Using Twitter Data and Sentiment Analysis to Predict Future Values of Cryptocurrencies

Ryan Walker

Abstract—This paper presents numerical schemes for gathering, processing and correlating sentiment data to actual cost for a given crypocurrency over a given unit of time in the interest in finding a time-lagged correlation.

1 Introduction

Sentiment analysis techniques have been used for stock market predictions in the past with mixed success [1]. We are of the opinion that for a technique like this to work there are three major requirements.

1) High volume of sentiment source;
2) Strong correlation to trader action and community opinion, and
3) Low quantity of non-deterministic value changing artifacts: news reports, earnings, company announcements, etc.

In most cases, points one and three are mutually exclusive, meaning if there is a high volume of people talking about a stock publicly $\frac{100k+}{Day}$, the company will typically be publishing earning reports, posting news, etc... These are important for investors, but cannot be accurately modeled as they are considered random artifacts.

As there are still news artifacts regarding Cryptocurrencies, they are less common and typically have less of an impact because unless there is a major problem with the algorithm, they are mostly subjective, unlike an earning report or other financial documents.

The aim of this paper is to attempt to validate what has been listed above, and use the data to make further informed buys and sells.

1.1 Gathering Sentiment Data

The main source of sentiment data was from twitter. The Python module *tweepy* [2] was used to gather tweets and bin them into cryptocurrencies of interest. The volume was anywhere from $1.2k\frac{tweets}{hr}$ to $24k\frac{tweets}{hr}$ depending on the currency of interest. For this paper we were mostly interest in BTC, as the tweet volume is the highest; in the past five months, we have gathered over 21 billion tweets on Bitcoin alone.

Once the data is gathered it is put into a time series dataset with a period of 30 seconds - from that point Python *NLTK* (Natural Language Toolkit) [3] is used to perform sentiment analysis to rate each tweet and make a net sum per unit time, as shown in Figure 1.

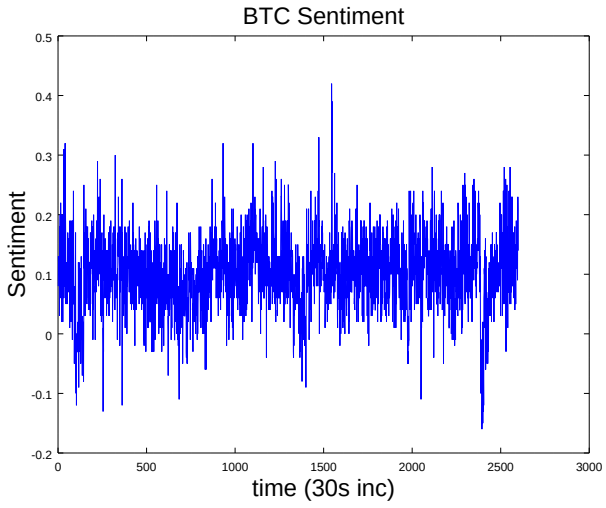

Fig. 1. Raw Sentiment, October 9th 2017

The filtered signal is shown in Figure 2 - it's possible to compare the sentiment and values time series.

