

Prediction of Cryptocurrency Market

Rareş Chelmuş¹, Daniela Gîfu^{1,2}, Adrian Iftene¹

¹„Alexandru Ioan Cuza” University, Faculty of Computer Science,

²Romanian Academy – Iaşi branch, Institute for Theoretical Computer Science
{rareş.chelmuş, daniela.gifu, adiftene}@info.uaic.ro

Abstract. In this paper, we present a set of experiments on predicting the rise or fall of a cryptocurrency using machine learning algorithms and sentiment analysis of the afferent media (online press mainly). The machine learning part is using the data of the currencies (the prices at a specific time) to predict in a mathematical sense. The sentiment analysis of the media (articles about a cryptocurrency) will influence the mathematical prediction, depending on the feeling created around the currency. The study can be useful for entrepreneurs, investors, and normal users, to give them a clue on how to invest. Furthermore, the study is intended for research regarding natural language processing and human psychology (deducting the influence of masses through media) and also in pattern recognition.

Keywords: machine learning, cryptocurrency market, online press corpus, price prediction, sentiment analysis

1 Introduction

Human history is filled with prediction attempts and with prophets [Nelson, 2000], but the ability to anticipate the future events of life evolution, especially in financial economics [Gifu and Cristea, 2012], is a complex task that can be solved using artificial intelligence (AI) techniques. Forecasting the market, the purpose of this paper remains an important challenge with a lot of implications for safety economy. Economists developed, through many studies and observations on graphs and data, diverse techniques to get an idea of how the prices will rise or fall (martingales) [Feller, 1971]. The accuracy of data retrieval depends on the understanding of the rational, emotional and physiological limiting factors towards the stock market. There are also various variables to consider as: the causal impact of media, investor's behavior, time-varying local economic conditions, history of prices, natural calamities, etc. They can systematically affect the fluctuation of the market. Some specialists in financial economics and big data consider that trying to predict and invest in a random manner can direct to same results. This assumption is known as the EMH (*Efficient Market Hypothesis*)¹ or the RWH (*Random Walk Hypothesis*) [Fama,

¹ Starting with Eugene Fama's PhD (1965), EMH becoming one of the most known theories in financial economics that confirm the connection between prices and public information.

1965; Jonathan Clarke *et al.*, 2001; Zunino *et al.*, 2012; Bariviera *et al.*, 2014; Marwala, 2015; Marwala and Hurwitz, 2017].

In this study, we have chosen the cryptocurrency market. *Why?* Because, Bitcoin (BTC), Ethereum (ETH), Ripple (XRP) and the rest of the cryptocurrencies and tokens represent the future of money. Moreover, in the last year, the market capitalization of cryptocurrency has grown from 16 billion of dollars (10 January 2017) to an outstanding peak of 829 billion of dollars (7 January 2018)². Because of the high trading volume, a lot of data (graphs and text) is gained everyday³ and is free for the entire public to be used by every possible mean. The best approach is proving to be a combination of mathematical models with psychological models because it has a projected growth rate on prediction.

The paper is organized as follows: section 2 shortly reviews the relevant literature on various approaches about the huge role of predicting the cryptocurrency in human life; section 3 presents materials (data set) and methods based on machine learning (ML) algorithms; the results are described in section 4; section 5 highlights a short discussion about this survey, and, finally, our conclusions are given in section 6.

2 Related Work

A collection of AI algorithms has been used to predict the fluctuation of a cryptocurrency. Some of them were developed through observation, like *mean reversion*, based on the assumption that the stocks will reach the average value in time [Spierdijk and Bikker, 2012; Dai *et al.*, 2013]. However, the problem still remains: you cannot really predict when and how long the prices will stay on the average. We know that stock prices fluctuate in a random manner [Granger, 1992]. It is the reason why stock investors use martingales to predict the future trend.

To forecast a linear graph, we have discovered the fact that linear regression could be intuitively used, being among the first attempts in predicting the course of a practical linear model [Xin, 2009]. This algorithm, one of the most well known algorithms in machine learning, was used in solving the prediction problem due to the nature of the data set used in the stock market (graph, CSV) [Nunno, 2017].

Another (supervised) machine learning algorithm used in reducing the risk of investing in the stock market is Random Forests. This model constructs a number of decision trees and gives as an output the class or mean prediction of the trees. The method has been proven effective in trading, but on a long-term period [Khaidem *et al.*, 2016] and on the stock market. It is known the fact the stock market is a more stable market than the crypto one.

