Using Deep Pre-Trained Language Models to Understand Investor Sentiment and Volatility for Cryptocurrencies

Philip Spalding
Rongbing Liang
Yupeng Liu

October, 2021

Abstract

Researches have developed multiple ways to quantify investor sentiment, from simple proxies to more refined survey based methods. The prevalence of social media data, however, presents a unique opportunity to observe sentiment in almost real time. We use Twitter data and RoBERTA, a popular deep pre-trained language model, to quantify investor sentiment for cryptocurrencies. We create sentiment indices that can be updated in close to real time, and evaluate their efficacy in explaining daily realized volatility for Bitcoin and Ether, two of the most popular cryptocurrencies in the market. We find that our sentiment indices have a significant negative relationship with volatility and that the RoBERTa model is highly effective in quantifying the sentiment of cryptocurrency related tweets without further domain specific fine tuning.

1. Introduction

Cryptocurrencies such as Bitcoin (BTC) and Ethereum[1] (ETH) have rapidly risen in popularity and are quickly becoming a mainstream asset class. Unlike more traditional assets, however, the underpinning of their valuations is hotly debated with some hoping that cryptocurrencies will supplant fiat currencies and others claiming they are effectively worthless. While the long term implications of cryptocurrencies is still far from clear, understanding their current dynamics is of vital importance for investors interested in gaining exposure to the crypto market. As a relatively new asset class, though, cryptocurrencies exhibit significantly higher volatility than many other asset classes, and understanding the drivers of such volatility should be of great importance to investors.

Unlike traditional assets, such as equities, cryptocurrencies do not have extensively studied factors that can help explain their returns. In fact, we see many cases where news and investor sentiment seems to completely drive the market. Elon Musk's tweets in particular have the tendency to move BTC and dogecoin, an alt coin originally created as a joke, prices dramatically[2]. Further,

[1] Technically, Ether is the main token on the Ethereum blockchain, but the the terms are usually used interchangeably when referring to the cryptocurrency.

[2] An article in Vox lays out an interesting timeline of Elon Musk's tweets and subsequent BTC returns: https://www.vox.com/recode/2021/5/18/22441831/elon-musk-bitcoin-dogecoin-crypto-prices-tesla

social media has recently shown to have a sizable impact on markets, including more established markets. In early 2021, the subreddit r/WallStreetBets triggered short squeezes in GameStop and AMC that sent prices skyrocketing and resulted in huge losses for many hedge funds[3]. Clearly, investors cannot ignore social media.

This paper examines the relationship between social media sentiment and BTC and ETH volatility. We use data from Twitter along with a deep language representation model, RoBERTa, to quantify Twitter sentiment and create sentiment indices. We then show that these sentiment indices explain some of the dramatic fluctuations in daily realized volatility, and we benchmark our models against traditional time series models. Section 2 reviews other research related to cryptocurrencies and deep language representation models. Section 3 discusses the data used in our empirical analyses, including the Twitter data collected. Section 4 discusses the construction of our sentiment indices and the RoBERTa model. Section 5 discusses our empirical findings and illustrates the efficacy of the sentiment indices in explaining daily realized volatility. Section 6 explores how sentiment evolved during the COVID-19 crash. Section 7 concludes and discusses areas for future research.

2. Background and Related Work

Market sentiment, the general prevailing attitude of investors, can be monitored in a number of ways. The VIX, often referred to as the "fear index," is commonly considered a proxy for investor sentiment, hinting at the close relationship between sentiment and volatility. The University of Michigan maintains a widely followed survey based consumer sentiment index[4]. We use deep natural language processing (NLP) models, which have exploded in popularity, in order to quantify cryptocurrency sentiment using Twitter data. While there are now a multitude of models, many of the most popular approaches utilize the transformer architecture originally proposed by (Vaswani et al., 2017). BERT (Devlin, Chang, Lee, & Toutanova, 2019) is a pre-trained language representation model (PLM) that uses a transformer network architecture and a masked language model pre-training objective. The pre-trained representation alleviates the need for heavily-engineered, task specific architectures, and the model can be fine tuned for a variety of different downstream tasks. PLMs received huge attention after BERT achieved state-of-the-art results on 11 NLP tasks and have been used in a variety of different disciplines. (Elwany, Moore, & Oberoi, 2019) fine tune a BERT model for classification tasks in the legal domain. (Araci, 2019) introduce FinBERT for tasks in the financial domain, obtaining state-of-the art results for two financial sentiment analysis datasets. (Liu et al., 2019) expand on BERT, finding that the original model was "significantly undertrained." The authors propose RoBERTa - which incorporates several novel improvements, including training the model longer, with bigger batches, over more data and training on longer sequences - that exceeds the performance of earlier post-BERT methods.

For our analyses, we use the Python Transformers library. Transformers is an open-source library with the goal of making these high capacity, PLMs available to the wider machine learning community (Wolf et al., 2020). The library has over 32 pre-trained models in 100 plus languages as well as a community driven model repository that includes thousands of implementations.

While many studies explore cryptocurrency returns, primarily BTC as it is by far the dominant coin in the market, other research focuses on volatility. (Dyhrberg, 2016) finds that BTC shares

[3]https://fortune.com/2021/01/29/gamestop-stock-how-much-hedge-funds-have-lost-sellers-losses-gme-steve-cohen-point72-andrew-left-citron-research-short-squeeze/

[4]http://www.sca.isr.umich.edu/

several characteristics with gold and the dollar and that it may be useful in risk management and ideal for risk averse investors. (Klein, Pham Thu, & Walther, 2018) challenge this suggestion, finding that the only similarity between gold and BTC is its asymmetric response in variance. (Conrad, Custovic, & Ghysels, 2018) find that S&P 500 realized volatility has a negative effect on long-term BTC volatility, an atypical finding when compared to other financial markets. While these results suggest that BTC may play an important role in a well diversified portfolio, investors must also consider BTC's high volatility compared to other assets, as shown in figure 1. In a paper more similar to ours, (Al-Khazali, Elie, Roubaud, et al., 2018) examine the asymmetric reaction of BTC volatility to positive and negative macroeconomic news surprises in several developed countries, including the U.S., Euro area, and Japan.

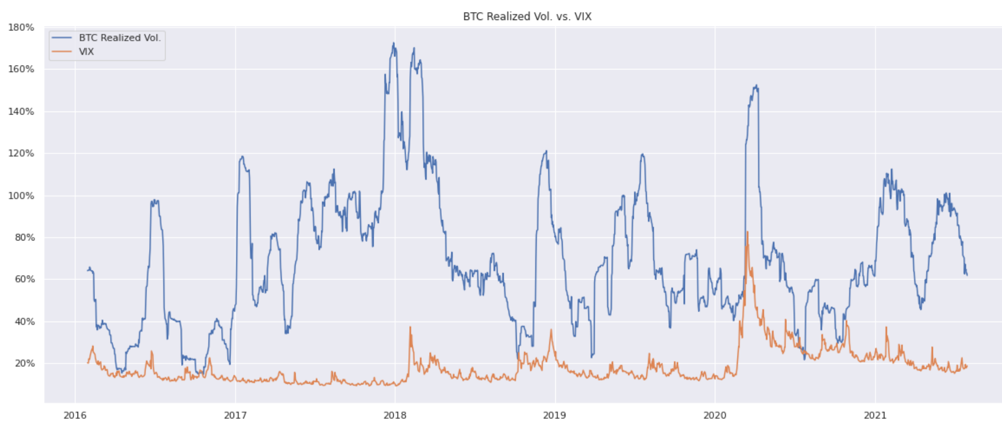

Figure 1: BTC annualized realized volatility over the preceding 30 days compared to the Chicago Board Options Exchange's volatility index, the VIX, which estimates the expected annualized change in the S&P 500 index over the following 30 days. Note the difference in scale.

While these and other studies primarily focus on historical realized volatility, (Woebbeking, 2021) uses options to construct a cryptocurrency volatility index that captures the market's expectation of future volatility, similar to the VIX. Given the relatively short period for which implied volatility estimates are available, we focus on daily realized volatility herein.

3. Data

While there are many social media networks, we elect to use Twitter data to measure cryptocurrency sentiment. There are over 500 million "tweets" per day on Twitter and, as of 2019, 22% of adults have used the platform[5]. In addition to everyday users, many politicians, celebrities, and domain experts use Twitter; as such, the messages included on the platform come from a broad and diverse population. Twitter data comes with two highly desirable features as well: (1) each tweet has a timestamp allowing us to precisely determine when the message was posted; and (2) many tweets contain "hashtags" that link similar tweets together, allowing users to find tweets with similar topics. Twitter data has been used in numerous studies, and there is even a Twitter based uncertainty index[6], developed by (Baker, Bloom, Davis, & Kost, 2019), that is highly correlated with the VIX.

