# Sentiment Analysis of Social Media as a predictor of Bitcoin price volatility

M.Machado, J. Singh

**Abstract** The primary objective of this project was to discover if sentiment analysis of social media could be used as a predictor for Bitcoin price volatility. The secondary objective was to attempt to discover if Bitcoin price volatility could be used as a predictor for the crypto currency market as a whole. Social media post on reddit.com were analyzed using the nltk.sentiment library and a data exploration of coin prices was conducted using python statistical libraries. We were able to establish a correlation between Bitcoin price fluctuations and the price fluctuations of the top 10 coins. We were also able to show a strong correlation between positive sentiments and value of bitcoin. The dataset used was small compared to the total data available, we believe that the results obtained warrant expanding the project to analyze exponentially larger dataset to confirm the findings.

## I. Introduction

In emerging markets, cryptocurrencies are a new and hyper-volatile asset class. Bitcoin was released as open source software in 2009. It is the first and premier cryptocurrency. Currently the cryptocurrency market size is approximately $700 billion dollars.

Every major cryptocurrency exchange offers Bitcoin trading pairs. Bitcoin is in the top in terms of being traded for other altcoins, market capitalization, trading volume, and general popularity. It has become the defacto gold standard for cryptocurrencies.

Anecdotally, volatility in the Bitcoin market is mirrored by the other cryptocurrency markets. A method to predict the swings in the market price of Bitcoin would be a useful tool to any investor entering or currently engaged in the market.

## II. Background

Peter and Gloor [1] used sentiment analysis to analyzes correlations and causalities between Bitcoin market indicators and Twitter posts and concluded that emotional sentiments rather mirror the market than that they make it predictable.

Glaser et al., 2014 [2] concluded that in the case of Bitcoin, price volatility is significantly influenced by the media coverage and positive sentiment.

Kim et al. [3] used sentiment analysis of comments and replies posted in online communities relevant to cryptocurrencies to create a prediction model to predict fluctuations in price levels. They were able to predicted fluctuations in the price of each cryptocurrency with approximately 8% accuracy gaps.

In none of the discovered research was sentiment analysis done using the social media platform, reddit.com. According to alexa.com [4], Reddit is the fourth most visited site in the United States. There are more than 90 subreddits that are focused on cryptocurrencies and cryptocurrency markets.

## III. Publicly available Datasets

Publicly available dataset of Reddit comments [5] is going to be analyzed for years 2017 in order to predict price volatility of bitcoin and other altcoins if applicable by using sentiment analysis and other data mining techniques.

Bitcoin historical data covering the years 2012 - 2018 is publicly available at Kaggle.com [6]. Every major cryptocurrency exchange such as Coinbase has an API to allow for the extraction of historical data [7]. Also identified as a resource is the reddit Data Extractor, a cross-platform GUI tool for downloading almost any content posted to reddit. [8].

## IV. Proposal

This project will primarily address two questions. One, is there a correlation between the price fluctuations of Bitcoin and other cryptocurrencies. And Two, can sentiment analysis of posts in reddit.com cryptocurrency subreddits predict price fluctuations in bitcoin markets.

Phase 1 of the project will be to conduct a data exploration of the cryptocurrency historical datasets to ascertain if a correlation can be drawn between Bitcoin price fluctuations and the other major cryptocurrencies. The comparison will be done using the top ten cryptocurrencies by market share. The part of the project will allow us to ascertain where the findings of the sentiment analysis is applicable only to Bitcoin or can be generalized to all cryptocurrencies.

Phase 2 will be to conduct a sentiment analysis of comments extracted from reddit.com and to plot the results against the historical price fluctuations studied in Phase 1. The initial part of Phase 2 will be to use a subset of the data collected. More specifically a one month time frame in the year 2017.

Phase 3 of the project will be to expand the techniques used in Phase 2 to incorporate the use of ever larger datasets. One of the identified datasets for reddit.com comments is in excess of 300GB.

## V. PHASE 1

The datasets used in Phase 1 were downloaded using the CryptoCompare public API [9]. We used the following Python wrapper available on GitHub: cryCompare [10] to simplfy the API usage. The top ten cryptocurrencies were chosen based by market cap as of Dec. 7, 2017, according to CoinMarketCap.com. The historical data was restricted to the last 90 days of the year 2017.

