacterInList = true; %character is in characterlist
            break;
        end
    end
    if ~isCharacterInList %keep character if not in characterlist
        processedText = [processedText lower(character)];
    end
end

end %function
```

## Sentiment analysis: Tone-dictionary (MATLAB)

```
%*************************************************************************
% Description: Sentiment analysis using tone-dictionary
%
% Prints: 1)Word count: number of words in text
%         2)Number of negative- and positive words
%         3)Tone score: the sentiment score of the text
%*************************************************************************

function toneAnalysis(text)
%preprocessing: remove characters such as # $ % & ' ( ) * etc.
words = preProcess(text);

%load negative and positive words from dictionary
negList = getWords('negative.txt');
posList = getWords('positive.txt');

%perform sentiment analysis
wordCount=0; pos = 0; neg = 0;
while words
    isPositiveWord = 0;
    [word, words] = strtok(words);
    if ~isempty(word)
    wordCount=wordCount+1;
    end
    for i = 1:size(posList,2)          %check for positive words
        if strcmp(word,posList{i})
            pos = pos+1;               %positive word found
            isPositiveWord = 1;
            break;
        end
    end
    if ~isPositiveWord                 %check for negative words
        for i = 1:size(negList,2)
            if strcmp(word,negList{i})
                neg = neg+1;           %negative word found
                break;
            end
        end
    end
end

%print out results
fprintf('Word count: \t\t%d\n', wordCount);
fprintf('Positive words: \t%d\n', pos);
fprintf('Negative words: \t%d\n', neg);
fprintf('Tone score: \t\t%.2f\n', (pos-neg)/(pos+neg));

end %function
```

## Sentiment analysis: Vader-dictionary (Python)

```python
#**************************************************************************
# Description: Sentiment analysis using vader-dictionary by implementing
#
#       1) Hutto and Gilbert (2014) program for scoring each sentence
#          found on https://github.com/cjhutto/vaderSentiment
#
#       2) Natural Language Toolkit NLTK for spliting sentences from text
#          found on https://www.nltk.org/index.html
#
# Prints: 1) Number of sentences of input text
#         2) Vader score: the sentiment score of the input text
#**************************************************************************

import nltk
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
from nltk.tokenize import sent_tokenize, word_tokenize

def vaderAnalysis(text):
    #define analyzer
    analyzer = SentimentIntensityAnalyzer()

    #store all sentences of input text
    sentences = []

    #store compound score of each sentence of input text
    compound = []

    #sentenize text and store in sentences
    for x in sent_tokenize(text):
        sentences.append(x)

        #iterate though every sentence in text
        i = 0

        #compute and store compound score of each sentence
        while (i<len(sentences)):
            k = analyzer.polarity_scores(sentences[i])
            compound.append(k['compound'])
            i = i+1

        #compute Vader score of input text. (Checks for division by 0.)
        if len(compound) != 0:
            vaderScore = sum(compound) / float(len(compound))
        else:
            vaderScore = 'Error: division by 0'

    #print out results
    print("Sentence count: " + str(len(sentences)))
    print("Vader score: " + str(vaderScore))
    return
```

# Sentiment analysis: Valence-dictionary (MATLAB)

```
%**********************************************************************
% Description: Sentiment analysis using valence-dictionary
%
% Prints: 1)Word count: number of words in text
%         2)Valence words: number of words found in dictionary
%         3)Valence score: the sentiment score of the text
%**********************************************************************

function valenceAnalysis(text)
%preprocessing: remove characters such as # $ % & ' ( ) * etc.
words = preProcess(text);

%load valence dictionary
dictionary = loadDictionary('dictionary.xlsx');

%perform sentiment analysis
wordCount = 0; valenceSum = 0; valenceWords = 0;
while words
    [word, words] = strtok(words);
    if ~isempty(word)
        wordCount = wordCount+1;
        [pos1, pos2] = searchPositions(word); %location region of the word
        for i = pos1:pos2
                if strcmp(word,dictionary{i,2}) %word found in dictionary
                    valenceWords = valenceWords + 1;
                    valenceSum = valenceSum + str2num(dictionary{i,3});
                    break;
                end
        end
    end
end