[5]https://pewrsr.ch/2VxJuJ3
[6]http://www.policyuncertainty.com/twitter_uncert.html

The volume of cryptocurrency related tweets has grown enormously as BTC and other cryptocurrencies have grown in popularity. Figure 2 presents the daily volume of tweets related to BTC, ETH, and cryptocurrencies in general[7]. Collecting all of the relevant tweets for the time period we examine was not possible due to Twitter API limitations. Therefore, we took two different approaches to collecting relevant tweets.

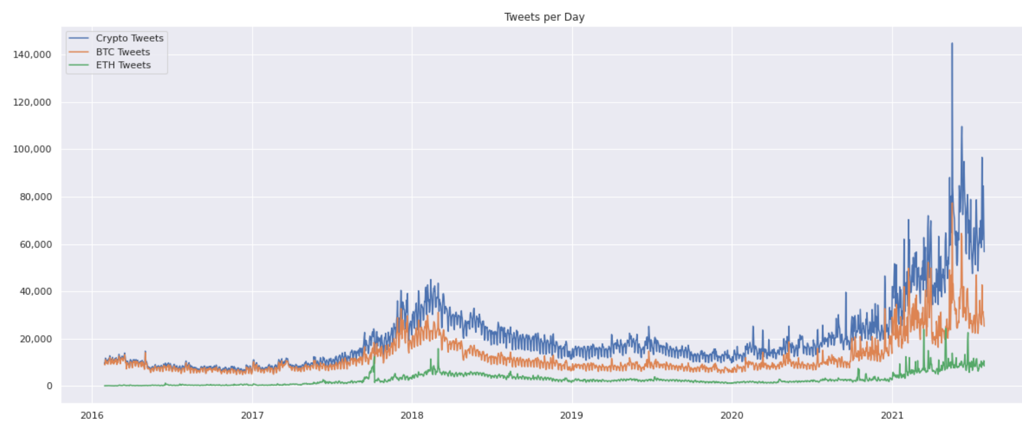

Figure 2: Daily tweet volume for BTC, ETH, and cryptocurrencies in general.

First, we identified 100 top "crypto influencers" and obtained all of their tweets between 2016 and July 2021[8]. It is reasonable to believe that these influencers have a larger impact on people's decision making given their prominence in the crypto community, and many of these accounts have large numbers of followers, making their messages more likely to be seen by the general public. The majority of the tweets from these accounts are solely related to cryptocurrencies or blockchain technology and many authors do not use hashtags, so we do not apply any hashtag or keyword filters for these tweets. We collected 847,540 tweets from these accounts. The volume of tweets from these accounts follows a trend similar to the total volume of cryptocurrency related tweets in figure 2. We refer to this dataset as the "Top 100" data.

While our first dataset includes many prominent voices in the crypto community, we also wanted to quantify the sentiment of the general public. In order to address the volume of cryptocurrency related tweets, we sampled approximately 2,400 tweets per day from Twitter. Unfortunately, the Twitter API does not have a way of randomly sampling tweets. To overcome this limitation, we queried 100 tweets during each hour of each day that included a cryptocurrency related hashtag[9]. The Twitter API returns tweets in reverse chronological order, so this approach results in the query returning the last 100 tweets for each hour queried. While this does not result in a completely random sample of tweets, spreading the queries out over the course of the day results in a more uniformly generated sample than querying 2,400 tweets at a single point in time. We collected a total of 4,829,264 tweets using this approach; due to Twitter API limits, it took approximately two days to collect the data. We refer to this dataset as the "Random Tweet" data.

[7]Includes all tweets with the following hashtags: #BTC, #Bitcoin, #ETH, #Ethereum, and #crypto. Note that hashtags are not case sensitive.

[8]Our list was composed of Twitter accounts originally identified by www.openbusinesscouncil.org/top-200-blockchain-influencers-authors and https://blockinfluence.com/our-view/the-top-100-crypto-and-blockchain-influencers-on-twitter.

[9]We utilized the following hashtags: #BTC, #Bitcoin, #ETH, #Ethereum, and #crypto.

Twitter data is highly unstructured, and, with millions of tweets, understanding the content of the data is non-trivial. We present two word clouds, one for the Top 100 dataset and one for the Random Tweet dataset, in figure 3. The word clouds present the most common words in each dataset, with the size of the word proportional to the frequency of the word's occurrence[10]. Bitcoin, crypto, and blockchain are among the most common words in both datasets. While the Random Tweet data includes notable mentions of Ethereum, the Top 100 dataset does not; even "AI" is more prominent in the Top 100 data than Ethereum. The word clouds make it clear that both datasets are primarily related to cryptocurrencies as desired. Appendix A includes further details on the composition and characteristics of both datasets.

Figure 3: Word clouds for the Top 100 (left) and Random Tweet (right) datasets.

We obtained daily price data for BTC and ETH from CoinDesk. We also obtained minute level price data for BTC and ETH from the Gemini Exchange[11]. Unlike other papers that estimate daily volatility using GARCH models and daily returns, we calculate daily realized volatility using the minute level price data and the two-scale realized volatility (TSRV) estimator proposed by (Zhang, Mykland, & Ait-Sahalia, 2003), which adjusts for microstructure noise in high frequency data. We use 1-minute and 5-minute log returns to calculate the TSRV estimator. Figure 4 shows the daily realized volatilities calculated from the minute level BTC and ETH price data.

4. Sentiment Indices

Constructing the sentiment indices from Twitter data is two step process: first, we score each tweet using a RoBERTa model fine tuned for sentiment analysis; then, we accumulate the scores of the tweets as an exponentially weighted moving average (EWMA) to determine the value of the sentiment index at any given point in time. The RoBERTa model, like the BERT model, uses a multi-layer bidirectional Transformer encoder[12] architecture, which was originally introduced by (Vaswani et al., 2017). The Transformer architecture is ubiquitous in the deep learning community, and has been applied to multiple tasks, from natural language processing to deep reinforcement learning[13]. The Transformer encoder consists of L layers. Each layer is composed of two sub-layers:

[10] We exclude common stop words (e.g., "the", "and","is", etc.) and some common Twitter specific terms such as account names and hyperlinks. All words are also lower cased.

[11] Data was obtained from https://www.cryptodatadownload.com/data/gemini/.

[12] The Transformer in (Vaswani et al., 2017) uses an encoder-decoder structure for a translation task while BERT and RoBERTa only use a Transformer encoder.

[13] Deepmind's AlphaStar, which learned how to play StarCraft II, uses a Transformer as part of its larger architecture: https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

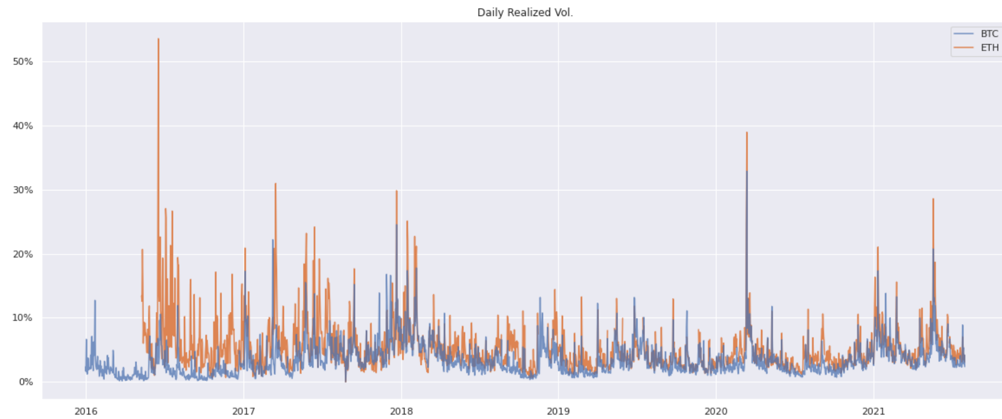

Figure 4: Daily realized volatility for BTC and ETH calculated using the TSRV estimator.

(1) a multi-head self-attention mechanism, with A self-attention heads, and (2) a feed forward layer of size H. A residual connection followed by layer normalization is included around both sub-layers. The RoBERTa model we use has $L = 12$ layers; each block has $A = 12$ self-attention heads and a hidden layer of size $H = 768$. We use a RoBERTa model pre-trained on over 55 million tweets and fine tuned for sentiment analysis on the TweetEval sentiment benchmark dataset (Rosenthal, Farra, & Nakov, 2017) (Barbieri, Camacho-Collados, Espinosa-Anke, & Neves, 2020); we did not perform additional fine tuning. The model can be found on the Transformers model repository[14]. Using a model trained on Twitter data ensures that the underlying language representation includes properties unique to Twitter data, such as the prevalence of account mentions and hyperlinks and the 140/280 character limit for tweets.

