Received: 27 April 2019 | Revised: 1 September 2019 | Accepted: 25 October 2019

DOI: 10.1111/exsy.12493

ORIGINAL ARTICLE

Expert Systems WILEY

Advanced social media sentiment analysis for short-term cryptocurrency price prediction

Krzysztof Wołk

Department of Multimedia, Polish-Japanese Academy of Information Technology, Warsaw, Poland

Correspondence
Krzysztof Wołk, Polish-Japanese Academy of Information Technology
Koszykowa 86, 02-008, Warsaw, Poland.
Email: kwolk@pja.edu.pl

Abstract

In recent years, the scrutiny of bitcoin and other cryptocurrencies as legal and regulated components of financial systems has been increasing. Bitcoin is currently one of the largest cryptocurrencies in terms of capital market share. Therefore, this study proposes that sentiment analysis can be used as a computational tool to predict the prices of bitcoin and other cryptocurrencies for different time intervals. A key characteristic of the cryptocurrency market is that the fluctuation of currency prices depends on people's perceptions and opinions, not institutional money regulation. Therefore, analysing the relationship between social media and web search is crucial for cryptocurrency price prediction. This study uses Twitter and Google Trends to forecast the short-term prices of the primary cryptocurrencies, as these social media platforms are used to influence purchasing decisions. The study adopts and interpolates a unique multimodel approach to analyse the impact of social media on cryptocurrency prices. Our results prove that people's psychological and behavioural attitudes have a significant impact on the highly speculative cryptocurrency prices.

KEYWORDS

cryptocurrencies, machine learning, sentiment analysis, social media, speculative models

1 | INTRODUCTION

Bitcoin (BTC), Ethereum (ETH), Electroneum (ETN), Ripple (XRP), ZEC Cash (ZEC), and Monero (XMR), crypto in short, are known as cryptocurrencies, an electronic form of currency used in digital transactions. Crypto is a decentralized form of currency transaction that takes place without an intermediary and can be circulated in the market through peer-to-peer networks. The crypto system was introduced to the market in 2008 by Satoshi Nakamoto (as the Bitcoin project) (Nakamoto, 2008). Crypto is different from traditional forms of currency transaction within the banking system as users engage in transactions without operation fees or any rules, regulations, or restrictions imposed by financial institutions that are often rife with fraud and corruption.

Bitcoin and other cryptocurrencies are among the fastest growing forms of digital transactions worldwide. At the beginning of 2017, the value of a bitcoin was $863 but it rose to around $17,000 by the end of the same year, which is an increase of approximately 2,000%. This massive and unprecedented rise captured global attention concerning digital currency transactions. Based on previous research, bitcoin possesses distinct characteristics not shared by traditional modes of currency transaction. This is because its price fluctuation depends on people's perception and opinions rather than institutional regulations. However, the value of crypto is volatile, which makes it a risky currency option.

Twitter is one of the most widely used social media platforms and reflects diverse perspectives from users worldwide. Twitter is a marketing tool for crypto because public discussion on cryptocurrencies is ongoing on social media channels. This chatter can be monitored and used to predict cryptocurrency prices based on public sentiment. Another website used for this purpose is Google Trends, a web search tool and prominent research platform. Google Trends provides data composed of relative search volume scores for a given search term during a given time interval.

Several researchers have used Google Trends data to predict the stock market (Alessandretti, ElBahrawy, Aiello, & Baronchelli, 2018; Nardo, Petracco-Giudici, & Naltsidis, 2016). This study analyses the correlation between the number of Tweets and crypto prices. In addition, this study examines the effect of web search data on crypto prices.

According to previous research, sentiment analysis is a useful technique for modelling the capital market and cryptocurrencies. Therefore, we can use sentiment analysis to predict crypto price changes at various time intervals using different computational and statistical models such as linear regression, boosting methods, and neural networks. Thus, we verify the significance of the coefficient of determination from Twitter and Google Trends data. Applying linear modelling to the number of Tweets and Google Trends data allows us to accurately predict the direction of crypto price changes (Abraham, Higdon, Nelson, & Ibarra, 2018).

To establish the usefulness of the data, we analyse only data that contain a certain set of keywords (cryptocurrency in full and its abbreviation). The underlying assumption is that sentiment correlates with the movement of the financial instrument, in this case, bitcoin. There is solid research to suggest this correlation exists. Many searches for bitcoin or associated keywords could indicate a reaction to current events or predict a future event.

The usefulness of sentiment analysis of Twitter and Google Trends concerning the price of bitcoin depends on whether there is a correlation between Twitter data and crypto price fluctuations. Additionally, can the prediction of a naive model for sentiment changes yield better output than random accuracy?

2 | PREVIOUS RELATED WORK

This study is based on broad research. Behavioural economists articulated that decisions on financial systems are influenced by emotional ethics and not by capital value alone. Dolan and Edlin (2002) supported this idea and argued that decision making is influenced by emotions. Therefore, tools such as sentiment analysis are useful in showing that the price of a commodity is impacted by values such as emotions as well as economic fundamentals (Panger, 2017).

Recent research noted that individuals' purchase decisions are influenced by the information found on websites and social media. Gallen Thomas' study showed that Twitter data analysis correlates with people's views on the price of cryptocurrency. The author also noted that social media sentiments have a significant impact on the emotional state of the final users of cryptocurrency. Bollen, Mao, and Zeng (2011) studied Twitter sentiment towards the stock market using neural networks and causality analysis to predict the cryptocurrency prices. The authors' results demonstrated the method's ability to predict changes in capital markets for almost 1 week.

Another study by Prosky, Song, Tan, and Zhao (2017) used tensor networks to formulate a model for learning. The study conducted sentiment analysis using Twitter data to verify a learning model and to investigate the relationships with various other stochastic events. Rather, Agarwal, and Sastry (2015) developed a hybrid model composed of a recurrent neutral network and a multiple linear regression and attempted to overcome the limitations associated with each. According to the researchers, these three methods are useful for the prediction of cryptocurrency prices, particularly bitcoin.

Nie and Ji (2014) showed that Twitter comments and Google Trends data were significant sources of information when predicting the prices of bitcoin and other cryptocurrencies. The authors found that social media data produced highly significant results with a low p value for search terms related to mining and blockchain, which are important aspects of cryptocurrencies (Stenqvist & Jacob Lönnö, 2017).

Karalevicius, Degrande, and De Weerdt (2018) used sentiment analysis of social media forums to predict intraday bitcoin prices and concluded that short-term price fluctuations could be predicted with some degree of accuracy, which diminished as time increased. This is significant for our research project because our timeframes are short, mostly 10 or 60 min long. What is more, we focus on many cryptocurrencies not only on the bitcoin. In addition, they apply lexicon-based (Jurek, Mulvenna, & Bi, 2015) sentiment analysis, whereas we propose machine learning based hybrid model. Garcia and Schweitzer (2015) showed that it is possible to use a combined strategy to predict bitcoin price using standard financial modelling techniques and social media signals. These signals included target words, sentiment, and other features that describe the changing environment of social media such as post frequency and comments. The researchers implemented a strategy that yielded a maximum of 32.29% daily gain, but with average of daily returns above 0.3%. Valence measures alone yielded a 0.1183% daily gain. With sufficient capital, at these rates, trading bitcoin could be profitable. The authors back tested their results, which lends credibility to their prediction model.

