Abstract

Last year, bitcoin gained attention of many investors – sophisticated, unsophisticated, institutional, recreational and otherwise. Due to an increased interest in this new market, analysis of cryptocurrency has become subject of many technical writings that aim to predict token prices in highly speculative market. One of the most popular techniques to predict volatility is the use of Natural Language Processing to categorize investor sentiment evaluated via tweets and sub reddits. Most papers yielded great results when cryptocurrency prices were high, when investors inundated social media about successes. But now that the prices have plummeted, and mining tokens is becoming increasingly difficult, this study suggests a phase of stability and decreased sensitivity to public emotions. The study goes over a model that collects publicly available tweets from the hashtag "cryptocurrency", analyzes the sentiment of tweets and makes a prediction of bitcoin's volatility.

Introduction

When regarding a financial commodity, the public confidence in a commodity is a core base of its value. Social media has served as platform to express opinions since their inception, and as such tapping into the open APIs provided of the likes of Facebook and Twitter, these biased pieces of information become available with a sea of meta-data. Bitcoin (BTC), the decentralized cryptographic currency, is similar to most commonly known currencies in the sense that it is affected by socially constructed opinions; whether those opinions have basis in facts, or not. Since the Bitcoin was revealed to the world, in 2009, it quickly gained interest as an alternative to regular currencies. As such, like most things, opinions and information about Bitcoin are prevalent throughout the Social Media sphere. In the paper Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns by Sul et al., 2.5 million tweets about S&P 500 firms were put through the authors own sentiment classifier and compared to the stock returns. The results showed that sentiment that disseminates through a social network quickly is anticipated to be reflected in a stock price on the same trading day, while slower spreading sentiment is more likely to be reflected on future trading days. Basing a trading strategy on these predictions are prospected to yield 11-15% annual gains. The paper Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis by Colianni et al., similarly analyzed how tweet sentiment could be used to impact investment decisions specifically on Bitcoin. The authors used supervised machine learning techniques that yielded a final accuracy of above 90% hour-by-hour and day-by-day. The authors point out that the 90% accuracy was mustered through robust error analysis on the input data, which on average yielded a 25% better accuracy. Colianni et al. together with Hutto and Gilbert both mentioned levels of noise in their dataset, and the former team got a significant reduction in error rates after cleaning their dataset for noise. The sentiments as well as the currencies price are analyzed on a short-term basis, disregarding how micro-blogging sentiment correlates to macro trends in a cryptocurrency or attempting to identify if they exist. Short term in this paper is defined to be within the 24h mark, (based on the findings of Colianni et al.). The sentiments classification is limited to the naivest binary form of positive or negative, not attempting to capture sentiment on a more complex emotional level. On the BTC side, the key value will be limited to an increase and decrease in price over specific time intervals, disregarding volume and other key metrics. Further, BTC transactions are collected for the BTC/USD currency pair, and only collected from the Coindesk exchange due to difficulty in finding open-source aggregated exchange data.