For the overreaction-driven price momentum mechanism [Daniel *et al.*, 1998; Hou *et al.*, 2009], media is reflected on the investors' decision behavior [Barber & Odean, 2008].

In general, sentiment analysis (SA) as a NLP (*Natural Language Processing*) approach is a popular predicting algorithm. Hilbert considers that the negative news increase the probability to lower the prices of stocks. On the other hand, the positive news could reflect a rise of the value of stocks [Hilbert *et al.*, 2014].

² cryptolization.com

³ coinmarketcap.com

In our work, the linear regression, random forest and random forest aided by sentiment analysis algorithms formed the basis for the prediction models of the cryptocurrency market described below.

3 Materials & Methods

In this section, we describe the data set and methods used to predict cryptocurrency market (here 3 cryptocurrencies, called: Ethereum, Bitcoin and Ripple), on which we will build our prediction models.

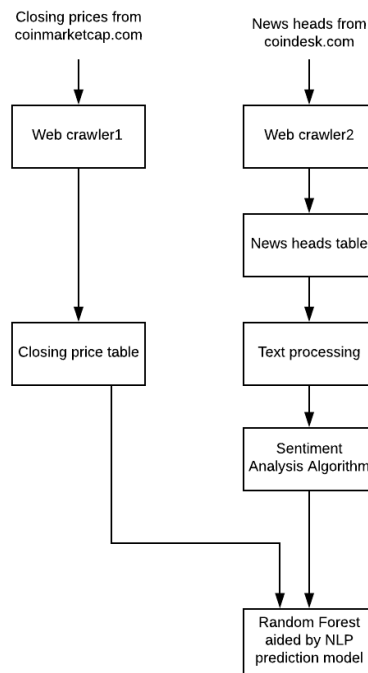


Fig. 1. Architecture of the cryptocurrency prediction

For these experiments, we preferred to choose cryptocurrencies that are found in the top 5 of the cryptocurrency market, because these coins are considered mature, meaning that there's a lot of trading activity every day (the data is diversified). The trading volume and, also the notoriety of the cryptocurrencies is direct proportional with the place held in the top. Moreover, the media closely follows how they fluctuate over time. The titles of articles have been analyzed by us using sentiment analysis classifiers from the NLTK python library. The algorithm will gather data on a period of 3 months, from 15 October 2017 to 05 January 2018. The reason for this short period of time is that the cryptomarket is still evolving and in the last 3 months in media spiked a huge flow of articles regarding cryptocurrencies.

3.1 Data Set

We test our system on two data collections: the closing price chart and the news heads table. Their differences permit us to improve the performance of our approach in types of cryptocurrency's fluctuations. This is also a great start for entrepreneurs, programmers and researchers to find better solutions on predictions, because using a site to get the news titles helps in avoiding hazard in data. Data could be scraped from other sites like Twitter, Reddit forums, YouTube and other media from where you can gather various information. However, the problem is finding the best source and cleaning the data set (the English used in forums and social media isn't all the time accurate and can have a lot of mistakes). Thus, for the moment, we will get our data from the best news site on crypto.

3.1.1. Text Data

The text data set (Table 1 contains only a sample of data) is used by the NLP component of the prediction algorithm and is formed of article titles from cryptocurrency news site⁴. Note that the articles, directed towards the coin in order to be predicted, will not appear daily. In this case, we approximate the missing scores using our formula (3), tailored for our needs. In addition, there is a possibility that many articles about the coin would appear in a single day.

Table 1. Statistical data.

Data	Title	N words of title
15/10/2017	Hours to go: How to Watch Ethereum's Fork as It Happens	11
16/10/2017	Bulls Take Breather? Bitcoin Slows as Price Struggles to Breach \$6,000	11
...
1/01/2018	What Will the Bitcoin Price Be in 2017?	8
02/01/2018	Is It Too Late To Buy Bitcoin? Video: \$1 Million? Bitcoin Sign Guy on Why It's Not Too Late to Buy	21
03/01/2018	RSK Beta Brings Ethereum-Style Smart Contracts Closer to Bitcoin	10
04/01/2018	Ripple Fever? Other Crypto Assets Are Outpacing Its 2018 Gain	10
05/01/2018	Ethereum Price Highs Overshadow New Wave of Tech Issues	9
Total		4473
Average		8.89

⁴ www.coindesk.com

Note that all text data set contains 420 (Bitcoin), 56 (Ethereum), and 22 (Ripple) articles.

