Bitcoin price fluctuation prediction using Twitter Sentiment Analysis

Jeevan Reddy Rachepalli
School of Informatics, Computing and Engineering
Indiana University, Bloomington.
jeereddy@iu.edu

ABSTRACT

Bitcoin is a decentralized electronic currency system, which brought an enormous change in the financial system after its creation by Satoshi Nakamoto. It stands for an IT innovation based on peer to peer networks and cryptographic protocols. Bitcoin is not managed or controlled by any government or any bank, because of its decentralized nature and electronic form. The main purpose or intention of Bitcoin is to facilitate the transactions of goods and services. Bitcoin has grown tremendously and has managed to attract large number of users and has gained huge popularity due to its frequent mention and propagation in the media. Due to its popularity, the Bitcoin price, which fluctuates constantly on real-time like a stock exchange, it is very curious to build a model that can predict of the price of Bitcoin on real-time using the social media data from the internet. If the prediction can be made at an acceptable accuracy level, it might be useful for investors, business persons, banks, organizations, etc that use Bitcoin for transactions.

The Internet has grown tremendously in the past decade and with the invention of communication platforms aka, social media like Twitter, Facebook, Instagram, Blogs etc., sharing knowledge and experiences has become easy. For example, hundreds of thousands of Twitter users generate huge volumes of tweets data every day related to Bitcoin. This huge data can be helpful to study the trends in Bitcoin using technologies like Machine Learning, Natural Language Processing, Time Series Analysis etc. using the continuously generated data from the social networking sites on real time. Since Bitcoin is a new phenomenon, the application of Machine learning and Deep learning with greater accuracy and speed using social media data is still new. While much research exists surrounding the use of different machine learning techniques for time series prediction, research in this area relating specifically to Bitcoin is lacking. This study might fill the gap by building Emoji based Sentiment analysis for effective capture of Twitter Sentiment.

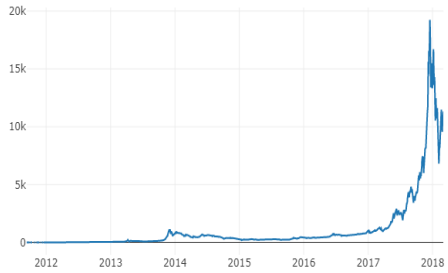
.

KEYWORDS

Sentiment Analysis, Bitcoin, Emoji, Price fluctuation, Time Series Analysis, Natural Language Processing

1. INTRODUCTION

Bitcoin has gathered a lot of attraction from people during the recent times and its price has been volatile-suddenly increased by multiple times and suddenly decreased. There is no efficient way of predicting the price of Bitcoin Price even though we dig deeper into the Blockchain. We know that people express their opinions and sentiments through online portals like Reddit, Twitter, Facebook and some other social media websites. Among these websites, Twitter is a perfect platform for us to do sentiment analysis on a small number of words that are used by a user. The tweet is not only a comprehensive way of expressing user's feeling but also helpful in understanding his view of a particular topic.



.

The above plot shows the Bitcoin price from January 2012- March 2018. We can observe that from January 2017- December 2017 there is an increasing trend of the price of Bitcoin and then suddenly dropped from December 2017- February 2018.

Our goal is to capture the Bitcoin Price fluctuation from the sentiment analysis of tweets. So, we have collected Twitter data from 01-October 2017 to 03-March-2018 because within this time period the Bitcoin price has varied a lot in terms of Price. So, we have collected all the tweets (around 2.5 Million) from Twitter within this time period.

We have gone for unsupervised learning lexicon-based tweet classification. In this paper, we have used the novel technology of capturing sentiment by the usage of emojis in a tweet. If the tweet does not contain an emoji, we have gone for 3 lexicon-based tweet polarity classifiers namely AFFIN Sentiment Classifier, TEXT BLOB Sentiment Classifier and VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Classifier.

2. RELATED WORK

There have been various works already conducted on predicting the Bitcoin price using sentiment analysis. Valence Aware Dictionary and sEntiment Reasoner (VADER) is used to classify a tweet polarity- it gives the sentiment classification as Positive, Negative or Neutral. This is an open source Python library licensed under MIT and can be used in analyzing the sentiment in social media posts. This VADER has outperformed 11 other tools for polarity classification especially on social media posts. Bitcoin Price point history was collected using the publicly available API CoinDeck (Can give Bitcoin price at 1-minute intervals). A Naïve model can predict the direction of Bitcoin fluctuation per unit time with 50% but with this VADER the author attained an accuracy of 83%. This model does not provide the magnitude of change of Bitcoin at a particular time.

