Reddit as a prediction tool for crypto-assets

Luis Antonio Loredo Camou†

Abstract Cryptocurrencies, such as Bitcoin and Ethereum, have recently become a conversation topic among the general population. This paper will explore the information available in Reddit regarding crypto assets. Unlike other social platforms, Reddit allows analyzing the general population sentiment while conveniently organizing 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 benefit from Reddit sentiment variables consistently, results are not statistically different from a benchmark. In contrast, returns present mixed forecasting results but show statistical differences from the proposed benchmark. We also offer evidence that the 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 15th, 2021, with over 5 million searches in the US. 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 its all-time high (Cointelegraph, 2021).

What are the incentives for someone to post and read financial message boards? Antweiler and Frank (2004) provide an overview of the existing literature. In particular, the authors focus on the theory by 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 among them 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 will focus on Reddit. Reddit is most well known for being 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 people 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 7th in the United States.

Several papers dealing with the impact of sentiment indices on returns of cryptocurrencies precede ours. Among these, we note research done by Kristoufek (2013) studying the effect of search trends in Wikipedia and Google for Bitcoin. Also, 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 previously. 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 - but not all - besides Bitcoin, commonly referred to as altcoins. In contrast, Ahn and Kim (2020) conclude, using posts from a Bitcoin forum that, contrary to future returns, volume and volatility are related to emotional factors.

Research using Reddit has also been done previously. In particular, Prajapati (2020) makes use of different lexical dictionaries to perform sentiment analysis over Reddit and Google News to predict the price of Bitcoin using data from January 1, 2018, to November 20, 2019. It is concluded that social sentiment captured through Reddit and Google News improves forecasts against past prices-only models.

Wooley et al. (2019) predicts, using 24 Reddit communities, the following three months of price directions (up vs down) using information from July 1, 2016, to July 24, 2018. It is concluded that predictions benefit when using sentiment variables against a lagged prices-only model.

To the best of our knowledge, no previous research has been done study-
ing the forecasting power of general use forums, such as Reddit, on volatility and returns and included Granger, Mariano-Diebold, and robustness tests. Further, no other study has performed detailed feature importance analysis as presented in this paper. Our research not only focuses on Bitcoin, as most research does, but also on altcoins.

This study finds 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 a HAR-RV model. In contrast, although mean squared error results for returns are mixed, these 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 falls along the lines of Antweiler and Frank (2004), who concludes that stock messages can be used to help predict volatility and returns for companies in the Dow Jones Industrial Average. Our work also relates to Engle et al. (2011), showing that public information arrival is related to increases in volatility.

Our work reinforces previous research showing the positive impact of sentiment indices on returns and volatilities of cryptocurrencies. In particular, our conclusions are similar to those found by 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 Prajapati (2020) that sentiment seems to improve Bitcoin price prediction, we achieve mixed results when expanding to other assets. However, we find evidence that information extracted from Reddit seems to consistently reduce volatility forecasting error for all assets studied.

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 the existence of 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, such as in Manela and Moreira (2017).

The document is organized as follows. Section 2 includes a description of
the data and sources. Section 3 motivates our model selection and methodology. Next, section 4 offers 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 presents our conclusions.

2. Data

There is a wide range of cryptocurrencies and tokens available. As of May 2, 2021, CoinMarketCap (2021a) reports the existence of 9,527 cryptos. We have decided to focus on five of them, described in the following paragraphs. The first four cryptocurrencies have been selected due to being the top 4 projects by market capitalization at the start of 2021, representing almost 85% of the total market capitalization (CoinMarketCap, 2021b). In particular, Bitcoin and Ethereum represent 70.68% and 10.79%, respectively. Finally, Dogecoin has been selected due to its internet popularity. The assets are presented by market capitalization.

**Bitcoin (BTC):** A cryptocurrency invented in 2008 by an unknown persona denominated Satoshi Nakamoto. Bitcoin (BTC) 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 (ETH).

