Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices

Conference Paper *in* SSRN Electronic Journal · October 2015

DOI: 10.2139/ssrn.2607167

5 authors, including:

George Giaglis
Athens University of Economics and Business
179 PUBLICATIONS **2,889** CITATIONS

SEE PROFILE

Dionisios N Sotiropoulos
University of Piraeus
47 PUBLICATIONS **88** CITATIONS

SEE PROFILE

USING TIME-SERIES AND SENTIMENT ANALYSIS TO DETECT THE DETERMINANTS OF BITCOIN PRICES

Complete Research

Georgoula, Ifigeneia, Athens University of Economics & Business, Athens, Greece, ifgeorgoula@aueb.gr

Pournarakis, Demitrios, Athens University of Economics & Business, Athens, Greece, pournadi@aueb.gr

Bilanakos, Christos, Athens University of Economics & Business, Athens, Greece, xmpilan@aueb.gr

Sotiropoulos, Dionysios, Athens University of Economics & Business, Athens, Greece, dsotirop@aueb.gr

Giaglis, M. George, Athens University of Economics & Business, Athens, Greece, giaglis@aueb.gr

Abstract

This paper uses time-series analysis to study the relationship between Bitcoin prices and fundamental economic variables, technological factors and measurements of collective mood derived from Twitter feeds. Sentiment analysis has been performed on a daily basis through the utilization of a state-of-the-art machine learning algorithm, namely Support Vector Machines (SVMs). A series of short-run regressions shows that the Twitter sentiment ratio is positively correlated with Bitcoin prices. The short-run analysis also reveals that the number of Wikipedia search queries (showing the degree of public interest in Bitcoins) and the hash rate (measuring the mining difficulty) have a positive effect on the price of Bitcoins. On the contrary, the value of Bitcoins is negatively affected by the exchange rate between the USD and the euro (which represents the general level of prices). A vector error-correction model is used to investigate the existence of long-term relationships between cointegrated variables. This kind of long-run analysis reveals that the Bitcoin price is positively associated with the number of Bitcoins in circulation (representing the total stock of money supply) and negatively associated with the Standard and Poor's 500 stock market index (which indicates the general state of the global economy).

Keywords: *Bitcoins; error correction; machine learning; sentiment analysis.*

1 Introduction

The Bitcoin is a digital currency which has recently emerged as a peer-to-peer payment system to facilitate transactions. It is not issued by any central bank or other financial institution but uses cryptographic methods and relies on an open-source software algorithm which verifies decentralized transactions and controls the creation of new Bitcoins. The large fluctuations of Bitcoin prices (especially within the year of 2013) and the huge increase in the capitalization of the associated market have given rise to a branch of the literature studying the factors which help to explain or predict the value of Bitcoins [1, 2, 3, 4]. In this paper, we study the dynamics governing the formation of Bitcoin prices by focusing on Twitter sentiment as an explanatory factor along with other economic and technological variables.

It has been argued in [4] that the price of Bitcoins is mainly driven by the interaction of supply and demand fundamentals (as it happens with other currencies or standard commodities). In this context, the impact of mining technology – which affects the production cost structure and thus the supply side of the market – on Bitcoin prices has been investigated by [3]. However, the supply of Bitcoins evolves according to a publicly known algorithm and the level of demand is not fully determined by the fundamentals of the underlying economy but also depends on expectations about future price

movements. Therefore, the standard economic theory might not adequately describe changes in Bitcoin prices and one should also take short-run speculative investment incentives or expectations into account. These expectations might be reflected in collective sentiment, thus raising the question of measuring public mood and studying its impact on the evolution of Bitcoin prices. In this context, [1] uses search queries on Google trends and Wikipedia (representing the degree of public recognition or interest in Bitcoins) as a proxy for public sentiment and finds a positive correlation between these measures and the price of Bitcoins.

We extend the above line of reasoning here by constructing a sentiment ratio for Twitter users on a daily basis. Several measures of public mood associated with online social media have been suggested in the literature to predict the movement of stock market indexes [5, 6, 7]. It has recently been argued that Twitter posts (related to Bitcoins) which express negative sentiments or uncertainty are negatively correlated with the price of Bitcoins [8]. In this paper, we perform sentiment analysis through the use of a state-of-the-art machine learning algorithm (namely Support Vector Machines). The econometric analysis of our time-series data implies that the Twitter sentiment ratio has a significantly positive impact on Bitcoin prices. The frequency of Wikipedia views and the level of mining difficulty (measured by the hash rate) are also positively associated but the exchange rate between the USD and the euro is negatively associated with the value of Bitcoins. For the cointegrated time series, the estimation of a vector error-correction model shows that the stock of Bitcoins has a positive long-run impact and the Standard & Poor's 500 index has a negative long-run impact on Bitcoin prices. Finally, we find that the price of Bitcoins adjusts to its long-run equilibrium value at a relatively high speed.

