Reddit as a prediction tool for crypto-assets

Luis Antonio Loredo Camou[†]

Abstract Cryptocurrencies, such as Bitcoin and Ethereum, have recently become a topic of conversation for the general public. This paper will explore the information available in Reddit regarding crypto assets. Unlike other social platforms, Reddit allows analysis of public opinion and conveniently organized information by topic. We study the benefit of sentiment variables derived from Reddit's crypto forums to forecast volatilities and returns. While volatility forecasts seem to consistently benefit from Reddit sentiment variables, results are not statistically different from a benchmark. In contrast, returns show mixed forecasting results but show statistical differences from their proposed benchmark. We find evidence that Reddit variables gain importance in market-wide and asset-specific events.

Keywords: Cryptocurrencies; Internet message boards; Machine learning; Forecasting returns; Volatility forecasting; Sentiment analysis.
JEL Code: G13, G14, G17.

1. Introduction

Cryptocurrencies have, once again, gained mainstream attention. According to Google Trends (2021), on April 18, 2021, the term *Bitcoin* ranked as the ninth trendiest search in the US, and *Dogecoin* ranked first on April 15, 2021, with over five million searches. Institutional interest has also risen recently. Companies such as Tesla, Square, and Microstrategy have acquired Bitcoin for their balance sheet, and everyday applications, like Paypal, now allow their customers to buy and sell cryptocurrencies (Markets Insider, 2021). This has been accompanied by increased social media activity, with mentions of Bitcoin on Twitter reaching an all-time high on February 9, 2021 (Cointelegraph, 2021).

What are the incentives to post and read financial message boards? The existing literature is reviewed by Antweiler and Frank (2004). In particular, they focus on the theory of DeMarzo et al. (2003), who introduce the concept of *persuasion bias*, under which individuals fail to account for possible repetition in the information they receive. This anomaly may happen if two individuals read the same piece of information and then discuss it between

Submitted on June 13, 2021. Revised on February 13, 2022. Accepted on February 20, 2022. Published online in March 2022. Editor in charge: Marcelo Fernandes.

[†]Bendheim Center for Finance, Princeton University, USA: lcamou@alumni.princeton.edu

themselves without revealing their source. While they may both believe they have heard the same information for a second time, they fail to account for repetition. Given this ability to influence people, it may be profitable to be well-connected in a community to increase the repeated information others receive. An increased sense of confidence in decisions may push people to read message boards.

In this paper, we focus on Reddit. Reddit is well-known as a collection of forums based on interests. A new community, called a *subreddit*, can be created about any topic if it complies with general rules. In particular, most crypto-projects have a dedicated subreddit where participants are free to join and share news, express opinions, and discuss ideas. This information sharing is done through *posts*. Each post is composed of a title set by the author and comments made by those who wish to discuss the submission. According to Alexa Internet (2021), as of April 13, 2021, Reddit is the 19th most popular website worldwide, and seventh in the United States.

Several previous authors deal with the impact of sentiment indices on returns of cryptocurrencies. Among them, Kristoufek (2013) studies the effect of search trends for Bitcoin in Wikipedia and Google. Naeem et al. (2021) and Anamika et al. (2021) use the Twitter Happiness Sentiment index and survey-based sentiment measures, respectively. Sentiment analysis using publicly-available textual information has also been used. Using lexical dictionaries, Karalevicius et al. (2018) identify Twitter posts as a predictor for the price of Bitcoin. Kraaijeveld and De Smedt (2020) reach similar conclusions for other cryptocurrencies besides Bitcoin, commonly referred to as *altcoins*. In contrast, Ahn and Kim (2020), using posts from a Bitcoin forum, conclude that, unlike future returns, volume and volatility are related to emotional factors.

Reddit has also received research attention. In particular, Prajapati (2020) uses different lexical dictionaries to perform sentiment analysis on Reddit and Google News to predict the price of Bitcoin using data from January 1, 2018, to November 20, 2019. He concludes that social sentiment captured through Reddit and Google News improves forecasts, compared against past prices-only models.

Analyzing twenty-four Reddit communities, Wooley et al. (2019) use information from July 1, 2016, to July 24, 2018 to predict three months of price directions (up vs down). They conclude that predictions benefit from using sentiment variables against a lagged prices-only model.

To the best of our knowledge, no previous researcher has studied the forecasting power of general-use forums, such as Reddit, on volatilities and returns and included Granger, Mariano-Diebold, and robustness tests. Further,

no one has performed detailed feature-importance analysis, as we present here. Our research focuses not only on Bitcoin, as most research does, but also on altcoins.

We find that although sentiment variables derived from Reddit seem to help reduce the mean-squared-error of our volatility predictions, these results are not statistically different from an HAR-RV model. In contrast, although mean-squared-error results for returns are mixed, they are consistently different from an in-sample-mean benchmark. Our variables gain relative importance around market-wide and asset-specific events such as market booms and class actions. We use natural language processing and machine learning tools to create sentiment variables and evaluate our results, assessing the impact of including our constructed variables through linear and nonlinear models. Our work follows along the lines of Antweiler and Frank (2004), who conclude that stock messages can help predict volatility and returns for companies in the Dow Jones Industrial Average. Our work also relates to that of Engle et al. (2011), showing that public information arrival relates to increased volatility.

This paper supports previous research showing the positive impact of sentiment indices on returns and volatilities of cryptocurrencies. In particular, our conclusions are similar to those of Ahn and Kim (2020) who show that while sentiment does help reduce forecasting error in volatility, the effect on returns is not clear. While our work coincides with that of Prajapati (2020), that sentiment seems to improve Bitcoin price prediction, we achieve mixed results when expanding our focus to other assets. However, we find evidence that information extracted from Reddit seems to consistently reduce volatility forecasting error for all assets studied, although further tests show these forecasts are not statistically different from a benchmark.

This work is of interest to investors, risk managers, regulators, and academics. From an investing and risk management perspective, the recollection of new features presents unique opportunities to understand and profit from market behavior. For example, accurate predictions can be used in portfolio rebalancing, options trading, and value-at-risk estimation. From a regulatory perspective, the impact of well-connected individuals may imply possible market manipulation. This implication is of particular interest in anonymous forums such as Reddit. Finally, from an academic perspective, mainstream media has proven beneficial to explain risk premia and above-average stock returns, as seen by Manela and Moreira (2017).

The paper is organized as follows. Section 2 includes a description of data and sources. Section 3 motivates our model selection and methodology. Section 4 performs an exploratory analysis of the subreddits for each asset, using

natural language processing tools. Section 5 presents our prediction results and the relative feature importance analysis. Section 6 offers our conclusions.

2. Data

A wide range of cryptocurrencies and tokens are available. As of May 2, 2021, CoinMarketCap (2021a) reports the existence of 9,527 cryptos. We have decided to focus on five of these, described in the following paragraphs. We select the first four cryptocurrencies which were the top four projects by market capitalization at the start of 2021, representing almost 85% of the total market value (CoinMarketCap, 2021b). In particular, Bitcoin and Ethereum account for 70.68% and 10.79%, respectively. Finally, we select Dogecoin due to its Internet popularity. We present these assets in order of their market capitalization.

Bitcoin (BTC): A cryptocurrency invented in 2008 by an unknown persona denominated Satoshi Nakamoto, Bitcoin uses peer-to-peer technology to operate in a decentralized manner. Transactions are verified through cryptography and recorded in a public distributed ledger called a blockchain.

Ethereum (ETH): Ethereum is a blockchain with *smart-contract* functionality. A smart contract is a computer program intended to execute automatically. Given this, decentralized finance, a movement to offer traditional financial instruments in a decentralized architecture, has made Ethereum the most actively-used blockchain (Bloomberg, 2021). The currency used in Ethereum is Ether.

Litecoin (LTC): Litecoin is a cryptocurrency based on Bitcoin. Litecoin differs by using a different cryptographic algorithm, more resistant to custom hardware.

Ripple (XRP): Ripple is a payment solutions company. Ripple makes use of its native cryptocurrency, known as XRP, to allow for prompt payments. This asset is of interest to us given a class action against Ripple in May 2018, given the unregistered sale of its XRP tokens, and in December 2020, two of its founders were sued by the SEC for selling these tokens.

Dogecoin (DOGE): Introduced on December 6, 2013, Dogecoin is a cryptocurrency based around the figure of the *Doge* meme, a Shiba Inu dog. Compared to other cryptocurrencies, it is focused on a fun and

welcoming community. In January 2021, in a movement influenced by the GameStop short squeeze, Dogecoin's price increased by 800% (CNBC, 2021). Further, in April 2021, a movement to raise its price pushed its value from $0.06 on April 7 to $0.40 on April 19. The currency has been mentioned by Elon Musk, Mark Cuban, Snoop Dogg, and Gene Simmons, among others.

Our data can be split into two major categories: financial and Reddit variables. In the following paragraphs, we describe each of these.

2.1 Financial variables

To obtain our prices, we use the Binance API. As of May 2, 2021, Binance is the largest crypto exchange by trading volume (CoinMarketCap, 2021c). Through the Binance API, we download all available prices for the given crypto-assets at a 5-minute frequency. The initial observation for each asset varies according to the first listing date in the exchange, as presented in Table 1.

Table 1
Initial price of each asset as extracted from Binance

name	start date
BTC	August 17, 2017
ETH	August 17, 2017
LTC	December 13, 2017
XRP	May 4, 2018
DOGE	July 5, 2019

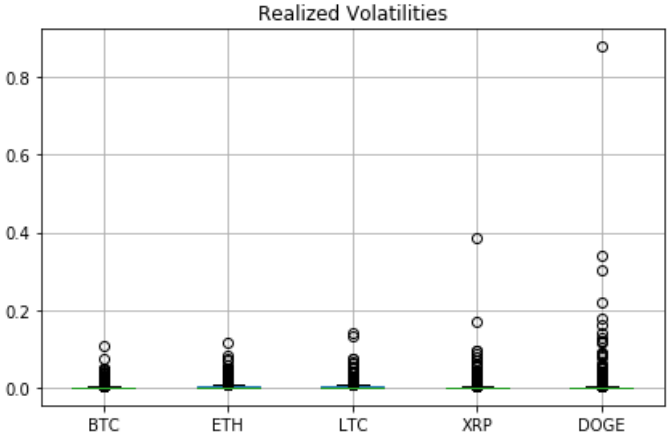
While resource such as Bitcoin Average (2022) aggregate several sources of information, they may suffer from volume inflation as exchanges try to gain visibility, as described by CoinMarketCap (2022). This is why we use only one reliable data source.

