University of St.Gallen

School of Management, Economics, Law, Social Sciences and International Affairs

The effect of Twitter activity on Bitcoin price fluctuation

Software Engineering for Economists
(7,610,1.00)

Alen Stepic - 11-475-258
Dimitrios Koumnakes - 10-613-370
Joël Sonderegger - 11-495-488
Severin Kranz - 13-606-355
Chi Xu - XX-XXX-XXX

Fall Term 2017

Supervisor:
Prof. Dr. Philipp Zahn
FGN HSG
Varnbüelstrasse 19
9000 St. Gallen

2nd January 2018

Abstract

This paper, examines the dynamics between the amount of twitter activity regarding bitcoin and the actual price fluctuation of bitcoin. For the analysis two sets of data have been collected. One reflecting twitter activity and the other one showing the bitcoin prices for the same period. Based on that an econometric model is used for the statistical analysis and the interpretation of the results.

INPUT ALEN @DIMI: ADD JUST KEY FINDING THE REST I LIKE

This work was created in the context of a programming course for economists at the university of St. Gallen. To not exceed the framework of this study we require extensive knowledge in econometrics and take concepts as known. The underlying goal was to get familiar with software engineering tools and project management. Therefore, the academic content of this work is neither completed nor concluding.

Contents

List of Figures

List of Abbreviation

VAR	Vector Autoregression
EMH	Efficient Market Hypothesis
etc	et cetera

1 Introduction

malkiel1970efficient introduced the Efficient Market Hypothesis (EMH), where they claim that under certain market conditions actual prices include all information. This hypothesis is broadly accepted in the financial world. The digitalisation leads to an increased network effect, where information can be share within second over the globe. Based on **mao2015quantifying** (**mao2015quantifying**, as cited in **shiller2015irrational**; **kahneman2013prospect**) the EMH fails to address the behavioural and emotional role of investors . The large fluctuation of the bitcoin prices in the last year (mostly increasing) and the fact that Twitter has become a very popular social network, where people interact, led to different research in the field of sentimental analysis with the focus on bitcoin and twitter. **mao2015quantifying**, claim that sentiment analysis using twitter tweets which contain the buzzword bullishness can indeed be used as an sentiment indicator for stock prices. As digitalization is expected to continue in the future, crypotocurrencies will hypothetical get more important. Thus, the topic is characterized as relevant.

1.1 Research Question and Goal of the Paper

As existing research focus mostly on the sentimental analyses of tweet content, this paper aims to examine existence of a correlation between the price development of bitcoin and the amount people talk about bitcoin on twitter. The goal should be reached by answering the following research question:
In what extend does the amount of twitter tweets about bitcoin has an effect on the price movement of bitcoin?

1.2 Methodology

To answer the above mentioned research question a scientific approach has been applied. The paper consists out of three parts. The first part contains an introduction into the topic by pointing out the relevance of the topic, the research question, methodology and the scope. This is conducted in chapter 1. The second part is characterized as the theory part. The theory aims to provide to the reader the necessary background information by an introduction of relevant bitcoin and twitter information in context of the paper. The conducted research is based on forward research and using relevant sources. Furthermore, it contains a short introduction into the Vector Autoregression approach (VAR) and the constrains in context with the paper. This is conducted in chapters 2 and 3. The third part contains the discussion of the results and a conclusion. The conclusion is based on the theory and the own conducted mathematical computation. Those computations are based on a data set, which was gathered by own coded python scripts for the timeperiod December 20th - 27th. Based on the short time of observation the results have to be interpreted carefully. The data set contains out of two data sources (1) twitter tweets which contain the buzzword "bitcoin" and (2) historical bitcoin prices. Those two sources has been aggregated with a separate python script. The output of aggregation was

further proceeded with stata to conduct the VAR on a daily basis. The third part of the paper is discussed in the chapters 3 and 4.