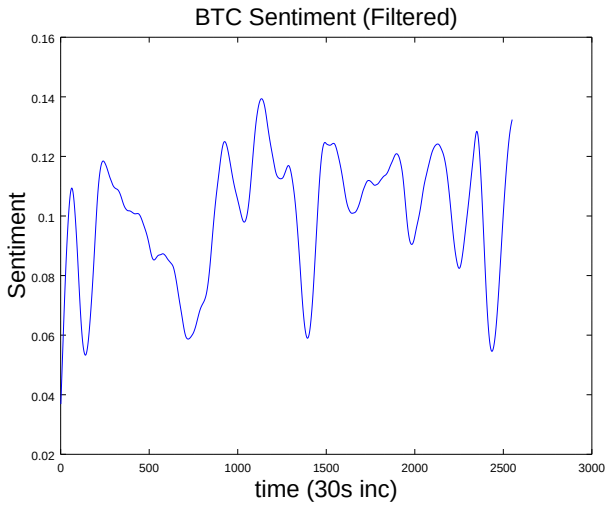

Fig. 2. Filtered Sentiment, October 9th 2017

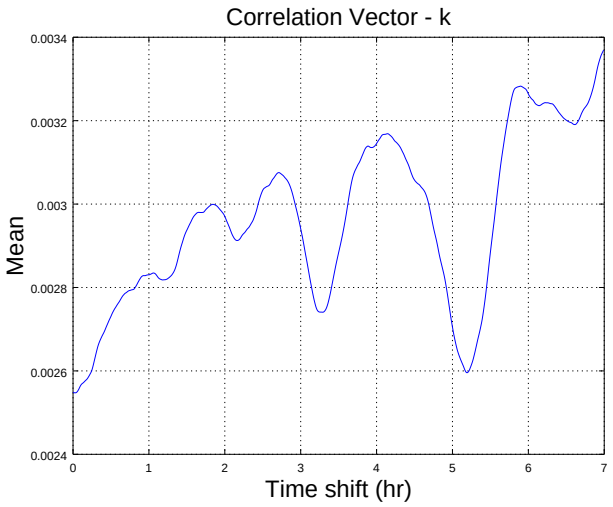

Fig. 3. Time Correlation vector k, October 7th 2017

1.2 Timewise Correlation

The timewise correlation was calculated using the correlation vector k,

$$k_j = \frac{\sum_{i=0}^{n} \dot{x}_i - \dot{y}_{i+j}}{n-j}, \quad j = 0, \ldots, n \qquad (1)$$

where x denotes the cost of the currency and y the sentiment.

The reasoning for using derivatives for the correlation is simply because the sentiment values are effectively a floating quantity – the magnitude of the sentiment is not believed to have any direct relationship with the magnitude of the cost. In other words we are looking for a correlation between the change in sentiment with the change in cost. This is justified by the success of our analysis

We look for local minima in of k as these points indicate the highest level of signal similarity. As seen in Figure 3, k has a global minima around 5.2hrs, which is defined as $\tau_L = 5.2hrs$ or the effective time lag between high values of changing sentiment and high values of changing cost.

We observe that the calculation of τ_L has been repeatable over a given unit of time. Figure 4 shows the vector sum of 15 k vectors from November 2nd to November 17th. In addition there are some interesting patterns appearing as local minima at $\sim 1.8hrs$ and $\sim 3.2hrs$.

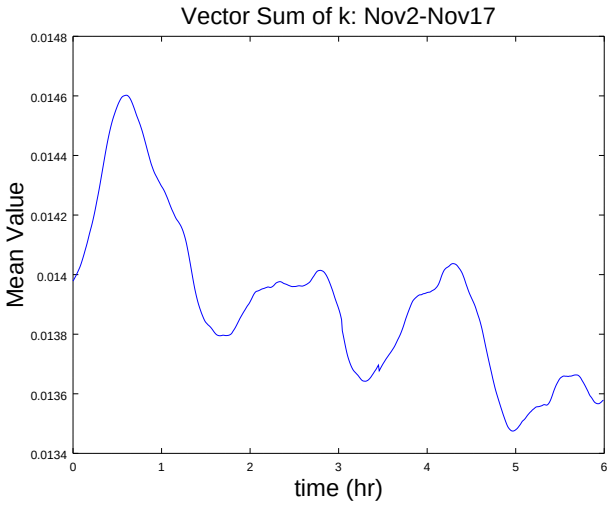

Fig. 4. Mean Time Correlation vector k, Nov2 - Nov17, 2017

Figure 5 shows the sentiment shifted forward in time by τ_L. The rates of high change and local minima and maxima are matched between the two datasets, see also figures 7 to 10.

1.3 Realtime Implementation

Moving into a real time implementation was simple enough; previously the data post processing was being done in Octave, for the real time implementation Python was used. The application followed this structure,