For each tweet, the RoBERTa model predicts positive, negative, or neutral sentiment. The head of the model is a softmax layer that outputs the model estimated probabilities of each sentiment classification. A final score is determined by taking the probability weighted average of the sentiment classifications where positive sentiment is +1, negative sentiment is -1, and neutral sentiment is 0. We also experimented with binary sentiment models that only predict positive or negative sentiment. However, because many classification models become highly "confident" in their predictions, with the probabilities of the output classification being very close to 1.0., these models rarely resulted in neutral sentiment measures. The approach used results in sentiment scores that are more evenly distributed between -1 and +1 rather than concentrated around -1 and +1. Appendix B includes examples of positive, negative, and neutral sentiment tweets, as scored by the RoBERTa model using the methodology outlined.

Once each tweet has a score, we construct the sentiment index as an EWMA:

$$x_t = \alpha s_t + (1 - \alpha) x_{t-1}$$

where x_t is the index value at time t, s_t is the score for the tweet that occurs at time t, and x_{t-1} is the previous value of the index. The index is updated each time a new tweet is observed in the data based on the tweet's timestamp. In practice, we could query tweets and update the index in an online fashion, allowing for close to real time measures of sentiment. The value of α

[14] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

is determined via a grid search. The dataset is split between a training set, comprising all data between 2016 and July 2020, and test set, comprising all data between August 2020 and July 2021. For each value of α in the grid, a sentiment index is calculated using the training set. Because tweets come in throughout the day, we resample the resulting sentiment index to a daily frequency by finding the last value for each day[15]. Daily volatility during the training period is then regressed on the calculated sentiment index, and the value of α that results in the highest R^2 is selected. The optimal value of α for the Top 100 dataset is $\alpha = 0.0006$, and the optimal value of α for the Random Tweet dataset is $\alpha = 0.0015$. We refer to the resulting sentiment indices as the "Sent100 Index" and "RandSent Index", respectively.

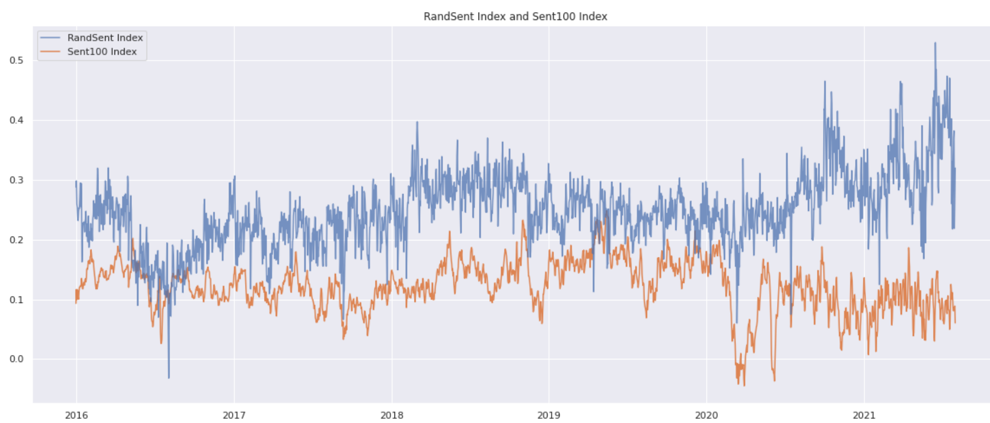

Figure 5: RandSent Index and Sent100 index.

Figure 5 illustrates the sentiment indices over both the training and test set periods. Interestingly, the value of the RandSent index is consistently higher than the value of the Sent100 index, suggesting that ordinary crypto investors/followers are more optimistic than the top 100 influences selected. Further, the correlation between the two indices is quite low, only 0.04, suggesting that the two indices contain different information. Both series also exhibit significant autocorrelation, though this is expected given their construction as EWMAs. Table 1 presents summary statistics for both indices over the 2016 through July 2021 period.

	RandSent Index	**Sent100 Index**
Mean	0.2478	0.1222
Median	0.2462	0.1238
Std. Dev.	0.0614	0.0424
Min	-0.0319	-0.0448
Max	0.5296	0.2482
Skew	0.3457	-0.5303

Table 1: Descriptive statistics of sentiment indices.

Figure 6 plots the distribution of sentiment scores for tweets in the Random Tweet dataset

[15]BTC and ETH trade throughout the day so both the sentiment index and the daily volatility values are as of midnight UTC.

on three different days for illustrative purposes. July 31, 2019 represents a standard day. The RandSent Index was relatively unchanged at 0.24 and neither BTC nor ETH prices experienced significant volatility. The distribution of sentiment scores is centered around 0, indicating that most tweets are sentiment neutral. The distribution is also skewed towards +1, as, in general, positive tweets outnumber negative tweets, explaining why the mean of the index is approximately 0.24, not 0.00.

On March 12, 2020, also known as "Black Thursday," U.S. stocks experienced the largest single-day percentage fall since the 1987 stock market crash as part of the market's response to the COVID-19 pandemic. BTC and ETH also experienced double digit losses over 25%, and the RandSent Index decreased from 0.16 to 0.06, its lowest point in multiple years. Surprisingly, the mean sentiment score for the day is still positive, approximately 0.07; however, there are far more negative tweets than on other, more standard days. Table 10 below includes 10 randomly selected negative tweets from the day; there are notable mentions of the market crash and COVID-19.

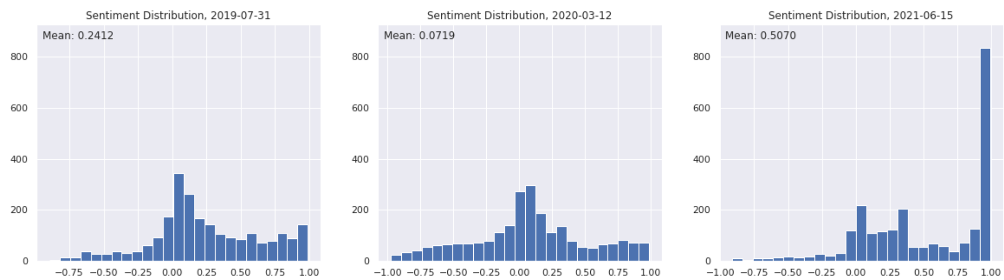

Figure 6: Distribution of sentiment scores for tweets in the Random Tweet dataset on three different days: July 31, 2019 (left), March 12, 2020 (middle), and June 15, 2021 (right).

By June 15, 2021, BTC and ETH had partially recovered from the May 2021 crash, and BTC had again breached the $40K level. The RandSent Index increased from 0.39 to 0.52. The tweets for the day are notably positive, with almost one-third having a score of +1; however, this is one of the most positive days in the dataset.

5. Empirical Results

In order to evaluate the efficacy of our crypto sentiment indices, we regress realized volatility on the sentiment indices along with a variety of different control variables. The regression model takes the form:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{c} \beta_j x_{j,t} + \epsilon_t$$

where y_t is the volatility at time t, ϕ_i are the autoregressive (AR) coefficients for the lagged values of y. x_j and β_j are the exogenous variables of interest, including the sentiment index values, and other control variables and their corresponding coefficients, respectively. We follow the same procedure as outlined above to resample the sentiment indices to a daily frequency. Like many financial instruments, BTC and ETH volatility exhibit high autocorrelation; given this persistence, we include lagged values of realized volatility in all regressions. All regressions are performed on the training set, which spans the 2016 through July 2020 period, and we evaluate each model's out-of-sample performance by freezing the model's coefficients and calculating the root mean squared error (RMSE) between the model's predicted values and the actual realized volatility for the test set, which spans the August 2020 to July 2021 period. All independent variables have been standardized to have mean 0 and standard deviation 1, so that regression coefficients can be interpreted as the sensitivity to a one standard deviation change in the variable value. We use White's heteroskedasticity-robust standard errors to calculate all t-statistics in the tables below. Table 2 presents the results for the RandSent Index and BTC realized volatility.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Intercept	0.0119***	0.0110***	0.0159***	0.0160***	0.0158***	0.0153***
	(7.6532)	(7.2269)	(10.1235)	(10.2398)	(9.9729)	(9.8650)
Realized Vol.L1	0.6421***	0.6700***	0.5215***	0.5197***	0.5240***	0.5404***
	(12.5938)	(13.2163)	(10.5778)	(10.6413)	(10.4898)	(10.9920)
RandSent Index	−0.0028***		−0.0073***	−0.0073***	−0.0069***	−0.0060***
	(−3.9248)		(−7.3340)	(−7.4213)	(−7.4745)	(−6.5498)
ΔRandSent Index		−0.0049***				−0.0018***
		(−6.4236)				(−3.0959)
Log(Crypto Tweets)			0.0092***	0.0091***	0.0089***	0.0083***
			(9.0904)	(9.0517)	(8.9195)	(8.2059)
VIX				0.0006		0.0006
				(0.6844)		(0.6662)
BTC Return					−0.0016	−0.0012
					(−1.4478)	(−1.1471)
BIC	−7.7237	−7.7601	−7.8549	−7.8516	−7.8567	−7.8556

Table 2: Regression results for the RandSent Index. The independent variable for all regressions is BTC realized volatility. All dependent variables are standardized to have mean 0 and standard deviation 1 for interpretability. t-statistics are shown below coefficient values. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. The BIC for each model is presented in the last row of the table.