**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 referenced by, among others, Elon Musk, Mark Cuban, Snoop Dogg, and Gene Simmons.

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 make use of 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 and is presented in Table 1.

<table>
<thead>
<tr>
<th>name</th>
<th>start date</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTC</td>
<td>August 17th, 2017</td>
</tr>
<tr>
<td>ETH</td>
<td>August 17th, 2017</td>
</tr>
<tr>
<td>LTC</td>
<td>December 13th, 2017</td>
</tr>
<tr>
<td>XRP</td>
<td>May 4th, 2018</td>
</tr>
<tr>
<td>DOGE</td>
<td>July 5th, 2019</td>
</tr>
</tbody>
</table>

While resources aggregating several sources of information are available, such as Bitcoin Average (2022), these may suffer from volume inflation as exchanges try to gain visibility as described in CoinMarketCap (2022). It is, therefore, that we make use of one single reliable data source.

The daily volatility measure for our assets will be constructed as seen in McAleer and Medeiros (2008). As recommended, to find a trade-off between accuracy and microstructure noise, we use the 5-minute close-to-close returns as our input. The procedure is as follows:
Figure 1
Boxplot for realized volatilities

Realized volatilities constructed as seen in McAleer and Medeiros (2008) using price history from the initial price observation specified in Table 1 to April 30, 2021.

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

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

A boxplot figure of the realized volatilities is provided in Figure 1. We also offer additional descriptive images for Bitcoin and Dogecoin in Figures A1 and A2 and make the plots available for other tickers upon request. These results are in line with the usual stylized facts of financial returns data. In line with 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, daily log-returns are estimated using 5-minute close-to-close prices and summarized as a boxplot in Figure 2. The prices and returns for Bitcoin and Dogecoin are presented in Figures A3 and A4. From here, we construct the cumulative 7-day returns.

2.2 Reddit Variables

As mentioned previously, our second source is Reddit. The start date of each subreddit and the number of submissions are presented in Table 2. The titles of each submission are recollected through the Pushshift Reddit dataset, as presented in Baumgartner et al. (2020), a big-data and analytics
Log-returns using price history from the initial price observation specified in Table 1 to April 30, 2021.

Table 2

<table>
<thead>
<tr>
<th>name</th>
<th>start date</th>
<th>number of submissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTC</td>
<td>September 9th, 2010</td>
<td>942,668</td>
</tr>
<tr>
<td>ETH</td>
<td>December 14th, 2013</td>
<td>165,897</td>
</tr>
<tr>
<td>LTC</td>
<td>October 14th, 2011</td>
<td>78,578</td>
</tr>
<tr>
<td>XRP</td>
<td>October 14th, 2009</td>
<td>73,097</td>
</tr>
<tr>
<td>DOGE</td>
<td>December 8th, 2013</td>
<td>56,9865</td>
</tr>
</tbody>
</table>

Information up to April 30th, 2021.

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-based sentiment analysis model 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 punctuations compared to other sentiment analysis models. This is the reason we use VADER over other alternatives.

The VADER procedure is now outlined. Using a pre-trained dictionary, each lexical feature in a corpus is assigned a valence score on a scale from
Table 3
Examples of punctuation, capitalization, degree modifiers, and shift in polarity under VADER

<table>
<thead>
<tr>
<th>sentence</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitcoin is a cryptocurrency.</td>
<td>+0.00</td>
</tr>
<tr>
<td>Bitcoin is a secure cryptocurrency.</td>
<td>+0.34</td>
</tr>
<tr>
<td>Bitcoin is a secure cryptocurrency!</td>
<td>+0.40</td>
</tr>
<tr>
<td>Bitcoin is a SECURE cryptocurrency.</td>
<td>+0.48</td>
</tr>
<tr>
<td>Bitcoin is a very secure cryptocurrency.</td>
<td>+0.40</td>
</tr>
<tr>
<td>Bitcoin is a secure cryptocurrency but a risky investment.</td>
<td>−0.13</td>
</tr>
</tbody>
</table>

−4 to +4, where −4 represents the most negative sentiment and +4 the most positive. This dictionary includes words and emoticons, acronyms, and slang. The results for the corpus 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:

\[ x = \frac{x}{\sqrt{x^2 + 15}} \]

A set of examples are available in Table 3.