There was another model proposed to study Bitcoin price based on the frequency of search queries on Google trends and Wikipedia. Taken search queries on Google trends (weekly) and Wikipedia (daily) to find the correlation between Bitcoin price and search results frequencies on Bitcoin-related words and found correlation up to 0.85. In order to find the polarity of the search query on Bitcoin, they have induced a dummy variable into the model to track the sentiment on the Bitcoin search. The major drawback of this model is that they have considered only an 8-hour trading duration per day whereas Bitcoin transactions occur on a continuous 24-hour window.

The correlations and causalities between the price of Bitcoin and emotional posts on Bitcoin in Twitter is another recent trend which is closely monitored by Analysts. It has been proven multiple times by behavioral economics that emotions affect individual behavior and decisions taken. Hence these words which unambiguously state the emotion of an individual through a tweet play a major role in this

analysis. Various blocks are formed based on the information gathered from these emotional words in tweets such as sum_positive_tweets (Sum of tweets on Bitcoin containing positive signals), sum_negaitve_tweets (Sum of tweets on Bitcoin containing negative signals) and sum_emotions (Sum of positive and negative tweets). Then a series of bivariate correlation analyses are performed between twitter sentiments and Bitcoin market indicators, Twitter sentiments and Bitstamp close price, Twitter sentiments and Bitstamp trading volume are performed to observe deeper underlying trends between them.

The Sentistrength tool is used to find the Positive or Negative strength of short texts. This tool is developed by a team of researchers from the United Kingdom. It takes care of non-standard grammar into account -especially useful for Twitter data. This tool has a specific weight for each word deciding the strength of Positive or Negative Sentiment of a short text. Collected around 1,924,891 tweets on Bitcoin between January 2015 and March 2015 (60 days) using Twitter streaming API with the time stamp on each tweet. The correlation between the Bitcoin price and the Positive mood tweets is limited to 0.35 only from this model because they have collected the tweets in a 2 months window. In order to improve this model, they should collect more number of tweets on Bitcoin.

The researcher has performed sentiment analysis on 101,377 ratings and reviews of the customers. The customers were classified using the logistic regression and naïve byes classifier on the bag of words model. The Naïve Bayes has given best accuracy of up to 97%. Classification performance based on both the ratings and sentiment analysis was done and it was observed that subjective versus objective instances was more difficult than polarity classification and the improvements in subjective classification were observed to positively impact sentiment classification. Further, the sentiment lexicon was created in an unsupervised manner and then determining the degree of positivity of a text unit via some function based on the positive and negative indicators, as determined by the lexicon, within it.

Predicting the fluctuation in the price of Bitcoin has gained paramount importance for market analysts since Bitcoin is the most traded form of cryptocurrency. A model was developed using Neural Networks to predict the fluctuation in the price of Bitcoin. The researchers developed a deep neural network model on the posts and comments collected for almost 3 years from Bitcoin online forum rather than just going for sentiment analysis and got an accuracy of 80%. This model has overfitted the data by taking 90% into training and 10% for testing this may give erroneous results if they have taken different Train/Test split.

3. DATA COLLECTION

We need a time stamp on the price of Bitcoin and on the tweets to do time series analysis. We have used APIs and a little bit of web scraping for the collection of tweets. We have collected tweets for 155 days within the above specified time period.

3.1 Bitcoin Price Data collection

We have used Quandl module API of Python for collecting the Bitcoin Price on that particular day.

Instead of collecting Bitcoin Price data from a single source. We are collecting The Bitcoin Price from 4 major websites- “BITSTAMP”, “COINBASE”, “ITBIT”, “KRAKEN” that tracks the Bitcoin Price. Some of these contain missing values. So, we are going to take the average of these to get a smoother curve.

3.2 Twitter Data collection

We have taken developer access for downloading the tweets from Twitter, but this has given trouble in downloading the tweets. So, we have used the Twitter

web scraping Python file provided by our Professor Vincent Malic to download the tweets from Twitter having a hashtag of Bitcoin. By doing that we have lost important information about a tweet like the location of the user, username, retweet count. Also, we got the tweets of Bitcoin in the languages other than English.