1.3 Scope

The paper's focus lies on the scientific discussion and answer of the research question mentioned above. The research question has consciously been design very tiny, as the focus of the course Software Engineering for Economists is the application of different tools, the documentation and reproducibility of the results, rather than writing a high quality scientific paper. As already mentioned above, the results have to be interpreted carefully because of the limited observation time of seven days. The approach and technical issues (like code scripts etc.) are not addressed. Those points are explained in the separate documentation paper.

2 Background and Data Collection

2.1 Bitcoin

After the global financial crisis, Satoshi Nakamoto (2008, p. 1) claimed, that a purely peer-to-peer version of electronic cash could bypass financial institutions as third parties for commercial transactions. Starting from his whitepaper, Bitcoin – the first electronic payment system that relies on the cryptographic concatenation of ongoing transactions (blockchain) has been developed (Nakamoto, 2008, p. 1).

Today, Bitcoin is the most popular of over 1300 so-called cryptocurrencies worldwide and increased in value over 1700 percent within the last year according to Coinmarketcap (https://coinmarketcap.com/). Even if the daily transaction volume increased rapidly over time (https://coinmarketcap.com/), this does not necessarily mean that Bitcoin is used for payment purposes. Even more, several authors argue that the Bitcoin price and transaction volumes are mainly pushed by speculative investments rather than actual usage as a currency (Corbet, Lucey & Yarovya, 2017; Forbes, 2017; Yermack, 2013). Kristoufek (2015) points out that there are other influencing factors such as usage in trade, money supply and price level, that influence the Bitcoin price in the long term. However, evidence shows that the level of Bitcoin prices are clearly driven by investors' interest in the digital currency. Kristoufek (2015) further explains the correlation is most evident in the long run. During explosive increases or rapid declines of the price higher investors' interest further boosts the movement in the direction (Kristoufek, 2015). These findings are in line with other researchers (see Garcia, Tessone, Mayrodiev & Perony, 2014; Kondor, Posfai, Csabai, &Vattay, 2014).

One core assumption of behavioural finance is that investors' interests influence their behaviour and therefore stock prices. (Mao et al., 2015, p. 3) A widely used approach to measure investors' interests is the sentiment analysis of Twitter data. Based on analysis of the influence of twitter bullishness on the stock market by Mao et al. (2015), this paper further elaborates on the concept by applying the approach to Bitcoin prices.

2.2 Twitter

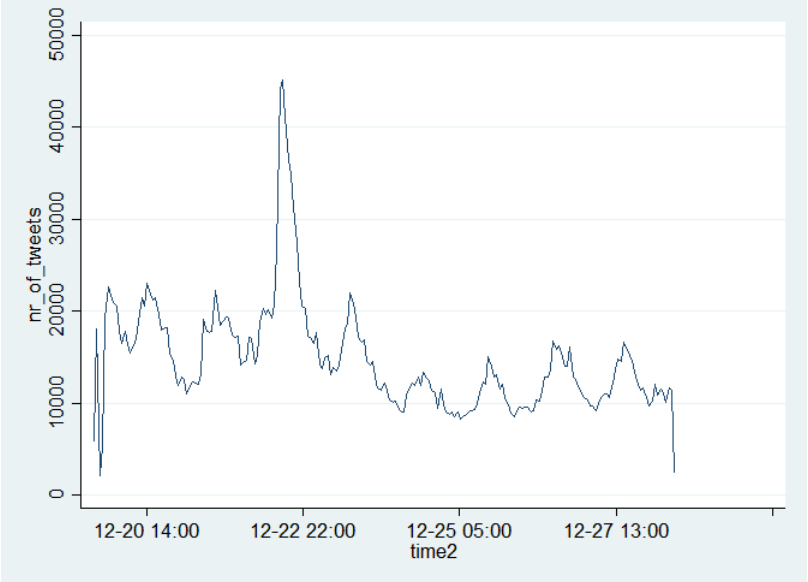
Outline: Growth of Social Media, Opinion Mining, Twitter -¿ micro-blogging platform, tweets, functionalities, etc. ; Why good for measuring opinions ? "gold mine"; Sentiment Analysis, What we do. Comment Severin: This part will contain the following steps and will be mostly based on Stenqvist et al. Predicting Bitcoin price fluctuation with Twitter sentiment analysis (sections 2.2, 2.3)