Once a score has been obtained for all titles, we will first count the number of submissions per day. We will then delete all observations with a VADER score equal to 0 before computing the average score for each day. We delete all zero observations to induce variation. In case of a missing observation, we will fill with 0, representing a lack of submissions and sentiment.

A summary of all our variables used is presented in Table A1. A correlation heatmap is presented in Figure A5.

3. Models and forecasts

Rolling windows are commonly used to keep the estimation window at a fixed length. Different estimation rolling window sizes have been used to forecast 1-day ahead prices and volatilities in 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 the WTI crude oil futures. However, as Pesaran and Timmermann (2005) note, although there is not a 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, without including too much data not to indicate the current regime, addressing 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 have decided to use LASSO and Ridge. These choices are due to their properties and wide application in the literature. Through the analysis of $l_1$ regularization, we will allow 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 will focus on a nonlinear model, Random Forest. The details of each model, as well as the benchmarks used, are outlined below.

### 3.1 Benchmarks

#### 3.1.1 Volatility: Heterogeneous Autoregressive model

Following Corsi (2009), we make use of 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 make use of the HAR(1,5,22) model for the prediction of 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 specification has been used previously to study Bitcoin volatility by Yu (2019) and Bouri et al. (2021).

Further, this model may be defined in logs:

$$\log(RV)_{t+1} = \beta_0 + \beta_1 \log(RV)_t + \beta_2 \log(RV)_{w,t} + \beta_3 \log(RV)_{m,t} + U_{t+1}$$
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 will use the log-model given the approximate normality of $\log(RV)$, given a more stable behavior. Further, we observe a stronger persistence in logs than in levels. This is in line with Andersen et al. (2001). We will need to apply the exponential function to our results to recover our forecast in levels. This transformation will avoid negative values since the exponential is always positive. We will default to the levels model if we have a zero observation.

### 3.1.2 Returns: Random Walk and in-sample mean

Previous work has been done on the efficiency of cryptocurrency markets. Most notably, Urquhart (2016) shows that, while on a 2010 to 2016 time-frame, Bitcoin prices seem inefficient on one-day ahead returns, when split from 2013 to 2016, evidence of efficiency appears. Following this, Apopo and Phiri (2021) expand the analysis for altcoins and weekly returns. The authors conclude that while one-day ahead returns are efficient, weekly returns show evidence against this.

For our returns, we will make 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:

$$E_t [P_{t+1}] = P_t$$

and therefore

$$E_t [r_{t+1}^{(7)}] = 0$$

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:

$$E_t [r_{t+1}^{(7)}] = \frac{1}{365} \sum_{i=0}^{364} r_{t-i}^{(7)}$$

This benchmark has been used before as a naive forecast for next period prediction in works 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 framework. First, Information Criterion provides a single value of penalization value to use, saving us computational time. Second, Information Criterion does not have to split data between a training and test set, which allows us to use all the information available at the time.

Our data is standardized before fitting the models. Our forecasting models can be represented as:

\[
\log(RV_{t+1}) = \beta_0 + \beta' \log(X_t) + \beta_d \log(RV_t) + \beta_w \log(RV_{w,t}) + \beta_m \log(RV_{m,t}) + U_{t+1}
\]

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

\[
r_{t+1} = \beta_0 + \beta' X_t + \beta_d r_t + \sum_{i=0}^{6} \beta_w r_{t-i} + \sum_{i=0}^{29} \beta_m r_{t-i} + U_{t+1}
\]

where \(r_{t+1}^{(7)}\) 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.

A 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. Once constructed, we compute
our final prediction using the average of these bootstrapped trees. In this way, we have created a more stable estimator called a Random Forest.