The rest of this paper proceeds as follows. Section 2 introduces the conceptual framework and describes the selected dataset. Section 3 suggests the methodology used to conduct the sentiment analysis and the set of econometric estimations. Section 4 derives the empirical results and discusses their implications. Section 5 concludes and provides directions for further research.

2 Theoretical Framework and Dataset

If we consider Bitcoin as a medium of exchange, then its price should be determined by standard supply and demand interactions [2, 4]. Fisher's [9] equation of exchange associated with the quantity theory of money stipulates MV=PT (where M is the nominal supply of money, V is the velocity of money circulation, P is the general price level and T is the size of the underlying economy). The nominal supply of Bitcoins is given by $M=P^BB$ (where P^B is the price of Bitcoins and B is the stock of Bitcoins in circulation), thus implying $P^B=PT/VB$. Therefore, the equilibrium price of Bitcoins (i.e. the price equalizing demand and supply) should be positively related to the general price level (P) and the size of the Bitcoin economy (T) but negatively related to the total stock of Bitcoins in circulation (B). In the same context, the mining difficulty can be used to measure the production cost of Bitcoins which affects the supply side of the market [3]. Furthermore, the level of demand for Bitcoins might be related to the general macroeconomic state of the global economy captured, for example, by alternative stock market indices [2]. However, Bitcoins are also treated as an investment asset whose demand could be affected by speculative behavior associated with expectations and public feelings about their future price movements. These feelings might be captured by the degree of public recognition and interest in Bitcoins measured by the number of search queries in Google trends and Wikipedia [1]. Alternatively, public feelings can be measured through the sentiment analysis of posts related to Bitcoins on social media such as Twitter [9].

On the grounds of this conceptual framework, we use time series data for eleven variables collected from 27 October 2014 to 12 January 2015 on a daily basis. All series have been transformed by taking natural logarithms to overcome the problems of many outliers and high skewness mainly associated with financial variables. The dependent variable is the price of Bitcoins given by the Bitstamp closing price (*bcp*) in USD. There are four independent variables representing the supply and demand fundamentals of the market: First, the stock of Bitcoins in circulation (*totbc*) represents the total

money supply. Second, the daily total number of unique transactions (*ntran*) describes the size of the Bitcoin economy. Third, the number of Bitcoin days destroyed for any given transaction (*bcdde*) measures the Bitcoin money velocity and is calculated by multiplying the number of Bitcoins in a transaction with the number of days elapsed since these coins were last spent. Fourth, the daily exchange rate (*exrate*) between the USD and the euro ($/€) represents the price level of the global economy. We also include the Standard & Poor's 500 stock market index (*sp*) as an independent variable representing the general state of the global economy. The level of mining difficulty is captured by the hash rate (*hashrate*) measuring the processing power of the Bitcoin network. All these series were downloaded by quandl.com. We also use three proxies for the degree of public recognition and interest in Bitcoins: First, the number of Bitcoin searches in Wikipedia (*wikiviews*) downloaded from bitcoinpulse.com. Second, the (normalized) number of search queries in Google (*googleviews*) retrieved from Google Trends. Third, the daily number of Twitter posts (*ntweets*) related to Bitcoins as collected in our database. The last explanatory variable of our model is the daily sentiment ratio (*sent*) associated with Twitter posts. The methodology for constructing the time series of this sentiment ratio is described in the next section. The full set of variables included in the model is listed in Table 1 below.

Table 1. The set of variables.

Name of variable	Description
bcp	Bitstamp daily closing price
totbc	Total daily number of Bitcoins in circulation
ntran	Total daily number of unique Bitcoin transactions
bcdde	Bitcoin days destroyed for any given transaction
exrate	Daily exchange rate between the USD and the euro ($/€)
sp	Standard & Poor's 500 stock market daily index
hash	Processing power required for the secure operation of Bitcoin network (in billions of hashes per second)
wiki	Daily number of Bitcoin search queries on Wikipedia
google	Daily number of Bitcoin search queries on Google
ntweets	Daily number of Twitter posts related to Bitcoins
sent	Daily sentiment ratio of Twitter posts related to Bitcoins

3 Methodology

3.1 Sentiment Analysis Methodology

We collected and analysed a set of over 2,125,243 tweets during a time period of 78 days, between October 27th 2014 and January 12th 2015. The data collection process focused on gathering tweets for the keywords "Bitcoin", "BTC" and "Bitcoins" along with their respective hashtags ["#Bitcoin", "#BTC", "#Bitcoins"]. This task was accomplished by parsing the official streaming API of Twitter with Python and MySQL for storing the data. Appropriate wrappers were deployed in our dedicated Ubuntu server to ensure the process against runtime errors during network downtime. The resulting dataset was subsequently submitted to a series of data clearing and pre-processing operations. The data

preparation process involved text tokenization into words, elimination of English stop-words and words with less than three characters, and stem extraction from each word. Therefore, the final version of our corpus was formed by a collection of purified documents where each document contained the text from a single tweet.