3.1.2. Graphs Data

The price chart (see Figure 2) is a table that consists of eight columns, but for this survey, we keep the date and the closing price of the cryptocurrency. The table has an entry for each day, on a period of approx. 3 months and the data will be fetched by a web crawler. Figure 3 presents a good image about the volatility of the cryptocurrency market.

Market Cap ▾

Trade Volume ▾

Trending ▾

Tools ▾

Search Currencies

Q

All ▾

Coins ▾

Tokens ▾

USD ▾

Next 100 →

View All




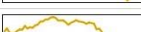


#	Name	Market Cap	Price	Volume (24h)	Circulating Supply	Change (24h)	Price Graph (7d)
1	 Bitcoin	\$196,978,854,505	\$11,717.80	\$19,004,600,000	16,810,225 BTC	13.45%	
2	 Ethereum	\$101,696,586,391	\$1,047.81	\$8,251,620,000	97,056,324 ETH	16.55%	
3	 Ripple	\$60,248,664,465	\$1.56	\$9,118,600,000	38,739,142,811 XRP *	46.66%	

Fig. 2. The prices distributions of the 3 most mature coins per time



Fig. 3. Price trends of Bitcoin, Ethereum, and Ripple over the past 3 months (October 15, 2017–January 05, 2018). Data source: coinmarketcap.com

3.2 Phases Methodology

The work flow (Fig. 1) starts with the web crawlers that gather data from specific sites to fill the tables. There are two crawlers: one that scrapes a site where the closing prices of the cryptocurrencies are held online in a table form⁵, and the other crawler gets the head of the news articles from a site specialized on cryptocurrency news⁶.

The next step is to pass the news titles through the SentimentIntensityAnalyzer module from the NLTK library⁷ in Python; this will be the sentiment analysis component. The title will be automatically annotated to be passed through a polarity detection algorithm regarding the feeling that is deducted. The process will give scores for each piece of text, consisting in two types of measures: positive and negative feeling score. The final score will be calculated by the formula:

$$final_score = (sum(positive_score) - sum(negative_score)) / number_of_articles \quad (1)$$

The sum (1) is used to solve the multiple articles problem, to answer at the question: *what sentiment is dominant?* Note that, in some days, media wrote more articles about those three cryptocurrencies. Averaging the sum of all positive scores minus the sum of the negative ones will result the overall sentiment on that day.

Table 2. SA scores of the media articles analyzed between 15.10.2017-05.01.2018.

Data	Scores
15/10/2017	0
16/10/2017	-0.157
16/10/2017	-0.149
16/10/2017	0
17/10/2017	-0.192
18/10/2017	0
18/10/2017	-0.208
18/10/2017	0
26/10/2017	0.426
26/10/2017	-0.216
...	...

We have three distinct situations (Table 2): 1) periods of time without data (e.g. between 2017-10-18 and 2017-10-26), 2) multiple articles in the same day (e.g. 2017-10-16), and 3) neutral articles (when the SA score is equal with 0). Because articles regarding a crypto coin do not appear every day, gaps in the table will occur.

Goel from Stanford University [Mittal and Goel, 2012] filled these blank spaces with the formula (2):

$$gap = (x + y) / 2 \quad (2)$$

⁵ <https://coinmarketcap.com>

⁶ <https://www.coindesk.com/>

⁷ <http://www.nltk.org/>

where x and y are table boxes filled with values, and between x and y are n number of blank spaces (period of days when nothing was published towards the analyzed cryptocurrencies).

Moreover, machine learning could be used to predict the values for the blank spaces. However, the machine learning algorithm won't behave well because of the anomalies in the graph (unexpected rise and fall in the graph, specific for the cryptomarket). Our approach is a modification to Goel's formula:

$$gap = (x + y) / (\sqrt{n + 1} + 1) \quad (3)$$

This alteration changes the behavior of the graphic, reflected in formula (3), having a more natural evolution from x to y . The algorithm responds better on our needs. There are three cases on how the graphic would behave (we will take $n = 5$) (Fig. 4, 5 and 6):