We now describe our hyperparameter selection process. While there have been proposals to adapt the Information Criterion approach to nonlinear models, the predominant method is to use Cross-Validation. However, there is an important property to highlight: we are working with time-series data, and the standard k-fold CV would break this time dependence. Instead, we implement $hv$- Cross-Validation, as proposed by Racine (2000). Compared to the standard k-fold Cross-Validation, $hv$- Cross-Validation 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 to our grid values, we reference Friedman et al. (2001). The 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, we select the following parameters:

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

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

In addition, 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. All other parameters will be left as default in sklearn’s RandomForestRegressor, given no a priori knowledge.

Once again, our data is standardized before fitting the models.

4. **Natural language processing**

This chapter will provide some basic natural language processing analysis to the titles recovered from Reddit from August 17, 2017, to April 30, 2021. Doing this, we hope to uncover similarities and differences among communities. We will discuss polarity, word frequencies, and the most used words. This will be based on Martin and Koufos (2018). We will end by introducing
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.

We start by looking at polarity. We categorize as positive all submissions with a VADER score higher than 0.25 and as negative any submission with a score lower than −0.25. Everything else is labeled as neutral. The results are displayed in Figure 3. It is noticeable that all subreddits follow a similar behavior. The percentage of titles with positive sentiment is around 25% for all communities. ETH stands out as having the most positive community, and XRP has the least percentage of positive titles. In contrast, Bitcoin seems to have a considerable difference above all other communities regarding negativity. The results for DOGE are fascinating since the project is focused on having a *fun and friendly internet currency* (Dogecoin, 2021). The results presented here show an average behavior across all labels for Dogecoin.

We now look to understand the topics discussed in each subreddit. We will focus on the top 10 words mentioned in each community to do this. 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 terms in all columns self-reference to the asset. Focusing on words such as *crypto*, *wallet*, or *blockchain*, we notice that all columns except the one for Dogecoin reference the technology. In particular, Ethereum stands out...
Table 4
Top 10 most-mentioned words by community

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

Stop words and numbers have been excluded. The period studied is August 17th, 2017 to April 30th, 2021.

for having most of its words be tech-related. In contrast, Ripple presents only one word and Dogecoin none. Similarly, in all columns except for the one for Ethereum, we notice a reference to price with words such as price, buy, or coinbase (a popular exchange). Dogecoin presents more price-related words than other assets. 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 misspelling of the word hold, not selling. This table is evidence that communities carry different information, and therefore the predicting abilities may be different.

While the previous table focused on only the top 10 words, it may be interesting to study all words used in each community. According to Zipf 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 people want to be understood (Zipf, 2016). We present Figure 4.

We notice that our statistics behave roughly as expected. It appears that Bitcoin, Litecoin, and Ripple keep a constant spread among them, implying a similar complexity in their way of speech. Ethereum seems to have a 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, Ripple, and Dogecoin. 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, for Bitcoin, we notice that its highest number of posts coincides with the 2017 bubble. Similarly, we notice an increase in the number of posts during the 2021 rally for all assets. We now focus on two assets in particular: Ripple and Dogecoin. For Ripple, we observe that the peak was reached 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). For Dogecoin, we have noticed extremely high numbers since the end of January 2021. Once again, this is related to the
Figure 5  
Bitcoin’s average daily sentiment under VADER  

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  

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.
Table 5
Granger causality results

<table>
<thead>
<tr>
<th>data</th>
<th>model</th>
<th>BTC</th>
<th>ETH</th>
<th>LTC</th>
<th>XRP</th>
<th>DOGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>VADER → Volatility</td>
<td>0.00</td>
<td>0.05</td>
<td>0.15</td>
<td>0.10</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Count → Volatility</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>VADER → Return</td>
<td>0.02</td>
<td>0.51</td>
<td>0.81</td>
<td>0.52</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>Count → Return</td>
<td>0.05</td>
<td>0.27</td>
<td>0.12</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>nonlinear</td>
<td>VADER → Volatility</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Count → Volatility</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>VADER → Return</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Count → Return</td>
<td>0.00</td>
<td>0.06</td>
<td>0.00</td>
<td>0.03</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Reported p-value from a Granger causality test with seven lags using each asset’s start date as described in Table 1 to April 30th, 2021. The linear model used is as defined in 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 50 neurons each.