We construct our assets' daily volatility measures following McAleer and Medeiros (2008). To make a trade-off between accuracy and microstructure noise, we follow the suggestion to use the 5-minute close-to-close returns as our input. The procedure is as follows:

$$RV_t = \sum_{i=1}^{288} r_{i,t}^2$$

where $r_{i,t}^2$ is the close-to-close intraday 5-minute return during day t.

Figure 1
Boxplot for realized volatilities



Note: Realized volatilities constructed following the method of McAleer and Medeiros (2008) using price history from the initial price observation specified in Table 1 to April 30, 2021.

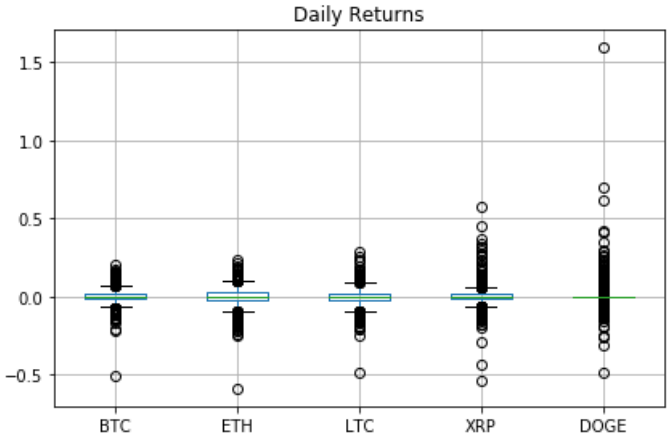
Figure 1 provides a boxplot figure of the realized volatilities. We offer additional descriptive images for Bitcoin and Dogecoin in Figures A1 and A2. The plots for other tickers are available upon request. These results are in line with the usual stylized facts of financial returns data. Similar to Andersen et al. (2001), we observe that our realized volatilities are highly right-skewed, approximately Gaussian-distributed when presented in logarithms, and finally, strongly temporally-dependent.

Similarly, Figure 2 summarizes daily log-returns, estimated using 5-minute close-to-close prices, in a boxplot. Figures A3 and A4 present prices and returns for Bitcoin and Dogecoin. From here, we construct the cumulative 7-day returns.

2.2 Reddit Variables

As we mentioned previously, our second source is Reddit. We present the start date of each subreddit and its number of submissions in Table 2. The titles of each submission are recollected through the Pushshift Reddit dataset, as presented by Baumgartner et al. (2020), a big-data and analytics project providing a copy of Reddit's comments and submissions. We then apply Valence Aware Dictionary for Sentiment Reasoning (VADER) analysis, as introduced by Hutto and Gilbert (2014). VADER is a lexicon and rule-

Figure 2
Boxplot for returns



Note: Log-returns using price history from the initial price observation specified in Table 1 to April 30, 2021.

Table 2
Creation date and number of submissions for each subreddit

name	start date	number of submissions
BTC	September 9, 2010	942,668
ETH	December 14, 2013	165,897
LTC	October 14, 2011	78,578
XRP	October 14, 2009	73,097
DOGE	December 8, 2013	569,865

Note: Information up to April 30, 2021.

based sentiment analysis model that is sensitive to polarity and intensity. In other words, VADER can detect if a text is positive or negative and assign an intensity score to it. VADER is specially-tuned to work for social media by understanding acronyms, emoticons, slang, and punctuation, compared to other sentiment analysis models. This is the reason we use VADER over other alternatives.

The VADER procedure uses a pre-trained dictionary to assign to each lexical feature in a corpus a valence score on a scale from −4 to +4, where −4 represents the most negative sentiment and +4 the most positive. This dictionary includes words, emoticons, acronyms, and slang. The results for the cor-

Table 3
Examples of punctuation, capitalization, degree modifiers, and shift in polarity under VADER

sentence	score
Bitcoin is a cryptocurrency.	+0.00
Bitcoin is a secure cryptocurrency.	+0.34
Bitcoin is a secure cryptocurrency!	+0.40
Bitcoin is a SECURE cryptocurrency.	+0.48
Bitcoin is a very secure cryptocurrency.	+0.40
Bitcoin is a secure cryptocurrency but a risky investment.	−0.13

Note: Comments are used to illustrate the methodology used and do not necessarily reflect the author's opinions.

pus are then summed up, and a set of heuristics are applied. These heuristics help determine the sentiment when a text includes punctuation, capitalization, degree modifiers, and shifts in polarity. Finally, the VADER sentiment score is normalized to be between −1 and +1 through

$$x = \frac{x}{\sqrt{x^2 + 15}}.$$

Table 3 provides a set of examples.

Once all titles have a score, we first count the number of submissions per day. Then, we delete all observations with a VADER score equal to 0, and compute the average score for each day. We delete the zero observations to induce variation. In case of a missing observation, we fill with 0, representing a lack of submissions and sentiment.

Table A1 presents a summary of all of our variables. Figure A5 shows a correlation heatmap.

3. Models and forecasts

Rolling windows are commonly used to keep estimation windows at a fixed length. Different estimation rolling-window sizes have been used to forecast one-day-ahead prices and volatilities of different assets. For example, Qiu et al. (2019) use a 1,000-observation window size for the Nasdaq index and its constituents, and Liu et al. (2018) use 1,300 and 1,500 window sizes for WTI crude oil futures. However, as Pesaran and Timmermann (2005) note, although there is no pre-defined window length to use, a shorter window should be considered if structural breaks may be present.

We fit each model on a 1-year rolling window. Given crypto assets' rapidly-changing market cycles, a year allows us to capture a breadth of market environments, including high-volatility and low-volatility periods. We avoid including an excessive amount of data, so as not to indicate the current regime, and address potential structural breaks in the sample. A 365-day window has been used previously by Kayal and Balasubramanian (2021) and Kristoufek (2018) to study Bitcoin's market efficiency.

For our penalized models, we use lasso and ridge, due to their properties and wide application in the literature. Analysis of l_1 regularization allows for sparsity. The l_2 regularization is often regarded as an excellent selection for forecasting, although it does not have feature selection properties. Finally, we focus on a nonlinear model, called random forest. We outline the details of each model, as well as the benchmarks used, below.

3.1 Benchmarks

3.1.1 Volatility: Heterogeneous autoregressive model

Following Corsi (2009), we use the heterogeneous autoregressive (HAR) model. In his paper, Corsi shows that the HAR model *"successfully achieves the purpose of reproducing the main empirical features of financial returns (long memory, fat tails, and self-similarity) in a very tractable and parsimonious way. Moreover, empirical results show remarkably good forecasting performance."*

In particular, we use the HAR(1,5,22) model to predict realized volatilities, defined as

$$RV_{t+1} = \beta_0 + \beta_1 RV_t + \beta_2 RV_{w,t} + \beta_3 RV_{m,t} + U_{t+1}$$

where

$$RV_{w,t} = \frac{1}{5}\sum_{i=0}^{4} RV_{i-1}$$

$$RV_{m,t} = \frac{1}{22}\sum_{i=0}^{21} RV_{i-1}.$$

This is the specification used to study Bitcoin volatility by Yu (2019) and Bouri et al. (2021).

Further, this model may be defined in logs:

$$\log(RV)_{t+1} = \beta_0 + \beta_1 \log(RV)_t + \beta_2 \log(RV)_{w,t} + \beta_3 \log(RV)_{m,t} + U_{t+1}$$

where

$$\log(RV)_{w,t} = \log\left(\frac{1}{5}\sum_{i=0}^{4} RV_{i-1}\right)$$

$$\log(RV)_{m,t} = \log\left(\frac{1}{22}\sum_{i=0}^{21} RV_{i-1}\right).$$

Whenever possible, we use the log-model with the approximate normality of $\log(RV)$, given its more stable behavior. Further, we observe a stronger persistence in logs than in levels. This follows Andersen et al. (2001). We apply the exponential function to our results to recover our forecast in levels. This transformation avoids negative values, since the exponential is always positive. We default to the levels model when we have a zero observation.

3.1.2 Returns: Random walk and in-sample mean

Assessing cryptocurrency market efficiency, Urquhart (2016) shows that, although in the time frame from 2010 to 2016, Bitcoin prices seem inefficient in one-day-ahead returns, when the period analyzed is split from 2013 to 2016, evidence of efficiency appears. Apopo and Phiri (2021) expand this analysis to altcoins and weekly returns. The authors conclude that while one-day-ahead returns are efficient, weekly returns show the contrary.

For our returns, we use two common benchmarks: the random walk and the in-sample mean. Under the random walk model, tomorrow's price is assumed to be equal to today's price:

$$\mathbb{E}_t\left[P_{t+1}\right] = P_t$$

and therefore

$$\mathbb{E}_t\left[r_{t+1}^{(7)}\right] = 0$$

where $r_{t+1}^{(7)}$ represents the next 7-day cumulative return.

For the in-sample mean, we take the average of the returns of our training sample, and use this as a prediction:

$$\mathbb{E}_t\left[r_{t+1}^{(7)}\right] = \frac{1}{365}\sum_{i=0}^{364} r_{t-i}^{(7)}.$$

This benchmark is used as a naive forecast for next-period prediction by Fleming et al. (2003) and Henriques and Sadorsky (2018).

3.2 Penalized regression models

For our penalized models, lasso and ridge, we select our penalization parameter using the Bayesian information criterion (BIC). There are several reasons to use an information criterion over a cross-validation (CV) framework. First, the use of information criterion provides a single penalization

value, saving computational time. Second, information criterion does not have to split data between training and test sets, which allows us to use all information available to us.