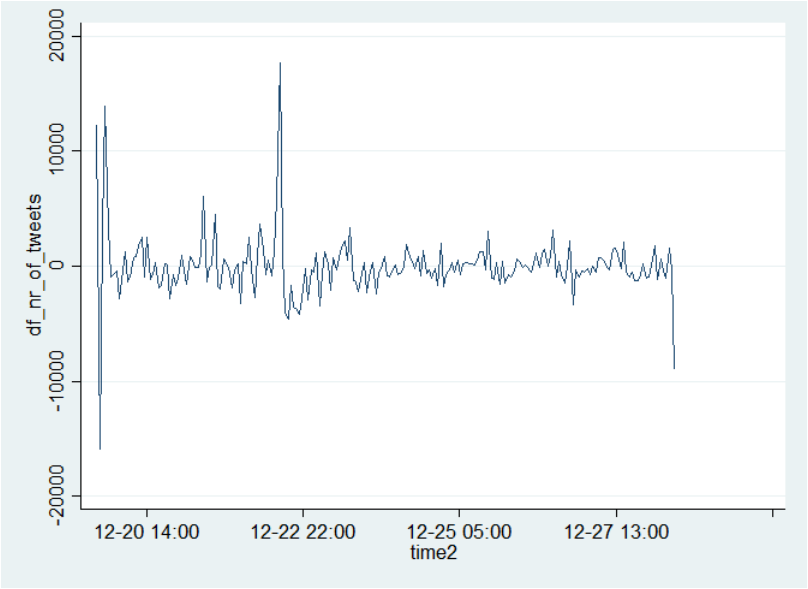
1) Introduction sentence -¿ growth of social media https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
2) Twitter as a "gold mine" to gather opinions -¿ What is twitter, Why is it valuable -¿ length of messages etc.
3) Sentiment Analysis -¿ what is a sentiment analysis
4) Sentiment Analysis of the term bitcoin based on twitter -¿ explain what we do