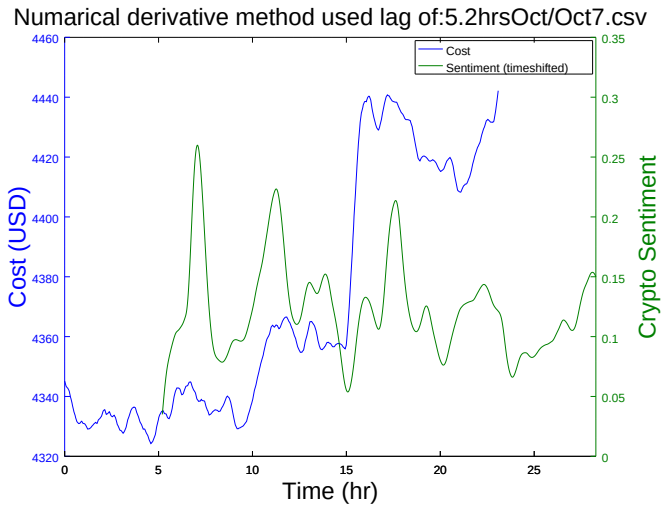

Fig. 5. Time Shifted Sentiment, October 9th 2017

1) $\tau = 10hrs$ of data is gathered.
2) Data is filtered, k and hence τ_L are calculated.
3) Buy and sell triggers are found, using the data from $\tau - \tau_L$, where τ is the current time.
4) Finally, Push notifications are sent to a mobile device using pushover [5].

The crux of the realtime application is finding the trigger points. happen in τ_L time. Is is expected that current high values of $\dot{y}$ correlate to high values of $\dot{x}$ in time τ_L. So high slope positive and negative values were taken and labeled as buy and sell triggers respectively. This is one possible method of finding the triggers, but further work should be done to understand if this is the best.

This realtime application is currently running and sending out advice for HFT (high frequency trading).

2 Future Work

2.1 Optimisation problem

The parameters in the real time application that are in need of optimisation.

- τ - the amount of data to collect in the real time app before starting the post operations.
- Number of buy/sell triggers, per τ.
- Stronger understanding of the Time Zone Inference (See appendix).
- Threshold for buy/sell triggers, what counts a 'high sentiment'.
- Filter cutoff frequency.

We could optimize each parameter using our existing dataset. This would be extremely computational expensive most likely require offloading to an HPC environment.

2.2 Machine Learning

It's possible that there are correlations in the dataset that are not distinguishable by humans. Treating the problem like a black box and using a deep neural net it might be possible to gain a deeper understanding of trends and better results. In addition, we believe we could also use Bayesian inference techniques.

2.3 Public Accessibility

All the work that has been done could be rolled into a pay to access platform to which users can subscribe to make informed trading decisions.

3 Conclusion

The crypto currency landscape is rapidly changing, as newly opened exchanges have facilitated access in the investment of decentralised currencies. This gives reason to believe that in addition to understanding the tech, it's also interesting to understand what people are thinking about the tech. This is important because the majority vote will determine the most widely adopted currency.

This was partially demonstrated in the work done above, but there's much more work that can be done. Analysis of other platforms other than just Twitter we believe it's possible to see an even better results. Aggregating a larger dataset could be used to understand the limitations of currencies and how volume spikes affect their network.

4 Appendix

4.1 Time Zone Inferance

There are also highly observable daily oscillations in tweet volume as seen in Figure 6. These can be used to infer time zones of maximum influence.