We standardize our data before fitting the models. Our forecasting models can be represented as

$$\begin{aligned}\log(RV_{t+1}) =& \beta_0 + \boldsymbol{\beta}^{'} \log(\boldsymbol{X}_t) + \beta_d \log(RV_t) \\ &+ \beta_w \log(RV_{w,t}) + \beta_m \log(RV_{m,t}) + U_{t+1}\end{aligned}$$

where RV_t is the realized volatility and X_t consists of the Reddit variables observed at time t. Similarly, for returns we use

$$r^{(7)}_{t+1} = \beta_0 + \boldsymbol{\beta}^{'} \boldsymbol{X}_t + \beta_d r_t + \sum_{i=0}^{6} \beta_w r_{t-i} + \beta_m \sum_{i=0}^{29} r_{t-i} + U_{t+1}$$

where $r^{(7)}_{t+1}$ is the next 7-day cumulative return and r_t the one-day observation.

3.3 Nonlinear models

While linear models are appealingly simple, a nonlinear treatment is usually needed to capture the underlying behavior. To explore the benefits of a nonlinear model, we will use a random forest.

Random forest is an ensemble method composed of *decision trees*. A decision tree model recursively partitions the input space into sequential choices. In this way, the feature space is divided into smaller regions. Once divided, the simple average of each region is computed and used as output. Given that the effect of an error in a split is propagated down to all of the splits below, decision trees are highly unstable. In other words, a slight change in the input may have drastic changes in the output (Andrews, 1986). Bootstrap aggregation, commonly called bagging, helps achieve more-stable estimators by reducing variance and leaving bias unchanged (Breiman, 1996).

To apply bagging to a regression tree model, we first sample with replacement from the data. Then, for the sample, we estimate a tree. We randomly select a subset of the original features to use as split variables for this tree. We let the tree grow until the desired tree depth is reached. We repeat this for as many bootstrap samples as desired. After we construct the trees, we compute our final prediction using the average of these bootstrapped trees. In this way, we create a more stable estimator called a random forest.

We now describe hyperparameter selection. While ways to adapt the information criterion approach to nonlinear models have been proposed, the predominant method uses cross-validation. However, the standard k-fold CV

would break time dependence, and it is important to remember that we are working with time-series data. So instead, we implement *hv*- CV, as proposed by Racine (2000). Compared to the standard k-fold CV, *hv*- CV introduces gaps between the training and test sets to tackle this dependence. Here, the test sets are chosen in a contiguous manner. In our code, we use five splits and discard one block before and after the test set.

We focus on two hyperparameters: 1) the number of trees (or the number of bootstrap samples) and 2) the tree depth (or the number of splits). As a starting point for our grid values, we refer to Friedman et al. (2001). These authors mention that the inventors of random forest recommend a default value of the number of variables used in each split equal to $\lfloor p/3 \rfloor$, where p is the number of features available, and a minimum node size of five when using random forest for regression.

Given this guidance, we select the following parameters:

Number of bootstrap samples: Usually, a higher number of trees is better to learn from the data, but this comes with a computational cost. We implement a value of 100.

Tree depth: Following Friedman et al. (2001), we implement using 3, 5, 8, and 12.

We limit the number of features to consider when looking for the best split to $\lfloor p/3 \rfloor$, where p is the number of features available. We leave all other parameters at their default in *sklearn's RandomForestRegressor*, given no *a priori* knowledge.

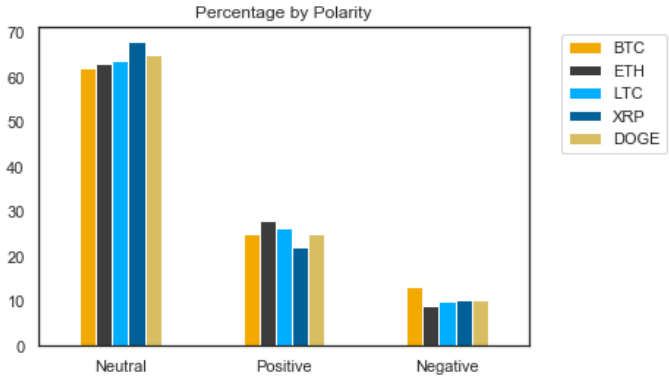
Once again, we standardize our data before fitting the models.

4. Natural language processing

Doing some basic natural language processing analysis to the titles recovered from Reddit from August 17, 2017, to April 30, 2021, we hope to uncover similarities and differences among communities. We discuss polarity, word frequencies, and the most-used words, following Martin and Koufos (2018). We end by introducing our average daily sentiment variable and the number of submissions per day, to be used in our prediction exercise.

We start by looking at polarity. We categorize as positive all submissions with a VADER score higher than 0.25, and as negative submissions with scores lower than −0.25. Everything else is labeled as neutral. Figure 3 displays the results. All subreddits follow a similar pattern. The percentage of titles with positive sentiment is around 25% for all communities. ETH

Figure 3
Percentage of titles by sentiment



Notes: Polarity is measured via VADER and then grouped into three buckets: negative (polarity below –0.25), neutral (–0.25 to 0.25), and positive (above 0.25). The period studied is August 17, 2017, to April 30, 2021.

stands out as having the most positive community, and XRP has the lowest percentage of positive titles. In contrast, Bitcoin seems to be notably different than all other communities regarding negativity. Results for DOGE are fascinating, since the project is focused on having a *"fun and friendly internet currency"* (Dogecoin, 2021). However, results show average behavior across all labels for Dogecoin.

We now look to understand the topics discussed in each subreddit. To do this, we focus on the top ten words mentioned in each community. To study words particular to these communities, we remove stop words, such as *"and," "is,"* and *"the"* through the *NLTK* package in Python. Additionally, we remove single-digit numbers.

Table 4 provides exciting insights. Our first observation is that the first term in all columns is a self-reference to the asset. Focusing on words such as *crypto*, *wallet*, or *blockchain*, we notice that all subreddits except the one for Dogecoin reference the technology. In particular, Ethereum stands out as having most of its words be tech-related. In contrast, Ripple has only one word, and Dogecoin none. Similarly, all columns, except for Ethereum's, have a reference to price with words such as *price*, *buy*, or *coinbase* (a popular exchange). Dogecoin presents more price-related words than other assets do. A final observation is that Dogecoin includes words referencing internet culture, such as *moon* and *hodl*. In crypto culture, prices are expected to make significant gains to reach the moon. Hodl is a common joke stemming from a

Table 4
Top 10 most-mentioned words by community

	Bitcoin	Ethereum	Litecoin	Ripple	Dogecoin
1	bitcoin	ethereum	litecoin	xrp	doge
2	btc	eth	ltc	ripple	dogecoin
3	crypto	blockchain	crypto	crypto	buy
4	buy	crypto	bitcoin	price	moon
5	new	new	buy	buy	hold
6	price	bitcoin	coinbase	fb	get
7	wallet	token	wallet	bitsecret	hodl
8	cryptocurrency	wallet	new	daily	go
9	blockchain	amp	price	discussion	let
10	coinbase	cryptocurrency	cryptocurrency	thread	sell

Notes: Stop words and numbers are excluded. The period studied is August 17, 2017, to April 30, 2021.

misspelling of the word *hold*, not selling. This table is evidence that communities carry different information, and therefore their predictive abilities may be different.

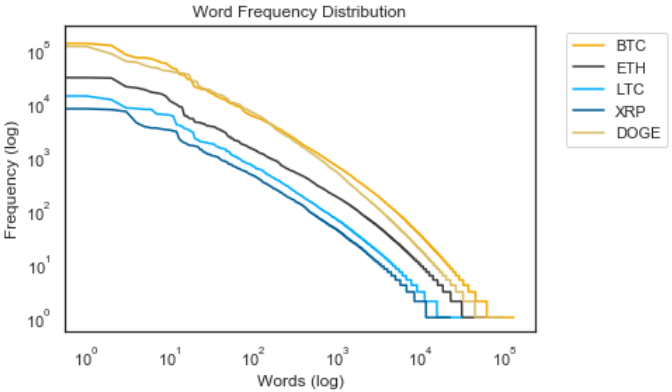
While the previous table focuses on only the top 10 words, it may be interesting to study all words used in each community. According to Zipf's law, the frequency of occurrence of a class is roughly inversely proportional to the rank of the class in the frequency list. Therefore, we would expect to see a roughly straight line by plotting the word frequency against its rank on a log-log scale. Most finite-size corpora follow a quasi-Zipfian law. There are several explanations for Zipf's law. One of them is the path of least effort: individuals use the most commonly-used words since they want to be understood (Zipf, 2016). Figure 4 presents word frequency distributions.

Our statistics behave roughly as expected. It appears that the subreddits for Bitcoin, Litecoin, and Ripple keep a constant spread among them, implying a similar complexity in their way of speech. Ethereum's subreddit seems to have more complex speech, as seen from more unique words being used. The opposite is true for Dogecoin. We presume this may be related to the insight uncovered in Table 4 and their commonly-discussed topics.

Finally, we present the two final plots: the daily average sentiment variable and the number of submissions in each forum. Figures 5 and 6 show the results for Bitcoin. All other plots are available in Figures A6 and A7.