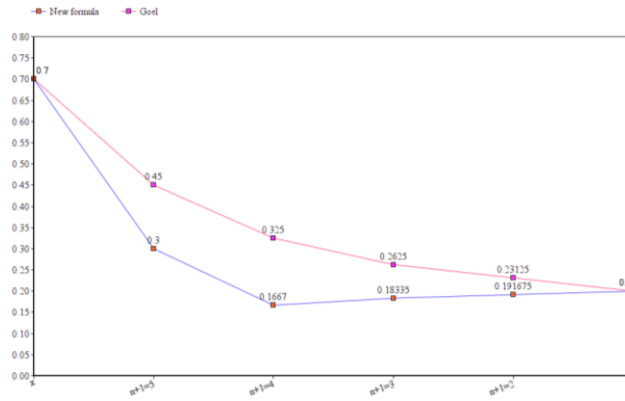


Fig. 4. Case 1: $x > y$

The formula (3), represented by the blue line, has a more natural evolution, because it falls and it also rises to reach the y value, despite Goel's equation, where it only falls.

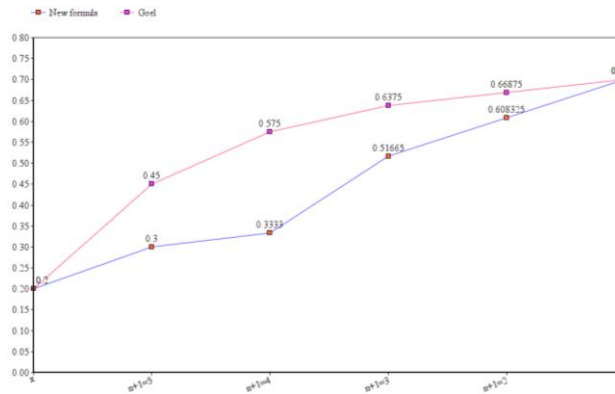


Fig. 5. Case 2: $x < y$

Note that the first rise is not as abrupt as Goel's formula ($0.45 > 2 \times x$) \rightarrow responds better according with the real cryptocurrency market fluctuations.

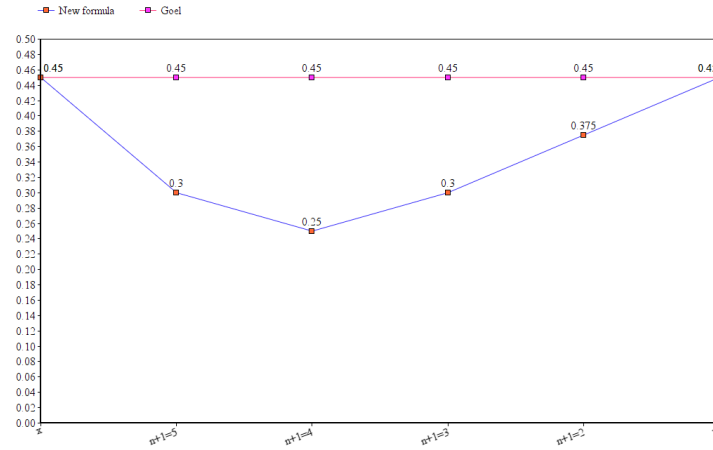


Fig. 6. Case 3: $x = y$

We can observe as well that there is a movement in the line even though x is equal with y . In Table 3, we used the formula (3), in order to complete the scores where we have no data.

Table 3. The final input for the ML algorithm

Data	Scores	Close
15/10/2017 00:00:00	0	336.6
16/10/2017 00:00:00	-0.153	333.38
17/10/2017 00:00:00	-0.192	317.08
18/10/2017 00:00:00	-0.208	314.32
19/10/2017 00:00:00	-0.0206	308.09
20/10/2017 00:00:00	0.00094	304.01
21/10/2017 00:00:00	0.0067	300.19
22/10/2017 00:00:00	0.00658	295.45
23/10/2017 00:00:00	0.00656	286.95
24/10/2017 00:00:00	0.00656	298.33
25/10/2017 00:00:00	0.02231	297.93
26/10/2017 00:00:00	0.105	0.105
...

Our method is based on three algorithms, using the scikit-learn Python library⁸: linear regression, random forests and random forests aided by sentiment analysis. In this way, we can compare the precision, the recall and the f-measure obtained with each of them, in order to analyze the potency of the NLP component.

⁸ <http://scikit-learn.org>

3.2.1. Linear Regression Model

As mentioned before, Linear Regression is a primitive model used in predicting stock prices and not very efficient (because of the high fluctuation of the market). The Linear regression formula (4) usually looks like:

$$y = b_0 + b_1 \times x \quad (4)$$

where y is a dependent value (the price we want to predict), b_0 is called intercept, x the independent variable (the regressor, the price we know) and b_1 is the slope of the line.