Initially, as we were collecting the data we thought that we are going to collect around 1.5 million tweets over 155 days as there were approximately 10,000 tweets per day but during the December to February the number of tweets per day has increased up to 25,000. So, we have collected around 2.5 million tweets over the period of 154 days.

4. PRE-PROCESSING OF DATA

4.1. Preprocessing of Bitcoin Price data

Out of the "Opening Price", "Closing Price", "High Price", "Low Price" of Bitcoin in a day we have selected only the "Close Price" in a day of Bitcoin. There were some missing values in the Bitcoin Price in some of the Bitcoin Source websites. We have replaced it with the other websites mean Bitcoin Price of that day. We have included another column i.e., Average Bitcoin price i.e., an average of all the 4 closing Bitcoin Prices in a day.

4.2. Preprocessing of Twitter Data

We have downloaded the Bitcoin-related Tweets for an individual month and stored it in a pickle file. We have loaded individual pickle file into Python and merged it together and stored it in a one single CSV file for future usage and easy accessibility. Deleted all the attributes except the Time Stamp and the Tweet. Cleaned all the URLs that are contained in a Tweet. Converted the URL cleaned Tweet into lower-case letters.

5. MODELS USED FOR SENTIMENT ANALYSIS

Used the standard sentiment analysis like AFFIN Sentiment Analysis, TextBlob Sentiment Analysis, Vader and compare their respected emojis (if they are present in the tweet).

5.1 AFFIN Sentiment Classifier

In AFFIN sentiment analysis, we use AFFIN, which is a list of English words, which are scored with an integer ranging between -5 and +5. In the twitter sentiment analysis, using AFFIN we would be evaluating the overall average sentiment/polarity score (positive or negative) for the extracted text from the tweets data.

5.2 TextBlob Sentiment Classifier

In the TextBlob sentiment analysis, by feeding the unique tweets, we obtain polarity as the output that ranges between -1 to +1. So, a tweet has Positive sentiment when it's polarity is greater than 0 and negative sentiment when it's polarity is lesser than 0. When the sentiment polarity is exactly 0 the tweet is said to have neutral polarity. The TextBlob also gives the Subjectivity of a tweet

5.3 VADER Sentiment Classifier

VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Classifier. This works just like the AFFIN as it also has word weights ranging from positive to negative but this VADER is designed in such a way that it is aware of the Social media jargon used by its' users.

5.4 Emoji Classifier

Happy Emojis

```
happy_set = set([":)",":-)","=)"])
```

Searched for above-shown kind of emojis in a tweet and classified a tweet as a Positive polarity tweet on Bitcoin

Sad Emojis

☹😕😠

```
sad_set = set([":(",":-(","=("])
```

Searched for above-shown kind of emojis in a tweet and classified a tweet as a Negative polarity tweet on Bitcoin

5.5 Getting Effective Sentiment of a tweet

As our dataset contains tweet, AFFIN Sentiment score, TextBlob Sentiment score, VADER sentiment score and the presence of Happy or Sad emojis in a tweet. For the AFFIN if the AFFIN score is positive we have given it as +1 and for negative it is -1 and zero remain as 0. For TextBlob we have compared the Positive and Negative sentiment scores if Positive sentiment is greater than Negative sentiment we have given it a +1 and -1 for the vice versa, if both are equal then it is given a 0. For VADER we followed the same technique of TextBlob.

We have classified the effective tweet polarity as +1 if a happy emoji is present in it and -1 if a sad emoji is present in it. Out of the 3 sentiment scores we have defined a function that returns 1 if a Happy emoji is present in a tweet and -1 if a sad emoji is present in a tweet. It also checks the sum of all sentiment scores of 3 classifiers. If it is greater than or equal to 1 then the effective polarity is given 1 and if it is less than or equal to -1 then the tweet is given an effective polarity of -1. For all other cases, it is given a 0.

Resampled the time series data with their mean on a particular day. So, on a day we have the Bitcoin average price and average Sentiment polarity on Twitter. As it is a time series data we can calculate the percentage change in the price of Bitcoin data with respect to the previous day. It can be Positive or Negative

6. ANALYSIS

6.1 AFFIN Sentiment Classifier

We have observed that the AFFIN Sentiment of a tweet maximum value achieved is 2.06 and minimum is -3.0. So, we have kept a threshold of Positive Polarity tweet > 1.7 for the highly Positive tweet and looked at the word cloud of that. Observed many positive words related to Bitcoin.