Kristoufek (2013) prediction model showed that Google Trends is one factor affecting the price of bitcoin, and Google Trends has a low p value with highly significant results. The study also used a vector autoregression technique that showed that Wikipedia information was also a good predictor that could be applied to produce a fair model for predicting bitcoin prices. Collected tweets related to the price of bitcoin and formulated a model that was useful in predicting the price of bitcoin (Kim et al., 2017). The study used Valence Aware Dictionary and sEntiment Reasoner to analyse the effect of each tweet and classify them as either positive, negative, or neutral. The study only retained tweets classified as negative or positive in their analysis.

3 | METHODS

In this study, we applied different predictive and descriptive models that are important for data analysis. The work used two predictive models that are essential for predicting the price of cryptocurrencies using Twitter sentiments and Google Trends data; that is, the least square linear regression (LSLR) and Bayesian ridge regression models. These models are embedded in Python language library SKLEARN.[1] This model is extensively discussed in Kuchibhotla, Brown, Buja, George, and Zhao (2018), who argued that highly dimensional data and methods have proliferated throughout the literature over the last two decades.

When data are expressed as a linear combination of a product of independent variables and a coefficient matrix, LSLR minimizes any necessary error that occurs. We use an array of independent and dependent variables to determine the coefficients. We express the relationship between the predicted value *Y* based on the coefficients and the inputs of the array *X* as the following:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{X}_j \hat{\beta}_j.$$

To calculate the better coefficient matrix, we use the following formula:

$$\beta = \mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}.$$

Another important technique used in our analysis is the Bayesian ridge regression modelled by Nie and Ji (2014), who claimed that future learning refers to learning to transform raw data into useful and analytical data for various purposes. Feature learning techniques can be either supervised or unsupervised, which commonly include auto-encoders, dictionary learning, restricted Boltzmann machine, and *k*-means among other approaches. Over the past few years, the restricted Boltzmann machine received increasing attention from researchers due to its ability to handle different types of data and its efficient learning method.

Bayesian ridge regression is similar to LSLR; however, it adds a lambda parameter to the input values that penalizes the beta coefficients and shifts them toward zero. Bayesian ridge regression returns a probabilistic model with a Gaussian parameter. MacKay (1992) described the Bayesian model with a Gaussian probability parameter in the following equation:

$$p(\lambda) = N\left(\alpha, \lambda^{-1} I_p\right).$$

Using the precision effects of the Gaussian parameter, we choose alpha and lambda for gamma distribution. To examine the default parameter in the model for alpha and lambda, we use 10−6. These values can be adjusted to the modelling data using the SKLEARN package. Bayesian ridge regression assigns coefficient values using the equation:

$$\beta = X\left(X^T X + \lambda I\right)^{-1} X^T Y,$$

where *I* resembles the identity matrix and the lambda term is applied across only the diagonal elements of the input array.

We also employed boosting algorithms, specifically, AdaBoost and gradient boosting. Typically, these boosting algorithms work by minimizing the error. The following equation illustrates this procedure:

$$E_T = \sum\nolimits_i E(f_{t-1}(x_i) + \alpha_t h(x_i)),$$

where *E* is the error during each iteration, and $\alpha * h(x)$ is the weak learner for the classifier function. Each result is also weighted. When implementing gradient boosting, the model applies steepest descent (or gradient descent) updating the model by computing the derivative of the residuals (loss) and a multiplier.

$$r = \frac{dL(y_i, F(x_i))}{dF(x_i)}.$$

$$m = \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + m\, h_m(x_i).$$

[1] http://scikit-learn.org/

TABLE 1 Model results for Bitcoin

Models	ME	R^2	T.s.
Support vector regression	1,357.482	.706722	−384.664
Stochastic gradient descent	8,706.509	.684718	−254.258
Gradient boosting model	1,370.471	.704768	−209.353
Multilayer perceptron neural network	1,382.774	.703728	−117.398
Least squares linear regression	2,000.951	.685587	395.1489
AdaBoost	1,986.594	.676574	−5.40525
Bayesian ridge regression	1,234.013	.720673	48.95157
Decision tree	8,791.874	.678843	359.0313
ElasticNet	2,280.217	.769856	313.6858
Hybrid (mean)	498.6117	.94169	151.6282

The information regarding crypto was retrieved from a web-based platform known as Crypto Compare,[2] which provides historical prices for various cryptocurrencies. During data processing, data were time indexed, concatenated, and averaged. Various data models, as detailed in our results section, were employed to make predictions.

As it was discussed earlier, we used the Valence Aware Dictionary and sEntiment Reasoner[3] sentiment analysis tool, which is remarkably sensitive to nuances in the text, such as punctuation, capitalization, negation, and amplification of lexicon values. Data processing was the most time-consuming aspect of the research besides variable transformation. All variables were included in the final model, given that they all had at least moderate correlation coefficients and there was no logical reason to exclude them given their potential predictive ability.

We used a bagging method for many different models to generate the final prediction. We collected the results from different categories and either summed the averages or identified the probability of their occurrence. We found that having an ensemble method of learning was beneficial for error reduction in a particular model. Comparing the linear regression and the ensemble method, we found that the latter performed better. Mean error and correlation coefficients were our measures of fit, and the other measures of fit indicate potential profit.

The correlation coefficient R^2 was calculated from the testing data set. In practical application, the full set of data minus the final target value should be trained. Only the final point or the last unknown price value should be predicted. As we were only interested in knowing how the final predicted value differed from the actual value, we introduced another measure of fit called ±T or dT, the error from our target value in dollars. This is the most useful measure of fit and establishes the potential to be profitable when trading crypto.

4 | EVALUATION AND RESULTS

The models used were support vector regression (Smola & Schölkopf, 2004), stochastic gradient descent (Bottou, 2010), gradient boosting model (Friedman, 2002), multilayer perceptron neural network (Salinca, 2017), least squares linear regression (Wold, Ruhe, Wold, & Dunn, 1984), AdaBoost (Collins, Schapire, & Singer, 2002), Bayesian ridge regression (Hoerl & Kennard, 2000), decision tree (Kohavi, 1996), ElasticNet (Zou & Hastie, 2005), and Hybrid, which is the mean of all of the models. We ran each model on test data to gather the relative strength index and mean error values. Tables 1–6 include a measure of fit for each model and show the results of the different methods used for bitcoin, Electroneum, Ethereum, Monero, Ripple, and Zcash, respectively, where R^2 is the correlation coefficient and ±T is the actual error when predicting the price on a brand new data point (the final interval). We sampled in 10 and 60-min shifts initially, and we chose the 10-min shift for the experiments as it resulted in less error overall in the hybrid model.