%print out results
fprintf('Word count: \t%d\n', wordCount);
fprintf('Valence words: \t%d\n', valenceWords);
fprintf('Valence score: \t%.2f\n', valenceSum/valenceWords);

end %function
```

# C Regression Results of Restricted Model

**Table C.1:** Regression results from the restricted model with *Return* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: $\text{Return}_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| $\text{Return}_t$ | -0.147 | (0.217) | 0.03 | -0.002 | (0.128) | 0.00 | -0.021 | (0.119) | 0.00 | 0.165 | (0.106) | 0.03 | 0.028 | (0.115) | 0.00 |
| $\text{Tone}_t$ | -0.065 | (0.047) | 0.05 | -0.026 | (0.037) | 0.01 | 0.028 | (0.031) | 0.01 | 0.104* | (0.062) | 0.06 | 0.073 | (0.068) | 0.01 |
| $\text{Vader}_t$ | 0.012 | (0.088) | 0.03 | -0.020 | (0.053) | 0.00 | -0.188** | (0.093) | 0.04 | 0.033 | (0.078) | 0.03 | -0.021 | (0.144) | 0.00 |
| $\text{Valence}_t$ | 0.010 | (0.074) | 0.03 | 0.007 | (0.062) | 0.00 | -0.128 | (0.102) | 0.01 | 0.111 | (0.090) | 0.04 | -0.110 | (0.123) | 0.01 |
| $\text{Articles}_t$ | -0.000 | (0.005) | 0.03 | -0.003 | (0.002) | 0.02 | 0.008** | (0.003) | 0.06 | 0.005 | (0.005) | 0.04 | 0.003 | (0.005) | 0.00 |
| $\text{Addresses}_t$ | 0.030 | (1.539) | 0.03 | 0.371 | (2.238) | 0.00 | 11.847*** | (1.897) | 0.21 | 1.420* | (0.762) | 0.08 | - | - | - |
| $\text{GoogleTrends}_t$ | 0.003 | (0.004) | 0.04 | -0.003** | (0.002) | 0.08 | 0.006 | (0.005) | 0.03 | 0.000 | (0.004) | 0.03 | 0.007 | (0.007) | 0.02 |
| $\text{Volatility}_t$ | 0.188 | (0.200) | 0.05 | -0.557 | (0.376) | 0.04 | 0.710 | (0.572) | 0.02 | 1.036* | (0.586) | 0.06 | 0.058 | (0.570) | 0.00 |
| $\text{TradeVolume}_t$ | 0.010 | (0.039) | 0.03 | -0.017 | (0.014) | 0.02 | 0.057 | (0.036) | 0.03 | 0.050 | (0.038) | 0.05 | -0.007 | (0.034) | 0.00 |
| $\text{TransactionVolume}_t$ | 0.026 | (0.061) | 0.04 | -0.001 | (0.042) | 0.00 | 0.045 | (0.050) | 0.01 | 0.044 | (0.035) | 0.05 | -0.004 | (0.006) | 0.00 |

**Table C.2:** Regression results from the restricted model with *Volatility* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: Volatility$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| Volatility$_t$ | 0.626*** | (0.231) | 0.41 | 0.371*** | (0.100) | 0.14 | 0.554*** | (0.089) | 0.30 | 0.456*** | (0.115) | 0.21 | 0.591*** | (0.095) | 0.36 |
| Tone$_t$ | -0.007 | (0.032) | 0.41 | 0.011 | (0.015) | 0.15 | 0.012** | (0.005) | 0.33 | -0.002 | (0.008) | 0.21 | 0.004 | (0.012) | 0.36 |
| Vader$_t$ | -0.011 | (0.065) | 0.41 | 0.007 | (0.021) | 0.14 | -0.023 | (0.022) | 0.32 | -0.007 | (0.020) | 0.21 | -0.004 | (0.030) | 0.36 |
| Valence$_t$ | 0.021 | (0.054) | 0.42 | 0.008 | (0.022) | 0.14 | -0.012 | (0.025) | 0.31 | 0.016 | (0.014) | 0.22 | -0.036 | (0.028) | 0.37 |
| Articles$_t$ | 0.002 | (0.003) | 0.42 | -0.002*** | (0.001) | 0.18 | 0.002*** | (0.001) | 0.42 | 0.001** | (0.001) | 0.24 | 0.000 | (0.001) | 0.36 |
| Addresses$_t$ | 0.522 | (0.609) | 0.43 | -0.245 | (0.708) | 0.14 | 1.237** | (0.540) | 0.36 | 0.348*** | (0.123) | 0.30 | - | - | - |
