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

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]. I reason that for a technique like this to work there are three major requirements.

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

In most cases, point 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 typically will typically be publishing earning reports, posting news, etc... Where are important for investors, but cannot be accurately modeled as they are considered artifacts.

Of course there are still news artifacts regarding Cryptocurrencies, but they are less common and typically have less of an impact because they are mostly subjective, unlike an earning report or other financial documents.

In this paper I'm going to outline the numerical techniques I used in to do bla bla bla...

1.1 Gathering Sentiment Data

As mentioned above the main source of sentiment data was from twitter. A Python module *tweepy* was used to gather tweets and then bin them into cryptocurrencies of interest, the volume was anywhere from $1.2k\frac{tweets}{hr}$ to $24k\frac{tweets}{hr}$ for each currency.

From that point is was the possible to use Python *NLTK* (Natural Language Toolkit) to rate each tweet and make a net sum per unit time. Figure 1 shows the output of what I have described above.

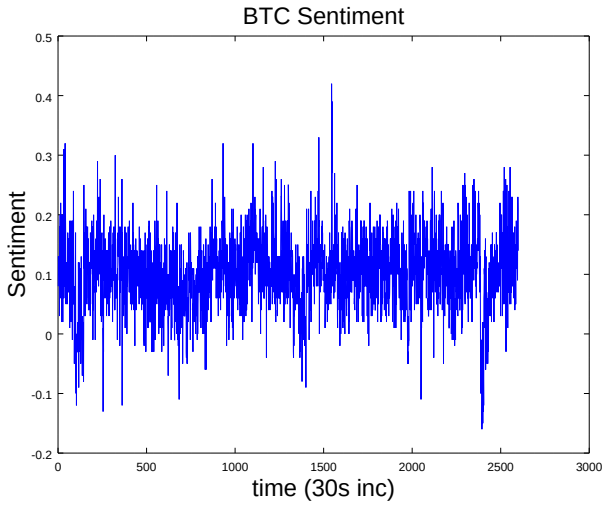

Fig. 1. Raw Sentiment, October 9th 2017

After a little filtering Figure 2. From here it was possible to compare to time series plots of the value.

1.2 Timewise Correlation

The timewise correlation was done by what I call the time correlation vector k which is defined as equation 1

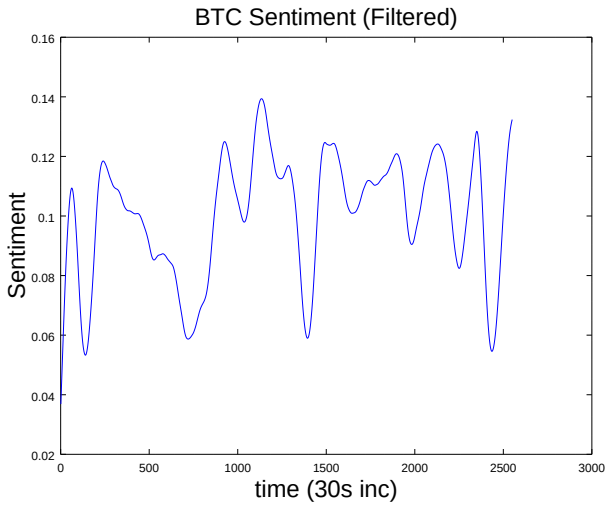

Fig. 2. Filtered Sentiment, October 9th 2017

$$k_j = \sum_{i=1}^{n} x_i - y_{i+j}, \text{ where j runs from 0 to n} \quad (1)$$

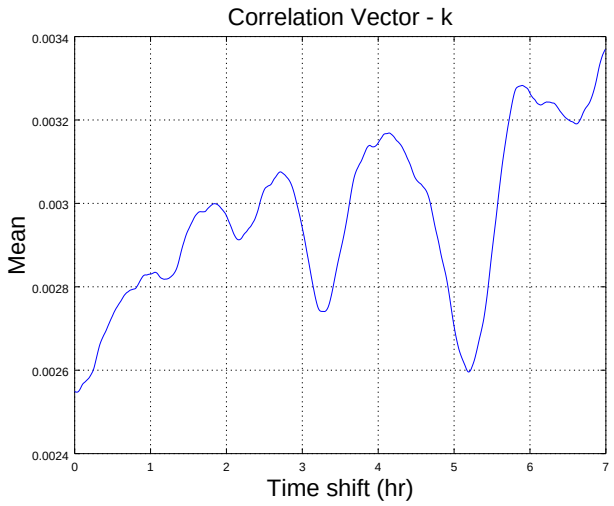

Fig. 3. Time Correlation vector k,October 7th 2017

Where $x = \frac{dCost}{dt}$ $y = \frac{dSent}{dt}$, the reasoning for using derivatives for the correlation is simply because the sentiment values are effectively a floating value - where the magnitude is not believed by myself to have any direct relationship to the magnitude of cost. However, my findings have lead me to believe that a changing value of sentiment will induce a changing cost.

k is a minimizing function where the lowest magnitude indicates the highest level of correlation. As seen in Figure 3, k has a global minima around 5.2hrs, which I define as $\tau_L = 5.2hrs$ or the effective 'Time lag' between high values of changing sentiment and high values of changing cost. Figure 4 shows the sentiment shifted forward in the by τ_L.

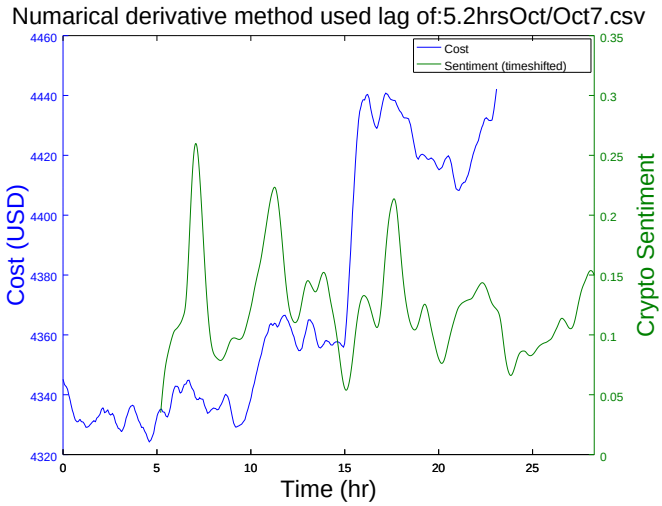

Fig. 4. Time Shifted Sentiment, October 9th 2017

It can be observed that rates of high change and local minima and maxima are matched between the two datasets.

1.3 Realtime Implementation

Write things here.

2 Conclusion

The conclusion goes here.

References

[1] Anshul Mitta and Arpit Goel, *Stock Prediction Using Twitter Sentiment Analysis*