We used the mean of the hybrid model for prediction. We added tweet frequency as a transformed variable and graphed it against the cryptocurrency prices. We found that tweets had a high inverse correlation with price: Unfavourable news led to an increase in tweet frequency. Figures 1–24 illustrate this finding, where we also compared Google Trends data with crypto data and tweet frequency data. The results imply a significant relationship between these entities.

5 | DISCUSSION

The analysis was performed by comparing sentiment analysis on Twitter data against crypto prices for a certain timeframe. Figures 1–24 show the comparison and experiments, performance of the models against the test data, results of these models for an interval of 10 min, and the

[2]https://www.cryptocompare.com/
[3]https://github.com/cjhutto/vaderSentiment

TABLE 2 Model results for Electroneum

Models	ME	R^2	T.s.
Support vector regression	0.004487	.946137	0.000601
Stochastic gradient descent	0.036371	.922049	−0.00082
Gradient boosting model	0.005131	.932071	0.000303
Multilayer perceptron neural network	0.008413	.935374	0.000488
Least squares linear regression	0.004657	.953642	0.001144
AdaBoost	0.009122	.927278	−0.00137
Bayesian ridge regression	0.006355	.898442	0.001081
Decision tree	0.033998	.952623	0.00044
ElasticNet	0.007629	.9364	−0.0008
Hybrid (mean)	0.001842	.99163	0.001

TABLE 3 Model results for Ethereum

Models	ME	R^2	T.s.
Support vector regression	75.8912	.964188	−17.1537
Stochastic gradient descent	364.9388	.966607	−12.644
Gradient boosting model	126.1168	.968592	−3.84195
Multilayer perceptron neural network	45.18472	.964776	−12.367
Least squares linear regression	126.2375	.963377	−3.38566
AdaBoost	73.22109	.969224	16.75794
Bayesian ridge regression	73.57855	.964501	−2.00044
Decision tree	615.5554	.961733	17.85768
ElasticNet	124.8832	.968191	−6.17257
Hybrid (mean)	16.02903	.994549	7.823218

TABLE 4 Model results for Monero

Models	ME	R^2	T.s.
Support vector regression	32.21549	.839515	−2.96132
Stochastic gradient descent	194.4944	.815937	−2.35604
Gradient boosting model	45.53463	.843566	−4.13285
Multilayer perceptron neural network	32.52362	.845643	8.301083
Least squares linear regression	26.08474	.862746	4.746065
AdaBoost	30.88287	.859943	2.887843
Bayesian ridge regression	50.84047	.84862	3.112188
Decision tree	196.3985	.859415	3.559582
ElasticNet	44.01616	.856563	−3.38396
Hybrid (mean)	13.79168	.978258	−8.01742

implementation of our model against the full data set for each cryptocurrency being studied. The last value in our predictions represents the true performance of the model when making a final prediction. Separately, we also show the results of the hybrid model. The best result gives an error of less than $6 when predicting the final price point; cryptocurrency volatility resulted in daily price swings that were much greater than our total error. This suggests that the model could be profitable.

TABLE 5 Model results for Ripple

Models	ME	R^2	T.s.
Support vector regression	32.21549	.839515	−2.96132
Stochastic gradient descent	194.4944	.815937	−2.35604
Gradient boosting model	45.53463	.843566	−4.13285
Multilayer perceptron neural network	32.52362	.845643	8.301083
Least squares linear regression	26.08474	.862746	4.746065
AdaBoost	30.88287	.859943	2.887843
Bayesian ridge regression	50.84047	.84862	3.112188
Decision tree	196.3985	.859415	3.559582
ElasticNet	44.01616	.856563	−3.38396
Hybrid (mean)	13.79168	.978258	−8.01742

TABLE 6 Model results for Zcash

Models	ME	R^2	T.s.
Support vector regression	32.21549	.839515	−2.96132
Stochastic gradient descent	194.4944	.815937	−2.35604
Gradient boosting model	45.53463	.843566	−4.13285
Multilayer perceptron neural network	32.52362	.845643	8.301083
Least squares linear regression	26.08474	.862746	4.746065
AdaBoost	30.88287	.859943	2.887843
Bayesian ridge regression	50.84047	.84862	3.112188
Decision tree	196.3985	.859415	3.559582
ElasticNet	44.01616	.856563	−3.38396
Hybrid (mean)	13.79168	.978258	−8.01742

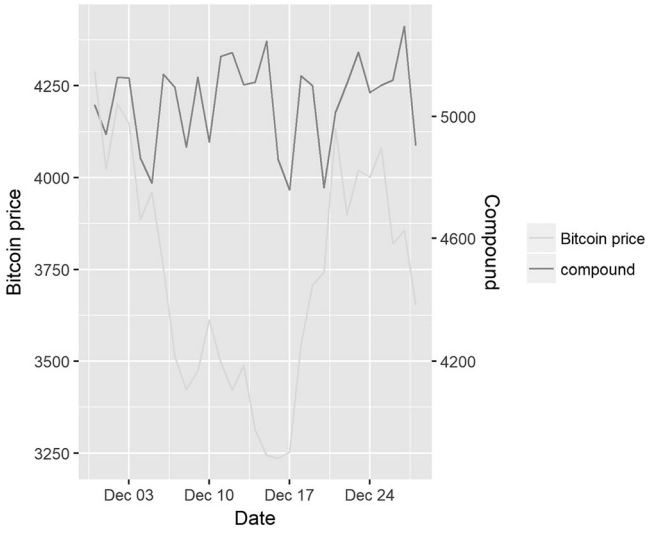

FIGURE 1 Bitcoin price versus number of tweets

Finally, we conducted an empirical experiment by trading $100 on the BitBay cryptocurrency exchange over 1 month. For this, we implemented a Python script that automatically gathered predictions every 10 min and took the recommended action if it was found profitable to buy, sell, or exchange cryptocurrency (after taking into account BitBay fees). Our bot conducted one to three transactions per day and, after a

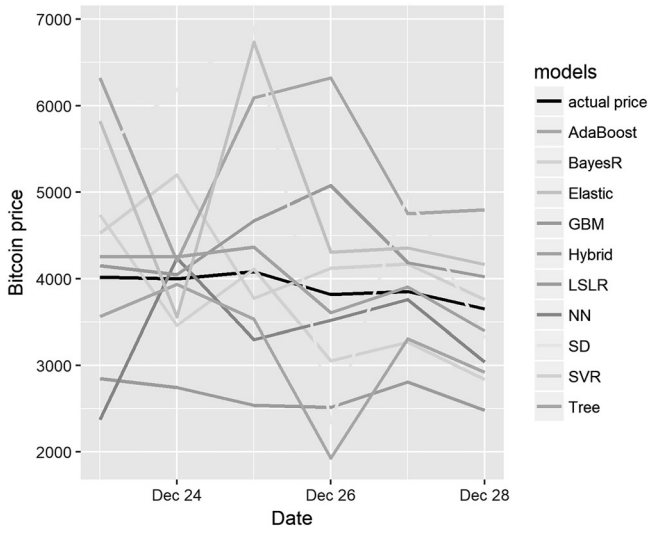

FIGURE 2 All models versus Bitcoin price. GBM, Gradient Boosting Model; LSLR, least square linear; NN, neural network regression; SD, Stochastic Gradient Descent; SVR, Support Vector Regression