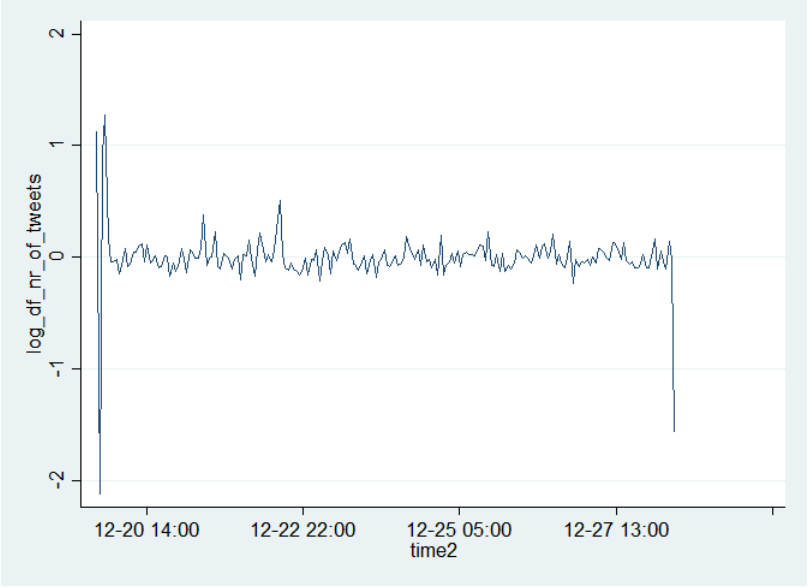
2.3 Data Collection

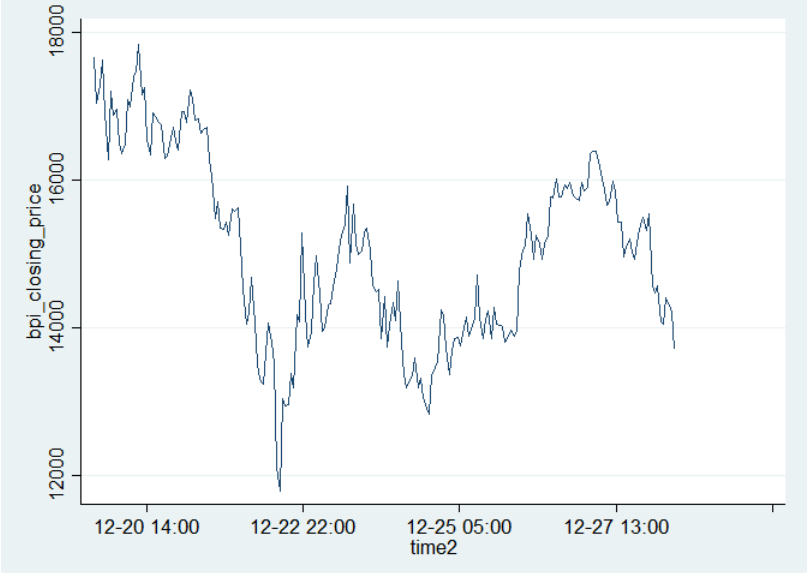
For our dataset, we used 2 different sources.
We first collected historical data on bitcoin prices to USD and secondly, we gathered twitter massages related to bitcoin. We fulfil this relationship by selecting tweets that contain the term "bitcoin" in the text message. The time period for both collected data sets was from 19.12.2017 20:00:00 UTC to 28.12.2017 10:00:00 in an hourly basis. We chose an hourly basis to be able to do a statistical analysis with a certain amount of observation for the restricted processing time of this homework.

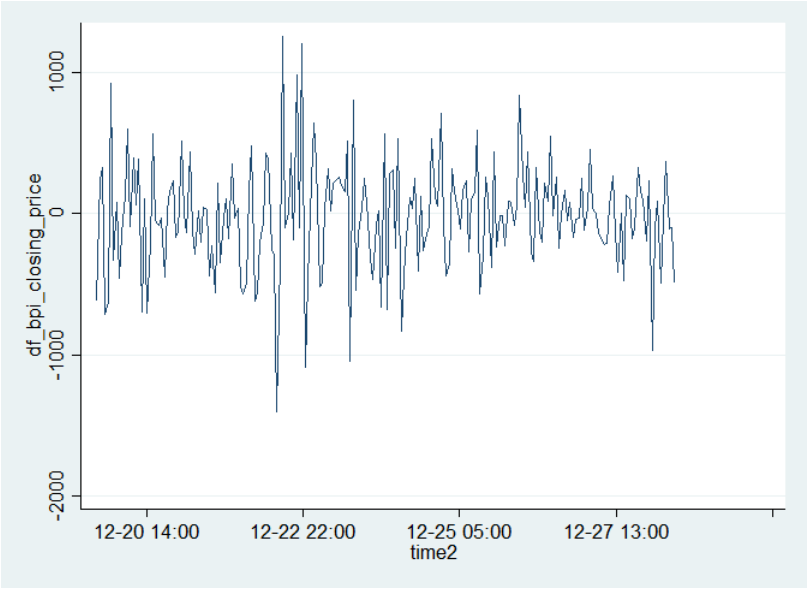
For the bitcoin prices, a publicly available API from CoinDesk was used (Source? See documentation?). The prices collected represent the final price level of bitcoin before every hour changes. For example, our first listed price of bitcoin was on 19.12.2017 at 19:59:59 and our last value on 28.12.2017 at 09:59:59. These values construct our first time series data set and are shown in the following graph.

