Data-X @ UC Berkeley

Predicting Influence of News Articles on Cryptocurrency Prices using Sentiment Analysis

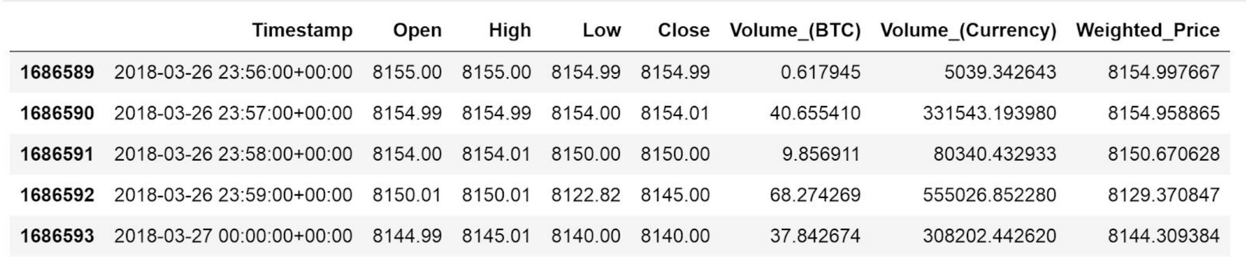
Sudarshan Gopalakrishnan, Parth Sanghvi, Johannes Kukula, Kenneth Choong, Srujay Korlakunta, Kshiti Bachlaus

1 Introduction

The project is to create a model that predicts the significance of an article, given its significance and the potential cryptocurrency trends it can lead to. The model is to serve as the back-end technology for Paradigm, a platform that automates Over-The-Counter (OTC) crypto trading for institutional crypto traders within a native chat application.

2 Data

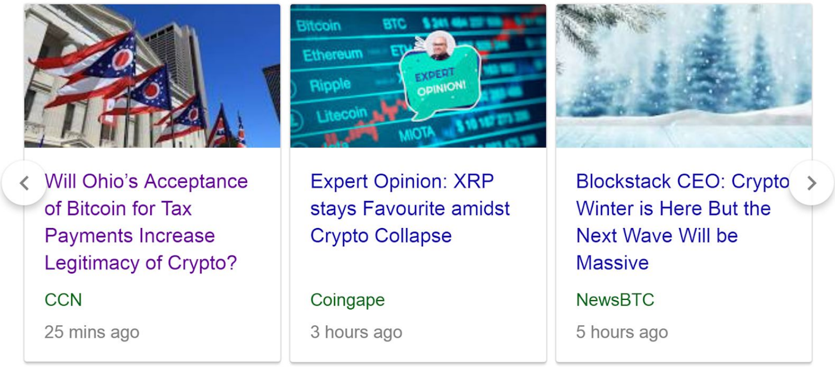
2.1 Minute by Minute Bitcoin Price data



	Timestamp	Open	High	Low	Close	Volume_(BTC)	Volume_(Currency)	Weighted_Price
1686589	2018-03-26 23:56:00+00:00	8155.00	8155.00	8154.99	8154.99	0.617945	5039.342643	8154.997667
1686590	2018-03-26 23:57:00+00:00	8154.99	8154.99	8154.00	8154.01	40.655410	331543.193980	8154.958865
1686591	2018-03-26 23:58:00+00:00	8154.00	8154.01	8150.00	8150.00	9.856911	80340.432933	8150.670628
1686592	2018-03-26 23:59:00+00:00	8150.01	8150.01	8122.82	8145.00	68.274269	555026.852280	8129.370847
1686593	2018-03-27 00:00:00+00:00	8144.99	8145.01	8140.00	8140.00	37.842674	308202.442620	8144.309384

- Source: Coinbase
- Information: Timestamp, open, high, low, close, volume, BTC price
- Timeframe: 4 years (2014-12-01 to 2018-03-27)

2.2 News Articles



Will Ohio's Acceptance of Bitcoin for Tax Payments Increase Legitimacy of Crypto?