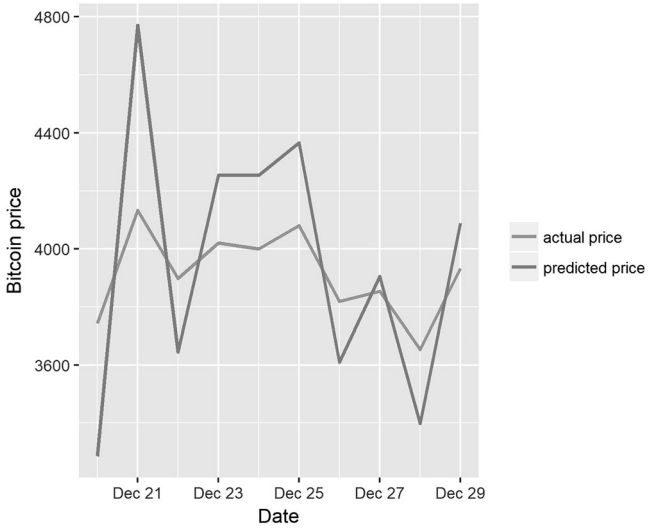

FIGURE 3 Bitcoin price versus predicted price

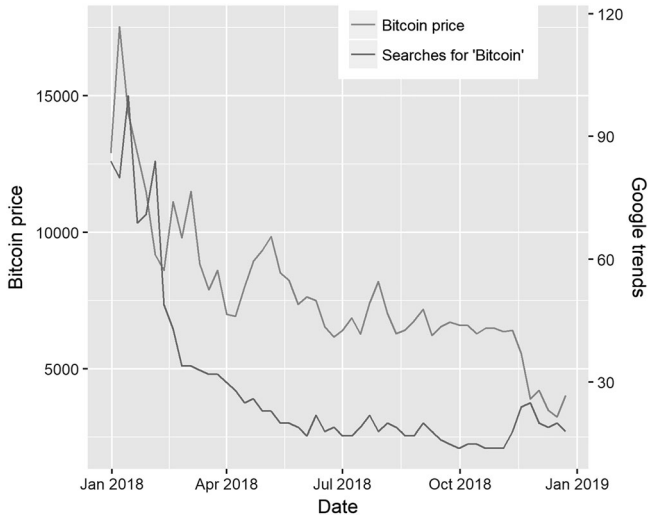

FIGURE 4 Bitcoin price versus Google trends

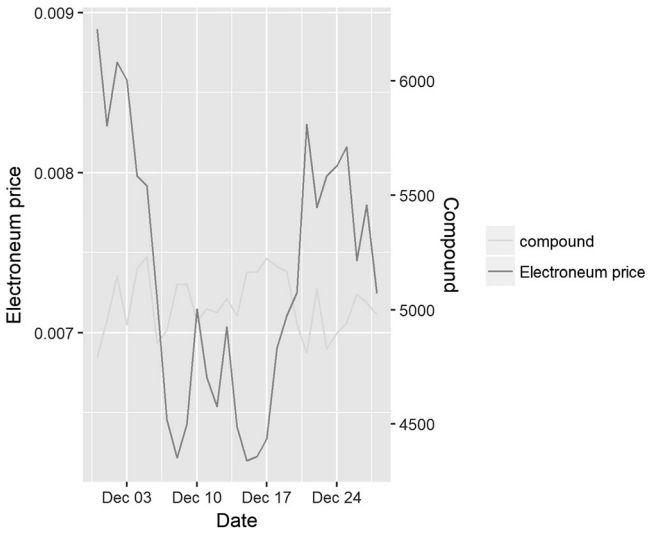

FIGURE 5 Electroneum price versus number of tweets

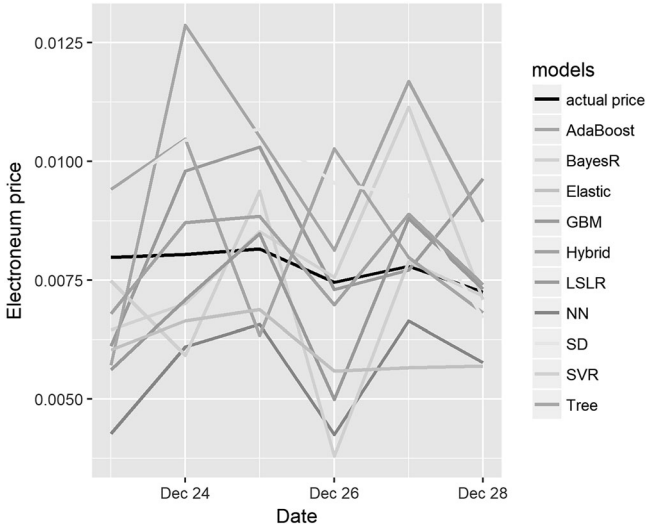

FIGURE 6 All models versus Electroneum price. GBM, Gradient Boosting Model; LSLR, least square linear regression; NN, neural network; SD, Stochastic Gradient Descent; SVR, Support Vector Regression