Before moving on, we will 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 y when already taking into account lagged values of y.

The test is usually seen in its linear form:

\[ y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_s y_{t-s} + \beta_1 x_{t-1} + \beta_s x_{t-s} + \epsilon_t \]

where: \( \epsilon_t \) 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 will follow the neural network procedure using a rather simplistic architecture for motivational purposes.

Our results are presented in Tables 5.

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 those 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 sen-
timent variables on the financial variables are somewhat mixed in the linear case, there are only two cases in which a \( p \)-value below 5\% is not reached in the nonlinear case.

Our results are in line with those presented by 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. In research by Naeem et al. (2021), the Twitter Happiness Index, a sentiment variable collected using lexical features, is compared against five different cryptocurrencies’ returns. 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? In Table 6, we see that there is indeed a bilateral relationship between our variables and find evidence of some unilateral relationships.

We now move to introduce our results for volatility and returns prediction.

5. Are Reddit variables useful to forecast?

5.1 Volatility

We present the prediction results for the one-day ahead volatility in Table 7. The values are obtained by computing the ratio of each model and the
benchmarks’s mean squared error. Our benchmark is a HAR(1,5,22) model as described in the Models and Forecasts section.

Our first observation is that nonlinear models such as Random Forests 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 Forests. Adding Reddit variables to a nonlinear model can have minor 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 compared have different accuracy levels. We carry out these tests over all our forecast dates. The results are presented in Table 8, and show 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 then compare the absolute value of each coefficient. For our Random Forest model, we will use the impurity-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 will divide each feature importance coefficient by the sum of all coefficients on that day. We present the results for Random Forest with Reddit variables in Figures 7 and 8.

From these plots, we first notice that the one-day lagged volatility seems

<table>
<thead>
<tr>
<th>data</th>
<th>model</th>
<th>BTC</th>
<th>ETH</th>
<th>LTC</th>
<th>XRP</th>
<th>DOGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>financial</td>
<td>HAR</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.47</td>
<td>0.70</td>
<td>0.87</td>
<td>0.92</td>
<td>0.57</td>
</tr>
<tr>
<td>financial + Reddit</td>
<td>OLS</td>
<td>1.15</td>
<td>1.03</td>
<td>1.05</td>
<td>3.12</td>
<td>2.91</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.99</td>
<td>0.95</td>
<td>0.91</td>
<td>2.66</td>
<td>1.97</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>0.96</td>
<td>0.93</td>
<td>0.97</td>
<td>2.74</td>
<td>2.29</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.46</td>
<td>0.63</td>
<td>0.69</td>
<td>0.81</td>
<td>0.54</td>
</tr>
</tbody>
</table>

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 30th, 2021. We use a one-year rolling window where the first 365 observations are used 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 results for volatility

<table>
<thead>
<tr>
<th>data</th>
<th>model</th>
<th>BTC</th>
<th>ETH</th>
<th>LTC</th>
<th>XRP</th>
<th>DOGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>financial</td>
<td>HAR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.39</td>
<td>0.56</td>
<td>0.75</td>
<td>0.57</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>0.06</td>
<td>0.16</td>
<td>0.26</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.90</td>
<td>0.50</td>
<td>0.47</td>
<td>0.24</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>0.76</td>
<td>0.39</td>
<td>0.44</td>
<td>0.25</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.41</td>
<td>0.49</td>
<td>0.47</td>
<td>0.13</td>
<td>0.46</td>
</tr>
<tr>
<td>financial + Reddit</td>
<td>OLS</td>
<td>0.06</td>
<td>0.16</td>
<td>0.26</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>0.90</td>
<td>0.50</td>
<td>0.47</td>
<td>0.24</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>0.76</td>
<td>0.39</td>
<td>0.44</td>
<td>0.25</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.41</td>
<td>0.49</td>
<td>0.47</td>
<td>0.13</td>
<td>0.46</td>
</tr>
</tbody>
</table>