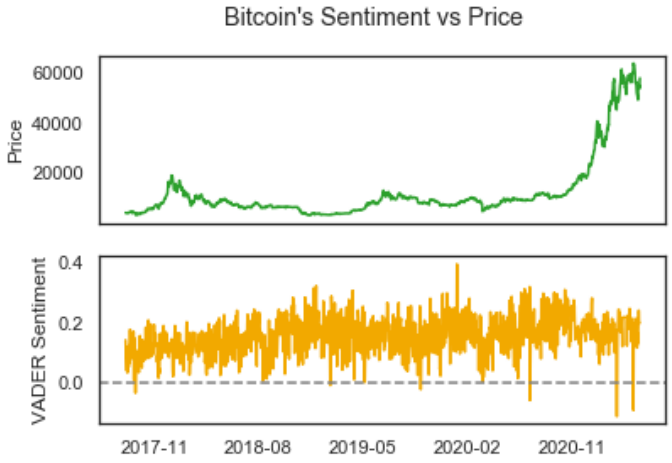
Figures 5 and A6 show that our variables move around a mean that slowly changes over time. For example, we notice a slowly increasing trend in the Bitcoin image. An even more exciting realization is that variability seems to cluster over time. This phenomenon is particularly noticeable in Litecoin,

Figure 4
Zipf's law



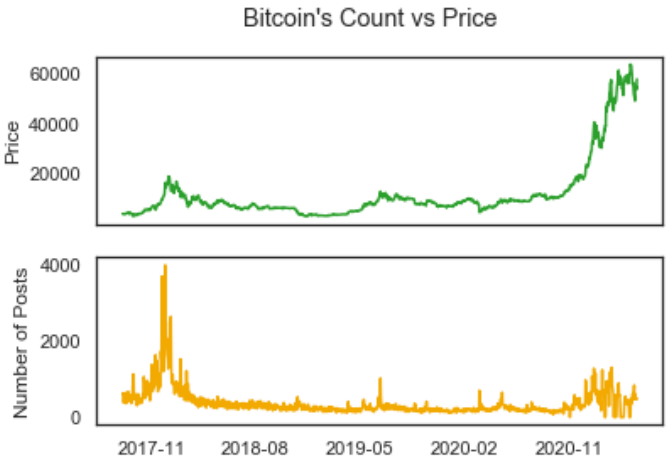
Notes: The frequency of every word used in the subreddit is plotted against its inverse rank in the frequency table. The period studied is August 17, 2017, to April 30, 2021.

Figure 5
Bitcoin's average daily sentiment under VADER



Notes: The top image presents the price in USD. The bottom image is the average daily VADER sentiment after removing all titles with a score equal to zero. Information used is from August 17, 2017, to April 30, 2021.

Figure 6
r/Bitcoin's count of daily submissions



Notes: The top image presents the price. The bottom image is the count of daily submissions in the community. Information used is from August 17, 2017, to April 30, 2021.

Ripple, and Dogecoin's subreddits. Focusing on Dogecoin, we see a sharp drop in variability, which coincides with its rise in internet popularity and the *WallStreetBets* movement against institutional investors. The WallStreetBets subreddit pumped the prices of several stocks heavily shorted by financial institutions to make them incur losses. The most significant asset affected was GameStop. Other assets affected included Dogecoin.

Moving on to Figures 6 and A7, we notice clear shocks around price moving periods. For example, Bitcoin's highest number of posts coincides with the 2017 bubble. Similarly, the number of posts increased during the 2021 rally for all assets. We now focus on two assets in particular: Ripple and Dogecoin. Ripple reached its peak on February 1, 2020, with over 2,500 submissions in a day. This event corresponds to a pump-and-dump scheme that pushed the price significantly (CoinDesk, 2021). Dogecoin has had extremely high numbers since the end of January 2021. Once again, this is related to the WallStreetBets movement, which made Dogecoin reach a peak more than ten times the one observed for Bitcoin.

Before moving on, we present the results of several *Granger causality* tests to motivate the following sections. The Granger causality test is a procedure to check if one time series is helpful to forecast another one. It is said that *x Granger causes y* if the lagged values of *x* are significant in predicting

Table 5
Granger causality results

data	model	BTC	ETH	LTC	XRP	DOGE
linear	**VADER → Volatility**	0.00	0.05	0.15	0.10	0.41
	Count → Volatility	0.00	0.00	0.00	0.00	0.00
	VADER → Return	0.02	0.51	0.81	0.52	0.37
	Count → Return	0.05	0.27	0.12	0.32	0.00
nonlinear	**VADER → Volatility**	0.00	0.00	0.00	0.00	0.00
	Count → Volatility	0.00	0.00	0.00	0.00	0.00
	VADER → Return	0.00	0.00	0.02	0.00	0.00
	Count → Return	0.00	0.06	0.00	0.03	0.07

Notes: Reported p-value from a Granger causality test with seven lags, using each asset's start date as described in Table 1 to April 30, 2021. The linear model used is as defined by Granger (1969). For the nonlinear procedure, we use an adapted Granger test using feed-forward neural networks, as described by Rosol et al. (2022). Our neural network is composed of two dense layers of fifty neurons each.

y when already taking into account lagged values of y.

The test is usually seen in its linear form:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_s y_{t-s} + \beta_1 x_{t-1} + \beta_s x_{t-s} + \varepsilon_t$$

where ε_t is white noise and $H_0 : \beta_1 = \cdots = \beta_s = 0$.

If we reject the null hypothesis with a significant p-value, x Granger causes y. Nonlinear versions have also been adapted, for example, Rosol et al. (2022) present a nonlinear test using feed-forward neural networks. We follow the neural-network procedure using a rather simplistic architecture for motivational purposes.

Table 5 shows our results. A first observation is that the number of submissions seems to have a strong signal on volatility, even in the linear case. Another observation is that the assets capable of taking signals linearly from sentiment variables to predict volatility seem to be the most mature: Bitcoin and Ethereum. Something similar happens for returns and Bitcoin. A final observation is an apparent benefit of using nonlinear models. Although the effects of our sentiment variables on the financial variables are somewhat mixed in the linear case, only twice does the p-value exceed 5% in the nonlinear case.

Our results are in line with those of Shen et al. (2019), who conclude that the number of tweets is a significant driver of realized volatility for Bitcoin. However, contrary to Twitter, Reddit does seem to be a statistically significant tool to predict Bitcoin returns. Naeem et al. (2021) compare the

Table 6
Inverse Granger causality results

data	model	BTC	ETH	LTC	XRP	DOGE
linear	**Volatility → VADER**	0.00	0.05	0.03	0.06	0.02
	Volatility → Count	0.00	0.48	0.00	0.00	0.00
	Return → VADER	0.01	0.12	0.52	0.03	0.40
	Return → Count	0.00	0.00	0.01	0.00	0.00
nonlinear	**Volatility → VADER**	0.00	0.00	0.00	0.00	0.00
	Volatility → Count	0.00	0.00	0.00	0.00	0.00
	Return → VADER	0.08	0.00	0.00	0.00	0.00
	Return → Count	0.00	0.00	0.00	0.00	0.00

Notes: Reported *p*-value from a Granger causality test with seven lags, using each asset's start date as described in Table 1 to April 30, 2021. The linear model used is as defined by Granger (1969). For the nonlinear procedure, we use an adapted Granger test using feed-forward neural networks as described by Rosol et al. (2022). Our neural network is composed of two dense layers of fifty neurons each.

Twitter Happiness Index, a sentiment variable collected using lexical features, against returns of five different cryptocurrencies. Our results coincide, since we also observe a significant nonlinear Granger causality between return and our VADER variable. However, we also observe linear Granger causality, in the case of Bitcoin.

It is also interesting to study reverse causality. We know the number of posts generates volatility, but could the volatility generate more posts? Table 6 indeed shows a bilateral relationship between our variables, as well as some unilateral relationships.

We now introduce our results for volatility and returns prediction.

5. Are Reddit variables useful to forecast?

5.1 Volatility

Table 7 presents prediction results for the one-day-ahead volatility. We obtain these values by computing the ratio of each model and the benchmark's mean-squared-error. Our benchmark is an HAR(1,5,22) model, as described in Section 3.

Our first observation is that nonlinear models such as random forest seem to benefit our prediction, even without the addition of the Reddit variables. Moving to the Reddit variables, we notice negative results for an OLS model, mixed results for our penalized regression models, and positive results for random forest. Adding Reddit variables to a nonlinear model can have minor

Table 7
Volatility results

data	model	BTC	ETH	LTC	XRP	DOGE
financial	**HAR**	1.00	1.00	1.00	1.00	1.00
	random forest	0.47	0.70	0.87	0.92	0.57
financial + Reddit	**OLS**	1.15	1.03	1.05	3.12	2.91
	lasso	0.99	0.95	0.91	2.66	1.97
	ridge	0.96	0.93	0.97	2.74	2.29
	random forest	**0.46**	**0.63**	**0.69**	**0.81**	**0.54**

Notes: Ratio between model and benchmark's mean-squared-error for rolling window out-of-sample predictions. The study period is from each asset's start date as described in Table 1 to April 30, 2021. We use a one-year rolling window with the first 365 observations as an initial calibration period. Forecasts are done on a one-step-ahead basis. For each asset, the model with the best performance is highlighted in bold.

Table 8
Diebold-Mariano test for volatility

data	model	BTC	ETH	LTC	XRP	DOGE
financial	**HAR**	-	-	-	-	-
	random forest	0.39	0.56	0.75	0.57	0.49
financial + Reddit	**OLS**	0.06	0.16	0.26	0.23	0.20
	lasso	0.90	0.50	0.47	0.24	0.15
	ridge	0.76	0.39	0.44	0.25	0.20
	random forest	**0.41**	**0.49**	**0.47**	**0.13**	**0.46**

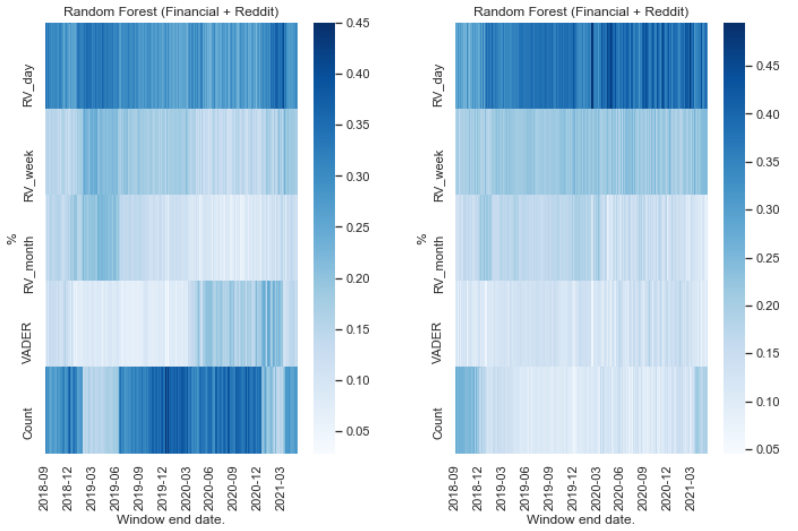
Notes: Diebold-Mariano test p-value results for each model against the financial HAR benchmark, across all available observation dates from Table 2 to April 30, 2021. For each asset, the model with the best performance in Table 7 is highlighted in bold.

to moderate benefits.