Methodology

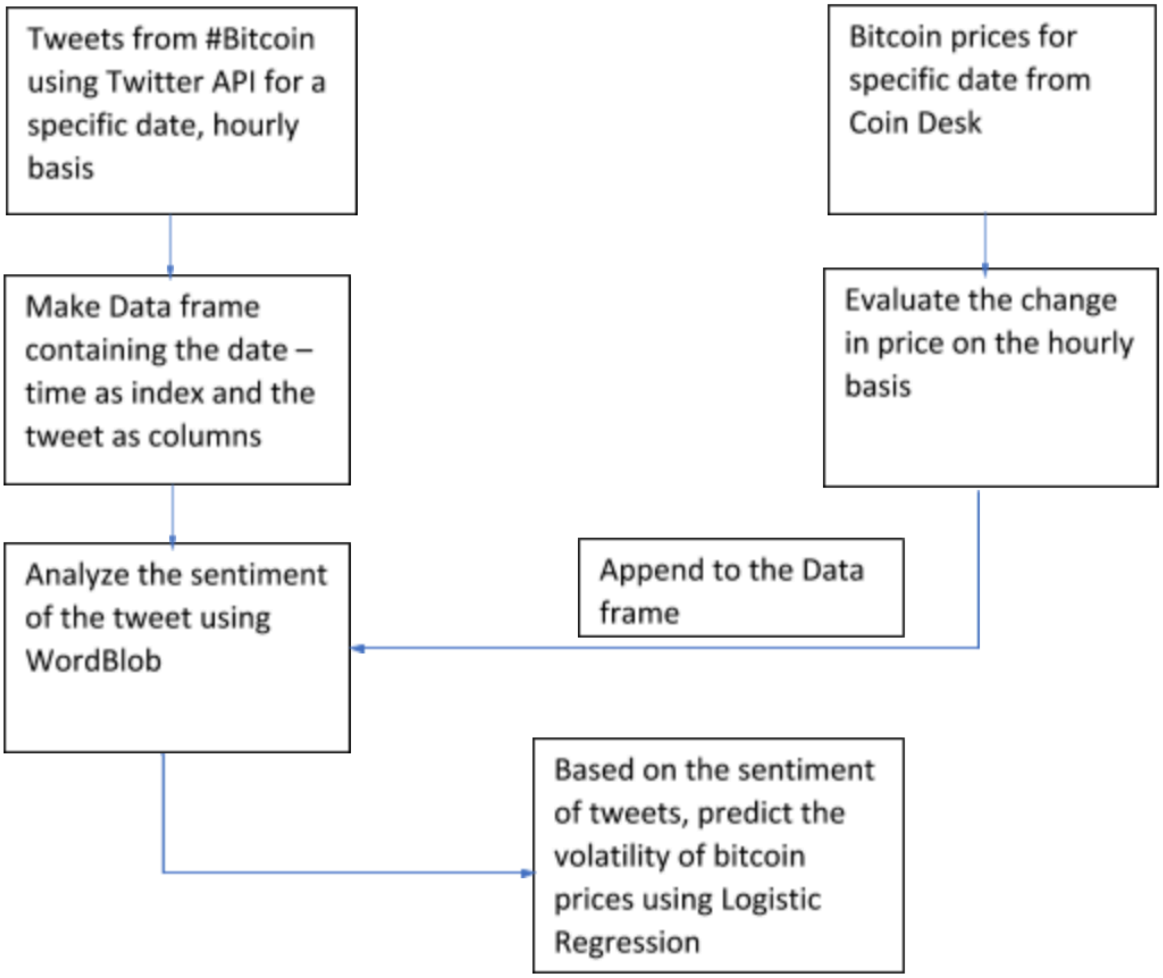

Tweets from #Bitcoin using Twitter API for a specific date, hourly basis
Make Data frame containing the date – time as index and the tweet as columns
Analyze the sentiment of the tweet using WordBlob
Bitcoin prices for specific date from Coin Desk
Evaluate the change in price on the hourly basis
Append to the Data frame
Based on the sentiment of tweets, predict the volatility of bitcoin prices using Logistic Regression

Figure 1. Flowchart describing the methodology undertaken un the study

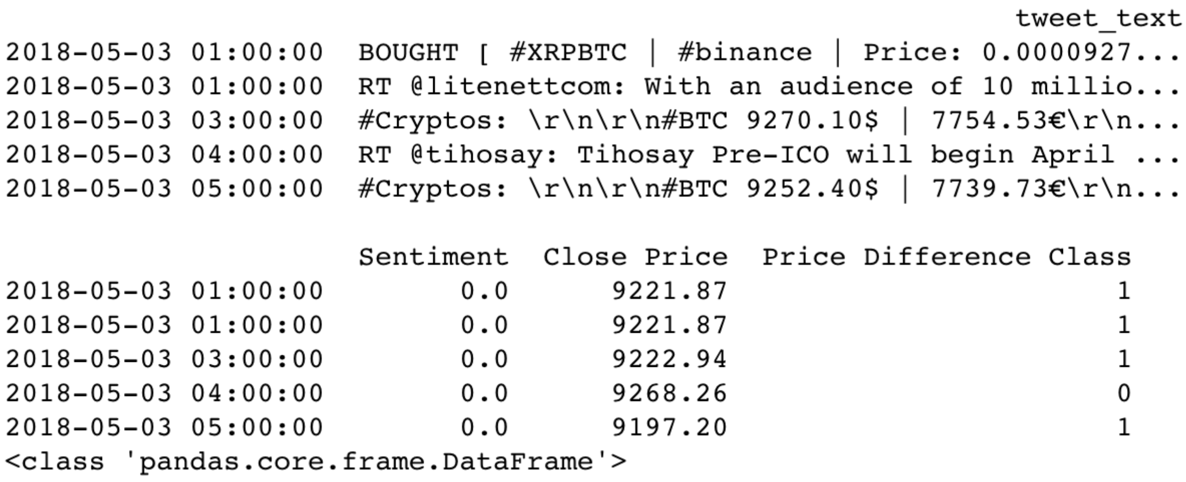
As described in the flow chart **Fig.1**, the twitter data was collected using the twitter API. The information extracted contained the following features.