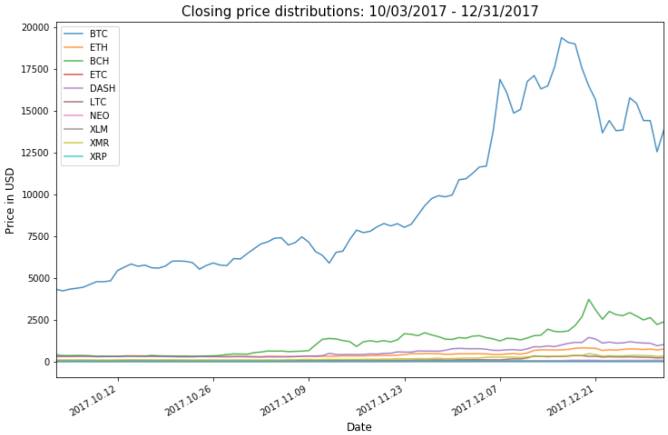
The coins are:

1. Bitcoin (BTC): $304.52 billion
2. Ethereum (ETH): $41.55 billion
3. Bitcoin Cash (BCH): $22.02 billion
4. IOTA (IOT): $11.59 billion
5. Ripple (XRP): $8.68 billion
6. Dash (DASH): $5.38 billion
7. Litecoin (LTC): $5.32 billion
8. Monero (XMR): $4.33 billion
9. Bitcoin Gold (BTG): $4.22 billion
10. Cardano (ADA): $2.78 billion

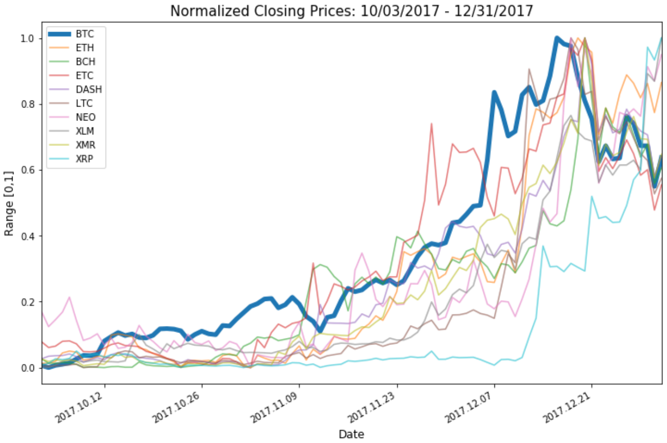
Because some coins didn’t start trading until late 2017 and has incomplete data sets for our purposes the following substitutions were made:

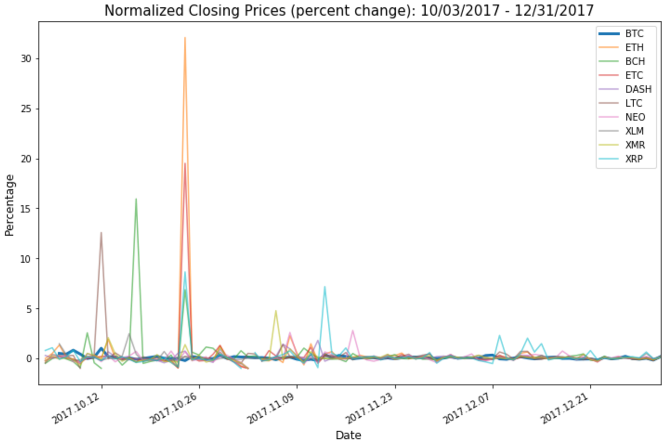
1. Ethereum Classic (ETC): $2.62 billion was substituted for IOTA (IOT): $11.59 billion
2. Stellar Lumens (XLM): $2.53 billion was substituted for Cardano (ADA): $2.78 billion
3. NEO (NEO): $2.23 billion was substituted for Bitcoin Gold (BTG): $4.22 billion

After the initial data retrieval pulling the date and OHLC prices, first step is to plot the closing price of our chosen crypto currencies and observe how the price has changed over time.