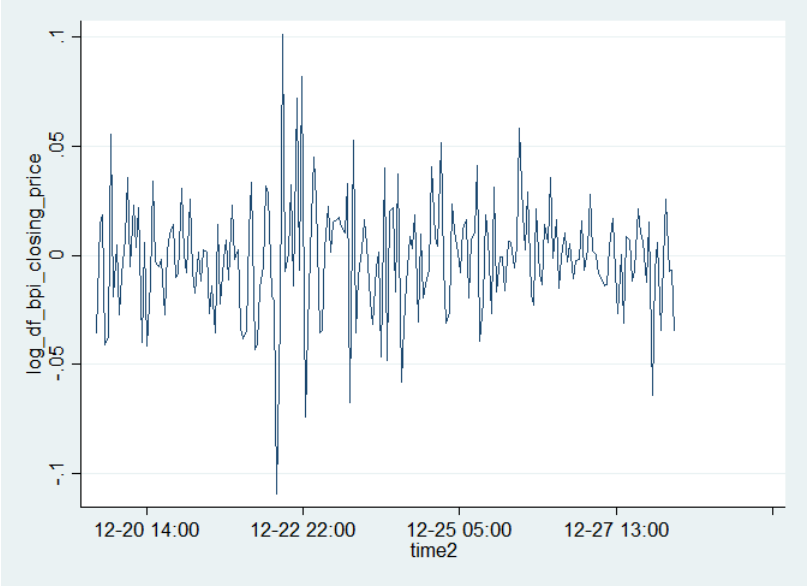



df_nr_of_tweets
20000
10000
0
-10000
-20000
12-20 14:00
12-22 22:00
12-25 05:00
12-27 13:00
time2


log_df_nr_of_tweets
2
1
0
-1
-2
12-20 14:00
12-22 22:00
12-25 05:00
12-27 13:00
time2


bpi_closing_price
18000
16000
14000
12000
12-20 14:00
12-22 22:00
12-25 05:00
12-27 13:00
time2


df_bpi_closing_price
1000
0
-1000
-2000
12-20 14:00
12-22 22:00
12-25 05:00
12-27 13:00
time2


log_df_bpi_closing_price
.1
.05
0
-.05
-.1
12-20 14:00
12-22 22:00
12-25 05:00
12-27 13:00
time2

3 Econometric Modelling and Results

As stated in the beginning, we want to examine the underlying relationship between twitter activity and bitcoin price fluctuations. From our data collection, we are left with two time series datasets representing two variables. A broadly used statistical method to simultaneously analyze multiple time series is the VAR (Vector Autoregression) approach. In this approach the endogenous variables are determined both by their own historical values and by the historical values of the other endogenous variables (**l“IeC −“”u”tkepohl2007new**).
To generate the econometric results that follow we used the statistical software STATA.

3.1 Stationarity

Before estimating a VAR model, the time series data must be checked for stationary. Thus, the means and variances are constant over time and the dataset do not show any trending behavior. Non-stationary data can lead to an inaccurate model which is undesirable. To test for stationarity, we use the Augmented Dickey-Fuller (ADF) test.
First, we check for stationarity of the number of tweets (variable name in data set: nr_of_tweets). Because there this variable may not be stationary we have also vreated the first difference of the number of tweets (variable name: df_nr_of_tweets) and the logarithmic first difference of the number of tweets (variable name: log_df_nr_of_tweets). The augmented Dickey-Fuller test for these variables gives the following results (via STATA):

```
dfuller nr_of_tweets, lags(0)
dfuller df_nr_of_tweets, lags(0)
dfuller log_df_nr_of_tweets, lags(0)

Dickey-Fuller test for unit root
                                        ——— Interpolated Dickey-Fuller ———
                     Test          1% Critical    5% Critical    10% Critical
                   Statistic          Value          Value          Value
——————————————————————————————————————————————————————————————————————————
 nr_of_tweets        -3.152          -3.475         -2.883         -2.573
——————————————————————————————————————————————————————————————————————————
 df_nr_of_tweets     -14.167         -3.475         -2.883         -2.573
——————————————————————————————————————————————————————————————————————————
 log_df_nr_of_tweets -17.583         -3.475         -2.883         -2.573
——————————————————————————————————————————————————————————————————————————
```