| GoogleTrends$_t$ | -0.004* | (0.002) | 0.44 | 0.000 | (0.000) | 0.15 | 0.001 | (0.001) | 0.32 | 0.000 | (0.001) | 0.21 | 0.002** | (0.001) | 0.40 |
| Return$_t$ | 0.033 | (0.056) | 0.42 | 0.102*** | (0.036) | 0.22 | -0.012 | (0.021) | 0.31 | 0.008 | (0.021) | 0.21 | 0.024 | (0.018) | 0.37 |
| TradeVolume$_t$ | 0.006 | (0.017) | 0.42 | 0.010* | (0.006) | 0.18 | 0.004 | (0.008) | 0.31 | 0.012* | (0.007) | 0.24 | 0.007 | (0.008) | 0.36 |
| TransactionVolume$_t$ | -0.011 | (0.023) | 0.42 | 0.028** | (0.014) | 0.20 | 0.011 | (0.008) | 0.32 | 0.008* | (0.005) | 0.23 | 0.000 | (0.001) | 0.36 |

**Table C.3:** Regression results from the restricted model with *TradeVolume* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: TradeVolume$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| TradeVolume$_t$ | 0.518*** | (0.097) | 0.31 | 0.667*** | (0.066) | 0.44 | 0.697*** | (0.071) | 0.48 | 0.796*** | (0.078) | 0.64 | 0.749*** | (0.079) | 0.59 |
| Tone$_t$ | 0.521*** | (0.177) | 0.36 | 0.037 | (0.193) | 0.44 | 0.081 | (0.075) | 0.48 | 0.238** | (0.094) | 0.66 | 0.180 | (0.183) | 0.60 |
| Vader$_t$ | 0.650* | (0.332) | 0.33 | 0.245 | (0.267) | 0.45 | -0.308 | (0.315) | 0.49 | 0.324 | (0.202) | 0.66 | 0.140 | (0.392) | 0.59 |
| Valence$_t$ | 0.601* | (0.363) | 0.33 | 0.412 | (0.370) | 0.46 | 0.002 | (0.363) | 0.48 | 0.530*** | (0.176) | 0.68 | 0.155 | (0.368) | 0.59 |
| Articles$_t$ | 0.008 | (0.015) | 0.31 | -0.039*** | (0.012) | 0.49 | 0.029*** | (0.008) | 0.54 | 0.010 | (0.008) | 0.65 | 0.005 | (0.008) | 0.59 |
| Addresses$_t$ | 3.150 | (3.510) | 0.32 | -1.353 | (10.338) | 0.44 | 22.203*** | (6.786) | 0.54 | 1.726 | (1.419) | 0.65 | - | - | - |
| GoogleTrends$_t$ | 0.002 | (0.010) | 0.31 | 0.002 | (0.005) | 0.44 | 0.038*** | (0.011) | 0.53 | -0.013 | (0.011) | 0.65 | 0.023* | (0.012) | 0.60 |
| Return$_t$ | 0.726* | (0.373) | 0.34 | 1.628*** | (0.620) | 0.48 | 0.447* | (0.244) | 0.50 | 0.416 | (0.276) | 0.66 | 0.553** | (0.267) | 0.61 |
| Volatility$_t$ | -0.239 | (0.555) | 0.31 | -3.989** | (1.585) | 0.46 | -4.025*** | (1.300) | 0.52 | -5.054*** | (1.434) | 0.70 | -3.277* | (1.928) | 0.61 |
| TransactionVolume$_t$ | 0.202 | (0.178) | 0.32 | 0.325 | (0.249) | 0.46 | 0.456*** | (0.133) | 0.54 | 0.108 | (0.096) | 0.65 | -0.005 | (0.018) | 0.59 |

**Table C.4:** Regression results from the restricted model with *TransactionVolume* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: TransactionVolume$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| TransactionVolume$_t$ | 0.653*** | (0.095) | 0.54 | 0.703*** | (0.080) | 0.50 | 0.811*** | (0.062) | 0.66 | 0.793*** | (0.066) | 0.66 | 0.496*** | (0.088) | 0.25 |
| Tone$_t$ | 0.250** | (0.115) | 0.57 | -0.028 | (0.104) | 0.50 | 0.025 | (0.060) | 0.66 | 0.147 | (0.166) | 0.67 | -0.734 | (1.082) | 0.26 |
| Vader$_t$ | 0.095 | (0.261) | 0.54 | 0.028 | (0.213) | 0.50 | -0.133 | (0.146) | 0.66 | 0.339* | (0.177) | 0.68 | 0.013 | (1.890) | 0.25 |
| Valence$_t$ | 0.311 | (0.256) | 0.55 | 0.079 | (0.222) | 0.50 | -0.103 | (0.170) | 0.66 | 0.457** | (0.199) | 0.68 | 1.204 | (1.857) | 0.26 |