CCN

25 mins ago

Expert Opinion: XRP stays Favourite amidst Crypto Collapse

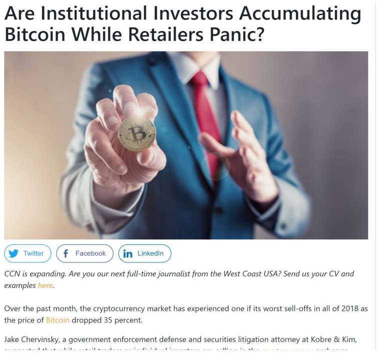
Coingape

3 hours ago

Blockstack CEO: Crypto Winter is Here But the Next Wave Will be Massive

NewsBTC

5 hours ago



Are Institutional Investors Accumulating Bitcoin While Retailers Panic?

CCN is expanding. Are you our next full-time journalist from the West Coast USA? Send us your CV and examples here.

Over the past month, the cryptocurrency market has experienced one if its worst sell-offs in all of 2018 as the price of Bitcoin dropped 35 percent.

Jake Chervinsky, a government enforcement defense and securities litigation attorney at Kobre & Kim,

We had 40,000 full news articles related to bitcoin that were scraped from the web . The articles were scraped from news website APIs and from sources such as CNBC, The Telegraph and Bitcoin Magazine to name a few. The information on each article gathered included the title of the article, the author, the source, the url, the content, and the timestamp.

2.3 Labelling the News Articles

To train our models using supervised learning, we needed to label our news articles as having made an impact on the Bitcoin price or not. All the articles are labeled with a sentiment of having a positive impact on price (+1), a neutral impact on price (0) or having a negative impact on price (-1). Over the course of the project, we used different methods and the following 3 sets of labelled data to train our machine learning models:

1. NTU Dictionary Corpus (Link)

   NTUSD-Fin provides various scoring methods including frequency, CFIDF, chi-squared value, market sentiment score and word vector for the tokens. Only the tokens appeared at least ten times and shown significantly difference between expected and observed frequency with chi-squared test are remained in our dictionary. The predetermined significance level is 0.05. The market sentiment score is calculated by subtracting the bearish PMI from the bullish PMI. There are 8,331 words, 112 hashtags and 115 emojis in the constructed dictionary, NTUSD-Fin[1].

2. 300 manually labelled articles (labeled as having a positive, neutral or negative impact)
3. Classified data labelled by previous team

3 Hypotheses

A hypothesis-driven exploratory data analysis was conducted in order to gain a better understanding of the problem statement and conduct further experimentation. The initial hypothesis was as follows:

- The Bitcoin market works like stock markets: if new information becomes available, prices adjust (more or less) immediately (Efficient Market Hypothesis, Fama 1998).

This hypothesis, however, has not proven to be true. As a result, we developed new hypotheses, moving to higher time lags and taking into account the sentiment of news instead of the “raw” information (as done in the first hypothesis):

- Instead, Bitcoin news reflects the overall sentiment of market participants on higher time lags.
- On higher time lags, Bitcoin sentiment predicts future prices to some extent.
- Since our sample shows persistent trends, we should be able to predict prices via autoregressive modeling (e.g. ARIMA, Box & Jenkins 1970).

Those three hypotheses proved to be true to some extent, enabling us to run the following strategy regarding our project task:

- We then can use a (weighted) average of the articles’ sentiments probabilities and predicted price movement from autoregressive models to get the scoring for each article.
  The scoring of each article should then reflect its likelihood of impact on price.

[1] Description of NTU Dictionary: Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen. 2018. NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications. In Proceedings of the 1st Financial Narrative Processing Workshop, 7 May 2018, Miyazaki, Japan.

However, since the ARIMA modeling showed that Bitcoin prices behave like a random walk, meaning that they are not predictable in an autoregressive way,, we developed our final modeling strategy (in the poster, it says we are using ARIMA in our final model. However, we did not use ARIMA in the final model due to the just mentioned argument):

In a first step, we predict sentiment of single news articles. In a second step we use the predicted sentiment in order to predict Bitcoin prices. In a last step, we translate the predicted prices into a scoring which should reflect the likelihood that a given article has a significant impact on Bitcoin price movement. The time frame to be used is daily Bitcoin price changes.

4 Methods

4.1 Sentiment Analysis using Recurrent Neural Networks

Long short-term memory (LSTM) in recurrent neural networks provide a more effective sentiment analysis, given that the LSTMs solve the vanishing gradient problem and hence are able to remember and process long-term contexts. Given the relative neutrality of financial articles, the ability to process long-term contexts would yield better results for executing sentiment analysis on our dataset.

We devised two different approaches to training the RNN model in order to understand each approach's sentiment predictive capacities.