From these result we can see that only the first variable in not stationary. This is because the absolute value of the test statistic (3.152) is smaller than the absolute value of the 1% critical value (3.475). The second variable is stationary ($14.167 > 3.475$) and from here onwards we are going to continuous the analysis by using this first difference transformetion of the variable number of tweets, which fulfils stationarity.

Secondly, we have to check for stationarity for our second variable, the bitcoin prices (variable name in data set: bpi_closing_price). Here we also included the first difference (variable name: df_bpi_closing_price) and the logarithmic first difference (variable name: log_df_bpi_closing_price) transformations. The ADF test for these variables gives the following results:

```
dfuller bpi_closing_price, lags(0)
dfuller df_bpi_closing_price, lags(0)
dfuller log_df_bpi_closing_price, lags(0)

Dickey-Fuller test for unit root
                                        ——— Interpolated Dickey-Fuller ———
                     Test          1% Critical    5% Critical   10% Critical
                  Statistic          Value          Value         Value
------------------------------------------------------------------------------
 bpi                -2.609          -3.475         -2.883        -2.573
------------------------------------------------------------------------------
 df_bpi            -16.303          -3.475         -2.883        -2.573
------------------------------------------------------------------------------
 log_df_bpi        -15.996          -3.475         -2.883        -2.573
------------------------------------------------------------------------------
```

These result also suggest to use the first difference transformation of the variable bitcoin prices, for stationarity to be fulfilled.

3.2 Lag-Specification

To select the optimal number of lags in our VAR regression we check for various information criteria. Information criteria are measuring the tradeoff between model fit and parsimony, giving use the optimal number of lag to use (**brandt˙williams˙2007**). The calculations of these criteria for our given data set are shown in the following table.

```
varsoc df_nr_of_tweets df_bpi_closing_price, maxlag(7)

Selection-order criteria
Sample:  12-20-2017 04:00:00 - 12-28-2017 10:00:00
                                        Number of obs      =       199
```

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-3266.9				6.4e+11	32.8533	32.8667	32.8864
1	-3252.58	28.654	4	0.000	5.7e+11*	32.7495*	32.7897*	32.8488*
2	-3249.55	6.0623	4	0.195	5.8e+11	32.7593	32.8262	32.9247
3	-3247.69	3.7128	4	0.446	5.9e+11	32.7808	32.8746	33.0125
4	-3244.1	7.1874	4	0.126	5.9e+11	32.7849	32.9054	33.0828
5	-3238.16	11.868*	4	0.018	5.8e+11	32.7654	32.9128	33.1295
6	-3237.12	2.0852	4	0.720	6.0e+11	32.7952	32.9693	33.2254
7	-3235.03	4.1679	4	0.384	6.1e+11	32.8144	33.0154	33.3109

```
Endogenous:  df_nr_of_tweets df_bpi_closing_price
 Exogenous:  _cons
```

As can be seen on the table, STATA calculates and presents various information criteria (AIC, HQIC, SBIC). The asterisk indicates the optimal lag lenght to use for the regression analysis. In this case all criteria suggest a lag lenght of 1.