3.1.1 Corpus Vectorization

A natural approach towards sentiment analysis is through a mathematical representation of the corpus via the employment of the standard Vector Space Model (VSM) originally introduced in [10]. The main idea behind VSM is to transform each document d into a vector containing only the words that belong to the document and their frequency by using the "bag of words" representation. According to VSM, each document is represented exclusively by the words it contains by tokenizing sentences into elementary term (word) elements losing the associated punctuation, order and grammar information. The underlying mathematical abstraction imposed by VSM entails a mapping which transforms the original purified document to its corresponding bag of terms representation. This transformation can be formulated by the following equation:

$$\varphi : d \to \varphi(d) = \left[tf\left(t_1, d\right), tf\left(t_2, d\right), \ldots, tf\left(t_M, d\right) \right] \in \mathbb{R}^M \quad (1)$$

where $tf(t_i, d_j)$ is the normalized frequency of the term t_i in document d_j given by:

$$tf\left(t_i, d_j\right) = f\left(t_i, d_j\right) / \max\left\{ f\left(t, d_j\right) : t \in d_j \right\} \quad (2)$$

where $f(t_i, d_j)$ is the absolute frequency of the term t_i in document d_j. Based on the adopted mathematical formulation for the fundamental notions of corpus and dictionary, a corpus D of n documents and a dictionary T of M terms may be represented by $D=\{d_1,d_2,\ldots,d_n\}$ and $T=\{T_1,T_2,\ldots,T_M\}$. Having in mind Eq.1 and the formal definitions of corpus and dictionary, the mathematical representation for corpus in the context of VSM can be done through the document-term matrix:

$$D = \begin{bmatrix} tf\left(t_1, d_1\right) & \cdots & tf\left(t_M, d_1\right) \\ \vdots & \ddots & \vdots \\ tf\left(t_1, d_n\right) & \cdots & tf\left(t_M, d_n\right) \end{bmatrix} \quad (3)$$

where N is typically quite large, resulting in a sparse VSM representation such that a few matrix entries are non-zero. In order to mitigate the effect related to the complete loss of context information around a term, we incorporated the term-frequency inverse document frequency (tf–idf) weighting scheme according to which each term t_i is assigned a weight of the form:

$$w_i = idf\left(t_i, D\right) = \log\left(|D| / |\{d \in D : t_i \in d\}| \right) \quad (4)$$

so that the relative importance of each term for the given corpus is taken into consideration.

3.1.2 Support Vector Machines

Sentiment analysis was conducted through the utilization of a state-of-the-art classifier, namely Support Vector Machines (SVMs). SVMs are non-linear classifiers that were initially formulated by [11], operating in higher-dimensional vector spaces than the original feature space of the given dataset.

Letting $S = \{(\vec{x_i}, y_i) \in \mathbb{R}^n \times \{-1, +1\}, \forall i \in [m]\}$ be the set of m training patterns with associated binary labels, such that -1 denotes the class of negative sentiment and +1 the class of positive sentiment, the learning phase of the SVMs involved solving the following quadratic optimization problem:

$$\min_{\vec{w},\vec{\xi},b} \frac{1}{2}\left\|\vec{w}^2\right\| + C\sum_{i=1}^{m}\xi_i$$

$$s.t.\ y_i\left(\left\langle \vec{w},\vec{x}_i\right\rangle + b\right) \geq 1-\xi_i,\ \ \xi_i \geq 0\ \forall i \in [m]$$

The previous primal optimization problem has a corresponding dual one which gives rise to a discrimination function of the form:

$$g(\vec{x}) = \sum_{i\in SV}^{m} \alpha_i^* y_i \left\langle \vec{x},\vec{x}_i\right\rangle + b^* \quad (5)$$

where $\{\alpha_i^*, i \in [m]\}$ and b^* denote the optimal solutions for the corresponding optimization variables and SV is the subset of training patterns associated with positive Lagrange multipliers. Given that the training patterns appear only in dot product terms of the form $\langle \vec{x_i}, \vec{x} \rangle$, a positive definite kernel function such as $K(\vec{u},\vec{v}) = \Phi(\vec{u})\Phi(\vec{v})$ can be employed in order to implicitly map the input feature space into a higher-dimensional vector space and compute the dot product. In this paper, we utilized the Gaussian kernel function defined by the following equation.