- Tweet ID
- Date and time
- Tweet Text

Following acquisition of raw data, sentiment polarity of each tweet text was evaluated using the textblob module. Simultaneously, the bitcoin prices were collected from the coindesk website.

Restricted by availability of real time twitter information, acquisition of bitcoin prices was from also restricted to a time interval from April 11th, 2018 to May 11th, 2018. The prices, just like the tweets, were collected on an hourly basis. Furthermore, the prices were labelled with following strategy

- P(t) - P(t+1) > 0: Price Difference Class = 1
- P(t) - P(t+1) <= 0: Price Difference Class = 0



```
                                                             tweet_text
2018-05-03 01:00:00  BOUGHT [ #XRPBTC | #binance | Price: 0.0000927...
2018-05-03 01:00:00  RT @litenettcom: With an audience of 10 millio...
2018-05-03 03:00:00  #Cryptos: \r\n\r\n#BTC 9270.10$ | 7754.53€\r\n...
2018-05-03 04:00:00  RT @tihosay: Tihosay Pre-ICO will begin April ...
2018-05-03 05:00:00  #Cryptos: \r\n\r\n#BTC 9252.40$ | 7739.73€\r\n...

                     Sentiment  Close Price  Price Difference Class
2018-05-03 01:00:00        0.0      9221.87                       1
2018-05-03 01:00:00        0.0      9221.87                       1
2018-05-03 03:00:00        0.0      9222.94                       1
2018-05-03 04:00:00        0.0      9268.26                       0
2018-05-03 05:00:00        0.0      9197.20                       1
<class 'pandas.core.frame.DataFrame'>
```

Figure 2. Dataframe indicating relevant features

For simplicity, a single dataframe was developed by margining sub dataframes that contained information from both coindesk, twitter and the pertinent sentiments of each tweet. An illustration of this dataframe can be seen in **Fig.2**

Results and Discussions

Through initial investigation of data, a majority of sentiments are neutral or 0 in polarity. Upon further investigation it was noted that most tweets in May 2018 contained images or charts indicating or naively predicting the prices of bitcoin. This type of dataset is beyond the scope of this project. Second most observed sentiment in the dataset was positive in nature at 33.5 % of tweets showing polarity greater than 0. Lastly, only 4% of tweets indicated negative sentiment with polarity less than 0. These features are best represented by the pie plot as shown in **Fig.3**

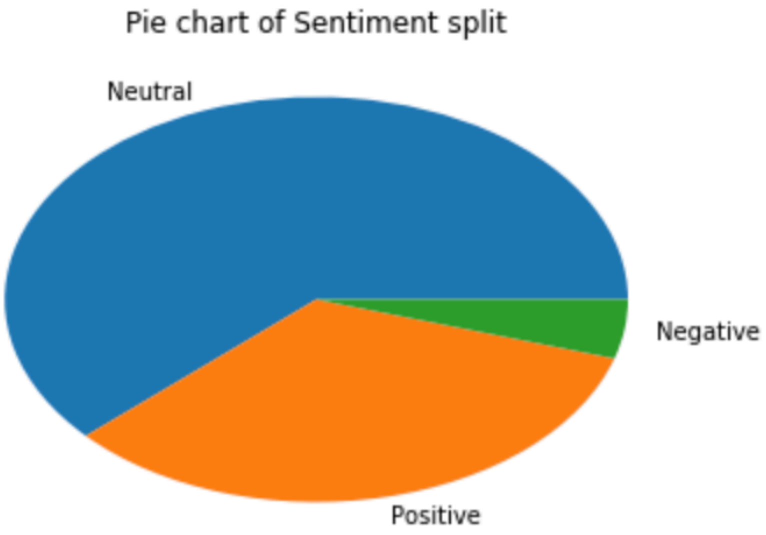


Figure 3. Pie chart of tweet sentiment polarity

The sentiment vs price class, as shown in **Fig.4** does not indicate an obvious correlation. For price class 1, the evaluated sentiments span all ranges i.e. negative to positive. For price class 0, the evaluated sentiments are less spaced out but, nonetheless, fall in all ranges of sentiments.