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

Figure 7
Random Forest’s volatility relative importance (1 of 2)

(a) Bitcoin
(b) Ethereum

The $x$-axis is the end date of our one-year rolling window, and $y$-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 up to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one step-ahead basis.
The $x$-axis is the end date of our one-year rolling window, and $y$-axis is the absolute value of the regressor divided by the sum of absolute values of all regressors. The study period is from each asset's start date as described in Table 1 up to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one step-ahead basis.
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. When focusing on our Reddit variables, we first notice that the number of posts seems to gain more importance than the average VADER score and becomes particularly significant at certain moments. We notice our variable becomes dominant for those assets present during the 2017 crypto bubble (Bitcoin, Ethereum, and Litecoin). It is interesting to notice how Ethereum does not rely on this variable as much as Bitcoin, although the study periods are the same. We presume this relates to Ethereum not discussing price as often as other forums, as seen in our natural language processing analysis. Interestingly, Bitcoin presents relatively high 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.

For XRP, we notice a strong effect of the number of submissions around May 2018, which coincides with a class action filed against Ripple for selling unregistered tokens. We also noticed an effect during December 2020 when Ripple and two of its executives were sued by the SEC for selling the unregistered securities. Finally, we notice a short but significant effect relating to the XRP pump-and-dump of February 2020. For Dogecoin, we notice a powerful effect during the last observed dates, which coincides with Dogecoin’s rise in internet popularity and euphoria.

Although the one-day-ahead forecasting period was chosen following previous literature, other specifications are possible. To test the benefits of different timeframes, 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 are presented in Figure 9 and show, on first inspection, that there does not seem to be a clear winner concerning the lags to use. Further, a Diebold-Mariano test, Table 8, shows that out of those forecasts which show an improvement, only Ethereum at six 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

We now present the prediction results for the 7-day returns in Table 9. Once again, the values are obtained by computing the ratio of each model and the benchmark’s mean squared error. The benchmark used here will be the Random Walk model described in the Models and Forecasts section. Ad-
The forecasting exercise using a Random Forest model with the architecture described in Section 3.3 is repeated but varying the number of periods that the sentiment variables are lagged. The forecasting exercise with one lag represents the base model in the current section. The 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 up to April 30, 2021.

Additionally, we will also present the results for the in-sample mean.

Our results are somewhat mixed when focusing on the models using Reddit variables. We first notice that 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 now perform the Diebold-Mariano test to check if the forecasts are statistically different from the benchmark. Notice that given that the random walk model presents no variance, we must use the in-sample model as our benchmark. The results of this analysis are presented in Table 10. These results are exciting as they show that, for three assets, there is a model using the Reddit variables that differs and outperforms the in-sample mean: BTC, XRP, and DOGE. Additionally, for Bitcoin and Dogecoin, we get models that outperform every other model in a statistically different way from the benchmark.

Once again, we present the feature importance analysis to better understand our results. These results are available in Figures 10 and 11.
Table 9
Returns results

<table>
<thead>
<tr>
<th>data</th>
<th>model</th>
<th>BTC</th>
<th>ETH</th>
<th>LTC</th>
<th>XRP</th>
<th>DOGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>benchmarks</td>
<td>Random Walk</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>In-sample mean</td>
<td>0.98</td>
<td><strong>0.98</strong></td>
<td>1.03</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>financial</td>
<td>OLS</td>
<td>1.00</td>
<td>1.00</td>
<td>1.03</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>1.02</td>
<td>1.03</td>
<td>1.07</td>
<td><strong>0.90</strong></td>
<td>0.80</td>
</tr>
<tr>
<td>financial + Reddit</td>
<td>OLS</td>
<td>1.01</td>
<td>1.00</td>
<td>1.03</td>
<td>1.02</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td><strong>0.97</strong></td>
<td>1.00</td>
<td>1.03</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>1.01</td>
<td>1.00</td>
<td>1.05</td>
<td>1.02</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.98</td>
<td>1.00</td>
<td>1.04</td>
<td>0.94</td>
<td><strong>0.74</strong></td>
</tr>
</tbody>
</table>