$$K(\vec{x},\vec{y}) = exp(-\|\vec{x}-\vec{y}\|^2/2\sigma^2) \quad (6)$$

3.1.3 Sentiment Classification

Our sentiment classification process was further divided into the corresponding training and testing stages. The training stage is an essential part of our methodology, since the application of SVMs on such a large amount of tweets requires a reasonable amount of labelled data (i.e. tweets already classified as positive, negative or neutral, based on a business perspective classification). This ensures that the SVM algorithm runs with accuracy, providing robust results that limit the amount of fault. These labelled data are in turn utilized by the SVM algorithm as a benchmark to score the number of tweets that are in scope of the sentiment exercise. In order to create a reasonable amount of labelled data, we manually labelled a set of collected tweets in terms of sentiment as positive (1), neutral (0) or negative (-1). The testing stage, on the contrary, aims at testing the accuracy and validity of the SVM algorithm on the largest subset of the dataset that was not previously classified.

In order to demonstrate the validity of the SVM algorithm for the sentiment classification problem, we adopted the standard 10-fold cross validation process on the previously labelled Tweets and measured the corresponding training and testing sentiment classification accuracy. Each fold involved splitting the complete set of pre-labelled samples into a 95% training data - 5% testing data ratio, where the fist subset of data instances was used to build the classifier and the latter for assessing its ability to infer the sentiment polarity of unseen data patterns. The scores of the SVM classifier are summarized in Table 2. The sentiment categorization for the rest of the unlabelled data patterns was conducted by exploiting the complete set of pre-labelled data instances so that the trained classifier accumulated the maximum amount of available knowledge for the problem of sentiment classification.

Table 2. Sentiment Classification Scores.

Performing 10-fold cross-validation on labeled data on 2 classes: (346L, 800L) | (48L, 800L) | (1L, 346L) | (1L, 48L) | (394L,)
Accuracy per fold: [0.9 | 0.875 | 0.9 | 1. | 0.875 | 0.875 | 0.87179487 | 0.87179487 | 0.89473684 | 0.89473684]

Mean Accuracy: 0.896 (+/- 0.073)
Precision per fold: [0.8974359 | 0.875 | 0.8974359 | 1. | 0.875 | 0.875 | 0.87179487 | 0.87179487 | 0.89473684 | 0.89473684]
Mean Precision: 0.895 (+/- 0.073)
Recall per fold: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Recall: 1.000 (+/- 0.000)
F-Score per fold: [0.94594595 | 0.93333333 | 0.94594595 | 1. | 0.93333333 | 0.93333333 | 0.93150685 | 0.93150685 | 0.94444444 | 0.94444444]
Mean F-Score: 0.944 (+/- 0.039)

3.2 Econometric Methodology

The econometric approach follows a number of steps associated with the analysis of time series data. The first step requires investigating whether the variables are stationary or not. A non-stationary series is integrated of order d (I(d)) if it becomes stationary by taking its differences d times. If we include a non-stationary series in a regression model, the ordinary least squares (OLS) estimators are not consistent and the standard statistical tests are not valid. Therefore, we might infer a statistically significant causal relationship between a pair of variables although such a relationship does not exist [12]. In order to avoid this problem of spurious regression, we conduct several stationarity tests for each series y_t. In particular, we start with the augmented Dickey-Fuller (ADF) test based on an autoregressive model of order one:

$$y_t = a_0 + a_1 y_{t-1} + \varepsilon_t \tag{7}$$

If $a_1=1$, the series has a unit root (i.e. it is non-stationary). We subtract y_{t-1} from both sides of (1) and define $\theta = a_1 - 1$, $\Delta y_t = y_t - y_{t-1}$ to get:

$$\Delta y_t = a_0 + \theta y_{t-1} + \varepsilon_t \tag{8}$$

If the series has a trend, we must explicitly include time as an explanatory variable in (7). Furthermore, we can add p lags of Δy_t to account for the dynamics of the process:

$$\Delta y_t = a_0 + \delta t + \theta y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + \varepsilon_t \tag{9}$$

The ADF test has a null hypothesis of a unit root ($\theta=1$) against the alternative of stationarity. In order to strengthen the validity of inferences, we also use the Phillips-Perron (PP) unit root test [13]. This can be viewed as a Dickey-Fuller test which has been made robust to serial correlation by using the Newey-West heteroscedasticity and autocorrelation-consistent (HAC) standard errors. Finally, we verify our results by also applying the KPSS test which has a null hypothesis of stationarity against the alternative of a unit root [14].