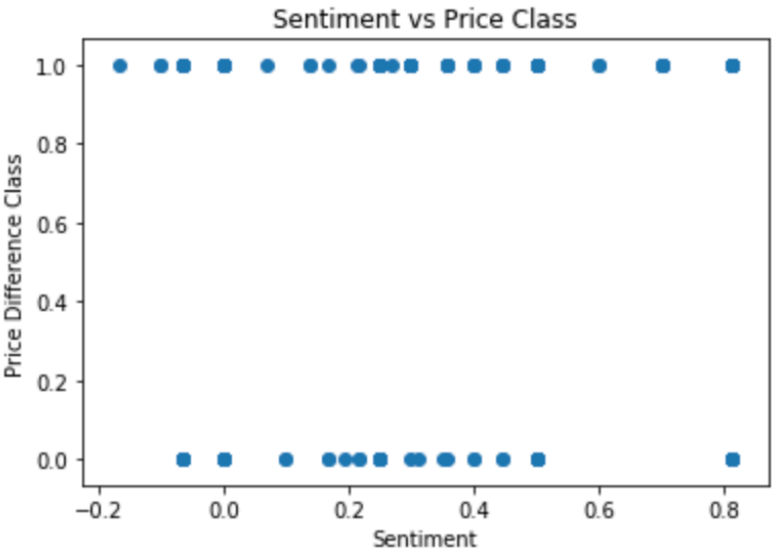
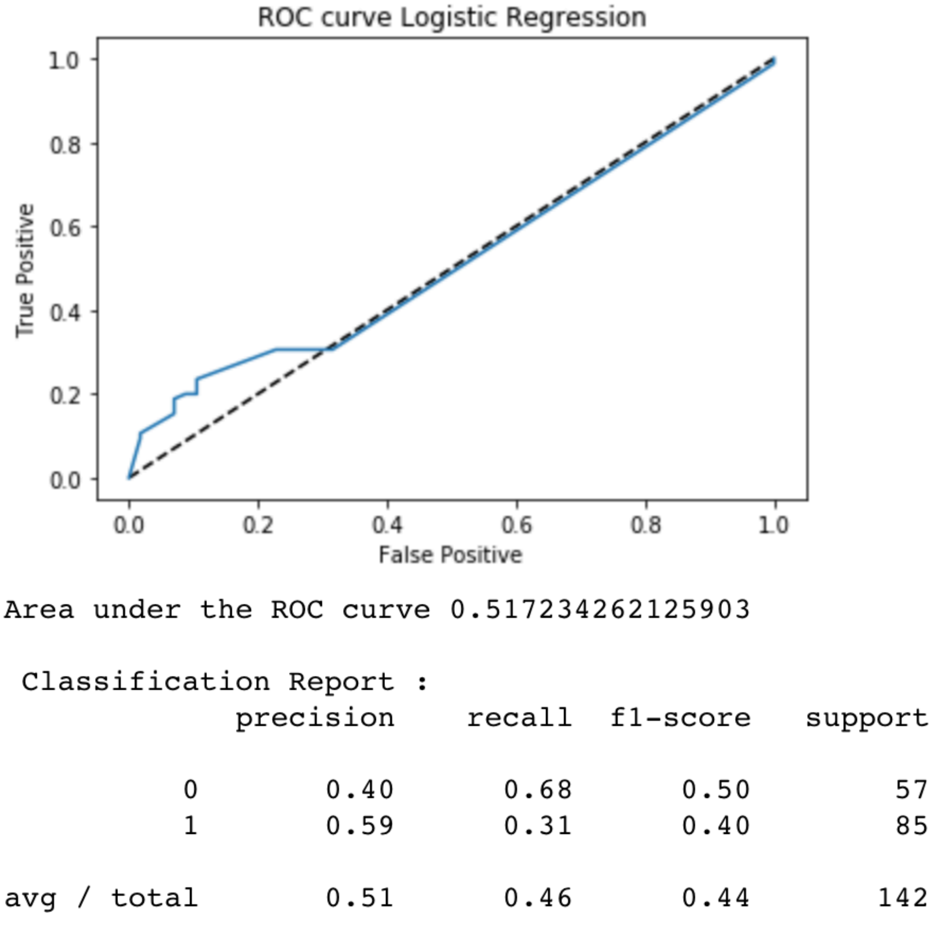


