**IDS 561**

**Big Data Analytics**

**Sentiment Analysis of Cryptocurrency Data**

**Team Members:**

**Mallika Trakanpasakul**

**Nandini Poddar**

**Kartikeya Sharma**

**Introduction and Background**

Cryptocurrency was first introduced in 1983, by the year 2009, the popular decentralized version, Bitcoin, created by pseudo developer Satoshi Nakamoto. By the year 2011, the need arose to regularize this growing market and the regulations had to be made to channelize the exchange and monitor the exchange activities.

In 2014, US IRS department ruled that Bitcoin, one of the forms of Cryptocurrency, will be treated as property for tax purposes. It is for this very reason, we see that the future of cryptocurrency is strong, and this project will be immensely helpful in capturing the sentiment around cryptocurrency and gauge market behavior.

In the longer run, it is believed that Bitcoin may well replace the conventional currency system and thus this project becomes even more important to understand.

**Objective**

There are multiple cryptocurrencies that are traded on numerous exchanges. However, the market is mainly driven by sentiment. In 2009, Bitcoin was available for 0.10 cents, however, by 2017 its value had risen over $8000. It is thus, a sincere attempt to gauge the market sentiment and predict the rise/fall in the currency value.

**Data Sources**

For the initial purpose, we intend to use Twitter as the main source of data. Twitter, as we are aware, is an online platform where users can share their views on any topic. It is this collective analysis of sentiment that we intend to gauge and then correlate this with the performance of cryptocurrencies.

Also, there are multiple news sources which also publish articles on their analysis and it can be associated with the Twitter analysis, to increase the model performance.

Also, we intend to look for a cause and effect relationship which will help in finding any triggering events that caused a stark change (positive/negative) in the currency performance.

We downloaded the Market Price of Bitcoin within the specified period from Blockchain.info to find a correlation between the actual public sentiment and the actual price of Bitcoin on a specific day.

**Techniques:**

1. We primarily used the Jupyter notebook(Python) for initial data cleaning to remove stopword (hyperlinks, abbreviations, non-alphabetic strings).
2. Later, *Pyspark* was used to creating the word cloud to find the top words that emerged after mining tweets.
3. Finally, to perform sentiment analysis we used the python library TextBlob, an extension of Natural Language Processing Toolkit (NLTK) was used to assign a score of -1, 0, 1 for negative, neutral, and positive tweets respectively.
4. To see which words, affect the sentiments score the most, we would like to visualize those words in a meaningful fashion. Therefore, we utilized word cloud package on *Pyspark* the concept of which is like the word count we have done in class
5. To do the analysis, we also extracted the Bitcoins stock price and do correlation and linear regression. Also, to understand how the sentiment and the stock price are correlated, we visualize it by utilizing Tableau

**Team Member contribution:**

While each team member was actively involved in every phase of the project, yet a SPOC was assigned to lead every phase and the roles can be classified as:

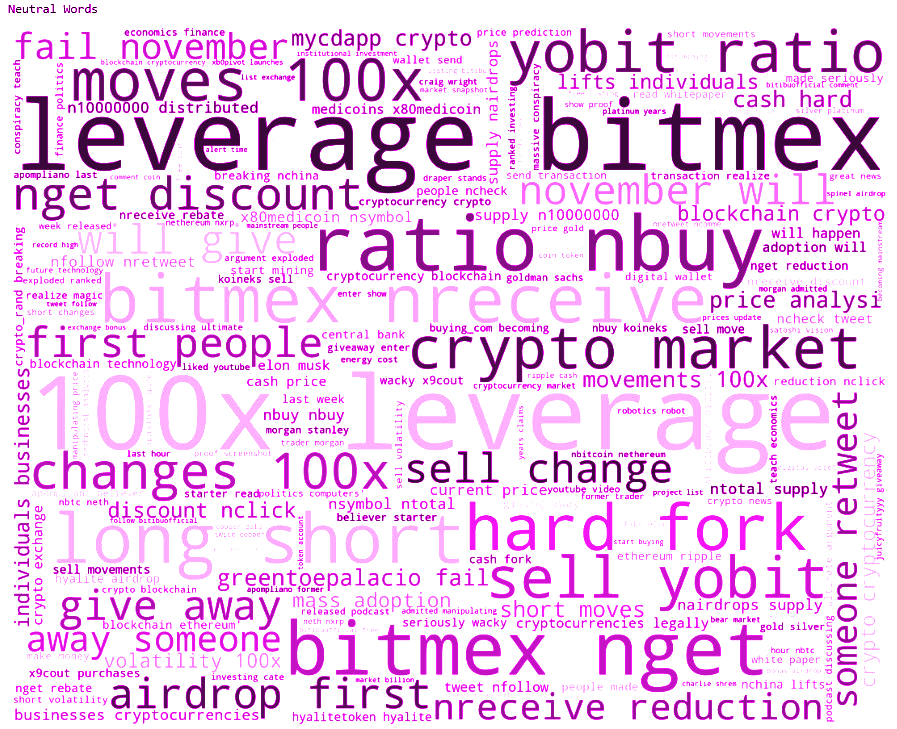
|  |  |
| --- | --- |
| Kartikeya | Data Retrieval from Twitter API and Sentiment Correlation |
| Nandini | Data Merging, Cleaning, and Transformation |
| Mallika | Word Count and Sentiment Analysis |

**Limitations**

As we are using TextBlob as our choice of the lexicon, we were limited to the vocabulary of TextBlob. In consequence, the vocabulary in the tweets might not have been included in the package. Moreover, it is also difficult to get emoji Unicode removed from the tweets. However, we cannot identify the sentiments based on the emoji. In today’s world, people tend to use emoji daily to express their emotions toward most of the situations. Lastly, although we downloaded tweets worth a week, it still gave us 1.3 Million tweets with the #bitcoin. Therefore, we have a very limited time window toward doing the analysis. Therefore, it would be too generalizing to summarize whether the tweets affect the Bitcoin price.

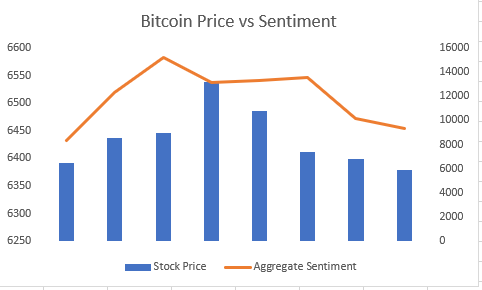
**Analysis**

Based on the word cloud below, we can see that the words seem similar between positive and negative sentiments.

****

**Sentiment vs Time of Day**

We have grouped the average sentiment by each day and find the correlation with the stock price from the market on the same days as those tweets we have extracted. Here is the result of the average sentiment and the real-time Bitcoin Market Price plotted in bar and line charts as follows:



From the above charts, we calculated the correlation analysis by using linear regression. We can see that we have got around 60 percent of the correlation between these two. However, having limitations mentioned above with regards to the exact method that might be more appropriate and the problem of decoding emoji Unicode, we might be able to improve the correlation

**Conclusion**

It is undeniable that sentiment essentially drives the market. Also, from the sentiment analysis shown above, we believe that this sentiment analysis can be another tool for investors to decide whether they should or should not jump in the market at a given time in accordance with the sentiment trend. In the foreseeable future, there will be more and more technology to help us improve this model and do better analysis.