A non-stationary series which is integrated of order d can be made stationary by taking its differences d times. However, the use of differences in a regression model does not allow determining potential long-run relationships between the variables. Therefore, we would like to use the levels of variables but also avoid the problem of spurious regression. If the variables are cointegrated then a regression model involving their levels yields consistent OLS estimators [12]. The series y_t, $x_{1t},\ldots,$ x_{kt} are cointegrated CI(d,b) if all of them are integrated of order d and there exists a linear combination of these series which is integrated of order d–b. The standard example involves d=b=1, implying that the

series are I(1) and there exists a linear combination $u_t = y_t - a_0 - a_1 x_{1t} - ... - a_k x_{kt}$ which is stationary. Then, there is a long-run relationship between the cointegrated variables given by:

$$u_t = y_t - a_0 - a_1 x_{1t} - ... - a_k x_{kt} + u_t \quad (10)$$

The vector $[1, -a_0, -a_1, ..., -a_k]$ is the cointegrating vector and might not be unique for the case of multiple (more than two) variables. The lagged series u_{t-1} is the error correction term measuring deviations from the long-run equilibrium. In this context, the second step of our econometric analysis involves testing for cointegration between the series which have the same order of cointegration. For the multivariate case, the cointegration test is based on Johansen's method calculating a trace statistic to specify the number of cointegrating vectors [15].

After conducting the cointegration tests, we proceed with two separate sets of estimations. On the one hand, we rely on several OLS regressions to identify short-run relationships between the price of Bitcoins and the set of independent variables. Of course, these regressions require the transformation of non-stationary series (by taking their differences) to render them stationary. On the other hand, for the series which are found to be cointegrated we build a vector error-correction model (VECM) to detect the existence of long-run relationships [16]. For ease of exposition, let us focus here on the case of two series x_t and y_t. If these series are cointegrated, there exist unique values of α_0 and α_1 such that $u_t = y_t - a_0 - a_1 x_t$ is stationary. If we think of y_t as the dependent variable and x_t as an exogenous regressor, the single-equation error-correction model is written as:

$$\Delta y_t = \beta_0 + \beta_1 \Delta x_t + \lambda u_{t-1} + \varepsilon_t = \beta_0 + \beta_1 \Delta x_t + \lambda (y_{t-1} - a_0 - a_1 x_{t-1}) + \varepsilon_t \quad (11)$$

The VECM extends (10) by allowing the joint evolution of x_t and y_t and by putting p lags on the right-hand side of both equations involved in the associated system:

$$\Delta y_t = \gamma_1 + \sum_{i=1}^{p} \delta_{1i} \Delta y_{t-i} + \sum_{i=1}^{p} \theta_{1i} \Delta x_{t-i} + \lambda_1 (y_{t-1} - a_0 - a_1 x_{t-1}) + \varepsilon_{1t} \quad (12)$$

$$\Delta x_t = \gamma_2 + \sum_{i=1}^{p} \delta_{2i} \Delta y_{t-i} + \sum_{i=1}^{p} \theta_{2i} \Delta x_{t-i} + \lambda_2 (y_{t-1} - a_0 - a_1 x_{t-1}) + \varepsilon_{2t} \quad (13)$$

Since the terms Δy_{t-i}, Δx_{t-i} and u_{t-1} are stationary, the OLS estimators of equations (12) and (13) are consistent. The coefficients δ_{ji} and θ_{ji} represent the short-run dynamics whereas α_0 and α_1 describe the long-run relationship between x_t and y_t. The parameters λ_1 and λ_2 contain information on the speed of adjustment to the long-run equilibrium by showing the correction of the previous period's disequilibrium error taking place in period t. The next section applies the above methodology and discusses the empirical results concerning the short-run and long-run determinants of Bitcoin prices.

4 Results

4.1 Stationarity and Cointegration

We mainly rely on the statistical program STATA but also use R and Matlab where necessary to verify our empirical results. First, we study the stationarity of each series by using the ADF, the PP and the KPSS tests. Since the ADF test yields ambiguous results concerning the stationarity of some independent variables, we turn to the PP unit root test which uses the Newey-West HAC standard errors to account for serial correlation. The results are summarized in Table 3 which shows the PP test statistic and the associated p-value for all series as well as for the first differences of non-stationary series. We conclude that the variables logbcp, logtotbc, logsp and logexrate are I(1), whereas for all other variables the null hypothesis of a unit root is rejected. These results are verified by the KPSS test having the opposite null hypothesis of stationarity.

Table 3. Phllips-Perron (PP) unit root tests.