3.2.2. Random Forests Model

Random forest works by creating a multiple of decisions trees trained by random samples of data from the data set. The result is an average or the vote of the majority of the outputs of the trees. In this case, we will only use the closing price table (only one feature). Thus the algorithm trains on 90% of the data (first 90% rows) and the last 10% will be used for test.

3.2.3. Sentiment Analysis with Random Forests Development

In this case, the algorithm will be trained on both tables (Table 2 and 3). According with two features (price and sentiment score) (Fig. 7) the decision trees are expected to be more ample, because of the new feature, and they will also use the same splitting scheme (first 90% for training and last 10% for testing).

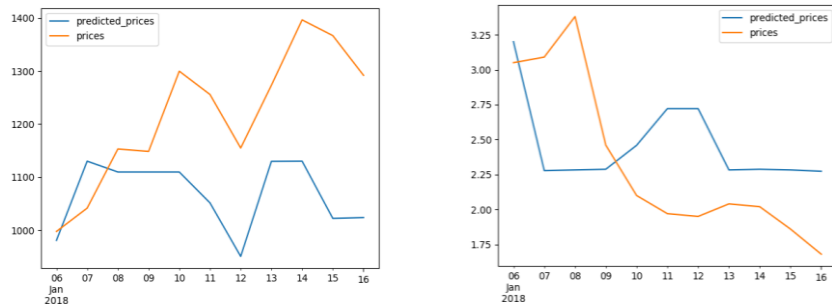


Fig. 7. The SA evolution for prices and predicted prices

The graphs above represent the output of the Random Forest, aided by Sentiment Analysis module. The orange line is the price evolution and the blue one represents the prediction of prices. The first chart is a prediction for the Ethereum data set, and the second graph was made on the Ripple data.

4 Results

From the entire set of data presented above, we looked at Precision, Recall and F-measure performance measures to evaluate our model for the prediction problem proposed in this paper (Table 4).

Table 4. The predictions measures of three models

Coin	Model	Precision	Recall	F-measure
BTC	Random Forest	75.40%	65.37%	66.37%
	Linear Regression	62.45%	60.08%	61.24%
	SA with Random Forests Development	75.23%	73.67%	74.44%
ETH	Random Forest	72.11%	60.33%	61.69%
	Linear Regression	62.71%	55.11%	56.85%
	SA with Random Forests Development	72.12%	71.55%	71.83%
XRP	Random Forest	68.23%	61.42%	61.32%
	Linear Regression	59.89%	54.78%	55.82%
	SA with Random Forests Development	68.45%	67.01%	67.72%

The results are promising. Still, we observe that through SA we gain better results, understanding that the feeling over media really influence the flow of the market. Furthermore, Random Forest works better using two features, not only one. The Linear Regression model had the worst performance on prediction, using this kind of data.

5 Discussion

This section describes the study results by running a series of three algorithms (Random Forests, Linear Regression, and SA with Random Forests Development). Our aim in these experiments was to compare three known techniques. In addition, our final goal was to define a new formula (3), by adding new missing data, in order to predict the fluctuation of a cryptocurrency. For this purpose, we compare the Precision and Recall measures.

As we observed the graphs of prediction, the nodes do not match the price very accurate. However, there's no need in matching exactly, more important is to predict the rise and fall of the price in the right time. The results showed that in the Ethereum prediction chart, a down spike occurred in the both lines at 12th January 2018. This reveals that is a great time to invest, and that the day before is perfect to sell the cryptocurrencies for profit. The behavior of the Ethereum prediction went well because the text data set is more than double then Ripple's media coverage. These points out the influence of the media on the investor's opinion on buying and selling on the cryptomarket. Only one problem we face: the lack of trusted media to train our algorithms.

6 Conclusions

This paper describes how machine learning algorithms and sentiment analysis of the afferent media are affecting the prediction process of cryptocurrency market. This method could become very helpful for entrepreneurs and people who would like to invest in such market and need a promising start.

Trying to predict a young market is not an easy task. The shortage of data is a big challenge. There are not many sites writing about cryptocurrency, making difficult the process of trusting. On the other hand, the multitude of unrefined sources (blogs, forums, twitter) gives us a hard time to identify feelings in comments; various heuristics need to be written in the code. Different sites may write about the same story, but can forge different feelings (humans tend to be subjective). The Efficient Market Hypothesis (EMH) seems to work here in some points of time, cryptocurrencies jumping on new levels of prices, becoming more and more valuable in a short time. This is the reason why the data is scraped only of the last three months and why Random Forest is used as the main algorithm (randomness seems to work).