We now perform the *Diebold-Mariano test* as presented by Diebold and Mariano (2002) to check if the forecasts are statistically different from the benchmark. Under the Diebold-Mariano test, the alternative hypothesis is that the two methods being compared have different accuracy levels. We carry out these tests over all our forecast dates. Table 8 shows that only one model using Reddit variables is statistically significant at a 10% level. However, this is due to worse performance.

To better understand our results, we proceed with a feature-importance analysis. Since the base estimator of most of our models is a linear regression, and we have standardized our variables, we can thus compare the absolute value of each coefficient. For our random forest model, we use the *impurity-*

Figure 7
Random forest's volatility relative importance (1 of 2)



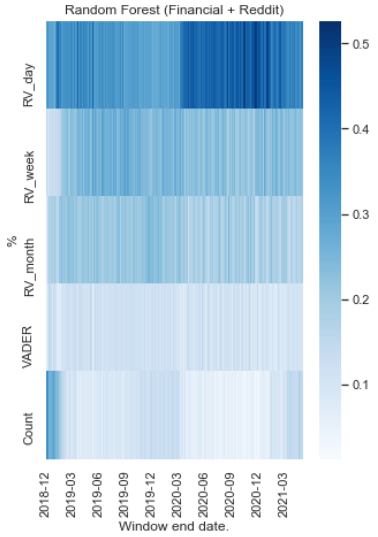
(a) Bitcoin (b) Ethereum

Notes: The *x*-axis is the end date of our one-year rolling window, and the *y*-axis is the absolute value of the regressor, divided by the sum of absolute values of all regressors for each date. The study period is from each asset's start date as described in Table 1 to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one-step-ahead basis.

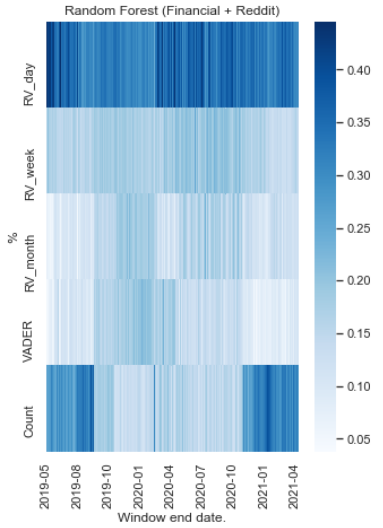
based feature importance. The impurity-based feature importance is the total reduction in the mean-squared-error brought by that feature. To standardize our results and make them comparable, we divide each feature-importance coefficient by the sum of all coefficients on that day. Figures 7 and 8 show the results for random forest with Reddit variables.

From these plots, we first notice that the one-day-lagged volatility seems to be the most consistently important feature across our analysis dates. We notice a significant decline in importance in the one-week and one-month average. Focusing on our Reddit variables, the number of posts seems to gain more importance than the average VADER score, and becomes particularly significant at certain moments. Our variable becomes dominant for those assets present during the 2017 crypto bubble (Bitcoin, Ethereum, and Litecoin). Ethereum relies on this variable less than Bitcoin does, although the study periods are the same. We presume this relates to the Ethereum subreddit not discussing price as often as other forums, as seen in our natural language processing analysis. Interestingly, Bitcoin presents relatively high

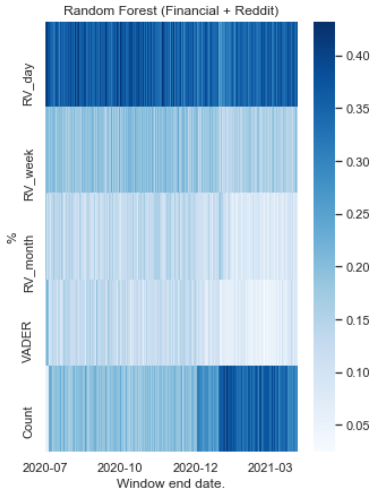
Figure 8
Random forest's volatility relative importance (2 of 2)



(c) Litecoin



(d) Ripple



(e) Dogecoin

Notes: The *x*-axis is the end date of our one-year rolling window, and the *y*-axis is the absolute value of the regressor, divided by the sum of absolute values of all regressors for each date. The study period is from each asset's start date as described in Table 1 to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one-step-ahead basis.

feature importance for the number of submissions during the relatively non-volatile period of 2018. This may be related to the uncertainty regarding the future of crypto-valuations at the time, making investors more reactive.

XRP's number of submissions shows a strong effect around May 2018, which coincides with a class action lawsuit filed against Ripple for selling unregistered tokens. Another effect during December 2020 occurred when the SEC sued Ripple and two of its executives for selling the unregistered securities. Finally, a short but significant effect relates to the XRP pump-and-dump of February 2020. Dogecoin's rise in internet popularity and euphoria occurred during the last observed dates, coinciding with a powerful effect of its number of submissions.

Although the choice of the one-day-ahead forecasting period follows previous literature, other specifications are possible. To test the benefits of different time frames, we perform a robustness test by trying different lags for our sentiment variables, from the one-day-lagged approach to a one-week lag. The results presented in Figure 9 show no clear winner concerning which lags to use. Further, a Diebold-Mariano test shows that out of those forecasts which show an improvement, only Ethereum at six lags and Dogecoin at 2, 3, and 7 lags are statistically different. This implies that forecasting windows should be chosen on a case-by-case basis. However, a one-step-ahead selection seems reasonable for a general analysis.

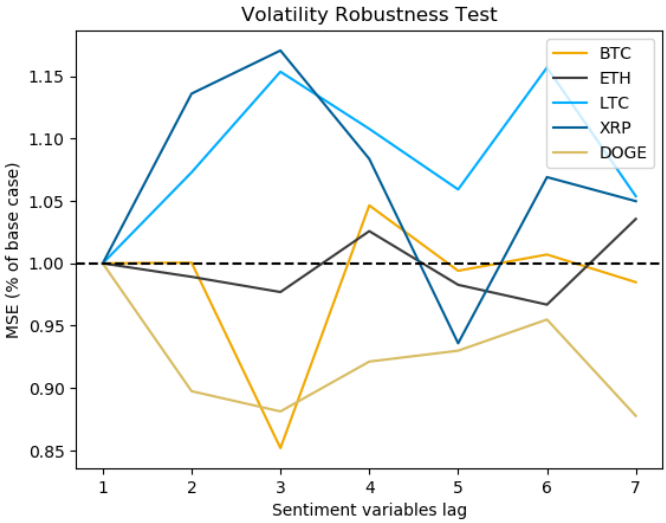
5.2 Returns

Table 9 shows prediction results for the 7-day returns. Once again, the values result from computing the ratio of each model and the benchmark's mean-squared-error. The benchmark used here is the random walk model described in Section 3. We also present results for the in-sample mean.

Our results are somewhat mixed for the models using Reddit variables. Models such as ordinary least squares and ridge fail to beat both benchmarks. Further, only two assets outperform the benchmarks and prices-only models: BTC and DOGE.

We use the Diebold-Mariano test to check if the forecasts are statistically different from the benchmark. Since random walk presents no variance, we must use the in-sample model as our benchmark. Table 10 shows the results of this analysis. These results are exciting; for three assets, a model using the Reddit variables differs from and outperforms the in-sample mean: BTC, XRP, and DOGE. Additionally, for Bitcoin and Dogecoin, we get models that outperform every other model while still statistically different from the benchmark.

Figure 9
Robustness test for volatility



Notes: The forecasting exercise using random forest with the architecture described in Section 3.3 is repeated, varying the number of periods by which the sentiment variables are lagged. The forecasting exercise with one lag represents the base model in the current section. Results are presented as a ratio of our base case. The study period runs from each asset's start date as described in Table 1 to April 30, 2021.

Table 9
Returns results

data	model	BTC	ETH	LTC	XRP	DOGE
benchmarks	**random walk**	1.00	1.00	**1.00**	1.00	1.00
	in-sample mean	0.98	**0.98**	1.03	0.99	0.94
financial	**OLS**	1.00	1.00	1.03	0.99	0.97
	random forest	1.02	1.03	1.07	**0.90**	0.80
financial + Reddit	**OLS**	1.01	1.00	1.03	1.02	1.24
	lasso	**0.97**	1.00	1.03	1.00	0.94
	ridge	1.01	1.00	1.05	1.02	1.21
	random forest	0.98	1.00	1.04	0.94	**0.74**

Notes: Ratio between model and benchmark's mean-squared-error for rolling-window out-of-sample predictions. The study period is from each asset's start date as described in Table 1 to April 30, 2021. We use a one-year rolling window with the first 365 observations as an initial calibration period. Forecasts are done on a one-step-ahead basis. For each asset, the model with the best performance is highlighted in bold.

Table 10
Diebold-Mariano test results for returns

data	model	BTC	ETH	LTC	XRP	DOGE
benchmarks	**random walk**	-	-	**-**	-	-
	in-sample mean	-	**-**	-	-	-
financial	**OLS**	0.00	0.00	0.00	0.00	0.00
	random forest	0.00	0.00	0.00	**0.00**	0.00
financial + Reddit	**OLS**	0.00	0.00	0.00	0.00	0.00
	lasso	**0.00**	0.00	0.00	0.00	0.04
	ridge	0.00	0.00	0.00	0.00	0.22
	random forest	0.00	0.00	0.00	0.00	**0.00**

Notes: Diebold-Mariano test p-value results for each model against the in-sample mean benchmark across all the available observation dates from Table 2 to April 30, 2021. For each asset, the model with the best performance in Table 9 is highlighted in bold.

Once again, the feature-importance analysis helps us better understand our results, as shown in in Figures 10 and 11.