Variable	PP-test statistic	p-value
logbc	-1.656	>0.1
d.logbc	-8.171	<0.01
logntran	-4.460	<0.01
logtotobc	-2.057	>0.1
d.logtotbc	-6.240	<0.01
logbcdde	-6.406	<0.01
logwiki	-7.995	<0.01
loggoogle	-6.096	<0.01
logntweets	-6.961	<0.01
Logsp	-2.875	>0.1
d.logsp	-7.213	<0.01
loghash	-8.615	<0.01
logexrate	-1.786	>0.1
d.logexrate	-9.499	<0.01
logsent	-6.635	<0.01

We proceed with the cointegration analysis to check whether there exists a long-run relationship between the four non-stationary series. For this purpose, we use Johansen's trace test for multivariate cointegration. The procedure starts with testing for zero cointegrating vectors and then accepts the first null hypothesis that is not rejected. As shown in Table 4, we reject the null hypothesis of no cointegration but fail to reject the null hypothesis of at most one cointegrating vector. Therefore, we accept that there is one cointegrating equation in the model.

Table 4. Johansen's test for cointegration.

Cointegrating vector	Trace statistic	5% critical value
0	53.3435	47.21
1	24.9293*	29.68
2	10.7067	15.41
3	1.7893	3.76

The short-run impact of independent variables on the price of Bitcoins is studied through the estimation of several OLS regression models in the next subsection. The long-run relationship between the cointegrated series is determined by using a vector error-correction model (VECM).

4.2 OLS Estimates

All the regressions run in this section use the Newey-West HAC standard errors. Non-stationary series have been transformed by taking their first differences to guarantee non-spurious results. We first consider a regression containing the full set of explanatory variables. Table 5 shows that the variables logwiki (representing the degree of public recognition or interest in Bitcoins) and loghash (measuring the mining difficulty) have a positive impact on Bitcoin price, whereas the effect of the exchange rate between the USD and the euro (reflecting the price level of the global economy) is negative. Even more interestingly, the sentiment ratio of Twitter users positively affects the price of Bitcoins. Since the removal of non-significant factors increases the overall F-statistic of the model, we keep only the

significant variables and add lags in the regression. Table 6 shows the results with three lags for each explanatory variable. Apart from own-price effects, we note that the second lag of Wikipedia views and the second lag of the Twitter sentiment index also have a significantly positive effect on current Bitcoin prices. The list of variables correlated with the price of Bitcoins in the short run is shown in Table 7.

Table 5. OLS regression with the full set of explanatory variables.

Regression with Newey-West standard errors — Number of obs = 76
maximum lag: 5 — F(10, 65) = 5.74
Prob > F = 0.0000

D.logbcp	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
logntran	-.0009874	.0378949	-0.03	0.979	-.0766687	.0746939
logtotbc D1.	-137.6733	148.6298	-0.93	0.358	-434.5074	159.1608
logbcdde	-.005857	.0042119	-1.39	0.169	-.0142688	.0025548
logwiki	.0202568	.007452	2.72	0.008	.0053742	.0351394
loggoogle	-.0237488	.0317945	-0.75	0.458	-.0872467	.0397491
logntweets	.000973	.022617	0.04	0.966	-.0441963	.0461423
logsp D1.	-.1239438	.6570604	-0.19	0.851	-1.436184	1.188296
loghash	.0072955	.002659	2.74	0.008	.001985	.0126059
logexrate D1.	-1.935874	1.053349	-1.84	0.071	-4.039557	.1678083
logsent	.6890404	.2568672	2.68	0.009	.1760412	1.20204
_cons	-.0670151	.4132169	-0.16	0.872	-.8922663	.7582361

Table 6. OLS regression with three lags for each significant explanatory variable.

Regression with Newey-West standard errors — Number of obs = 73
maximum lag: 5 — F(19, 53) = 10.37
Prob > F = 0.0000

D.logbcp	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
logbcp LD.	.2486992	.0907913	2.74	0.008	.066595	.4308033
L2D.	-.2584576	.0808432	-3.20	0.002	-.4206085	-.0963067
L3D.	-.0360947	.1621652	-0.22	0.825	-.3613569	.2891675
logwiki --.	.0197162	.0068254	2.89	0.006	.0060263	.0334062
L1.	-.000763	.0088237	-0.09	0.931	-.0184612	.0169351
L2.	.017811	.008851	2.01	0.049	.000058	.0355639
L3.	-.0001393	.0083484	-0.02	0.987	-.0168841	.0166054
loghash --.	.0088932	.0034975	2.54	0.014	.0018782	.0159082
L1.	.0071992	.0058986	1.22	0.228	-.0046319	.0190304
L2.	-.0105373	.0044826	-2.35	0.022	-.0195282	-.0015464
L3.	.001057	.0046555	0.23	0.821	-.0082807	.0103947
logexrate D1.	-1.674214	.7754822	-2.16	0.035	-3.229635	-.1187941
LD.	1.54933	1.329604	1.17	0.249	-1.117517	4.216177
L2D.	-.1552209	.7811126	-0.20	0.843	-1.721934	1.411492
L3D.	-.9975245	.8014653	-1.24	0.219	-2.60506	.6100113
logsent --.	.8069225	.283992	2.84	0.006	.2373067	1.376538
L1.	-.5240742	.3189217	-1.64	0.106	-1.16375	.1156016
L2.	.8888849	.5158606	1.72	0.091	-.1458004	1.92357
L3.	-.2830255	.3074419	-0.92	0.361	-.8996759	.3336249
_cons	-.3719047	.2401851	-1.55	0.127	-.8536549	.1098455

Table 7. Short-run influencers of Bitcoin prices.