Until the market will get in a mature phase and media will be more and more involved in the cryptomarket, the best approach in solving the gap problem is to approximate the sentiment score, using a formula. Therefore the algorithm will better tuned, even though, big spikes (low or high) in the graph (stock or cryptocurrencies) still remain a hard problem to solve.

Acknowledgments: This survey was published with the support of the grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI – UEFISCDI, project number PN-III-P2-2.1-BG-2016-0390 and by a grant of the Romanian Ministry of Research and Innovation CCCDI-UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818 within PNCDI III.

References

1. Barber, B. M., and Odean, T.: All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. In: *Review of Financial Studies*, 21, 785–818 (2008).
2. Bariviera, A. F., Zunino, L., Guercio, M. B., Martinez, L. B., and Rosso, O. A.: Revisiting the European sovereign bonds with a permutation-information- theory approach. *Eur. Phys. J. B.* 86: 509. doi:10.1140/epjb/e2013-40660-7 (2014).
3. Clarke, J., Jandik, T., and Mandelker, G.: The efficient markets hypothesis. In: *Robert C. ARFFA, ed. Expert Financial Planning: Investment Strategies from Industry Leaders*. New York: Wiley, Chapter 9, pp. 126-141 (2001).
4. Dai, Y., and Zhang, Y.: *Machine Learning in Stock Price Trend Forecasting*. Stanford University (2013).
5. Daniel, K., Hirshleifer, D., and Subrahmanyam, A.: Investor psychology and security market under- and overreactions. In: *Journal of Finance*, 53, 1839–1885 (1998).
6. Fama, E.: The Behavior of Stock Market Prices. *Journal of Business*. 38: 34–105. doi:10.1086/294743 (1965).

7. Feller, W.: Martingales. In: *An Introduction to Probability Theory and Its Applications*, Vol. 2, New York: Wiley, pp. 210-215 (1971).
8. Gîfu, D. and Cristea, D.: Public discourse semantics. A method of anticipating economic crisis presented at the Exploratory Workshop on Intelligent Decision Support Systems for Crisis Management, 8-12 May 2012, Oradea, Romania. In: *International Journal of Computers, Communications and Control*, see, I. Dzitac, F.G. Filip, M.-J. Manolescu (eds.), vol. 7/5, Agora University Editing House, pp. 829-836 (2012).
9. Granger, C. W. J.: Forecasting stock market prices: Lessons for forecasters. In: *International Journal of Forecasting* 8, North-Holland, 3-13 (1992).
10. Hilbert, A., Jacobs, H., and Müller, S.: Media Makes Momentum. *Review of Financial Studies*, 27(12), 3467–3501 (2014).
11. Hou, K., Peng, L., and Xiong, W.: A Tale of Two Anomalies: The Implications of Investor Attention for Price and Earnings Momentum, SSRN 976394 (2009).
12. Khaidem, L., Saha, S., and Dey, S. R.: Predicting the direction of stock market prices using random forest. In: *Applied Mathematical Finance*, 1-20 (2016).
13. Marwala, T.: Impact of Artificial Intelligence on Economic Theory – via arXiv.org (2015).
14. Marwala, T., and Hurwitz, E.: *Artificial Intelligence and Economic Theory: Skynet in the Market*. London: Springer (2017).
15. Mittal, A. and Goel, A.: *Stock prediction using twitter sentiment analysis*. Stanford University, CS229 (2012).
16. Nelson, R.: *Prophecy: A History of the Future - The Rex Research Civilization Kit*, <http://www.rexresearch.com/prophist/phfcon.htm> (2000).
17. Nunno, L.: *Stock Market Price Prediction Using Linear and Polynomial Regression Models* (2017).
18. Spierdijk, L., and Bikker, J. A.: *Mean Reversion in Stock Prices: Implications for Long-Term Investors* (2012).
19. Zunino, L., Bariviera, A. F., Guercio, M. B., Martinez, L. B., and Rosso, O. A.: On the efficiency of sovereign bond markets. *Phys. A Stat. Mech. Appl.*, 391: 4342–4349. doi:10.1016/j.physa.2012.04.009 (2012).
20. Xin, Y.: *Linear Regression Analysis: Theory and Computing* (2009).