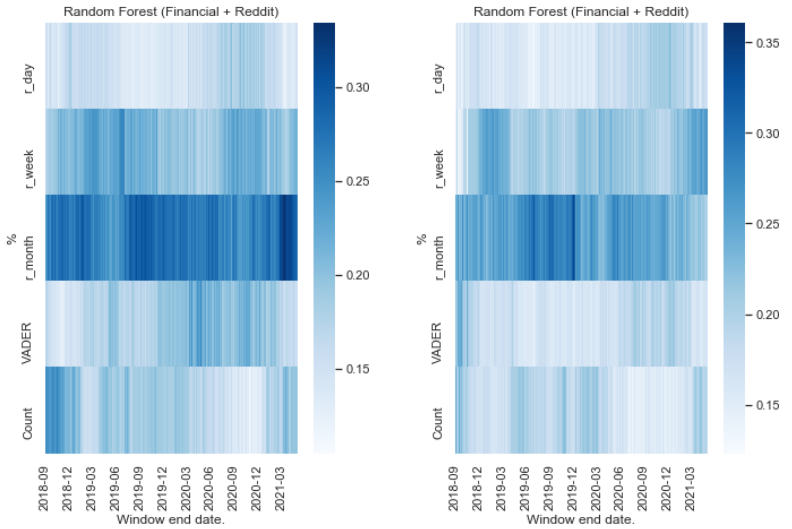
For those models able to beat the benchmark, unlike the volatility results, our most important financial feature is the monthly return. This implies that investors look at the returns from a more extended time frame when deciding to deploy capital. With Reddit variables, again, the variable for number of submissions can capture market-wide events such as the crypto bubble seen in Bitcoin's subfigure. Additionally, the two periods with legal actions for Ripple are also captured. Interestingly, the VADER variable is significant for Ripple. We presume it may be caused by increased uncertainty after the 2018 class action, as seen in Figure A6. Finally, the 2021 Dogecoin rally is captured in the number of submissions variable.

Similar to our analysis of volatility, we now perform a robustness test for returns. Figure 12 presents the results. There seems to be no clear winner, concerning the number of lags to use, although it may initially seem that a higher number of lags is preferred. However, a Diebold-Mariano test shows that only Ethereum at four lags is statistically different. Again, although forecasting windows should be chosen on a case-by-case basis, a one-step-ahead selection seems like a good choice for general analysis.

6. Conclusion

We study the applicability of variables extracted from Reddit to forecast next-day volatilities and 7-day returns. We first embark on an exploratory analysis using natural-language-processing tools to better understand our in-

Figure 10
Random forest's weekly returns relative importance (1 of 2)



(a) Bitcoin (b) Ethereum

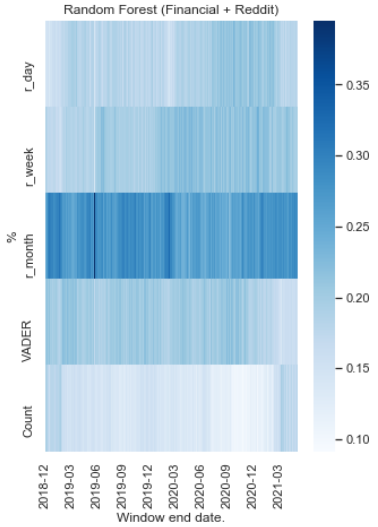
Notes: The *x*-axis is the end date of our one-year rolling window, and the *y*-axis is the absolute value of the regressor, divided by the sum of absolute values of all regressors for each date. The study period is from each asset's start date, as described in Table 1 to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one-step-ahead basis.

puts. Subreddits related to each crypto currency reveal dissimilar topics and markedly different levels of complexity. Further, sentiment seems to differ among communities.

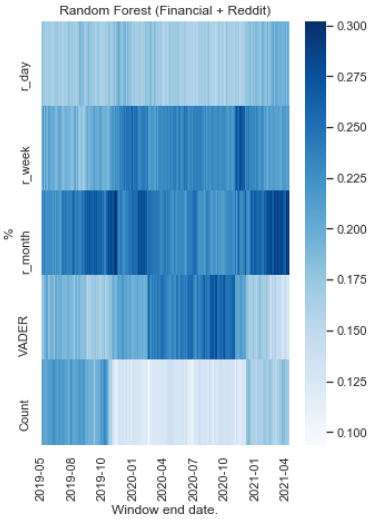
The results for our prediction exercise show that while sentiment variables seem to reduce the forecasting error for volatility, they generally fail to statistically differ from the proposed benchmark. In contrast, while our forecasts for returns have mixed results in reducing forecasting error, they are statistically different from the benchmark. These results are in line with those of Antweiler and Frank (2004), who conclude that stock messages help predict market volatility. For Bitcoin, we arrive at the same results as Prajapati (2020) and Karalevicius et al. (2018), that sentiment seems to improve price prediction. However, in expanding on their work, we agree with Ahn and Kim (2020) and Kraaijeveld and De Smedt (2020) that while sentiment seems to help reduce forecasting error in volatility, the effect is mixed for returns.

The Reddit variables seem to capture market-wide events such as the

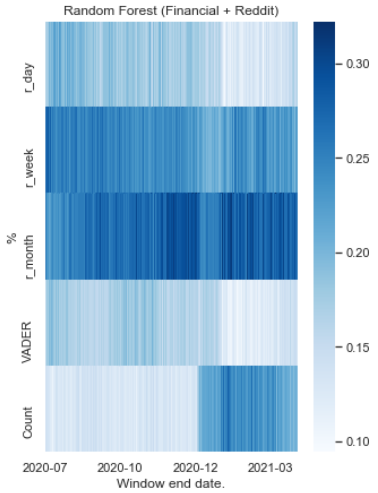
Figure 11
Random forest's weekly returns relative importance (2 of 2)



(a) Litecoin



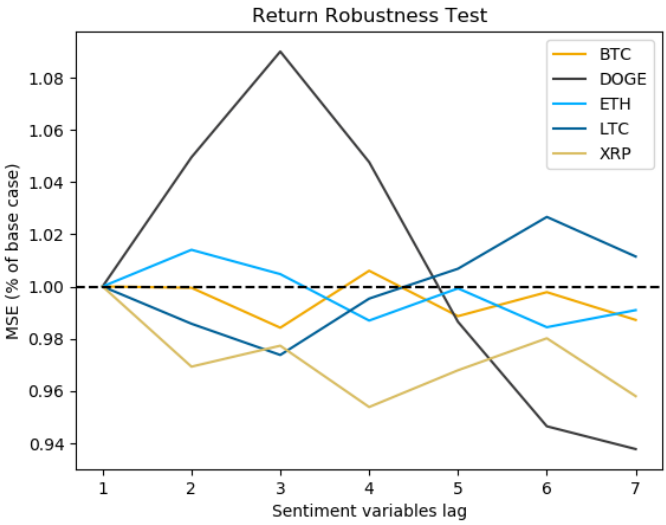
(b) Ripple



(c) Dogecoin

Notes: The *x*-axis is the end date of our one-year rolling window, and the *y*-axis is the absolute value of the regressor, divided by the sum of absolute values of all regressors for each date. The study period is from each asset's start date, as described in Table 1 to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one-step-ahead basis.

Figure 12
Robustness test for returns



Notes: The forecasting exercise using random forest with the architecture described in Section 3.3 is repeated, varying the number of periods by which the sentiment variables are lagged. The forecasting exercise with one lag represents the base model in the current section. Results are presented as a ratio of our base case. The study period is from each asset's start date as described in Table 1 to April 30, 2021.

crypto bubble and asset-specific events such as lawsuits, internet popularity, and pump-and-dump schemes through a feature-importance analysis. This is consistent with Engle et al. (2011), who find public information arrival to be related to increased volatility and volatility clustering.

References

Ahn, Y. and Kim, D. (2020). Emotional trading in the cryptocurrency market, *Finance Research Letters* p. 101912.

Alexa Internet (2021). Alexa internet, inc. Competitive Analysis, Marketing Mix and Traffic for reddit.com. Accessed May 2nd, 2021.
URL: https://www.alexa.com/siteinfo/reddit.com

Anamika, Chakraborty, M. and Subramaniam, S. (2021). Does sentiment impact cryptocurrency?, *Journal of Behavioral Finance* pp. 1–17.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Ebens, H. (2001). The dis-

tribution of realized stock return volatility, *Journal of financial economics* **61**(1): 43–76.

Andrews, D. W. (1986). Stability comparison of estimators, *Econometrica: Journal of the Econometric Society* pp. 1207–1235.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards, *The Journal of finance* **59**(3): 1259–1294.

Apopo, N. and Phiri, A. (2021). On the (in) efficiency of cryptocurrencies: have they taken daily or weekly random walks?, *Heliyon* **7**(4): e06685.

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

Bitcoin Average (2022). Bitcoinaverage. Accessed Jan 26th, 2022.
URL: https://bitcoinaverage.com/

Bloomberg (2021). Bloomberg. Ethereum Becoming More Than Crypto Coder Darling, Grayscale Says. Accessed May 2nd, 2021.
URL: https://www.bloomberg.com/news/articles/2020-12-04/ethereum-becoming-more-than-crypto-coder-darling-grayscale-says

Bouri, E., Gkillas, K., Gupta, R. and Pierdzioch, C. (2021). Forecasting realized volatility of bitcoin: The role of the trade war, *Computational Economics* **57**(1): 29–53.