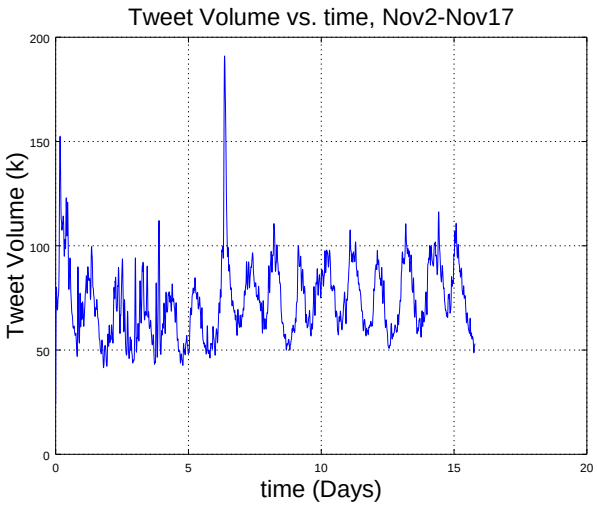

Fig. 6. Tweet volume, Nov2 - Nov17, 2017

4.2 Additional Data

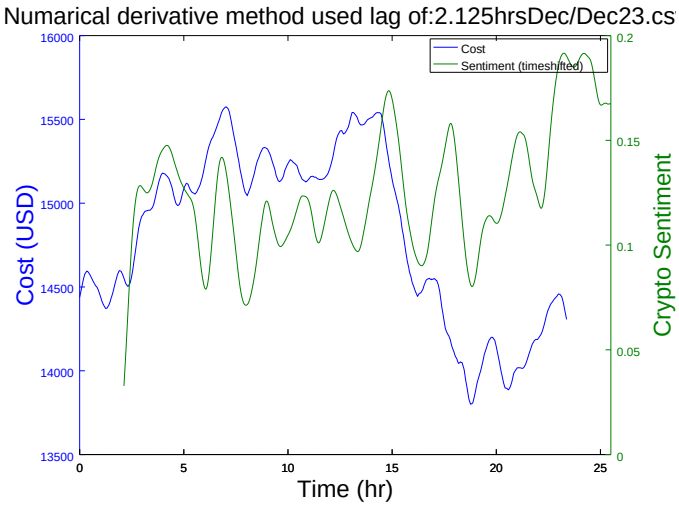

Fig. 7. Dec23

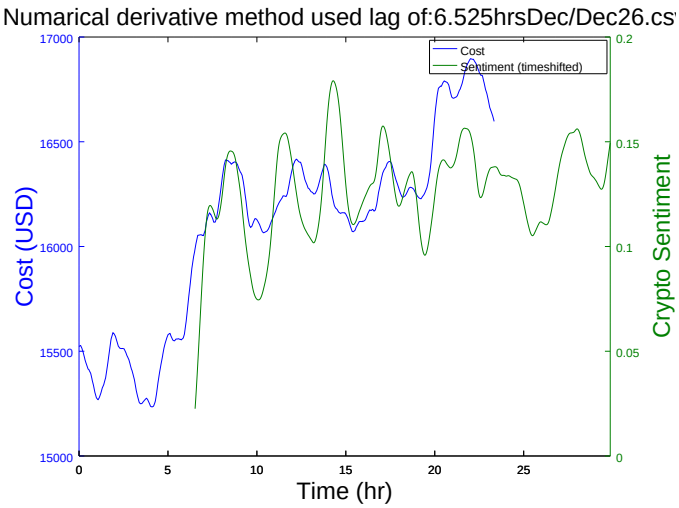

Fig. 8. Dec26

REFERENCES

[1] Anshul Mitta and Arpit Goel, *Stock Prediction Using Twitter Sentiment Analysis*
[2] https://github.com/tweepy/tweepy
[3] http://www.nltk.org/
[4] https://www.gnu.org/software/octave/
[5] https://pushover.net/