Both the RandSent Index level and first difference coefficients are negative and significant, though relatively small, when we only include these variables as exogenous regressors in the models. However, when we add the log of total daily crypto tweet counts (see figure 2 above), the magnitude of the coefficient for the RandSent Index level increases from -0.0028 to -0.0073; a one standard deviation - approximately 0.05 for the training period - decrease in sentiment level results in a 0.0073 increase in volatility per model (3). Tweet count is a suppressor variable: a variable that increases the magnitude of the regression coefficient of another variable (or set of variables) by its inclusion in the regression equation (MacKinnon, Krull, & Lockwood, 2000). Sentiment and realized volatility are negatively correlated while tweet counts and realized volatility are positively correlated. Further, sentiment and tweet counts are positively correlated, so excluding tweet counts in the model introduces an omitted variable bias and the model under estimates the magnitude of the coefficient for the RandSent Index level. To avoid this bias, both the RandSent Index level and the log of total daily crypto tweet counts are included as regressors in the subsequent regression models.

We add the VIX as a control variable to the regression equation in model (4) in order to understand how equity market volatility may effect BTC volatility. (Conrad et al., 2018) find a negative relationship between realized S&P 500 volatility and long-term BTC volatility. We do not find a significant relationship between daily BTC realized volatility and the VIX, and the inclusion of the VIX as a regressor does not have any effect on the coefficients of the other independent variables.

A natural question to ask is whether sentiment is simply a proxy for returns. It is easy to imagine cases where positive (negative) sentiment follows positive (negative) returns and other cases where positive (negative) returns follow positive (negative) sentiment. In practice, both cases are likely to occur. We add the daily BTC return as a control variable to the regression equation in model (5). We do not find a significant relationship between returns and realized volatility, and the estimated coefficient is smaller than the coefficients for the other covariates. Most notably, the inclusion of returns as a regressor only has a minor effect on the coefficients of the other independent variables, indicating that sentiment contains additional information relevant to volatility that is not included in returns.

Table 3 presents the results for the Sent100 Index and BTC realized volatility. The results are largely similar to the RandSent Index regressions, but the strength of relationship between sentiment and realized volatility is smaller. We, again, find that excluding crypto related tweet counts as a regressor introduces an omitted variable bias that suppresses the negative relationship between sentiment and realized volatility.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Intercept	0.0123***	0.0117***	0.0148***	0.0148***	0.0148***	0.0148***
	(7.9083)	(7.3874)	(9.0839)	(9.0523)	(9.0046)	(8.9334)
Realized Vol.L1	0.6298***	0.6487***	0.5551***	0.5542***	0.5550***	0.5545***
	(12.3546)	(12.4298)	(10.6704)	(10.5885)	(10.5437)	(10.4288)
Sent100 Index	−0.0023***		−0.0031***	−0.0034***	−0.0030***	−0.0032***
	(−3.3666)		(−4.4062)	(−5.4594)	(−4.5769)	(−5.2629)
ΔSent100 Index		−0.0011*				−0.0008
		(−1.9065)				(−1.4172)
Log(Crypto Tweets)			0.0057***	0.0058***	0.0056***	0.0057***
			(7.6698)	(7.4666)	(7.5592)	(7.3146)
VIX				−0.0007		−0.0007
				(−0.6143)		(−0.7088)
BTC Return					−0.0026**	−0.0026**
					(−2.2303)	(−2.2009)
BIC	−7.7173	−7.7083	−7.7794	−7.7758	−7.7920	−7.7857

Table 3: Regression results for the Sent100 Index. The independent variable for all regressions is BTC realized volatility. All dependent variables are standardized to have mean 0 and standard deviation 1 for interpretability. t-statistics are shown below coefficient values. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. The BIC for each model is presented in the last row of the table.

We experimented with various other covariates and regression specifications over the course of our research, including how our sentiment indices correlate with returns. We regressed daily BTC returns on the RandSent Index level, but did not find a significant relationship. However, regressing daily returns on the first difference of the index results in a coefficient of 0.009 (t-statistic of 8.7930), and an R^2 of 0.06, suggesting that Twitter sentiment also has a significant positive relationship with returns. However, we leave further examination of this as an area for future research.

The α values (for the EWMA) for the sentiment indices are selected to maximize the magnitude of the correlation between the index's level and the realized volatility over the training set period. In order to evaluate the indices' and model's efficacy out-of-sample, we calculate the model's RMSE error for the 1-year test period, which spans August 2020 through July 2021, and compare this RMSE to two baseline models. The baseline models selected are AR(1) and AR(4) models. The AR(1) model was selected as a simple baseline. The AR(4) model was selected as it produced the smallest out-of-sample RMSE of all the various lagged values considered on the validation set[16]. We

[16]The validation set is a subset of the training set, including all data between August 2019 and July 2020. Models

also look at AR(4) models that include the same regressors as the AR(1) models presented in Tables 2 and 3 above. The coefficients for the exogenous variables for the AR(4) models are effectively unchanged from the values above, so we do not present additional regression results tables. Tables 4 and 5 present the in-sample and out-of-sample RMSE for the two sentiment indices and for the AR(1) and AR(4) baselines.

	RandSent		**Sent100**	
	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
(1)	0.0208	0.0193	0.0209	**0.0184**
(2)	0.0204	0.0184	0.0210	0.0186
(3)	0.0194	0.0164	0.0202	0.0210
(4)	0.0194	0.0164	0.0202	0.0210
(5)	0.0194	**0.0161**	**0.0200**	0.0203
(6)	**0.0193**	0.0164	**0.0200**	0.0204
AR(1)	0.0210	0.0187	0.0210	0.0187

Table 4: In-sample and out-of-sample RMSE for each BTC volatility model along with the AR(1) baseline. Models (1) through (6) correspond to models (1) through (6) in tables 2 and 3 above.

	RandSent		**Sent100**	
	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
(1)	0.0201	0.0183	0.0203	**0.0177**
(2)	0.0198	0.0175	0.0203	**0.0177**
(3)	0.0190	0.0158	0.0198	0.0196
(4)	0.0190	0.0158	0.0198	0.0197
(5)	**0.0189**	**0.0155**	**0.0196**	0.0189
(6)	**0.0189**	0.0157	**0.0196**	0.0190
AR(4)	0.0203	0.0178	0.0203	0.0178

Table 5: In-sample and out-of-sample RMSE for each BTC volatility model along with the AR(4) baseline. Models (1) through (6) correspond to models (1) through (6) in tables 2 and 3 above with 4 lagged values as regressors instead of 1.

The RandSent Index performs notably better than the Sent100 Index out-of-sample. The RandSent Index's RMSE for model (5), which has both the lowest BIC and performs the best out-of-sample, is 13% and 14% lower than the corresponding AR(1) and AR(4) baselines, respectively. The Sent100 Index's best out-of-sample RMSE, however, is only 2% and 1% lower than the corresponding AR(1) and AR(4) baselines, respectively, and the models with the lowest RMSE in-sample do not perform better than the baselines out-of-sample. The RandSent Index is designed to capture the overall sentiment of crypto investors/followers whereas the Sent100 Index is focused on a more curated sample of highly visible influencers. The RandSent Index's improvement over the Sent100 Index suggests that the sentiment of the broader investor base has a larger impact on volatility than the sentiment of individuals that are prominent in the crypto space. However, additional research is needed to fully understand this distinction as there may be other highly influential individuals that were either not included in our Top 100 dataset or that simply do use Twitter as a means of communication. Our results do support the hypothesis that volatility is, in part, driven by social media sentiment, which can itself be quite volatile.