Variable	Effect on Bitcoin price
Wikipedia views	Positive
Hash rate	Positive
Sentiment ratio	Positive
USD/EUR exchange rate	Negative

4.3 Vector Error-Correction Model

The long-run relationship between the cointegrated variables is now determined through a VECM with four lags (based on the information criteria). The first part of Table 8 shows the short-run dynamics as well as the speed of adjustment to the long-run equilibrium. The second lag of *logsp* has a negative impact and the lagged difference of *logexrate* is now found to have a positive impact on the price of Bitcoins. It can be seen that 31.25% of the gap between the Bitcoin price in period t-1 and its equilibrium value tends to be reversed in period t. In other words, if the Bitcoin price is too high then it falls back towards its equilibrium level relatively quickly. The second part of Table 8 shows the coefficients of the cointegrating equation. An increase in the stock of Bitcoins leads to an increase in the Bitcoin price (contrary to our expectations), while an increase in the Standard and Poor's 500 stock market index (showing an improvement in the state of the global economy) negatively affects the price of Bitcoins in the long run. The last result potentially reflects the fact that investment in stocks and investment in Bitcoins are treated as substitutes. Table 9 shows the set of variables having a long-run effect on Bitcoin prices.

Table 8. Vector error-correction model with four lags.

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D_logbcp						
_ce1						
L1.	-.3125402	.1048414	-2.98	0.003	-.5180255	-.1070549
logbcp						
LD.	.3203674	.1287212	2.49	0.013	.0680785	.5726564
L2D.	-.208652	.1417118	-1.47	0.141	-.486402	.069098
L3D.	-.0429716	.137881	-0.31	0.755	-.3132135	.2272702
L4D.	-.1119016	.1336133	-0.84	0.402	-.373779	.1499757
logtotbc						
LD.	-226.9944	239.6149	-0.95	0.343	-696.631	242.6423
L2D.	-119.2837	213.636	-0.56	0.577	-538.0026	299.4353
L3D.	-235.5137	204.9175	-1.15	0.250	-637.1446	166.1172
L4D.	-18.90374	202.1341	-0.09	0.925	-415.0792	377.2718
logsp						
LD.	.6393147	.8691665	0.74	0.462	-1.06422	2.34285
L2D.	-1.832229	.913669	-2.01	0.045	-3.622988	-.0414707
L3D.	-.8660657	.9259208	-0.94	0.350	-2.680837	.9487058
L4D.	-.5406169	.8668753	-0.62	0.533	-2.239661	1.158427
logexrate						
LD.	3.698214	1.287785	2.87	0.004	1.174203	6.222226
L2D.	-.5359462	1.395253	-0.38	0.701	-3.270591	2.198699
L3D.	.177157	1.337561	0.13	0.895	-2.444415	2.798729
L4D.	-.0301283	1.141858	-0.03	0.979	-2.268128	2.207872
_cons	.0005305	.0728064	0.01	0.994	-.1421673	.1432284

```
Cointegrating equations

Equation           Parms     chi2     P>chi2
-------------------------------------------
_ce1                   3   88.02165   0.0000
-------------------------------------------

Identification:  beta is exactly identified

                 Johansen normalization restriction imposed
------------------------------------------------------------------------------
        beta |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_ce1         |
      logbcp |          1          .       .       .            .           .
    logtotbc |    15.9535   4.122128    3.87   0.000     7.874277    24.03272
       logsp |  -4.210831   .7492988   -5.62   0.000     -5.67943   -2.742233
   logexrate |  -.5230419   1.434541   -0.36   0.715    -3.334691    2.288607
       _cons |  -236.1674          .       .       .            .           .
------------------------------------------------------------------------------
```

Table 9. Long-run influencers of Bitcoin prices.