| Articles$_t$ | 0.012 | (0.012) | 0.55 | -0.008 | (0.007) | 0.50 | 0.007 | (0.006) | 0.67 | 0.011 | (0.014) | 0.66 | 0.113 | (0.086) | 0.27 |
| Addresses$_t$ | -0.618 | (1.978) | 0.54 | -3.850 | (4.376) | 0.50 | 9.339 | (5.709) | 0.68 | 1.522 | (1.694) | 0.66 | - | - | - |
| GoogleTrends$_t$ | 0.016 | (0.011) | 0.56 | 0.003 | (0.003) | 0.50 | -0.002 | (0.008) | 0.66 | -0.011 | (0.012) | 0.66 | 0.149* | (0.079) | 0.29 |
| Return$_t$ | 0.604*** | (0.181) | 0.58 | 0.994*** | (0.285) | 0.57 | 0.630*** | (0.157) | 0.72 | 0.865*** | (0.272) | 0.72 | 2.362 | (1.639) | 0.27 |
| Volatility$_t$ | 0.536 | (0.480) | 0.55 | -1.883** | (0.892) | 0.52 | -2.097*** | (0.772) | 0.69 | -1.366 | (1.047) | 0.67 | 10.066 | (7.352) | 0.27 |
| TradeVolume$_t$ | 0.004 | (0.109) | 0.54 | -0.033 | (0.039) | 0.50 | -0.045 | (0.072) | 0.66 | 0.251** | (0.102) | 0.68 | 0.892 | (0.574) | 0.28 |

**Table C.5:** Regression results from the restricted model with *Tone* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: $\text{Tone}_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| $\text{Tone}_t$ | -0.079 | (0.091) | 0.01 | 0.006 | (0.128) | 0.00 | 0.833*** | (0.058) | 0.70 | 0.223** | (0.109) | 0.05 | 0.190* | (0.115) | 0.04 |
| $\text{Articles}_t$ | -0.016* | (0.009) | 0.05 | 0.003 | (0.006) | 0.00 | 0.013** | (0.006) | 0.72 | 0.000 | (0.008) | 0.05 | 0.017** | (0.008) | 0.10 |
| $\text{Addresses}_t$ | -3.680** | (1.696) | 0.09 | 10.186 | (8.163) | 0.02 | 0.904 | (3.467) | 0.70 | 0.312 | (0.982) | 0.05 | - | - | - |
| $\text{GoogleTrends}_t$ | 0.000 | (0.006) | 0.01 | 0.004* | (0.002) | 0.01 | 0.001 | (0.006) | 0.70 | -0.013 | (0.009) | 0.07 | -0.005 | (0.008) | 0.04 |
| $\text{Return}_t$ | 0.240 | (0.221) | 0.02 | 0.280 | (0.327) | 0.01 | 0.238 | (0.145) | 0.71 | 0.469** | (0.199) | 0.13 | 0.067 | (0.223) | 0.04 |
| $\text{Volatility}_t$ | -0.750*** | (0.257) | 0.07 | 0.557 | (0.801) | 0.00 | 0.920 | (0.759) | 0.71 | -0.172 | (1.156) | 0.05 | -0.922 | (0.846) | 0.05 |
| $\text{TradeVolume}_t$ | -0.088 | (0.059) | 0.05 | 0.010 | (0.048) | 0.00 | -0.006 | (0.040) | 0.70 | 0.065 | (0.050) | 0.06 | -0.001 | (0.054) | 0.04 |
| $\text{TransactionVolume}_t$ | -0.039 | (0.080) | 0.01 | 0.047 | (0.116) | 0.00 | -0.036 | (0.064) | 0.70 | 0.035 | (0.047) | 0.05 | 0.006 | (0.011) | 0.04 |

**Table C.6:** Regression results from the restricted model with *Vader* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: Vader$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| Vader$_t$ | 0.192* | (0.112) | 0.03 | 0.251** | (0.101) | 0.06 | 0.486*** | (0.135) | 0.23 | 0.240** | (0.119) | 0.06 | 0.075 | (0.117) | 0.01 |
| Articles$_t$ | -0.005 | (0.005) | 0.05 | -0.009 | (0.006) | 0.09 | 0.000 | (0.003) | 0.23 | -0.010** | (0.005) | 0.09 | 0.001 | (0.004) | 0.01 |
| Addresses$_t$ | -0.841 | (0.658) | 0.05 | 5.568 | (5.535) | 0.08 | -3.320* | (1.743) | 0.25 | -2.284*** | (0.671) | 0.15 | - | - | - |