Tables 6 and 7 present our regression results for ETH for the RandSent Index and Sent100 Index, respectively. The results are directionally the same: there is a significant negative relationship

were fit on the training set data through July 2019 and evaluated using the validation set. The selected models - the AR(1) and AR(4) models - were then refit on the entire training dataset.

between sentiment and volatility, the RandSent Index's coefficient is larger in magnitude than the Sent100 Index, and we do not see a significant relationship between ETH volatility and the VIX. There are also notable differences, though. While excluding crypto related tweet counts as a regressor for the RandSent Index models introduces an omitted variable bias that suppresses the negative relationship between sentiment and realized volatility, we do not see this phenomenon for the Sent100 Index and ETH volatility. Interestingly, the coefficient for the RandSent Index is larger in magnitude, approximately -0.01, than the coefficient for the BTC volatility models, suggesting that there is a stronger relationship between the level of the RandSent Index and ETH volatility than BTC volatility. Further, we see a significant negative relationship between ETH returns and ETH volatility; while this is present for BTC as well, the relationship is not always statistically significant.

	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.0246***	0.0213***	0.0248***	0.0248***	0.0246***	0.0243***
	(7.6632)	(6.6317)	(7.8141)	(7.8361)	(7.6972)	(7.4986)
Realized Vol.L1	0.5386***	0.6025***	0.5269***	0.5269***	0.5321***	0.5378***
	(8.3421)	(9.1432)	(8.1895)	(8.1926)	(8.2010)	(8.1288)
RandSent Index	−0.0073***		−0.0102***	−0.0102***	−0.0093***	−0.0088***
	(−6.6064)		(−6.9226)	(−6.9658)	(−6.6547)	(−5.9738)
ΔRandSent Index		−0.0054***				−0.0012
		(−4.6071)				(−1.0658)
Log(Crypto Tweets)			0.0057***	0.0057***	0.0050***	0.0048***
			(4.2899)	(4.3810)	(3.9266)	(3.6372)
VIX				−0.0003		−0.0004
				(−0.2720)		(−0.3723)
ETH Return					−0.0039***	−0.0037***
					(−3.4086)	(−3.2739)
BIC	−6.8645	−6.8433	−6.8842	−6.8795	−6.8949	−6.8866

Table 6: Regression results for the RandSent Index. The independent variable for all regressions is ETH realized volatility. All dependent variables are standardized to have mean 0 and standard deviation 1 for interpretability. t-statistics are shown below coefficient values. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. The BIC for each model is presented in the last row of the table.

	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.0225***	0.0218***	0.0224***	0.0225***	0.0224***	0.0225***
	(7.0825)	(6.8829)	(7.0623)	(7.0645)	(7.0584)	(7.0426)
Realized Vol.L1	0.5777***	0.5919***	0.5790***	0.5759***	0.5785***	0.5752***
	(8.9032)	(9.1106)	(8.9459)	(8.8115)	(8.9065)	(8.7519)
Sent100 Index	−0.0031***		−0.0032***	−0.0040***	−0.0030***	−0.0038***
	(−3.3061)		(−3.3221)	(−4.1798)	(−3.3150)	(−3.9115)
ΔSent100 Index		−0.0013				−0.0004
		(−1.3797)				(−0.3819)
Log(Crypto Tweets)			0.0010	0.0012	0.0008	0.0011
			(0.9681)	(1.2245)	(0.7643)	(1.0657)
VIX				−0.0017		−0.0019
				(−1.1786)		(−1.3698)
ETH Return					−0.0053***	−0.0053**
					(−4.2261)	(−4.2096)
BIC	−6.8252	−6.8178	−6.8213	−6.8188	−6.8453	−6.8387

Table 7: Regression results for the Sent100 Index. The independent variable for all regressions is ETH realized volatility. All dependent variables are standardized to have mean 0 and standard deviation 1 for interpretability. t-statistics are shown below coefficient values. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively. The BIC for each model is presented in the last row of the table.

Finally, we present the in-sample and out-of-sample RMSE for ETH volatility and the two sentiment indices in tables 8 and 9 for the AR(1) and AR(4) models, respectively. Again, the RandSent Index performs better than the Sent100 Index out-of-sample. The RandSent Index's best out-of-sample RMSE is 16% lower than the AR(1) and AR(4) baselines. The Sent100 Index's best out-of-sample RMSE is 7% and 6% lower than the AR(1) and AR(4) baselines, respectively. Overall, both indices are able to explain some of the variation in daily realized volatility for both BTC and ETH and perform well both in-sample and out-of-sample.

	RandSent		**Sent100**	
	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
(1)	0.0320	0.0244	0.0326	0.0226
(2)	0.0323	0.0217	0.0328	0.0224
(3)	0.0316	0.0202	0.0326	0.0227
(4)	0.0316	0.0202	0.0326	0.0226
(5)	**0.0314**	**0.0190**	0.0322	**0.0209**
(6)	**0.0314**	0.0191	**0.0321**	**0.0209**
AR(1)	0.0328	0.0225	0.0328	0.0225

Table 8: In-sample and out-of-sample RMSE for each ETH volatility model along with the AR(1) baseline. Models (1) through (6) correspond to models (1) through (6) in tables 6 and 7 above.

	RandSent		**Sent100**	
	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
(1)	0.0313	0.0231	0.0317	0.0220
(2)	0.0314	0.0211	0.0318	0.0218
(3)	0.0309	0.0195	0.0317	0.0223
(4)	0.0309	0.0196	0.0317	0.0223
(5)	**0.0306**	**0.0184**	0.0313	**0.0206**
(6)	**0.0306**	**0.0184**	**0.0312**	**0.0206**
AR(4)	0.0318	0.0219	0.0318	0.0219

Table 9: In-sample and out-of-sample RMSE for each ETH volatility model along with the AR(4) baseline. Models (1) through (6) correspond to models (1) through (6) in tables 6 and 7 above with 4 lagged values as regressors instead of 1.

6. Periods of High Volatility

It is common for relationships between financial variables (e.g., prices, interest rates, volatilities, etc.) to change during times of market stress. During periods of high volatility in the stock market, correlations between assets regularly increase as compared to calmer market conditions. In order to understand how the relationship between sentiment and volatility fluctuates over time, we use a Kalman filter to estimate time varying coefficient for the RandSent Index. The state-space representation used is as follows:

$$\begin{array}{lll} \textbf{Measurement Equation:} & y_t = \alpha_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \beta_t x_t + \gamma_t c_t + \epsilon_t & \epsilon_t \sim N(0, \sigma_\epsilon) \\ \textbf{State Equations:} & \alpha_t = \alpha_{t-1} + u_t & u_t \sim N(0, \sigma_u) \\ & \beta_t = \beta_{t-1} + v_t & v_t \sim N(0, \sigma_v) \\ & \gamma_t = \gamma_{t-1} + w_t & w_t \sim N(0, \sigma_w) \end{array}$$

Here, x_t is the level of the RandSent Index at time t and c_t it the log of daily crypto tweet counts; β_t and γ_t are the time varying coefficients for the variables. We use two lagged values of volatility for the state-space representation as it resulted in better convergence of the model; however, the results for β_t are consistent with the model that only included one lagged value of volatility. Figure 7 presents the regression coefficient for the RandSent Index from model (3) in table 2 above, which includes the log of daily crypto tweet counts as a regressor, and the smoothed coefficient estimate from the Kalman filter.

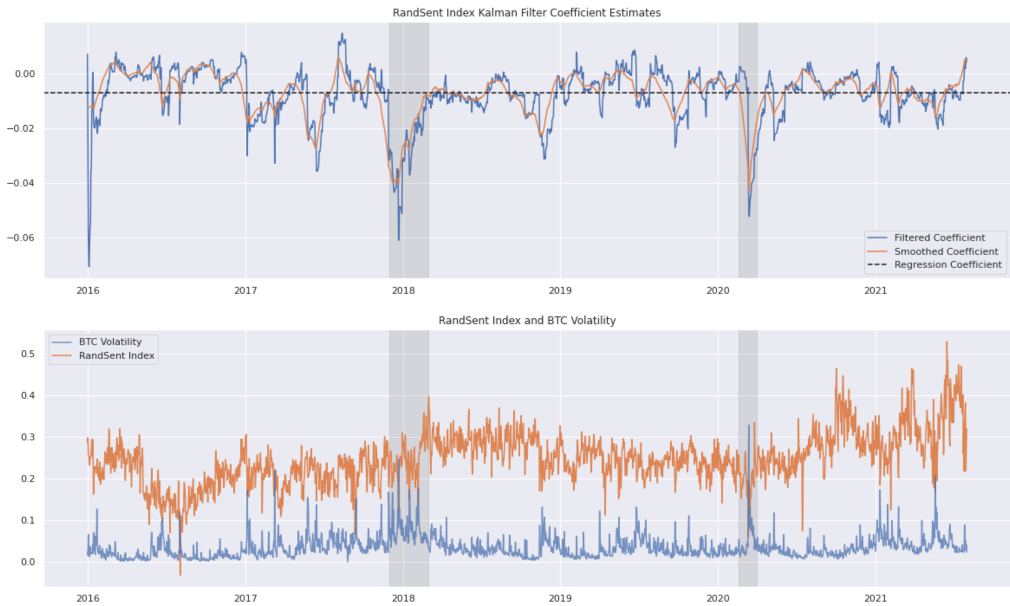

Figure 7: (Top) RandSent regression coefficient estimate and smoothed Kalman filter coefficient estimate. (Bottom) RandSent Index and daily BTC realized volatility. The 2017/2018 crypto crash and 2020 COVID-19 crash are highlighted for reference.