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 up to April 30th, 2021. We use a one-year rolling window where the first 365 observations are used 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

<table>
<thead>
<tr>
<th>data</th>
<th>model</th>
<th>BTC</th>
<th>ETH</th>
<th>LTC</th>
<th>XRP</th>
<th>DOGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>benchmarks</td>
<td>Random Walk</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>In-sample mean</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>financial</td>
<td>OLS</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td><strong>0.00</strong></td>
<td>0.00</td>
</tr>
<tr>
<td>financial + Reddit</td>
<td>OLS</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td><strong>0.00</strong></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td><strong>0.00</strong></td>
</tr>
</tbody>
</table>

Diebold-Mariano test p-value results for each model against the in-sample mean benchmark across all the available observation dates from Table 2 up to April 30, 2021. For each asset, the model with the best performance in Table 9 is highlighted in bold.
Focusing on those models able to beat the benchmark, our first observation is that, differently from the volatility results, our most important financial feature is the monthly return. This seems to imply that investors look at the returns in a more extended timeframe when deciding to deploy capital. When focusing on our Reddit variables, we notice, once again, that our number of submissions variable can capture market-wide events such as the crypto bubble seen in Bitcoin’s subfigure. Additionally, the two periods concerning 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, for Dogecoin, we confirm that the 2021 rally is captured in the number of submissions variable.

Similar to what was done for volatility, we now perform a robustness test for our returns. The results are presented in Figure 12. As observed, there does not seem to be a clear winner concerning the lags to use, although at first, it could seem that a higher number of lags is preferred. However, a
Figure 11
Random Forest’s weekly returns relative importance (2 of 2)

(a) Litecoin
(b) Ripple
(c) Dogecoin

The x-axis is the end date of our one-year rolling window, and 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 up to April 30, 2021. The first 365 observations are used as an initial calibration period. Forecasts are done on a one step-ahead basis.
The forecasting exercise using a Random Forest model with the architecture described in Section 3.3 is repeated but varying the number of periods that the sentiment variables are lagged. The forecasting exercise with one lag represents the base model in the current section. The 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 up to April 30, 2021.

Diebold-Mariano test shows that only Ethereum at four lags is statistically different. Once again, although the 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 studied 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 understand our inputs better. We unveil that different subreddits seem to discuss different topics in different complexities through these tools. Even more, the sentiment seems to differ among communities. We then present the results for our prediction exercise.

Our results show that while sentiment variables seem to reduce the forecasting error for volatility, they fail to be statistically different from the benchmark proposed. In contrast, while our forecasts for returns have mixed re-
results at reducing the forecasting error, they are statistically different from the proposed benchmark. These results are in line with those by Antweiler and Frank (2004), where it is concluded 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, when expanding on their work, we agree with Ahn and Kim (2020) and Kraaijeveld and De Smedt (2020) that while sentiment seems to help reduce the forecasting error in volatility, the effect is mixed for returns.

We also notice that the Reddit variables seem to capture market-wide events such as the 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) that public information arrival is related to increases in volatility and volatility clustering.

References


URL: https://bitcoinaverage.com/

URL: https://www.bloomberg.com/news/articles/2020-12-04/ethereum-becoming-more-than-crypto-coder-darling-grayscale-says


URL: https://www.cnbc.com/2021/01/29/dogecoin-cryptocurrency-rises-over-400percent-after-reddit-group-talks-it-up.html

URL: https://www.coindesk.com/xrp-pump-fails-to-materialize-as-price-crashes-40-from-days-high