| GoogleTrends$_t$ | -0.000 | (0.004) | 0.03 | -0.005* | (0.003) | 0.10 | -0.008*** | (0.003) | 0.28 | -0.009 | (0.008) | 0.08 | -0.010*** | (0.004) | 0.09 |
| Return$_t$ | 0.121 | (0.091) | 0.05 | -0.008 | (0.221) | 0.06 | 0.080 | (0.096) | 0.24 | 0.251* | (0.144) | 0.10 | -0.092 | (0.089) | 0.02 |
| Volatility$_t$ | -0.369*** | (0.116) | 0.09 | -1.336* | (0.688) | 0.11 | -0.325 | (0.372) | 0.24 | -1.676*** | (0.582) | 0.11 | -1.711*** | (0.509) | 0.14 |
| TradeVolume$_t$ | -0.019 | (0.024) | 0.04 | -0.041 | (0.030) | 0.09 | -0.079*** | (0.026) | 0.28 | -0.037 | (0.028) | 0.07 | -0.039 | (0.031) | 0.03 |
| TransactionVolume$_t$ | 0.002 | (0.029) | 0.03 | -0.067 | (0.077) | 0.08 | -0.079** | (0.035) | 0.27 | -0.060** | (0.030) | 0.08 | -0.002 | (0.006) | 0.01 |

**Table C.7:** Regression results from the restricted model with *Valence* as dependent variable using explanatory variables sampled one week prior. The lagged dependant variable is included in every regression. Every row reports the regression result for one variable. In each group of three columns, the first presents the coefficients, the second presents the robust standard errors and the third presents the $R^2$. The different sample periods cover weekly data from 2013 to 2019. The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

| | **Dependent variable: Valence$_{t+1}$** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bitcoin | | | | | | | | | Ether | | | Litecoin | | |
| | 17/11/2013 - 08/08/2015 | | | 09/08/2015 - 29/08/2017 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | | 30/04/2017 - 13/01/2019 | | |
| | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ | Coef. | SE | $R^2$ |
| Valence$_t$ | 0.181 | (0.111) | 0.03 | 0.141 | (0.097) | 0.02 | 0.405*** | (0.112) | 0.16 | 0.072 | (0.082) | 0.01 | 0.159 | (0.109) | 0.03 |
| Articles$_t$ | -0.002 | (0.005) | 0.03 | -0.006 | (0.007) | 0.04 | 0.000 | (0.002) | 0.16 | -0.002 | (0.004) | 0.01 | -0.001 | (0.003) | 0.03 |
| Addresses$_t$ | 0.213 | (0.465) | 0.03 | 1.936 | (5.299) | 0.02 | -3.233 | (2.277) | 0.18 | -0.579 | (0.552) | 0.01 | - | - | - |
| GoogleTrends$_t$ | -0.001 | (0.003) | 0.03 | -0.005 | (0.004) | 0.06 | -0.006** | (0.003) | 0.20 | -0.010 | (0.006) | 0.04 | -0.007** | (0.004) | 0.08 |
| Return$_t$ | 0.124 | (0.102) | 0.04 | -0.056 | (0.255) | 0.02 | 0.120 | (0.090) | 0.18 | 0.312** | (0.147) | 0.09 | -0.185*** | (0.057) | 0.08 |
| Volatility$_t$ | -0.225 | (0.147) | 0.05 | -1.468** | (0.736) | 0.08 | -0.263 | (0.306) | 0.17 | -0.810 | (0.661) | 0.02 | -1.949*** | (0.443) | 0.21 |
| TradeVolume$_t$ | -0.009 | (0.025) | 0.03 | -0.076*** | (0.029) | 0.11 | -0.062*** | (0.024) | 0.21 | 0.007 | (0.028) | 0.01 | -0.066*** | (0.024) | 0.11 |
| TransactionVolume$_t$ | 0.015 | (0.031) | 0.03 | -0.084 | (0.079) | 0.04 | -0.046 | (0.030) | 0.18 | 0.000 | (0.029) | 0.01 | -0.012** | (0.005) | 0.09 |

# Bibliography


Aalborg, H. A., Molnár, P., de Vries, J. E., 2018. What can explain the price, volatility and trading volume of bitcoin? Finance Research Letters.

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.

Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., Manicardi, S., 2016. A comparison between preprocessing techniques for sentiment analysis in twitter.

Balcilar, M., Bouri, E., Gupta, R., Roubaud, D., 2017. Can volume predict bitcoin returns and volatility? a quantiles-based approach. Economic Modelling 64, 74–81.

Bijl, L., Kringhaug, G., Molnár, P., Sandvik, E., 2016. Google searches and stock returns. International Review of Financial Analysis 45, 150–156.

Blau, B. M., 2017. Price dynamics and speculative trading in bitcoin. Research in International Business and Finance 41, 493–499.

Bukovina, J., Marticek, M., et al., 2016. Sentiment and bitcoin volatility.

Chuen, K., Lee, D., Guo, L., Wang, Y., 2017. Cryptocurrency: A new investment opportunity?

Ciaian, P., Rajcaniova, M., Kancs, d., 2016. The economics of bitcoin price formation. Applied Economics 48 (19), 1799–1815.

De Franco, G., Hope, O.-K., Vyas, D., Zhou, Y., 2015. Analyst report readability. Contemporary Accounting Research 32 (1), 76–104.

Dickerson, A., 2018. Algorithmic trading of bitcoin using wikipedia and google search volume.

Dyhrberg, A. H., 2016. Bitcoin, gold and the dollar–a garch volatility analysis. Finance Research Letters 16, 85–92.

Figa-Talamanca, G., Patacca, M., 2018. Does market attention affect bitcoin returns and volatility?

Garcia, D., Schweitzer, F., 2015. Social signals and algorithmic trading of bitcoin. Royal Society open science 2 (9), 150288.

Garman, M. B., Klass, M. J., 1980. On the estimation of security price volatilities from historical data. Journal of business, 67–78.

Ghiasvand, A., 2018. Investigating cryptocurrencies: return, exchange volume and volatility with investor's attention and investor sentiment: an empirical analysis.

Glaser, F., Zimmermann, K., Haferkorn, M., Weber, M. C., Siering, M., 2014. Bitcoin-asset or currency? revealing users' hidden intentions. Revealing Users' Hidden Intentions (April 15, 2014). ECIS.

Guo, L., Shi, F., Tu, J., 2016. Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science 2 (3), 153–170.

Henry, E., 2008. Are investors influenced by how earnings press releases are written? The Journal of Business Communication (1973) 45 (4), 363–407.

Heston, S. L., Sinha, N. R., 2016. News versus sentiment: predicting stock returns from news stories; 2016.

Hutto, C. J., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Jerdack, N., Dauletbek, A., Divine, M., Hult, M., Carvalho, A., 2018. Understanding what drives bitcoin trading activities. arXiv preprint arXiv:1901.00720.

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

Kim, N., Lučivjanská, K., Molnár, P., Villa, R., 2019. Google searches and stock market activity: Evidence from norway. Finance Research Letters 28, 208–220.

Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., Kim, C. H., 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS one 11 (8), e0161197.