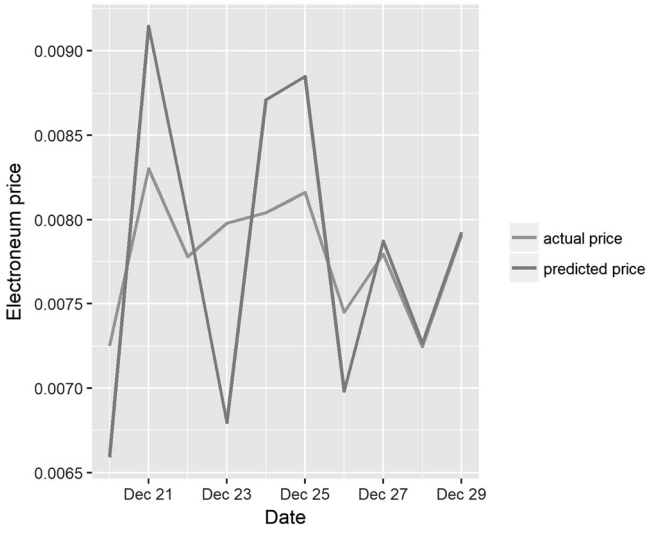

FIGURE 7 Electroneum price versus predicted price

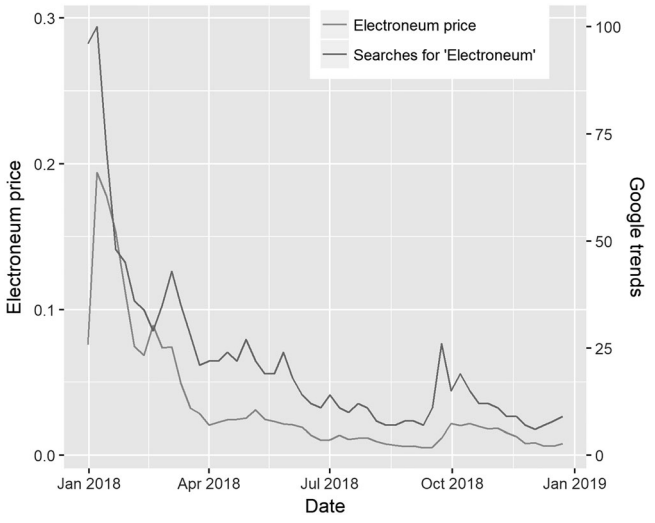

FIGURE 8 Electroneum price versus Google trends

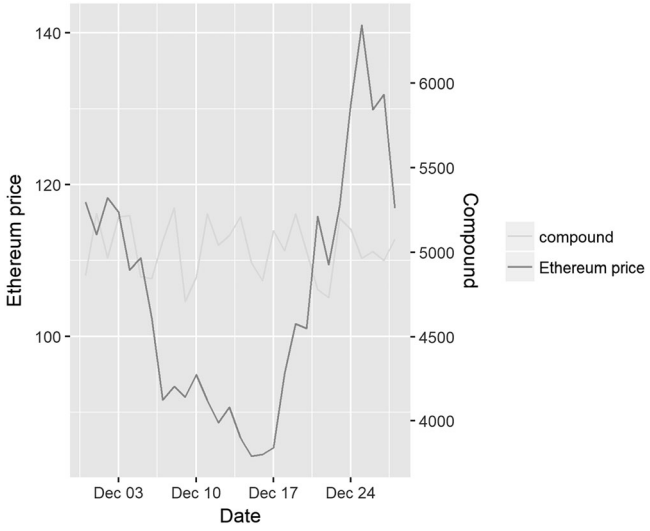

FIGURE 9 Ethereum price versus number of tweets

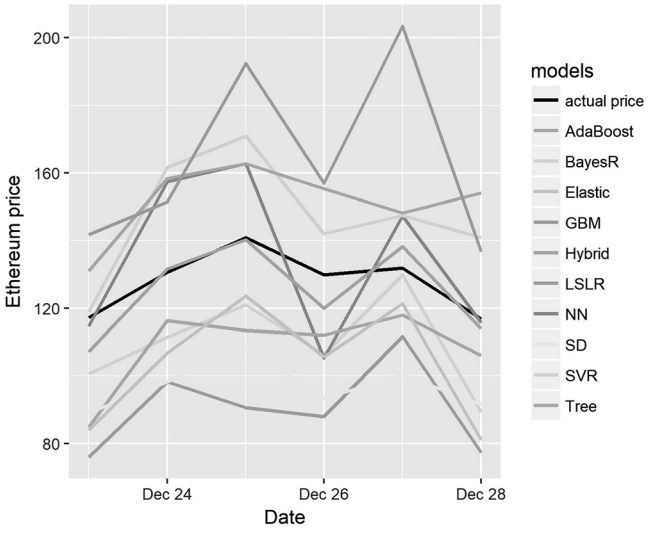

FIGURE 10 All models versus Ethereum price. GBM, Gradient Boosting Model; LSLR, **l**east square linear regression; NN, neural network; SD, Stochastic Gradient Descent; SVR, Support Vector Regression