While the Kalman filter estimate fluctuates around the regression coefficient, there are two notable periods where the estimate becomes increasingly negative. These large fluctuations correspond with two periods of significant volatility in the crypto market: the December 2017/January 2018 crypto crash and the broader market crash due to COVID-19 in March 2020. During these periods, the Kalman filter estimate spikes, plummeting to approximately -0.05: a one standard deviation (0.05) decrease in the RandSent Index is predicted to increase volatility 5% compared to the 0.7% predicted by the regression model. This phenomenon suggests that during turbulent market conditions, investor sentiment has a much stronger effect on volatility than during calmer periods. It is likely that high volatility results in lower investor sentiment which, in turn, results in further market volatility. (Schmeling, 2009) finds that the impact of sentiment on stock returns is higher for countries that are culturally more prone to herd-like behavior and overreaction. As of 2021, only about 10% of investors invest in cryptocurrencies[17], and it is possible that this group of investors exhibits such herd-like behavior during times of market stress, amplifying the importance of sentiment. However, more research is required to understand the causal relationship(s) between sentiment and volatility.

Given the almost unprecedented volatility at the beginning of the COVID-19 pandemic, we

[17]https://www.cnbc.com/2021/08/24/1-in-10-people-invest-in-cryptocurrencies-many-for-ease-of-trading.html

examine how the RandSent Index evolved during March 2020 in detail. During this period, investor (and consumer) sentiment in almost all markets fell dramatically, so we can use this period to validate that the RandSent index is quantifying investor sentiment related to cryptocurrencies as expected. Figure 8 plots the RandSent Index, daily BTC realized volatility, and BTC price between December 2019 and May 2020. By March 12, 2020, when BTC prices crash and realized volatility spikes, the RandSent Index decreased to approximately 0.05, down from 0.3 at the beginning of the year and at its lowest point in over 3-years.

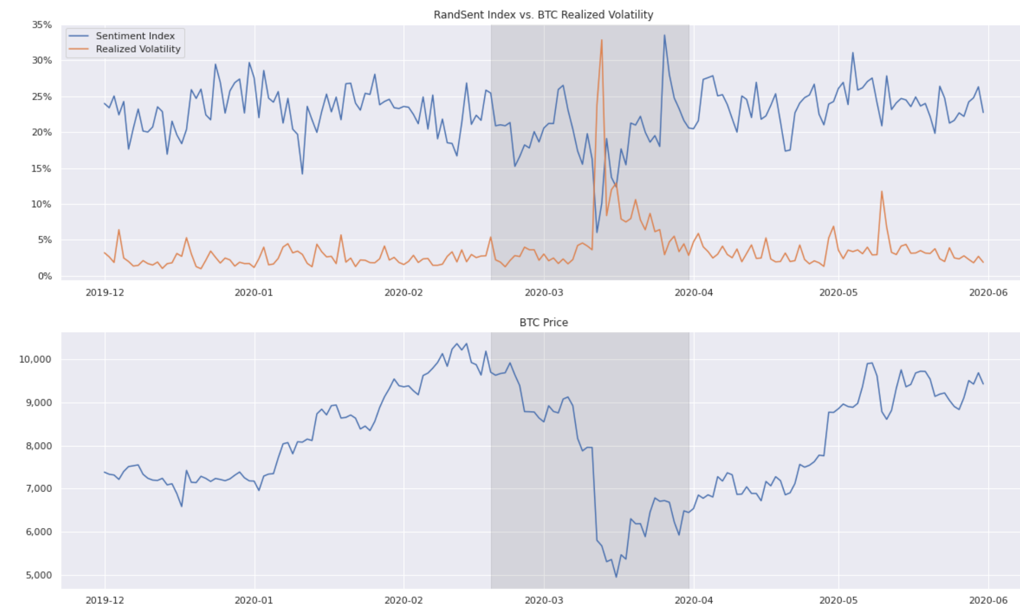

Figure 8: (Top) RandSent Index and daily BTC realized volatility. (Bottom) BTC price. The broad market crash, resulting from the COVID-19 pandemic, between February 19, 2020, and March 31, 2020, is highlighted for reference.

While tweets generally skew positive, we see a notable shift in the daily sentiment distribution on March 12, 2020, with more evenly distributed positive and negative tweets (see figure 6 above). Table 10 presents 10 randomly sampled negative tweets for the day to illustrate the kind of tweets that result in the index's decline. Tweets presented have been processed to exclude account names, hyperlinks, and emojis. Account names and hyperlinks have been replaced with "@user" and "http", respectively, and emojis have been replaced with their description. We have also excluded especially explicit tweets. As expected, the content of tweets is dominated by the effects of COVID-19 on BTC and other financial markets. The Twitter data collected and RoBERTa model used to score the sentiment of tweets effectively quantifies the sentiment of crypto markets during one of the largest market crashes in recent history.

Score	Cleaned Tweet
-0.7811	Bitcoin has proven no haven for investors, plunging 30% from its most recent peak in Febr as global markets were roiled by increasing headwinds, chart @user http
-0.7724	cryptocurrency markets (btc eth) have crashed along with the stock markets. I would have predicted the opposite, that investors would see crypto is a safe haven during a collapse of financial systems. Much like gold. So why did cryptos tank even worse than than dow?
-0.8716	Wrath of Corona, who would have thought that it will have such a devastating effect on world economy. CoronavirusPandemic Covid19 Bitcoin BTC
-0.7737	Nothing prepares you for losing a huge percentage of your net worth in 3 days. Nothing. I've been through a lot and let me tell you, this is nuts. However, still haven't sold any stocks, crypto or Gold/Silver. Maybe too shocked? Maybe 2 ballsy? man_shrugging_medium-light_skin_tone $link $btc $dag Stocks Crypto
-0.7594	Bitcoin plunges 26%, its sharpest selloff in 7 years – MarketWatch http bitcoin
-0.9232	all that low is in dream you was sold by idiots talking in absolutes....shattered! btc
-0.8675	Bitcoin price crashes majorly to fall below $6,000 as coronavirus wreaks havoc on global markets. http Cryptonews
-0.7730	This is the scariest movie that you can't watch in theatres. Only in crypto $BTC BTC Bitcoin http
-0.8021	bitcoin is infected with the Corona virus for GEMINI:BTCUSD by mrbors http http
-0.8067	Bitcoin and the Terrible, Horrible, No Good, Very Bad Day - coronavirus RIP-myportfolio stonks memes steem steemtweets esteem http

Table 10: 10 random negative tweets from March 12, 2020. Tweets are sampled randomly from tweets with scores less than -0.5.

7. Conclusion

Deep NLP models, such as BERT and RoBERTa, are especially suited to extract information from highly unstructured text data. We demonstrate the efficacy of these models in the financial domain by creating two Twitter based sentiment indices, the RandSent Index and Sent100 Index, that explain part of the variation in the daily realized volatility of BTC and ETH returns. We find a significant negative relationship between sentiment and realized volatility that strengthens during times of market turbulence, including during the COVID-19 induced market crash. There are, however, multiple areas to expand on in future research. While the RoBERTa model employed is trained on a significant amount of Twitter data, it is not further fine tuned on cryptocurrency specific data. Fine tuning the sentiment model with cryptocurrency tweets may result in more accurate sentiment scores that incorporate more domain specific information. Additionally, the focus of our paper is on daily realized volatility, but tweets occur throughout the day and there may be additional relationships between sentiment and returns and/or volatility at a higher frequency.

References

Al-Khazali, O., Elie, B., Roubaud, D., et al. (2018). The impact of positive and negative macroeconomic news surprises: Gold versus bitcoin. *Economics Bulletin*, *38*(1), 373–382.