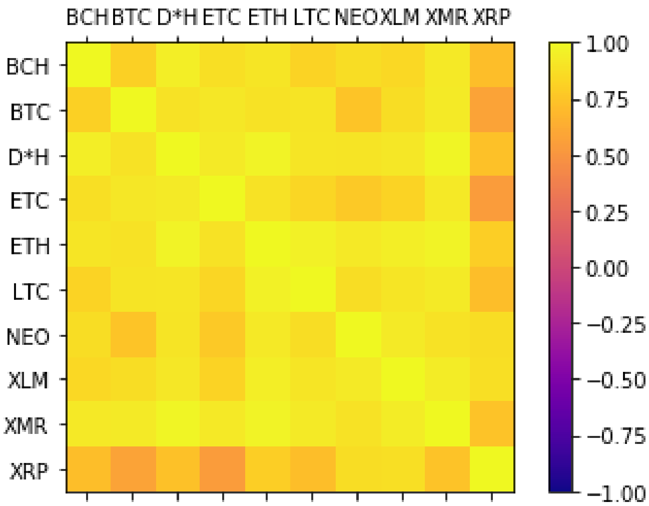
This chart was not very expressive, second step is to plot the normalized closing price of our chosen crypto currencies and observe if this gives us a better intuition about the relationship between coins.





The third step is to plot the percent change of the normalized closing prices of our chosen crypto currencies and observe if this gives us a better intuition about the relatioship between coins.

Correlation gives an indication of how related the changes are between two variables. If two variables change in the same direction they are positively correlated. If the change in opposite directions together (one goes up, one goes down), then they are negatively correlated. The next step was to create a Correlation Matrix Plot.



If we observe the second row of the correlation matrix, which displays the correlation between Bitcoin (BTC) and all the other coins, we see a strong correlation exists.

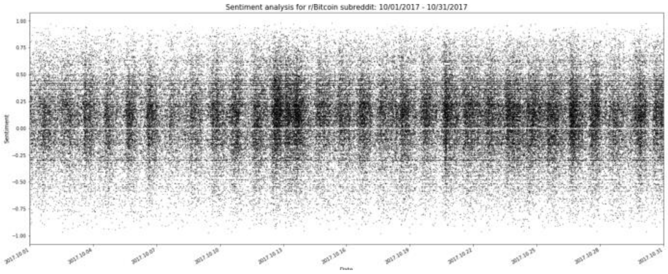
| BTC | BCH | DASH | ETC | ETH | LTC | NEO | XLM | XMR | XRP |
|---|---|---|---|---|---|---|---|---|---|
| 1.0000 | 0.8051 | 0.8969 | 0.9183 | 0.8933 | 0.8997 | 0.7511 | 0.8730 | 0.9363 | 0.5814 |

If we look at the raw data for the BTC matrix row, we see that all values are greater than 50% signaling a strong positive correlation. This is the current state of the data exploration. The archived data sets and the Jupyter notebook used in this data exploration have been made publicly available [11].

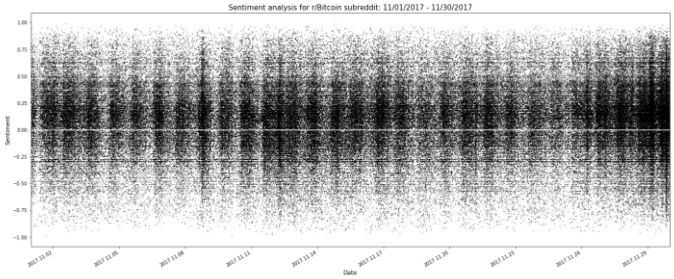
## VI. Phase 2

Phase 2 consists of conducting sentiment analysis of comments extracted from reddit.com and to plot the results against the historical price fluctuations studied in Phase 1. First, we extract sentiments from r/Bitcoin subreddit using SentimentIntensityAnalyzer function from nltk.sentiment library. We then plotted a simple point graph for comments (that aren't neutral) for the months of October, November and December 2017.