URL: https://coinmarketcap.com/historical/20210103/

URL: https://coinmarketcap.com/

URL: https://coinmarketcap.com/rankings/exchanges/

URL: https://blog.coinmarketcap.com/2020/05/29/coinmarketcap-revamps-market-pairs-ranking-to-empower-users-against-volume-inflation/


URL: https://dogecoin.com/


URL: https://trends.google.com/trends/trendingsearches/daily?geo=US


**URL:** https://www.learndatasci.com/tutorials/sentiment-analysis-reddit-headlines-pythons-nltk/


A. Additional tables and figures

Figure A1
Summary statistics for Bitcoin’s realized volatility

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.
Summary statistics for Dogecoin’s realized volatility

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

Bitcoin's Daily Returns


Figure A4
Summary statistics for Dogecoin’s daily returns

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

<table>
<thead>
<tr>
<th></th>
<th>count</th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BTC Return</strong></td>
<td>1353.00</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.50</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.20</td>
</tr>
<tr>
<td><strong>BTC Volatility</strong></td>
<td>1353.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>BTC VADER</strong></td>
<td>1353.00</td>
<td>0.16</td>
<td>0.06</td>
<td>-0.11</td>
<td>0.12</td>
<td>0.16</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>BTC Count</strong></td>
<td>1353.00</td>
<td>369.40</td>
<td>339.51</td>
<td>0.00</td>
<td>198.00</td>
<td>263.00</td>
<td>406.00</td>
<td>4012.00</td>
</tr>
<tr>
<td><strong>ETH Return</strong></td>
<td>1353.00</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.59</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.03</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>ETH Volatility</strong></td>
<td>1353.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>ETH VADER</strong></td>
<td>1353.00</td>
<td>0.24</td>
<td>0.09</td>
<td>-0.13</td>
<td>0.17</td>
<td>0.24</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>ETH Count</strong></td>
<td>1353.00</td>
<td>88.73</td>
<td>49.92</td>
<td>0.00</td>
<td>53.00</td>
<td>74.00</td>
<td>120.00</td>
<td>310.00</td>
</tr>
<tr>
<td><strong>LTC Return</strong></td>
<td>1235.00</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.59</td>
<td>-0.03</td>
<td>0.00</td>
<td>0.03</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>LTC Volatility</strong></td>
<td>1235.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>LTC VADER</strong></td>
<td>1235.00</td>
<td>0.27</td>
<td>0.09</td>
<td>-0.13</td>
<td>0.17</td>
<td>0.24</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>LTC Count</strong></td>
<td>1235.00</td>
<td>30.47</td>
<td>70.54</td>
<td>0.00</td>
<td>53.00</td>
<td>74.00</td>
<td>120.00</td>
<td>310.00</td>
</tr>
<tr>
<td><strong>XRP Return</strong></td>
<td>1093.00</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.59</td>
<td>-0.03</td>
<td>0.00</td>
<td>0.03</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>XRP Volatility</strong></td>
<td>1093.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>XRP VADER</strong></td>
<td>1093.00</td>
<td>0.27</td>
<td>0.09</td>
<td>-0.13</td>
<td>0.17</td>
<td>0.24</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>XRP Count</strong></td>
<td>1093.00</td>
<td>26.93</td>
<td>90.13</td>
<td>0.00</td>
<td>53.00</td>
<td>74.00</td>
<td>120.00</td>
<td>310.00</td>
</tr>
<tr>
<td><strong>DOGE Return</strong></td>
<td>666.00</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.59</td>
<td>-0.03</td>
<td>0.00</td>
<td>0.03</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>DOGE Volatility</strong></td>
<td>666.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>DOGE VADER</strong></td>
<td>666.00</td>
<td>0.27</td>
<td>0.09</td>
<td>-0.13</td>
<td>0.17</td>
<td>0.24</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>DOGE Count</strong></td>
<td>666.00</td>
<td>685.49</td>
<td>3030.26</td>
<td>0.00</td>
<td>53.00</td>
<td>74.00</td>
<td>120.00</td>
<td>310.00</td>
</tr>
</tbody>
</table>

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).
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.