Araci, D. (2019). *Finbert: Financial sentiment analysis with pre-trained language models.*

Baker, S. R., Bloom, N., Davis, S. J., & Kost, K. J. (2019, March). *Policy news and stock market volatility* (Working Paper No. 25720). National Bureau of Economic Research. doi: 10.3386/w25720

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of findings of emnlp.*

Conrad, C., Custovic, A., & Ghysels, E. (2018). Long- and short-term cryptocurrency volatility components: A garch-midas analysis. *Journal of Risk and Financial Management*, *11*(2). doi: 10.3390/jrfm11020023

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Dyhrberg, A. H. (2016). Bitcoin, gold and the dollar – a garch volatility analysis. *Finance Research Letters*, *16*, 85-92. doi: https://doi.org/10.1016/j.frl.2015.10.008

Elwany, E., Moore, D., & Oberoi, G. (2019). *Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding.*

Klein, T., Pham Thu, H., & Walther, T. (2018). Bitcoin is not the new gold – a comparison of volatility, correlation, and portfolio performance. *International Review of Financial Analysis*, *59*, 105-116. doi: https://doi.org/10.1016/j.irfa.2018.07.010

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.*

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention science*, *1*(4), 173–181.

Rosenthal, S., Farra, N., & Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 502–518).

Schmeling, M. (2009). Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance*, *16*(3), 394-408.

Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Woebbeking, F. (2021). Cryptocurrency volatility markets. *Digital Finance*. doi: 10.1007/s42521-

021-00037-3

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics.

Zhang, L., Mykland, P. A., & Ait-Sahalia, Y. (2003, November). *A tale of two time scales: Determining integrated volatility with noisy high frequency data* (Working Paper No. 10111). National Bureau of Economic Research. doi: 10.3386/w10111

Appendix A Twitter Data

The Top 100 dataset includes 847,540 tweets from 100 prominent voices in the crypto community. Unfortunately, other relevant Twitter data for these authors, such as number of followers, average number of likes, etc. are not available on time series basis from the Twitter API, only as of the date queried. We explored the distribution of some of these other useful data points for the 100 authors selected as part of our selection process. Table 11 presents summary statistics related to the tweets of the 100 top crypto influencers selected for inclusion in the Top 100 dataset.

	Num. Tweets	**Followers**	**Retweets**	**Likes Per Tweet**	**Replies**
Mean	8,533	374,618	547	209	19
Median	5,282	192,568	245	72	6
Std. Dev.	9,370	551,710	1,005	381	36
Min	93	485	6	0	0
Max	32,492	3,363,504	7768	2,808	290

Table 11: Summary statistics related to the tweets of the 100 top crypto influencers selected for inclusion in the Top 100 dataset. Likes per tweet represents the average likes per tweet for each author.

On average, these authors have hundreds of thousands of followers, illustrating their prominence in the crypto community. We see significant variation in the number of tweets per author and number of likes per tweet that authors receive. Interestingly, some of the most prominent authors - i.e., those with the most followers - have relatively few tweets that receive large numbers of likes; on the other hand, some of the less influential authors have an disproportional share of the total tweets, but have fewer followers and receive fewer likes. Unsurprisingly, there is a positive relationship between likes and followers - it is easier to get more likes if you already have a large number of people following your tweets.

The Random Tweet dataset contains approximately 2,400 tweets per day, for a total of 4,829,264 tweets between 2016 and July 2021. While the number of tweets per day is relatively constant[18], the number of authors, i.e., unique Twitter accounts posting about cryptocurrencies, grows significantly, from approximately 500 in 2016 to close to 2,000 in 2021, as shown in figure 9. This increase mirrors the growth in total tweet counts in figure 2, and suggests that not only are certain people tweeting more about cryptocurrencies, but the crypto community and those interested in cryptocurrencies is growing rapidly as well.

The Random Tweet dataset was compile by querying tweets with the following hashtags: #BTC, #Bitcoin, #ETH, #Ethereum, and #crypto. Hashtags are not case sensitive, so the "Bitcoin" hashtag also includes all tweets with the "bitcoin", "BITCOIN", etc. hashtags. While approximately 81% of the tweets in the Random Tweet dataset have a BTC related hashtag, only 16% have an Ethereum related hashtag. While most of the authors in the Top 100 dataset do not use hashtags, we see a similar, albeit less dramatic, pattern in that dataset as well: 17% of tweets in the Top 100 dataset have a BTC related hashtag while only 6% of have an Ethereum related hashtag. This is not entirely surprising given BTC's prevalence as the first and most popular cryptocurrency, but it does suggest that (1) far more people tweet about BTC than other coins and (2) our RandSent

[18]There are some days with less than 2,400 tweets because there were less than 100 tweets during certain hours of the day.

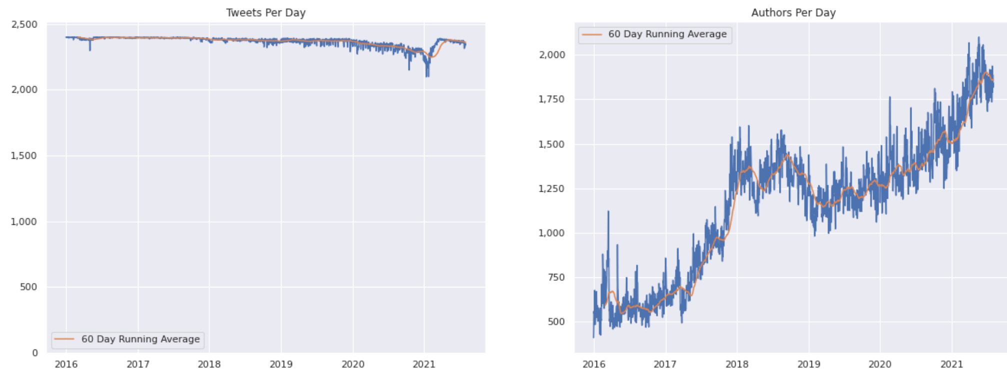

Figure 9: Tweet counts (left) and author counts (right) per day for the Random Tweet dataset.

Index is skewed towards BTC. However, it is notable that there is a significant relationship between the RandSent Index and ETH daily realized volatility, which suggests that many cryptocurrencies are influenced by BTC news and sentiment.

Text data requires various levels of preprocessing before it can be used by machine learning or statistical models. Twitter data has a number of unique features that require Twitter specific preprocessing as well. Figure 10 illustrates the initial cleaning process that we apply to all tweets, and we list the steps in detail below:

1. Retweets, re-postings of tweets that originate from other users, begin with “RT@username” (where username represents the original author’s Twitter handle). We remove all such retweet identifiers as they do not directly convey information about the sentiment of a tweet.

2. Next, the “#” symbol indicates that a word is a hashtag. However, the “#” symbol itself is simply a way to create a hashtag and does not convey additional information, so we remove it from all tweets as well.

3. Many tweets include emojis, but the RoBERTa model we use was not trained with emoji tokens, so the emoji itself is not recognizable to the model. Emojis, however, often convey information that is vital to the sentiment of the tweets. Therefore, rather than simply removing the emojis, we replace them with their descriptive text, which is usable by the RoBERTa model. For example, we replace the rocket emoji (see figure 10), with the text “rocket”.

4. Tweets regularly contain other users’ Twitter handles or links to external sites. The RoBERTa model we used includes two special tokens, “@user” and “http”, which represent Twitter handles and external links, respectively. We replace all specific handles and links with these special token so that they can be used by the RoBERTa model.

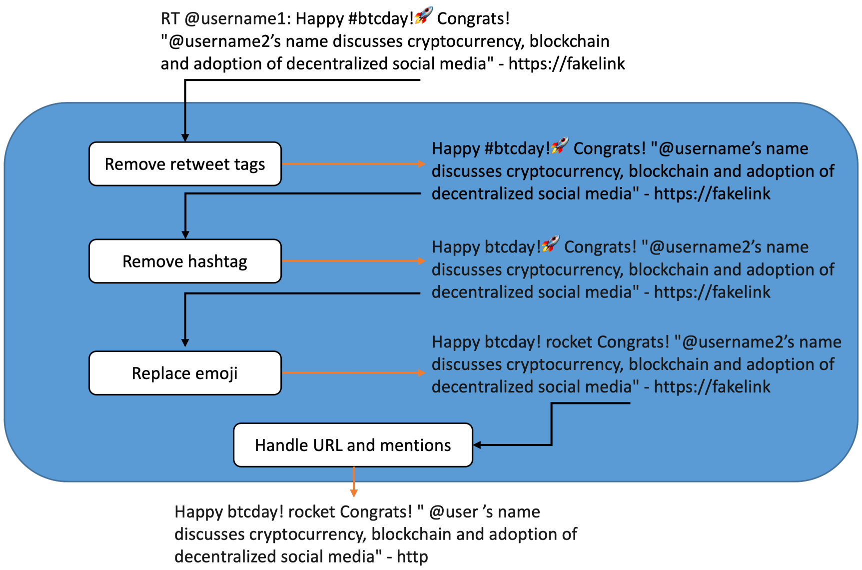

Figure 10: Tweet specific parsing applied to each tweet in the two datasets.