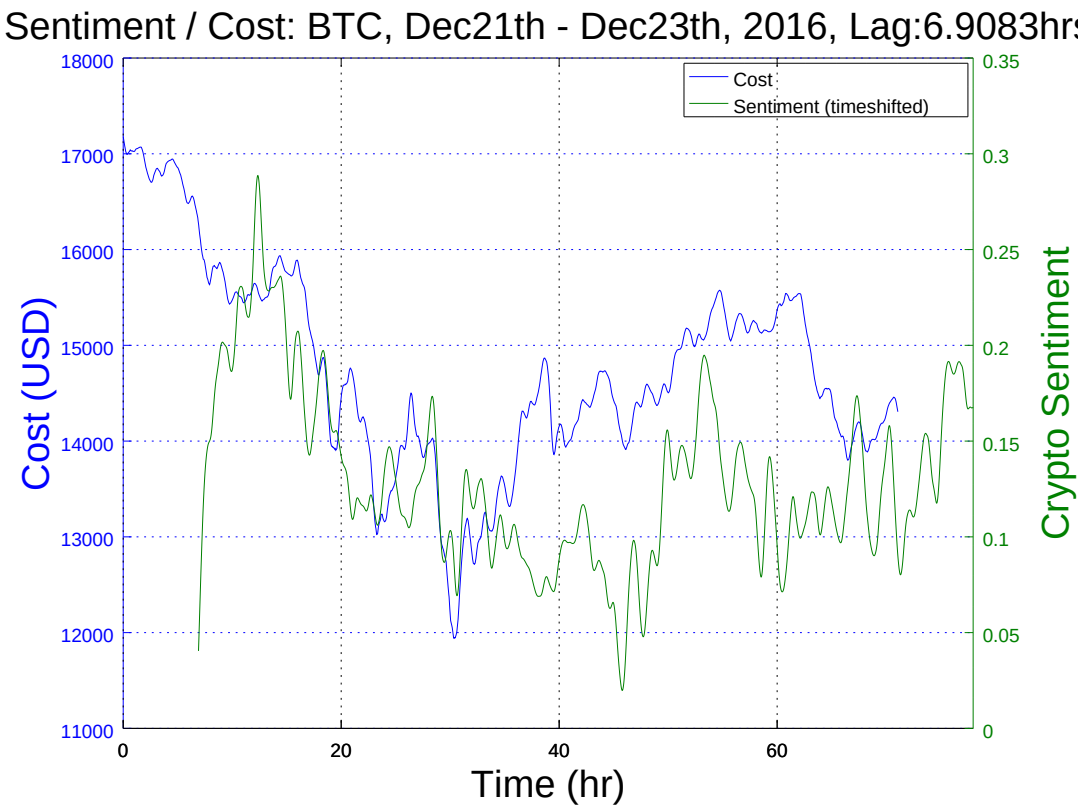

Fig. 9. Bitcoin 'crash', Dec21 - Dec23, 2017

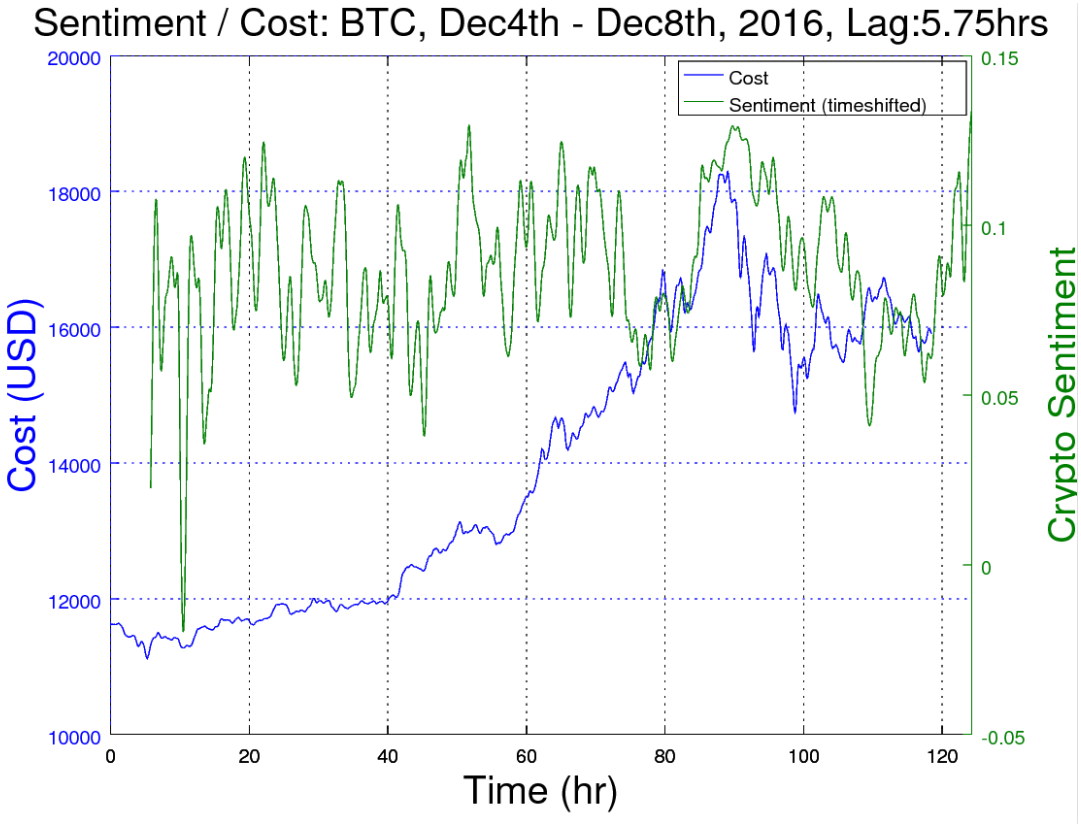

Fig. 10. Four day trend - Dec4 - Dec8, 2017