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

CNBC (2021). Cnbc. Reddit frenzy pumps up Dogecoin, a cryptocurrency started as a joke. Accessed May 2nd, 2021.
URL: https://www.cnbc.com/2021/01/29/dogecoin-cryptocurrency-rises-over-400percent-after-reddit-group-talks-it-up.html

CoinDesk (2021). Coindesk. XRP Pump Fails to Materialize as Price Crashes 40% From Day's High. Accessed May 2nd, 2021.
URL: https://www.coindesk.com/xrp-pump-fails-to-materialize-as-price-crashes-40-from-days-high

CoinMarketCap (2021a). Coinmarketcap. Historical Snapshot - 03 January 2021. Accessed May 2nd, 2021.
URL: https://coinmarketcap.com/historical/20210103/

CoinMarketCap (2021b). Coinmarketcap. Today's Cryptocurrency Prices by Market Cap. Accessed May 2nd, 2021.
URL: https://coinmarketcap.com/

CoinMarketCap (2021c). Coinmarketcap. Top Cryptocurrency Spot Exchange. Accessed May 2nd, 2021.
URL: https://coinmarketcap.com/rankings/exchanges/

CoinMarketCap (2022). Coinmarketcap blog. Accessed Jan 26th, 2022.
URL: https://blog.coinmarketcap.com/2020/05/29/coinmarketcap-revamps-market-pairs-ranking-to-empower-users-against-volume-inflation/

Cointelegraph (2021). Cointelegraph. Bitcoin's Twitter-volume spikes to new all-time highs on Elon pump. Accessed May 2nd, 2021.
URL: https://cointelegraph.com/news/bitcoin-s-twitter-volume-spikes-to-new-all-time-highs-on-elon-pump

Corsi, F. (2009). A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics* **7**(2): 174–196.

DeMarzo, P. M., Vayanos, D. and Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions, *The Quarterly journal of economics* **118**(3): 909–968.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy, *Journal of Business & economic statistics* **20**(1): 134–144.

Dogecoin (2021). Dogecoin. Official website. Accessed May 2nd, 2021.
URL: https://dogecoin.com/

Engle, R. F., Hansen, M., Lunde, A. et al. (2011). And now, the rest of the news: Volatility and firm specific news arrival, *Unpublished Working Paper, CREATES* .

Fleming, J., Kirby, C. and Ostdiek, B. (2003). The economic value of volatility timing using "realized" volatility, *Journal of Financial Economics* **67**(3): 473–509.

Friedman, J., Hastie, T., Tibshirani, R. et al. (2001). *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Google Trends (2021). Google trends. Trending Searches. Accessed May 2nd, 2021.
URL: https://trends.google.com/trends/trendingsearches/daily?geo=US

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica: journal of the Econometric Society* pp. 424–438.

Henriques, I. and Sadorsky, P. (2018). Can bitcoin replace gold in an investment portfolio?, *Journal of Risk and Financial Management* **11**(3): 48.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

Karalevicius, V., Degrande, N. and De Weerdt, J. (2018). Using sentiment analysis to predict interday bitcoin price movements, *The Journal of Risk Finance* .

Kayal, P. and Balasubramanian, G. (2021). Excess volatility in bitcoin: extreme value volatility estimation, *IIM Kozhikode Society & Management Review* p. 2277975220987686.

Kraaijeveld, O. and De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices, *Journal of International Financial Markets, Institutions and Money* **65**: 101188.

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

Kristoufek, L. (2018). On bitcoin markets (in) efficiency and its evolution, *Physica A: statistical mechanics and its applications* **503**: 257–262.

Liu, J., Ma, F., Yang, K. and Zhang, Y. (2018). Forecasting the oil futures price volatility: Large jumps and small jumps, *Energy Economics* **72**: 321–330.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns, *Journal of Financial Economics* **123**(1): 137–162.

Markets Insider (2021). Markets insider. A new wave of institutional interest has boosted bitcoin. Here are the key players getting involved, from Morgan Stanley to Tesla. Accessed May 2nd, 2021.
URL: https://markets.businessinsider.com/currencies/news/bitcoin-btc-institutional-interest-cryptocurrencies-wall-street-tesla-microstrategy-jpmorgan-2021-3-1030194067

Martin, B. and Koufos, N. (2018). Sentiment analysis on reddit news headlines with python's natural language toolkit (nltk). Accessed May 2nd, 2021.
URL: https://www.learndatasci.com/tutorials/sentiment-analysis-reddit-headlines-pythons-nltk/

McAleer, M. and Medeiros, M. C. (2008). Realized volatility: A review, *Econometric Reviews* **27**(1-3): 10–45.

Naeem, M. A., Mbarki, I., Suleman, M. T., Vo, X. V. and Shahzad, S. J. H. (2021). Does twitter happiness sentiment predict cryptocurrency?, *International Review of Finance* **21**(4): 1529–1538.

Pesaran, M. H. and Timmermann, A. (2005). Small sample properties of forecasts from autoregressive models under structural breaks, *Journal of Econometrics* **129**(1-2): 183–217.

Prajapati, P. (2020). Predictive analysis of bitcoin price considering social sentiments, *arXiv preprint arXiv:2001.10343* .

Qiu, Y., Zhang, X., Xie, T. and Zhao, S. (2019). Versatile har model for realized volatility: A least square model averaging perspective, *Journal of Management Science and Engineering* **4**(1): 55–73.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation, *Journal of econometrics* **99**(1): 39–61.

Rosol, M., Młyńczak, M. and Cybulski, G. (2022). Granger causality test with nonlinear neural-network-based methods: Python package and simulation study, *Computer Methods and Programs in Biomedicine* .

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

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

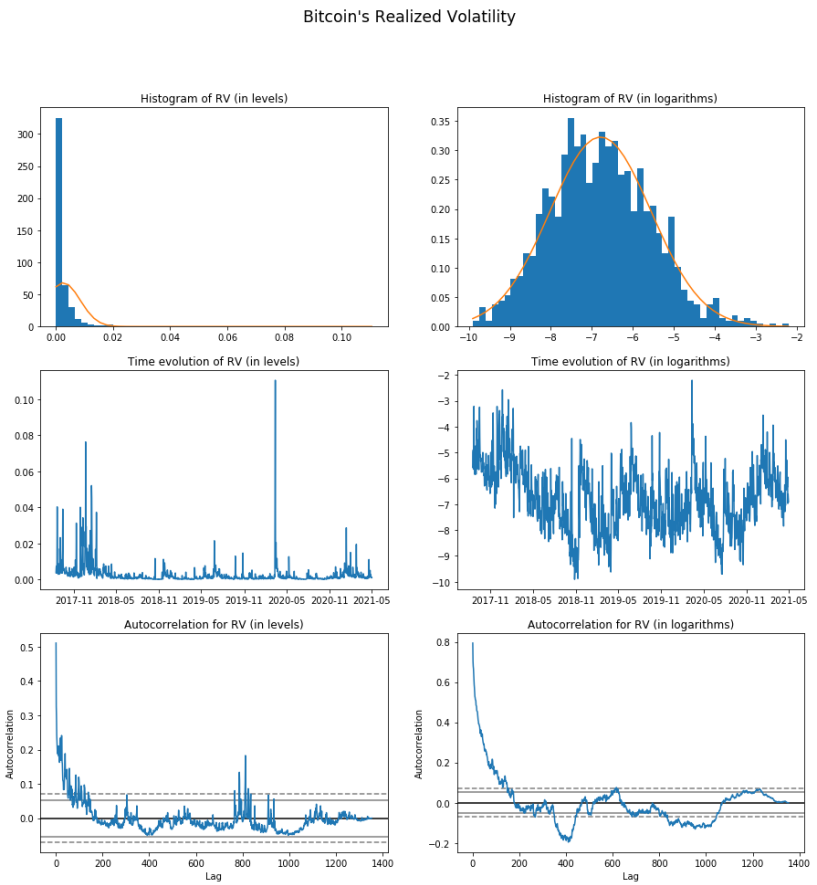
Wooley, S., Edmonds, A., Bagavathi, A. and Krishnan, S. (2019). Extracting cryptocurrency price movements from the reddit network sentiment, *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, IEEE, pp. 500–505.

Yu, M. (2019). Forecasting bitcoin volatility: The role of leverage effect and uncertainty, *Physica A: Statistical Mechanics and Its Applications* **533**: 120707.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*, Ravenio Books.

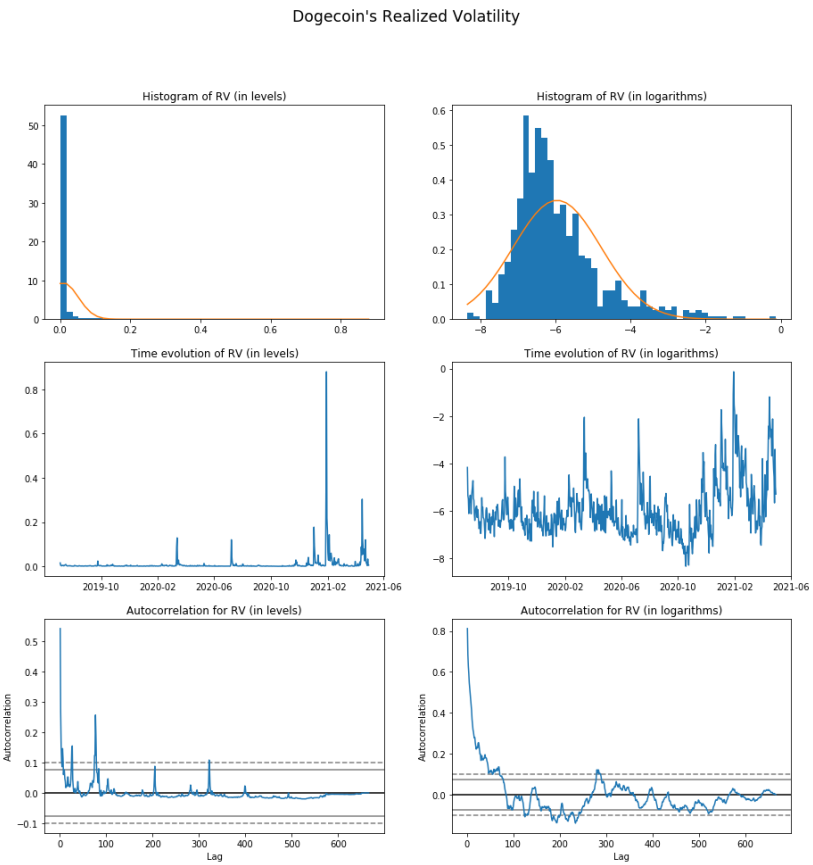
A. Additional tables and figures

Figure A1
Summary statistics for Bitcoin's realized volatility



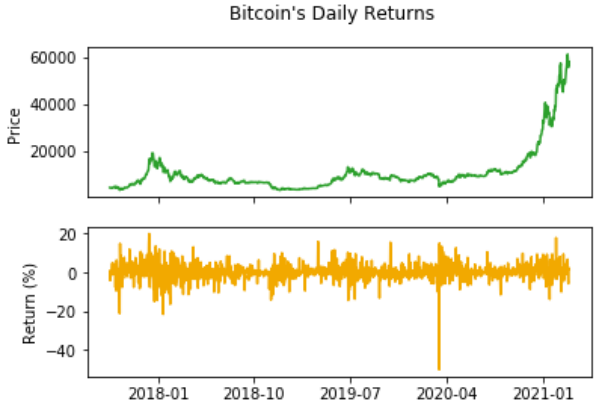
The yellow line in the first row denotes a normal distribution. Realized volatilities are constructed following McAleer and Medeiros (2008), using price information from August 17, 2017, to April 30, 2021.