Kjærland, F., Khazal, A., Krogstad, E., Nordstrøm, F., Oust, A., 2018. An analysis of bitcoin's price dynamics. Journal of Risk and Financial Management 11 (4), 63.

Koutmos, D., 2018. Bitcoin returns and transaction activity. Economics Letters 167, 81–85.

Kristoufek, L., 2013. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. Scientific reports 3, 3415.

Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5 (1), 1–167.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66 (1), 35–65.

Loughran, T., McDonald, B., 2014. Measuring readability in financial disclosures. The Journal of Finance 69 (4), 1643–1671.

Mai, F., Shan, Z., Bai, Q., Wang, X., Chiang, R. H., 2018. How does social media impact bitcoin value? a test of the silent majority hypothesis. Journal of Management Information Systems 35 (1), 19–52.

Matta, M., Lunesu, I., Marchesi, M., 2015a. Bitcoin spread prediction using social and web search media., 1–10.

Matta, M., Lunesu, I., Marchesi, M., 2015b. The predictor impact of web search media on bitcoin trading volumes 1, 620–626.

Mc Laughlin, G. H., 1969. Smog grading-a new readability formula. Journal of reading 12 (8), 639–646.

McAteer, C., 2014. Twitter sentiment analysis to predict bitcoin exchange rate.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal 5 (4), 1093–1113.

Meland, M., Øyen, V., 2017. Explaining bitcoin's price fluctuations.

Miller, B. P., 2010. The effects of reporting complexity on small and large investor trading. The Accounting Review 85 (6), 2107–2143.

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., Delen, D., 2012. Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Molnár, P., 2012. Properties of range-based volatility estimators. International Review of Financial Analysis 23, 20–29.

Nakamoto, S., et al., 2008. Bitcoin: A peer-to-peer electronic cash system.

Nasir, M. A., Huynh, T. L. D., Nguyen, S. P., Duong, D., 2019. Forecasting cryptocurrency returns and volume using search engines. Financial Innovation 5 (1), 2.

Polasik, M., Piotrowska, A. I., Wisniewski, T. P., Kotkowski, R., Lightfoot, G., 2015. Price fluctuations and the use of bitcoin: An empirical inquiry. International Journal of Electronic Commerce 20 (1), 9–49.

Sovbetov, Y., 2018. Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. Journal of Economics and Financial Analysis 2 (2), 1–27.

Steadman, I., 2013. Wary of bitcoin? a guide to some other cryptocurrencies.

Steinert, L., Herff, C., 2018. Predicting altcoin returns using social media. PloS one 13 (12), e0208119.

Stenqvist, E., Lönnö, J., 2017. Predicting bitcoin price fluctuation with twitter sentiment analysis.

Tetlock, P. C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of finance 62 (3), 1139–1168.

Uysal, A. K., Gunal, S., 2014. The impact of preprocessing on text classification. Information Processing & Management 50 (1), 104–112.

Warriner, A. B., Kuperman, V., Brysbaert, M., 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior research methods 45 (4), 1191–1207.

Wood, G., et al., 2014. Ethereum: A secure decentralised generalised transaction ledger.