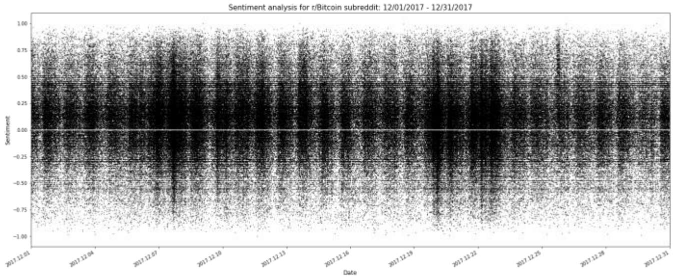
*October 2017*



*November 2017*

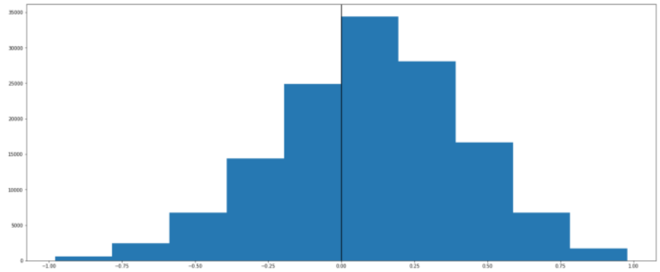


*December 2017*

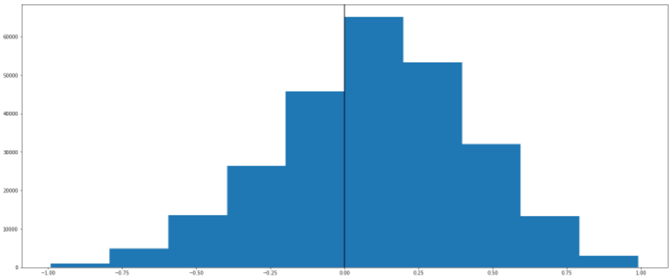


Next, we look at the most common values for sentiments using 1D histograms during these months. This graph is useful for discovering in which range most of the values lie. Sentiments which are more negative will lie on the left side and ones that are more positive on the right.

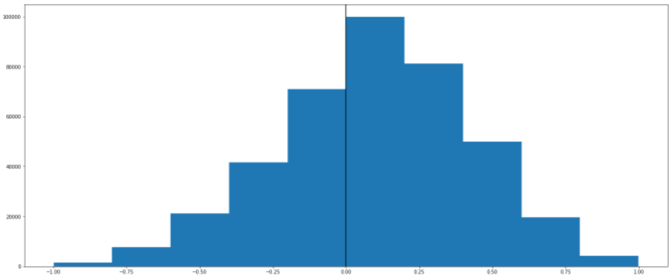
*October 2017*



*November 2017*

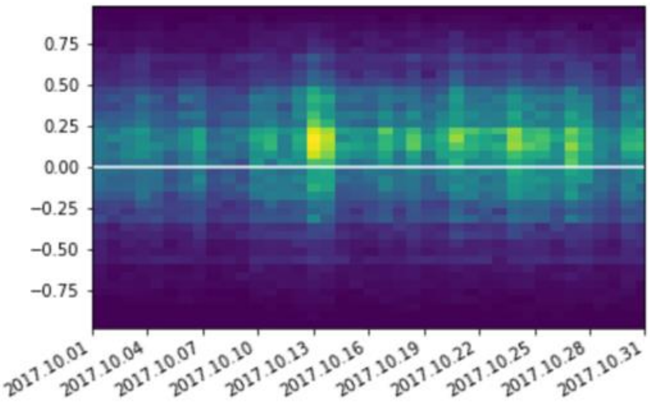


*December 2017*

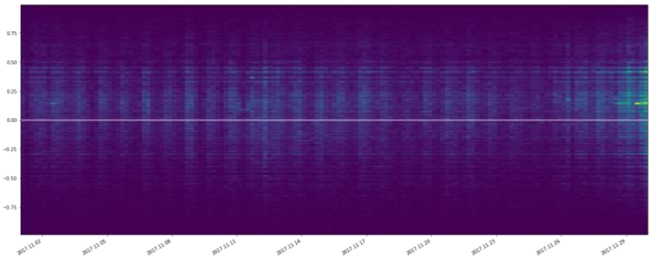


Finally, we look at the correlation matrix ("heatmap") of where most of these sentiments lie during these months.