Figure A2
Summary statistics for Dogecoin's realized volatility



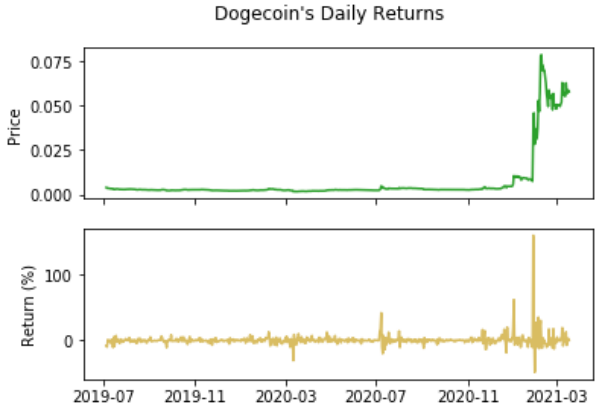
The yellow line in the first row denotes a normal distribution. Realized volatilities are constructed following McAleer and Medeiros (2008), using price information from July 5, 2019 to April 30, 2021.

Figure A3
Summary statistics for Bitcoin's daily returns



Price expressed in USD. Price information from August 17, 2017, to April 30, 2021.

Figure A4
Summary statistics for Dogecoin's daily returns



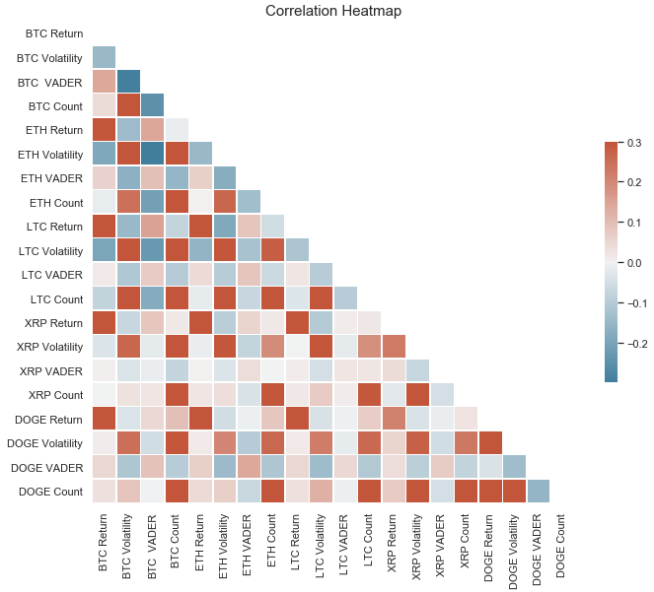
Price expressed in USD. Price information from July 5, 2019, to April 30, 2021.

Table A1
Summary table for all variables used

	count	mean	std	min	25%	50%	75%	max
BTC Return	1353.0	0.00	0.04	−0.50	−0.02	0.00	0.02	0.20
BTC Volatility	1353.0	0.00	0.01	0.00	0.00	0.00	0.00	0.11
BTC VADER	1353.0	0.16	0.06	−0.11	0.12	0.16	0.19	0.39
BTC Count	1353.0	369.40	339.51	0.00	198.00	263.00	406.00	4012.00
ETH Return	1353.0	0.00	0.05	−0.59	−0.02	0.00	0.03	0.23
ETH Volatility	1353.0	0.00	0.01	0.00	0.00	0.00	0.00	0.12
ETH VADER	1353.0	0.24	0.09	−0.13	0.17	0.24	0.29	0.54
ETH Count	1353.0	88.73	49.92	0.00	53.00	74.00	120.00	310.00
LTC Return	1235.0	−0.00	0.05	−0.49	−0.03	−0.00	0.03	0.28
LTC Volatility	1235.0	0.00	0.01	0.00	0.00	0.00	0.00	0.14
LTC VADER	1235.0	0.27	0.20	−0.74	0.15	0.27	0.39	0.91
LTC Count	1235.0	30.47	70.54	0.00	8.00	15.00	27.00	1043.00
XRP Return	1093.0	0.00	0.06	−0.54	−0.02	0.00	0.02	0.58
XRP Volatility	1093.0	0.00	0.02	0.00	0.00	0.00	0.00	0.39
XRP VADER	1093.0	0.19	0.19	−0.62	0.08	0.19	0.30	0.86
XRP Count	1093.0	26.93	90.13	0.00	9.00	13.00	25.00	2569.00
DOGE Return	666.0	0.01	0.10	−0.49	−0.02	0.00	0.02	1.59
DOGE Volatility	666.0	0.01	0.04	0.00	0.00	0.00	0.00	0.88
DOGE VADER	666.0	0.32	0.17	−0.56	0.23	0.31	0.43	0.79
DOGE Count	666.0	685.49	3030.26	0.00	14.00	19.00	31.00	43592.00

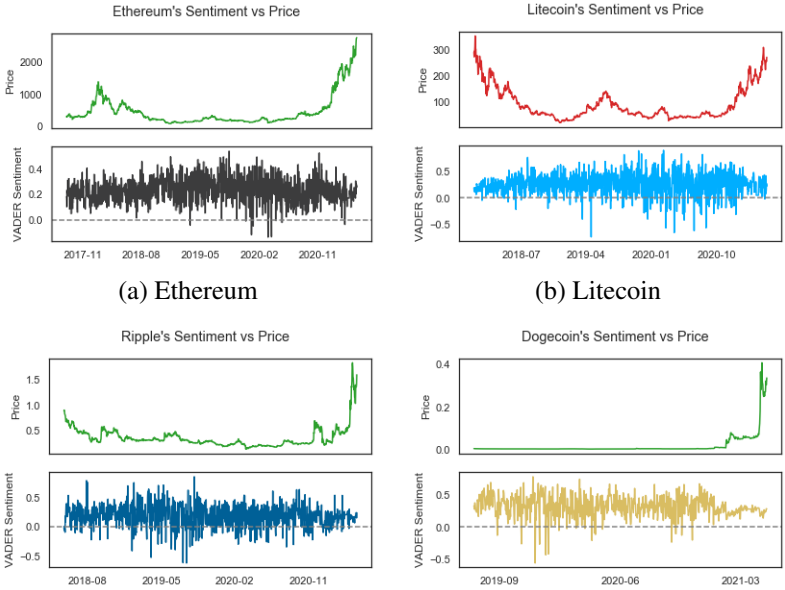
A VADER suffix represents the average daily sentiment using the Valence Aware Dictionary for Sentiment Reasoning methodology after removing zero-sentiment observations for each asset. A count suffix represents the total number of submissions per day in each asset's subreddit. Return and Volatility suffixes represent log-returns and realized volatilities following McAleer and Medeiros (2008).

Figure A5
Correlation plot for dataset



A VADER suffix represents the average daily sentiment using the Valence Aware Dictionary for Sentiment Reasoning methodology after removing zero-sentiment observations for each asset. A count suffix represents the total number of submissions per day in each asset's subreddit. Return and Volatility suffixes represent log-returns and realized volatilities following McAleer and Medeiros (2008).

Figure A6
Average daily sentiment under VADER

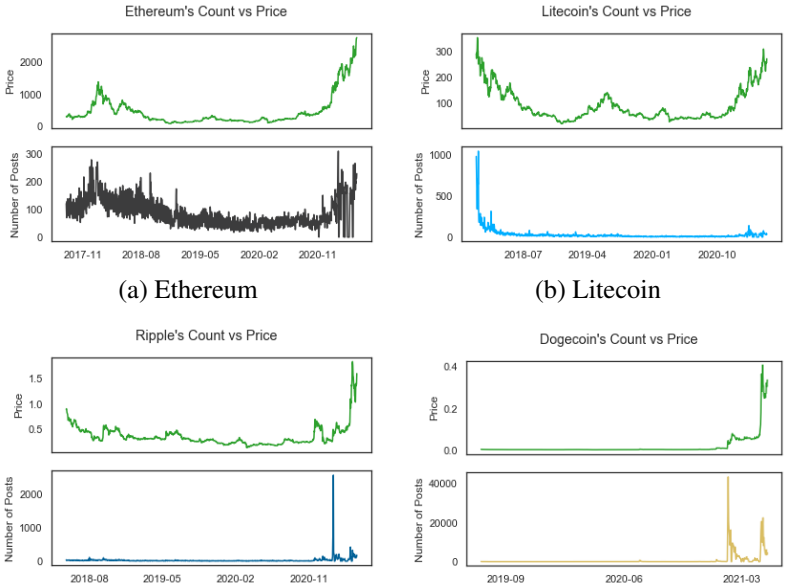


(a) Ethereum (b) Litecoin

(c) Ripple (d) Dogecoin

The top image presents the price in USD. The bottom image is the average daily VADER sentiment after removing all titles with a score equal to zero. Information used is from each asset's start date as described in Table 1 to April 30, 2021.

Figure A7
Count of daily submissions



(a) Ethereum (b) Litecoin

(c) Ripple (d) Dogecoin

The top image presents the price in USD. The bottom image is the count of daily submissions in each community. Information used is from each asset's start date as described in Table 1 to April 30, 2021.