> Approach 1: Training using Hand-Labelled Data
> A data-set of 300 articles (150 positive and 150 negative articles) was read and manually (by human) assigned sentiment in order to create a training set for supervised learning. RNN was then trained and the articles in our database were assigned a probability of sentiment between 0 and 1 using softmax regression, which were then translated to a qualitative sentiment.
>
> Approach 2: Training using NTU Word Corpus
> The RNN was trained using word sets for positive and negative sentiments in financial news articles and the dataset of articles was assigned sentiment. RNN was then trained and the articles in our database were assigned a probability of sentiment between 0 and 1 using softmax regression, which was then translated to a qualitative sentiment. The RNN will be then testing using a data-set of 300 articles (150 positive and 150 negative articles) will be read and manually (by human) assigned sentiment in order to create a training set for supervised learning.

Approach 1 had an accuracy of 78% for predicting the sentiment label in the article dataset created by the previous team. However, Approach 2 had a mediocre accuracy of 48%. Our key insights from this approach were that there was not a lot of overlap between the input dictionary and the text in the article. Furthermore, the use of a single token to train the RNN model did not provide enough context to create strong predictive capacity.

4.2 Trends Analysis for Prediction of Articles' Significance based on Bitcoin Price Trend

In order to predict price trends in the Bitcoin market, models based on decision trees (random forest, gradient boosting with decision trees) were chosen. Regarding the predictors of those models, we had to restrict them to features associated with news articles, since the predicted price trend should reflect the content in news articles and not be based on other factors unrelated news articles (e.g. properties of the time series, like mean, standard deviation, etc.). We chose the predicted sentiment of the last step as our main predictor, however, as an “insurance” against having low predictive accuracy in the predicted sentiment, we additionally used a vectorized word count of the top 100 words as an additional feature. Using those two type of features, we predicted Bitcoin price changes on a daily basis (that is, the price 24 hours after a given article was published compared to the price at the time when a given article was published) using random forest and gradient boosting.

5 Testing Through Scoring Indicator to Predict the Significance of Article on Price

In a final step, we used the min-max scaling of the absolute values of predicted prices from the previous step to create a scoring indicator. However, instead of using the minimum and maximum, we defined the values at the 3% percentile and 97% percentile of our distribution of the absolute of our predicted price changes as references to scale, making our approach more robust to outliers in the predictions.

The reason that we used the absolute values for the scaling is that this indicator should only reflect the likelihood of having a significant impact on price regardless of the direction of price change. The implicit assumption in using min-max scaling is that we define significance in terms of strength of price change. We multiplied the obtained values from min-max scaling by a factor of 10 in order to have our scoring in a range of 0 (meaning not significant at all) to 10 (meaning highly significant).

6 Results

We see a random pattern in a range from approximately -0.05 to 0.05. However, if we look at more extreme price changes above abs(0.05), we see a pretty strong correlation between predicted and actual price changes. This indicates that the model found a signal in news articles to predict daily price changes.

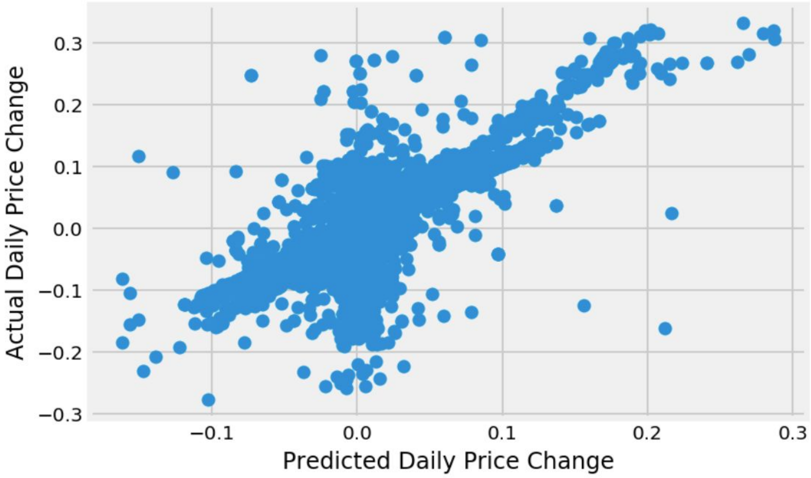
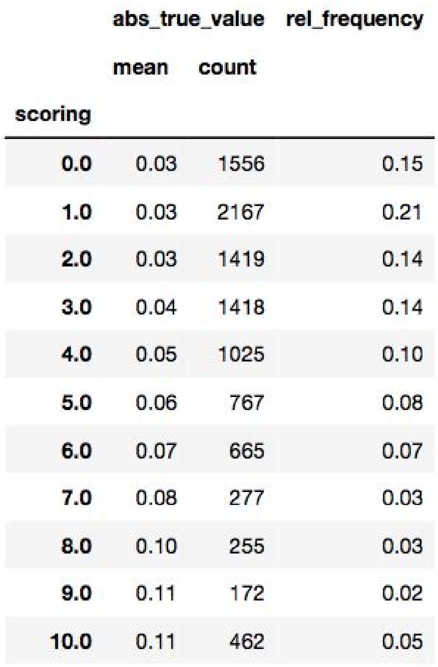