Unlike many other NLP models, the RoBERTa model requires relatively little text preprocessing. Many other models require[19] that stop words be removed, words be stemmed or lemmed to their roots, and various size n-grams created. We do not need to apply any of these steps for the RoBERTa model. The RoBERTa model uses Byte-Pair Encoding (BPE), introduced by (Sennrich, Haddow, & Birch, 2016), that encodes words as a sequence of subwords, allowing the model to handle the large vocabularies common in natural language corpora. BPE is a frequency-based character concatenating algorithm: it starts with two-byte characters as tokens and, based on the frequency of n-gram token pairs, it includes additional, longer tokens. For example, if the letters "e" and "r" are frequently together in the language, a new token, "er" is added to the vocabulary. Next, if "h" and "er" are frequently together, "her" is added to the vocabulary. The algorithm goes on until it reaches the desired vocabulary size. We defer to (Sennrich et al., 2016) for a more complete description of the BPE algorithm. As such, once we apply the Twitter specific text preprocessing steps above, the resulting clean text can be directly fed to the RoBERTa model.

[19] Technically none of these steps are required, but they do result in more robust, tractable models in practice.

Appendix B RoBERTa Sentiment Scores

Table 12 presents 10 positive tweets, as scored by the RoBERTa model, from the Random Tweet dataset. +1 is the maximum positive score and -1 is the maximum negative score; a score of 0 corresponds to a sentiment neutral tweet, which has no impact on the value of the sentiment indices. Quite a few of the tweets are actually about other cryptocurrencies besides BTC or ETH: the first tweet is about Pi and the eighth tweet is about another initial coin offering (ICO). Other tweets are related to the status the cryptocurrency market: the second tweet is about BTC market capitalization, the last tweet is clearly about BTC prices going up, and the seventh tweet is actually about certain technical indicators for BTC. However, there are also tweets that we would not expect to have any bearing on market sentiment: the ninth tweet is related to non-fungible tokens and is only included in the dataset due to its Ethereum hashtag. While there is clearly noise in the data, which is expected given the highly unstructured nature of tweets, the RoBERTa model does a good job of determining which tweets are positive.

Score	**Cleaned Tweet**
0.5149	Pi is a new digital currency being developed by a group of Stanford PhDs. For a limited time, you can join the beta to earn Pi and help grow the network... To join Pi, follow this link http and use my username (Pesamob) as your invitation code. Bitcoin http
0.5959	Market cap just crossed $100B http bitcoin blockchain fintech
0.7894	TechCrunch Hackathon Reveals The Power Of Hashgraph: Fast, Usable & Decentralized Apps Created In... xbt btc http
0.9074	Join this project to get wonderful profit @user airdrop crypto BSCGem http
0.5074	Made this in under 2 hours. Catch these pips. http forexbeginners learntotrade makemoneyonline forextraders earnandlearn cellphonecash crypto iminprofit http
0.8696	http on blockchain genetic medicine ecological solution. One of the hottest in 2018!!! Use my referral link: http and follow @user to claim your free GENE tokens. Airdrop ICO tokens erc20 Crypto bitcoin Genechainplus
0.6258	Will do detail bitcoin chart later But $btc needs to hold this channel to continue. Stochastic also bottoming out nicely. Second chart of $btc CME new gap The old CME gap is at $9670 it's in my timeline. Will drop further chart later today. http
0.9852	It's one of the best ICO I've ever seen. That's actually first-class ICO-project. Have a look at this Crypto Blockchain ARAW arawtoken icosale ecommerce
0.6885	Look what I found! watercolor.eth collectible http rarible ethereum nonfungible digitalasset nft via @user ensdomain
0.7844	BTCis go up thumbs_up thumbs_up thumbs_up thumbs_up thumbs_up Bitcoin

Table 12: 10 random positive tweets. Tweets are sampled randomly from tweets with scores greater than 0.5.

Table 13 presents 10 negative tweets, as scored by the RoBERTa model, from the Random Tweet dataset. Like the positive tweets, we see a mix of highly relevant tweets and less relevant tweets. Tweets five and seven, while negative, are about elephants and an Ethiopian runner, respectively. While tweet five has an Ethereum hashtag it is unclear as to why it is included in the tweet. Tweet seven has an ETH hashtag, which is most likely related to Ethiopa, not Ethereum. However, other tweets are clearly about cryptocurrencies: the first tweet is about losing a significant amount of BTC; the fourth tweet is about a common scam related to cryptocurrencies; and the tenth tweet is about the taxation of BTC miners. Again, the RoBERTa model does a good job at scoring clearly negative tweets.

Score	**Cleaned Tweet**
-0.5004	'I Forgot My PIN': An Epic Tale of Losing $30,000 in Bitcoin — WIRED http
-0.5778	Francesco Molinari is still searching for the game he lost at Augusta National http play_button http Bitcoin Sportsbook http
-0.8666	Fiat is the shitcoin BTC
-0.5169	Common scam/spam script in the Bitcoin Crypto sphere is: "I lost all my money on *exchange name* but thanks to *some website/email* I recovered all my losses"
-0.6991	Elephants chained and beaten but BBC won't apologise for using animals in show http phaajan phajaan CrosscutFest DefendPressFreedom Ethereum Fly-EaglesFly @user
-0.6634	This is bad... cryptomemes altcoinalerts cryptomeme ICO BTC ETH meme crypto bitcoin ethereum hodl cryptocurrencies cryptocurrency bitcoinnews cryptonews crowdsale stellar investments presale DAPP tokensale finance hodl-gang publicsale bountyhunters http
-0.9204	Another casualty.. Ethiopia's Tesfaye Abera has dropped out..He doesn't look well.. Marathon Rio2016 Olympics ETH
-0.6637	Bitcoin continues to face growing pains amid another sell-off The world's leading cryptocurrency dropped below $8,000 on Friday, about 60 percent off its all-time high of nearly $20,000 in mid-December. http
-0.5430	Arresting the MD of @user I believe is strictly uncalled for. Country like India needs to support the innovation. Others wise we might loose this wave of development too. cryptocurrency bitcoin inda MyOpinion dyor
-0.8767	Adding taxation on miners is yet another nail in the coffin for BCH. Taking away fees from miners, isn't a great way to compensate them for the upcoming reward halving. BitCoin is an economic system and BCH actors are economically illiterate. http

Table 13: 10 random negative tweets. Tweets are sampled randomly from tweets with scores less than -0.5.

Table 14 presents 10 neutral tweets, as scored by the RoBERTa model, from the Random Tweet dataset. Interestingly, these tweets are quite a bit shorter than the positive and negative example tweets above, suggesting that shorter messages may, in general, be less sentiment charged. Quite a few of these tweets are simply status messages related to various crypto prices or indicators, which are not inherently positive or negative without additional context; as such, the scores appear reasonable. Tweet three seems quite negative (and it does have a slightly negative sentiment score), but, again more information is needed to fully understand its implications. We would not expect any of the tweets presented - with potentially the exception of tweet three - to have a significant impact on investor sentiment, so the RoBERTa model's neutral scores appear quite accurate.

Score	Cleaned Tweet
0.0517	Bitcoin Price Watch; Breakout and Intra-Range Update http fintech bitcoin
0.0783	Bitcoin Mining Contract 3.5 TH/s +/- 10% for 7 days http bitcoin http
-0.0513	SEC to Bitcoin Investment Trust: 'cease and desist' http btc bitcoin
0.0817	Top 5 cryptocurrencies Alert Time: 2019-10-19 03:40:02 Bitcoin: $7,960.545 Ethereum: $172.908 XRP: $0.292 Tether: $1.005 BitcoinCash: $212.257 binance $USD $EUR $QRL $BTCUSD http
0.0440	Antminer S9 13.5 TH/s Bitcoin Miner http bitcoin crypto antminer bitmain
0.0249	Peer Review on Signal Private Messenger: Hello r/crypto, Can anyone point me to any peer reviewed ar... http bitcoin
0.0892	bitcoin dogecoin monetaryunit startcoin dogecoin miners faucets business we have 5... http http
-0.0346	The Hodl Index score for 01/12/2017 is: 1 http bitcoin
-0.0219	linux crypto Re: [PATCH] crypto: caam - select DMA address size at runtime http
0.0993	Buy Bitcoin and choose OVER 300 payment methods http os

Table 14: 10 random neutral tweets. Tweets are sampled randomly from tweets with scores between -0.1 and 0.1.