3.3 Regression Model

After determining the optimal lag gength the next step is to build the vector autoregression model (VAR model). For this analysis, we use a basic unrestricted VAR model which consists of two endogenous variables, T which repesents the variable df_nr_of_tweets and B which represents bpi_closing_price. As defined in the previous section the selected otpimal time lag is 1. Given these information, the model equations can be written as follows:

$$\begin{aligned} T_t &= c_1 + a_{11}T_{t-1} + a_{12}B_{t-1} + \epsilon_1 \\ B_t &= c_2 + a_{21}T_{t-1} + a_{22}B_{t-1} + \epsilon_2 \end{aligned} \tag{1}$$

This model shows that the current value of our endogenous variavles is determined by its own past values, the past values of the other variable and an error term. In our research approach we ask the question, if the amount of tweets mentioning bitcoin in the past, has an effect on the bitcoin prices in the periods that follow. This question is pictured in the second equation of this model.

After running the above VAR regression in STATA we get following results:

VARIABLES	(T_t) df_nr_of_tweets	(B_t) df_bpi_closing_price
L.df_nr_of_tweets (T_{t-1})	0.00999 (a_{11})	0.00463 (a_{21})
	(0.0693)	(0.0104)
L.df_bpi_closing_price (B_{t-1})	-0.675 (a_{12})	-0.125* (a_{22})
	(0.469)	(0.0707)
Constant	-88.34 (c_1)	-18.48 (c_2)
	(177.9)	(26.79)
Observations	205	205

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The equation of our VAR model are displayed vertically in this table. As said before the relevant equation in the model is the second one and it is listed on the second column (B_t) in the above table. The coefficient of 0.00463 (a_{21}) implies that an increase of tweets by 1 increases the bitcoin price by 0.005 $. This effect is very small and statistically insignificant (p value > 0.1). This means that based on our data set the amount of tweets have no significant effect on the price development of bitcoin.

3.4 Granger Causality

To assess the reliability (causality) of the previous results we use the Granger causality test. This test confirms if the statements by the regression are valid. The following table shows the results of the Granger causality test

vargranger

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
df_nr_of_tweets	df_bpi_closing_~e	2.0732	1	0.150
df_nr_of_tweets	ALL	2.0732	1	0.150
df_bpi_closing_~e	df_nr_of_tweets	.19677	1	0.657
df_bpi_closing_~e	ALL	.19677	1	0.657

The Null hypothesis here is... In the second row the chi2 is greater then the Prob value which allows as to reject the Null hypothesis and conclude that the regression results are valid...

...the Null hypothesis in our case is that lagged number of Tweets does not cause changes in

bitcoin prices. As the Table shows, we can reject the Null hypothesis because the probability is around 1%. Therefore, we accept the alternative hypothesis, which says that the lagged number of tweets does cause changes in bitcoin prices. How these changes look exactly we have discussed in the previous section. The granger causality confirmed that the effect is valid.

4 Conclusion

(Insert text)

5 References

6 Declaration of Authorship

We hereby declare,

- that we have written this thesis without any help from others and without the use of documents and aids other than those stated above;
- that we have mentioned all the sources used and that we have cited them correctly according to established academic citation rules;
- that we have acquired any immaterial rights to materials we may have used such as images or graphs, or that we have produced such materials ourself;
- that the topic or parts of it are not already the object of any work or examination of another course unless this has been explicitly agreed on with the faculty member in advance and is referred to in the thesis;
- that we will not pass on copies of this work to third parties or publish them without the University's written consent if a direct connection can be established with the University of St.Gallen or its faculty members;
- that we are aware that our work can be electronically checked for plagiarism and that we hereby grant the University of St.Gallen copyright in accordance with the Examination Regulations in so far as this is required for administrative action;
- that we are aware that the University will prosecute any infringement of this declaration of authorship and, in particular, the employment of a ghostwriter, and that any such infringement may result in disciplinary and criminal consequences which may result in our expulsion from the University or us being stripped of our degree.

.....
Dimitrios Koumnakes - 10-613-370

.....
Severin Kranz - 13-606-355

.....
Joël Sonderegger - 11-495-488

.....
Alen Stepic - 11-475-258

.....
Chi Xu - XX-XXX-XXX

By submitting this academic term paper, we confirm through my conclusive action that we are submitting the Declaration of Authorship, that we have read and understood it, and that it is true.