Variable	Effect on Bitcoin price
Number of Bitcoins	Positive
S&P's 500 index	Negative

5 Conclusion

This paper tried to shed light on the factors determining the price of Bitcoins in the short-run as well as in the long-run. We built an empirical model incorporating multiple economic and technological variables but also extended the existing literature by taking Twitter sentiment into account. Specifically, we used a state-of-the-art machine learning algorithm (Support Vector Machines) to measure the sentiment ratio of Twitter users concerning Bitcoins on a daily basis. After dealing with issues of stationarity and cointegration, we estimated several regression models indicating that our Twitter sentiment ratio has a positive short-run impact on Bitcoin prices. In other words, evidence shows that measurements of collective mood based on the appropriate sentiment analysis can help to predict short-run movements in the value of Bitcoins. Furthermore, the price of Bitcoins has been found to be positively affected by the number of Wikipedia search queries. This implies that a higher degree of public recognition or interest in Bitcoins increases their market price. Similarly, an increase in the hash rate has a positive effect on Bitcoin prices. This is hardly surprising, since the hash rate indicates the mining difficulty or marginal production cost of Bitcoins and thus normally exerts an upward pressure on their price. On the contrary, our estimations revealed a negative short-run relationship between the price of Bitcoins and the exchange rate between the USD and the euro. To the extent that this exchange rate represents the general level of prices, its inverse relationship with the value of Bitcoins contrasts the prediction of Fisher's equation associated with the quantity theory of money. For the set of cointegrated variables, we estimated a VECM to identify the underlying long-run relationships. The analysis revealed that the stock of Bitcoins has a positive long-run impact on their price. This is also a counter-intuitive result, since the number of Bitcoins in circulation measures the total supply of money which would be expected to have a negative effect on Bitcoin prices. The Standard and Poor's 500 index was found to have a negative impact on Bitcoin prices in the long run, implying that stocks and Bitcoins are treated as substitutes by investors. More specifically, a decrease in the Standard and Poor's 500 index induces investors to sell their stocks and substitute them for Bitcoins. Finally, the speed at which the price of Bitcoins adjusts to its long-run equilibrium value is relatively high. In particular, about one half of the deviation between the current and the equilibrium level of Bitcoin prices is already corrected within the next period.

The empirical model developed above can be improved and extended in multiple ways. First of all, a larger dataset should be used to check whether the conclusions reached here remain valid or not.

Second, a vector autoregressive (VAR) model might be used (instead of simple OLS regressions) to study the short-run dynamics of price formation. Finally, alternative sentiment indices for social media users might be constructed (by applying the appropriate algorithmic processes) to explore their short-run and long-run effectiveness in explaining the price of Bitcoins. These extensions are left for future research.

Acknowledgments. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Programs: Thalis – Athens University of Economics and Business – Software Engineering Research Platform; Aristeia – Athens University of Economics and Business – Herding Behavior and Asymptotic Learning in Electronic Social Media – Sociomine.

References

1. Kristoufek, L.: Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet area. Scientific Reports 3 (3415), 1-7 (2013)
2. Ciaian, P., Rasjcaniova, M., d'Artis, K.: The Economics of Bitcoin Price Formation. Working Paper (2014)
3. Li, X., Wang, C., Wang, Q.: Exploring the Determinants of Bicoin Exchange Rate. Working Paper (2014)
4. Buchholz, M., Delaney, J., Warren, J., Parker, J.: Bits and Bets, Information, Price Volatility and Demand for Bitcoin. Working Paper (2012)
5. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2, 1, 1-8 (2011)
6. Gilbert, E., Karahalios, K.: Widespread Worry and the Stock Market. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)
7. Zhang, X., Fueres, H., Gloor, P.: Predicting Stock Market Indicators through Twtter 'I Hope It is Not as Bad as it Fears'. Procedia – Social and Behavioral Sciences 26, 56-62 (2011)
8. Kaminski J., Gloor, P.: Nowcasting the Bitcoin Market with Twitter Signals. Working Paper (2014)
9. Fisher, I.: The Purchasing Power of Money, New-York, Macmillan (1911)
10. Salton, G., Wong, A., Yang, C. S.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18, 11, 613–620 (1975)
11. Vapnik, V.: The Nature of Statistical Learning Theory, Springer, New York (1995)
12. Engle, R.F., Granger, C.W.J.: Co-integration and Error Correction Representation, Estimation and Testing. Econometrica 55, 251-276 (1987)
13. Phillips, P.C.B., Perron, P.: Testing for a Unit Root in Time Series Regression. Biometrica 75, 335-346 (1988)
14. Kwiatkowski, D., Phillips, P, Schmidt, P., Shin, Y.: Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. J. of Econometrics 54, 159–178 (1992)
15. Johansen, S.: Likelihood-Based Inference in Cointegrated Vector Autorogressive Models. Oxford: Oxford University Press (1995)
16. Enders, W.: Applied Econometric Time Series. John Wiley & Sons, Hoboken, NJ (2003)