Fig. 1: Regression showing the predicted price changes on a test set of our best model (random forest)

Those predictions are then transformed using min-max scaling:



	abs_true_value		rel_frequency
	mean	count	
scoring			
0.0	0.03	1556	0.15
1.0	0.03	2167	0.21
2.0	0.03	1419	0.14
3.0	0.04	1418	0.14
4.0	0.05	1025	0.10
5.0	0.06	767	0.08
6.0	0.07	665	0.07
7.0	0.08	277	0.03
8.0	0.10	255	0.03
9.0	0.11	172	0.02
10.0	0.11	462	0.05

*Fig. 2: Table showing the mean absolute daily price change,
the frequency and relative frequency per scoring*

We see that the mean absolute price change increases as the scoring increases. Also, by looking at the relative frequencies, we see that approximately 50% of all articles have a scoring of 2 or lower, 82% of articles have a scoring of 5 or lower, and the articles with a scoring of 6 or higher have a total relative frequency of 18%.

7 Conclusion

The final model we produced gives a rating on an arbitrary scale to each article that is processed through it. The rating given to each article helps analysts and financial traders to sort through the plethora of articles posted every second in order to identify the articles that according to our model could significantly affect the price of cryptocurrency assets.

In addition to the significance of each article, the model also assigns the rating while accounting for the direction of price change. That is, the rating provided by the articles conveys whether the price of the Bitcoin cryptocurrency will change in the positive direction, or the negative.

The model was designed to be pipelined directly into the existing chat platform of our industry mentor, Paradigm. The chat platform allows users to trade based on script based technology. Our model, which is to be integrated along with a dashboard will allow Paradigm customers to benefit from real time information on news articles and their relevance to cryptocurrency assets.

8 Limitations and Next Steps

The fact that out RNN was trained using data from a previous team, also means that there are certain biases carried forward in our model. This was necessary because we required labeled datasets in order to have a semi-supervised approach.

If however, we had access to a labeled dataset which was labeled by professional analysts or some sort of a credible entity, there would be a greater reliance achieved by the model.
The final step in this project would be to create a model with a significantly high accuracy and reliance that will enable financial entities to eliminate the need for analysts to intervene and act upon the articles that are rated significantly by the model.

A completely reliant model such as this would enable non-professionals to understand the impact of news and sentiment simply based on the user interface of the product.

The model can further be prepared to deal with all articles available on the internet. The final model should then be able to sort relevant articles along with rating them. Also one of the assumptions we worked with this semester was that only articles related to cryptocurrencies have a relevant effect.

8 Acknowledgments and References

Mentorship of Paradigm, a platform that automates Over-The-Counter (OTC) crypto trading for institutional crypto traders within a native chat application.

Project built on top of pre-existing work from our mentor, Elias Castro's, Data-X Spring 2018 project.

Box, G. E. P, and GM Jenkins (1970) Time Series Analysis, Forecasting, and Control. *Francisco Holden-Day*.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance1. *Journal of financial economics*, *49*(3), 283-306

Seth, Shobhit. "Basics of Algorithmic Trading: Concepts and Examples." *Investopedia*, Investopedia, 22 Oct. 2018, www.investopedia.com/articles/active-trading/101014/basics-algorithmic-trading-concepts-and-examples.asp.

Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen. 2018. NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications. In Proceedings of the 1st Financial Narrative Processing Workshop, 7 May 2018, Miyazaki, Japan.