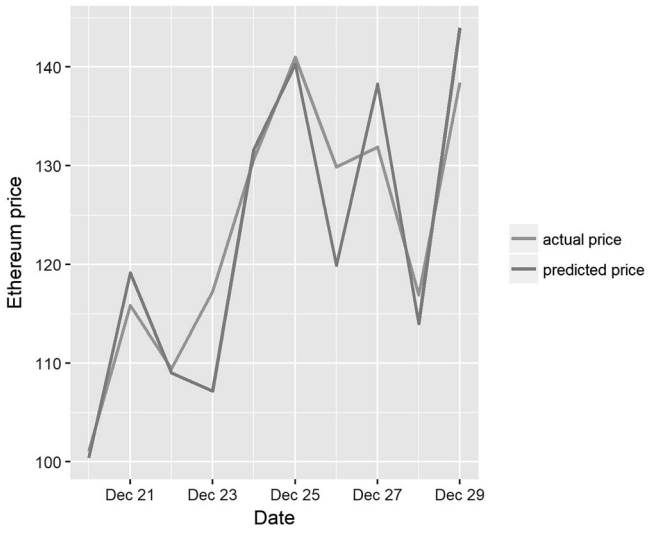

FIGURE 11 Ethereum price versus predicted price

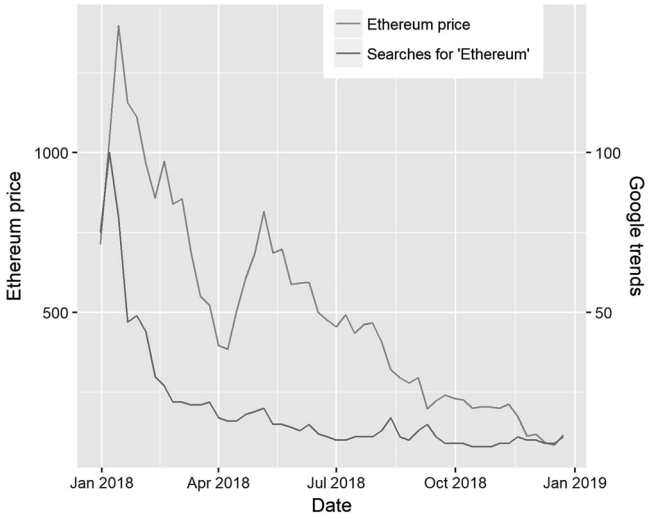

FIGURE 12 Ethereum price versus Google Trends

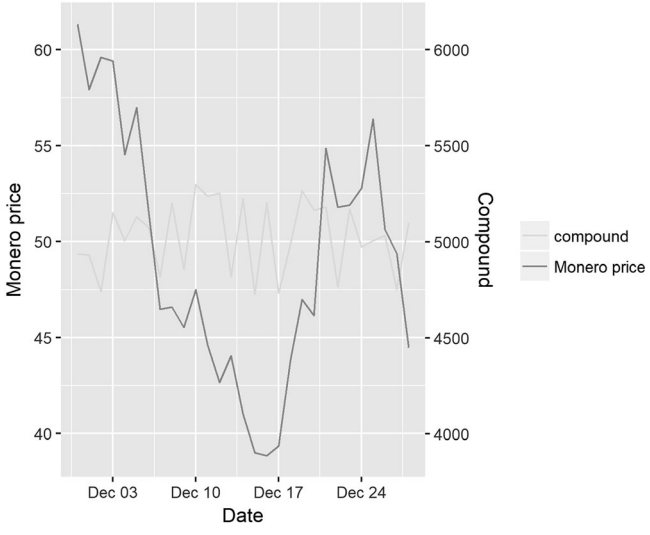

FIGURE 13 Monero price versus number of tweets

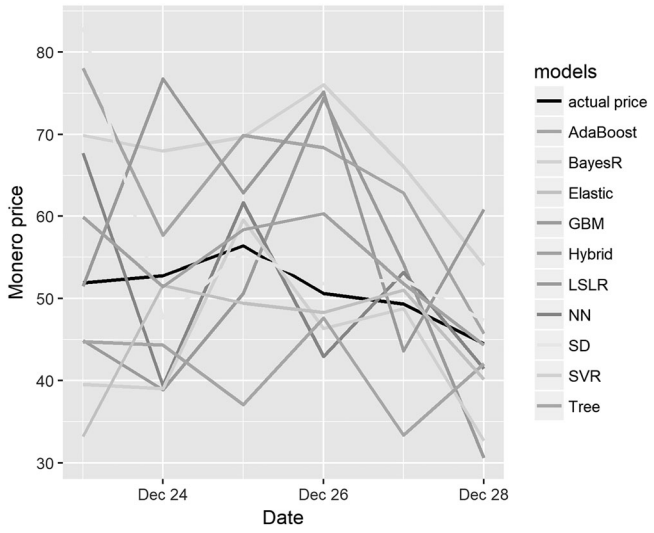

FIGURE 14 All models versus Monero price. LSLR, **l**east square linear regression

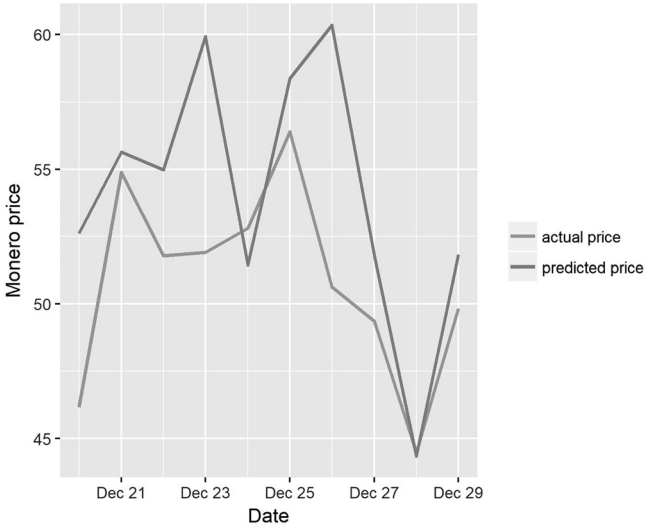

FIGURE 15 Monero price versus predicted price

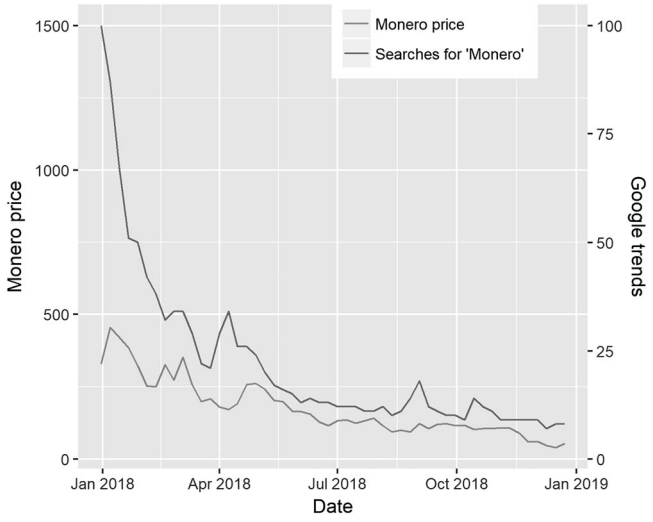

FIGURE 16 Ethereum price versus Google Trends

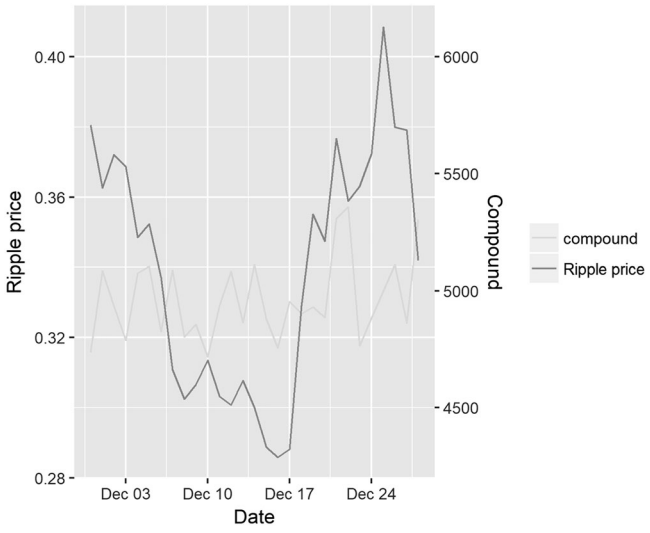

FIGURE 17 Ripple price versus number of tweets

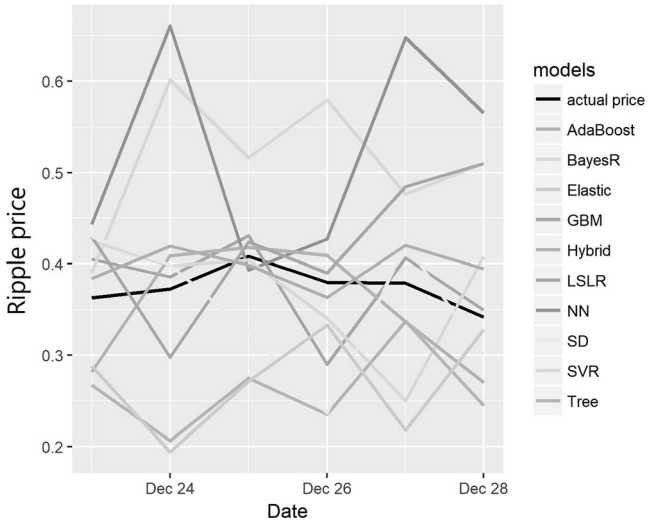

FIGURE 18 All models versus Ripple price. GBM, Gradient Boosting Model; LSLR, **least** square linear regression; NN, neural network; SD, Stochastic Gradient Descent; SVR, Support Vector Regression