Figure 4. Price difference class vs sentiment

The dataset contained visible class imbalance since the number of negative sentiments were outweighed by the number of positive and neutral sentiments. In order to address this, the logistic regression was performed by balancing the weights for all three possible cases-negative, positive and neutral sentiment polarity.

The logistic regression model trained on the data had the performance characteristics as shown in **Fig.5**. Of course, the performance of this model can be drastically improved by including more tweets from different cryptocurrency handles and bitcoin data for an elongated timeline, but the scope of this project contained only development of workflow to acquire real time tweets and to demonstrate the trend of the said crypto currency token at the time of investigation and not beyond.



```
Area under the ROC curve 0.517234262125903

 Classification Report :
             precision    recall  f1-score   support

          0       0.40      0.68      0.50        57
          1       0.59      0.31      0.40        85

avg / total       0.51      0.46      0.44       142
```

Figure 5. Model performance characteristics

As mentioned earlier in the abstract, most papers, thesis and projects have concluded some sensitivity of bitcoin prices to public sentiments. The timing of these projects is something to consider. Most publications were rolled out at the time when bitcoin prices had crossed the $

15000 mark and public inundated the social media with tweets about cryptocurrency market, making bitcoin very valuable and causing a ‘mob effect’. Incidentally, another set of such publications came out when mining cryptocurrency was easy, and the market was esoteric by nature. Consequentially, only a specific set of people who had vested interests in the crypto token spoke publicly about bitcoins, and as a byproduct of it, influenced the prices. Moreover, since the tokens were easy to mine, the investors who also happened to be miners and influencers, could increase the supply of tokens that they spoke for.

The results of this model suggest that public sentiment alone is not sufficient to make prediction of bitcoin volatility, for the month of May 2018. This finding is in lines with the following article https://www.forbes.com/sites/petertchir/2018/05/28/i-would-be-shocked-if-bitcoin-prices-werent-manipulated/#37c8cc632be9, which also concludes that it’s not the sentiment of the mass that affects the bitcoin prices, rather the sentiment of high net worth individuals and miners that directly impact the pricing and trade volumes. This falls under the case of predicting a "Black Swan" event that is difficult to do so with micro blogging data.

Conclusion

1. For the month of May 2018, there seems to be no visible correlation between pricing of a bitcoin and the public sentiment
2. By continually investigating the data by running on this model for coming months, it should be possible to make a comment about stability of bitcoin token and the cryptocurrency market, at large.
3. A portion of this project has been shared with the technical team at Centareum, Singapore to evaluate use case of sentiment analysis to categorize feedback for effective marketing. Results pending.

References

1. Hong Kee Sul, Alan R Dennis, and Lingyao Ivy Yuan. Trading on twitter: Using social media sentiment to predict stock returns. Decision Sciences, 2016.
2. Stuart Colianni, Stephanie Rosales, and Michael Signorotti. Al- gorithmic trading of cryptocurrency based on twitter senti- ment analysis. 2015. URL http://cs229.stanford.edu/ proj2015/029_report.pdf.
3. Collin Thompson. How does the blockchain work (for dum- mies) explained simply. URL https://medium.com/the- intrepid-review/how-does-the-blockchain-work- for-dummies-explained-simply-9f94d386e093.
4. Inc Twitter. Api overview — twitter developers, . URL https:dev.twitter.com/overview/api.
5. Bitcoin price index api - coindesk, . URL http://www. coindesk.com/api/.
6. Tweepy, . URL http://www.tweepy.org/.