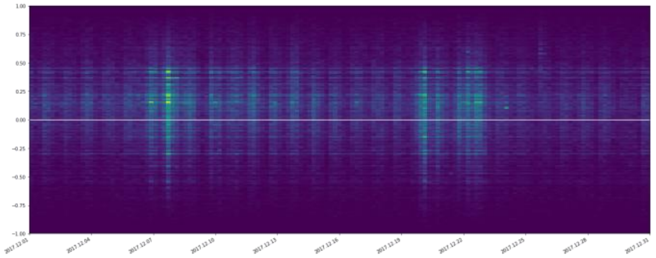
*October 2017*



*November 2017*



*December 2017*



November and December graphs are denser because we have more data on them. The most common sentiments being expressed are slightly negative to slightly positive. But there is a higher correlation between the prices and positive sentiments.

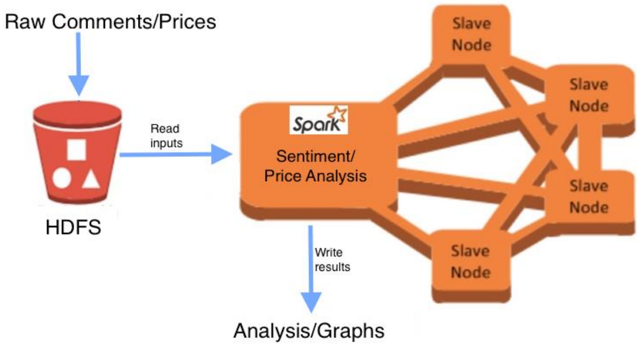
For example, when we see at the sentiments on certain dates (e.g. 10/13/2017, 11/30/2017, 12/8/2017) and we compare the price on those days (from Phase I), we see that the values are appreciating a lot as positive sentiments increase. This provides us with a strong correlation between positive sentiments and value of bitcoin. The code for this exploration is public [12].

Future work will include expanding the dataset to include more months and utilizing a machine learning algorithm like gradient boosting for sentiment extraction.

## VII. Phase 3

In phase 3, we productionized Phase 1 and Phase 2 using Amazon Web Services platform. The AWS technologies Hadoop HDFS filesystem and Apache Spark were used. The HDFS file system was used to store the large raw files and Spark for doing sentiment/price analysis. Final graphs were produced using pyplot.

*Overview of cloud environment*



Future work will include incorporating more months from reddit datasets. One of the potential candidate we found had comments is more than 300GB. We should also have an evaluation plan to evaluate the quality of predictor.

## References

[1] Kaminski, Jermain and Peter A. Gloom. "Nowcasting the Bitcoin Market with Twitter Signals." CoRR abs/1406.7577 (2014): n. pag.USA: Abbrev. of Publisher, year, ch. *x*, sec. *x*, pp. *xxx–xxx.*

[2] *Glaser, Florian and Haferkorn, Martin and Weber, Moritz Christian and Zimmermann, Kai, How to Price a Digital Currency? Empirical Insights on the Influence of Media Coverage on the Bitcoin Bubble (April 29, 2014). MKWI 2014 (Paderborn) & Banking and Information Technology, Vol.15, No. 1, 2014. Available at SSRN: https://ssrn.com/abstract=2430653*

[3] *Kim YB, Kim JG, Kim W, Im JH, Kim TH, Kang SJ, et al. (2016) Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. PLoS ONE 11(8): e0161197. https://doi.org/10.1371/journal.pone.0161197*

[4] *https://www.alexa.com/topsites/countries/US*

[5] *https://www.reddit.com/r/datasets/comments/65o7py/updated_reddit_comment_dataset_as_torrents/*

[6] *https://www.kaggle.com/mczielinski/bitcoin-historical-data*

[7] *https://developers.coinbase.com/docs/wallet/guides/price-data*

[8] *https://github.com/NSchrading/redditDataExtractor*

[9] *https://www.cryptocompare.com/api/#introduction*

[10] *https://github.com/stefs304/cryCompare*

[11] *https://github.com/fiddlemike/CS498_Project*

[12] *https://github.com/jaideep2/CS498_Project/blob/master/CS498_Project_Team31-Phase_2.ipynb*