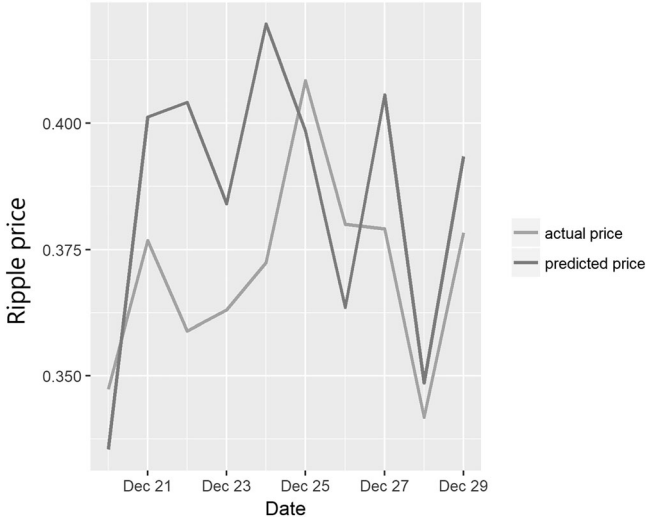

FIGURE 19 Ripple price versus predicted price

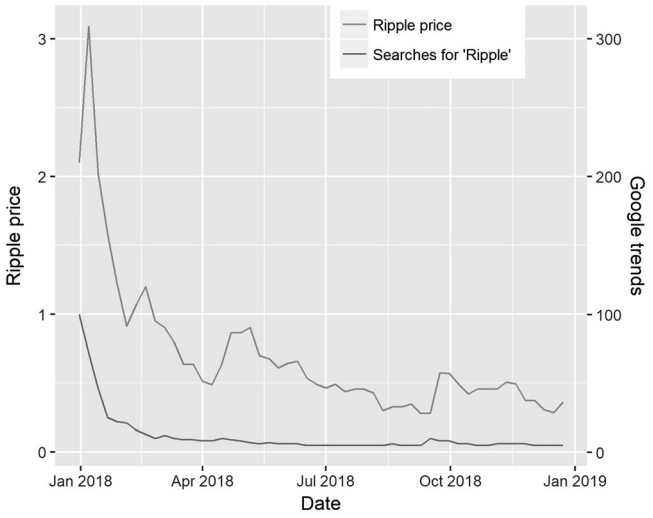

FIGURE 20 Ripple price versus Google Trends

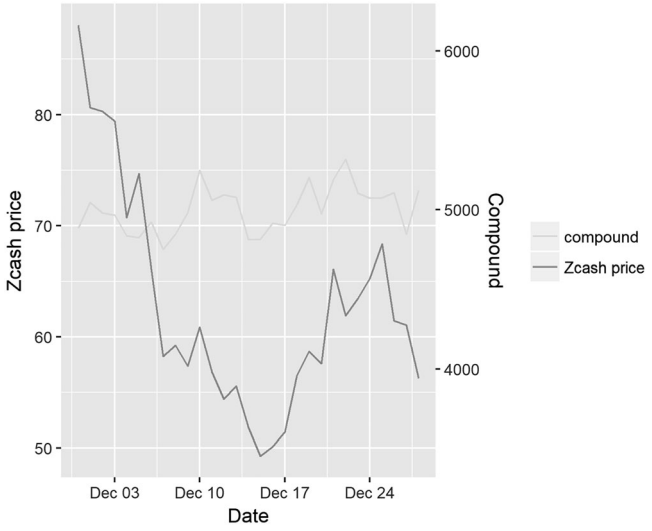

FIGURE 21 Zcash price versus number of tweets

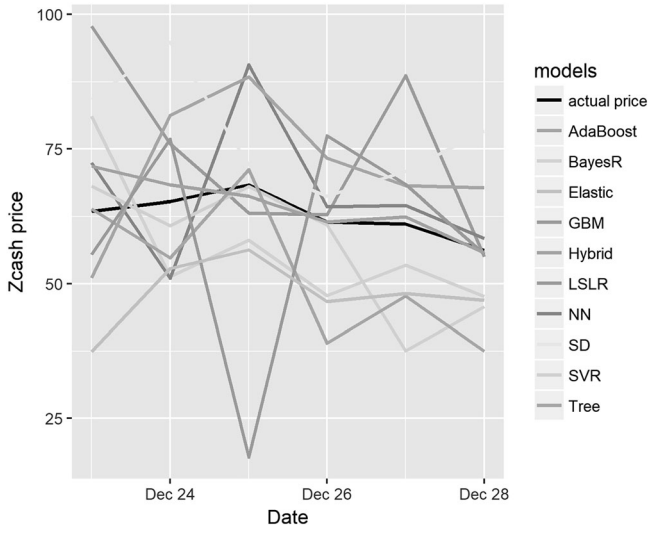

FIGURE 22 All models versus Zcash price. GBM, Gradient Boosting Model; LSLR, **lea**st square linear regression; NN, neural network; SD, Stochastic Gradient Descent; SVR, Support Vector Regression

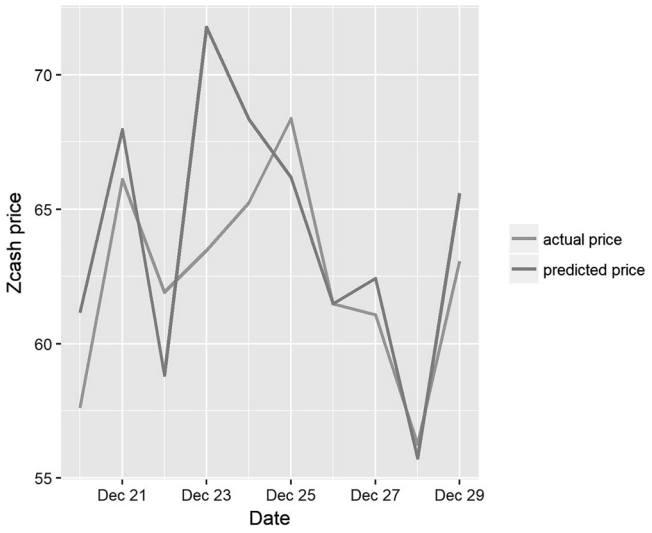

FIGURE 23 Zcash price versus predicted price

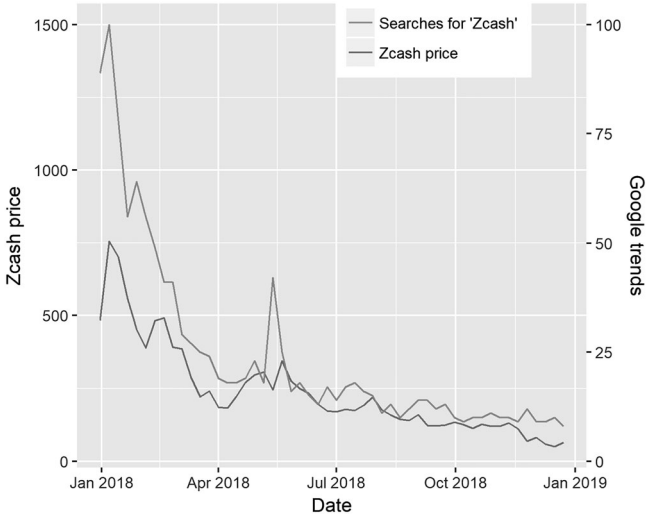

FIGURE 24 Zcash price versus Google Trends

month, our account balance stood at $114.82. In contrast, when we used KryptoBot, a well-known tool for cryptocurrency trading, we managed to convert $100 into $102.45 within the same period. Cryptocurrency is known to be highly volatile, and the testing period can heavily impact the performance. Our testing period was December 2018, when the prices were quite stable. If we would apply in this period, a basic buying and hold strategy, by buying crypto for $100 on December 1, and selling it on December 31, we would end with $92 using bitcoin, $117 using Ethereum, $94 using Ripple (XRP), and $71 using ZEC. This indicates that our proposed method is more profitable and stable, particularly considering that the cryptocurrency market has recently been down.

6 | CONCLUSIONS

We conclude that cryptocurrency price fluctuations depend heavily on social media sentiment and web search analytics tools such as Google Trends. Twitter sentiments regarding future cryptocurrency prices tend to be positive as many people tweet about cryptocurrencies even if their prices go down. However, it is difficult to predict cryptocurrency prices because of their volatile nature in the current market. Bank regulations, political risk, and regulatory agencies cause substantial fluctuations in the currency. Our results show that our hybrid model achieved consistently good results even with blind test data. We find a combination of Google Trends data and general negative sentiments (including weighted sentiments) to be the most powerful predictors; negative news carries a larger weight as shown by the correlation values during our data exploration phase. We, therefore, recommend a hybrid model to help alleviate some of the deficiencies of any one model and prove that people's psychological and behavioural attitudes have a significant impact on speculative cryptocurrency prices.

Finally, we have shared our solution as a Python tool on the GitHub repository. The script can perform several functions: It can be customized to any currency type; allows for custom windows to group tweets, average sentiments, and Google Trends data; allows a custom number of tweets to be extracted; connects to the Google Trends, CryptoCompare, and Twitter application programming interfaces to identify search trends, currency prices, and tweets with specified currency keywords; performs sentiment analysis on tweets using all the models described in this study; builds models using Google Trends and sentiment analysis; predicts future prices; gives recommendations for each model; and sums up the number of buy/sell/hold recommendations for the group.

ACKNOWLEDGEMENTS

None

CONFLICT OF INTEREST

None

ORCID

Krzysztof Wołk https://orcid.org/0000-0001-5030-334X

REFERENCES

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1.

Alessandretti, L., ElBahrawy, A., Aiello, L. M., & Baronchelli, A. (2018). Machine learning the cryptocurrency market. Available at SSRN 3183792.

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*, 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier, & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2604-3_16

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*, 253–285. https://doi.org/10.1023/A:1013912006537

Dolan, P., & Edlin, R. (2002). Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? *Journal of Health Economics*, *21*, 827–843. https://doi.org/10.1016/S0167-6296(02)00011-5

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Garcia, D., & Schweitzer, F. (2015). Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science*, *2*, 150288. https://doi.org/10.1098/rsos.150288

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*, 80–86. https://doi.org/10.2307/1271436

Jurek, A., Mulvenna, M. D., & Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, *4*(1), 1–13. https://doi.org/10.1186/s13388-015-0024-x

Karalevicius, V., Degrande, N., & De Weerdt, J. (2018). Using sentiment analysis to predict interday bitcoin price movements. *The Journal of Risk Finance*, *19*, 56–75. https://doi.org/10.1108/JRF-06-2017-0092

Kim, Y. B., Lee, J., Park, N., Choo, J., Kim, J.-H., & Kim, C. H. (2017). When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PLoS ONE*, *12*, e0177630. https://doi.org/10.1371/journal.pone.0177630

Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining* (pp. 202–207). Portland: AAAI Press.

Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, *3*, 3415. https://doi.org/10.1038/srep03415

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., & Zhao, L. (2018). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. arXiv preprint, Computing Research Repository, arXiv:1802.05801.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, *4*, 415–447. https://doi.org/10.1162/neco.1992.4.3.415

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system, https://bitcoin.org/bitcoin.pdf

Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys*, *30*(2), 356–369. https://doi.org/10.1111/joes.12102

Nie, S., & Ji, Q. (2014). Feature learning using Bayesian linear regression model. 22nd International Conference on Pattern Recognition, IEEE, doi:https://doi.org/10.1109/ICPR.2014.267

Panger, G. T. (2017). Emotion in social media (Doctoral dissertation, UC Berkeley).

Prosky, J., Song, X., Tan, A., & Zhao, M. (2017). Sentiment predictability for stocks. arXiv preprint, Computing Research Repository, arXiv:1712.05785.

Rather, A. M., Agarwal, A., & Sastry, V. N. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, *42*, 3234–3241. https://doi.org/10.1016/j.eswa.2014.12.003

Salinca, A. (2017). Convolutional neural networks for sentiment classification on business reviews. arXiv preprint, Computing Research Repository, arXiv: 1710.05978.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Stenqvist, E., & Jacob Lönnö, J. (2017). *Predicting bitcoin price fluctuation with Twitter sentiment analysis*. (Bachelor's thesis). Retrieved from https://kth.diva-portal.org/smash/get/diva2:1110776/FULLTEXT01.pdf. (Order No. 12345)

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The collinearity problem in linear regression: The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*, 735–743. https://doi.org/10.1137/0905052

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x

AUTHOR BIOGRAPHY

Krzysztof Wołk received the Ph.D. degree in computer science engineering from the Polish-Japanese Academy of Information Technology, where he is currently an Associate Professor with the Cathedral of Multimedia. He conducts research related to natural language processing and machine learning based on statistical methods and neural networks and deep learning. He eagerly takes up IT challenges and engages in interesting interdisciplinary projects, in particular related to HCI, UX, medicine, and psychology. He was a Lecturer with the Warsaw School of Photography and as an IT Trainer. His specialties as a Teacher are primarily deep learning, machine learning, natural language processing, computational linguistics, multimedia, HCI, UX, mobile applications, HTML 5, Adobe applications, and server products from Apple and Microsoft. As far as, his didactic work is concerned, he leads classrooms with the Faculty of Computer Science, with the New Media Art Department, and with the Polish-Japanese Academy of Information Technology. He also used to lead classes and lectures at the Warsaw School of Photography and Graphic Design. He was the Technical Supervisor of B.A. and M.A. diplomas, a Diploma Reviewer, and a Reviewer for scientific conferences and journals. He is certified Microsoft, Apple, Adobe, w3schools, and EITCA specialist. He is the author of scientific monographs and specialized IT books related to the administration of servers and multimedia. He is also an Editor of the portal in4.pl, pclab.pl, and e-biotechnologia.pl portals and the author of training materials and guides. on p.11.

How to cite this article: Wołk K. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*. 2020;37:e12493. https://doi.org/10.1111/exsy.12493