Similarly, we have kept a threshold of Negative Polarity tweet < -1.5 for the highly Negative tweet and looked at the word cloud of that. Observed many Negative words related to Bitcoin.

6.2 TEXTBLOB Sentiment Classifier

We have observed that the TEXTBLOB Sentiment of a tweet maximum value achieved is 1.o and minimum is -1.0. So, we have kept a threshold of Positive Polarity tweet > 0.7 for the highly

Positive tweet and looked at the word cloud of that. Did not observe many positive words related to Bitcoin as that of AFFIN.

Similarly, we have kept a threshold of Negative Polarity tweet < -0.7 for the highly Negative tweet and looked at the word cloud of that. Did not observe many negative words related to Bitcoin as that of AFFIN.

6.3 VADER Sentiment Classifier

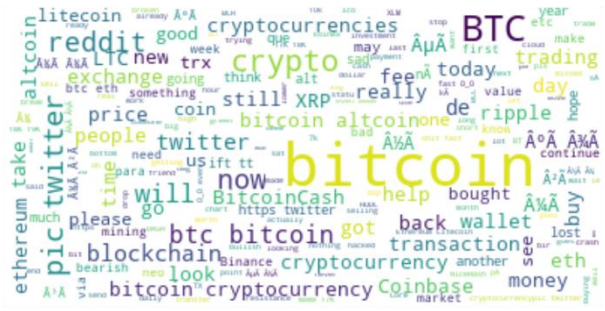
We have observed that the Positive VADER Sentiment of a tweet maximum value achieved is 0.93 and negative VADER Sentiment reached a value of 1.0. So, we have kept a threshold of Positive Polarity tweet > 0.8 for the highly Positive tweet and looked at the word cloud of that. Observed many positive words related to Bitcoin.

Similarly, we have kept a threshold of Negative Polarity tweet > 0.8 for the highly Negative tweet and looked at the word cloud of that. Observed many negative words related to Bitcoin.

6.4 Happy and Sad EMOJIS in a tweet

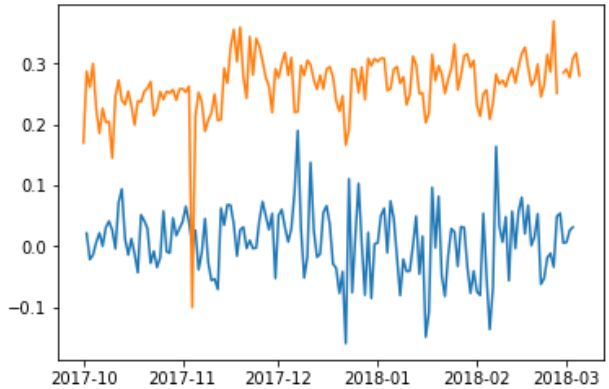
Out of 2.5 Million tweets only a less percentage i.e., 10,300 tweets have Happy Emojis in them and 440 tweets have sad Emojis in them. We have made Word clouds for Happy Emoji content tweets and we have observed only a little number of words that are positively related to Bitcoin as given below.

Made Word clouds for Sad Emoji content tweets and we have observed only a little number of words that are negatively related to Bitcoin as given below.



7. RESULTS AND FUTURE WORK

Observed that the Bitcoin percentage change and the Twitter sentiment are positively correlated. The Orange color graph is the scaled percentage change of Bitcoin over a day. The Blue color graph is the effective sentiment polarity of Twitter on Bitcoin within a given day. We can also say that there is a time shift between the Bitcoin Price fluctuation and Twitter Sentiment on Bitcoin. The Twitter sentiment lags in the Bitcoin Price fluctuation within a time frame. That's why our correlation between these two is around 0.17.



If we adjust the time lag there is a high chance that the correlation between our 2 variables will increase significantly. We can conclude that the Bitcoin Price fluctuation has been reactively caught by the Twitter users and not in a proactive manner. That means the increase in Bitcoin Price has a positive Sentiment effect in the Twitter after some days and not on beforehand and vice versa. So, through this study, we cannot predict the tomorrow's price fluctuation of Bitcoin instead we can catch the Bitcoin Price fluctuation through the Twitter sentiment after some days in Twitter.

We have bluntly classified tweets based on the Happy Emoji tweets and Sad Emoji tweets as Positive polarity tweet and Negative Polarity tweet on Bitcoin respectively. But we did not take sarcasm into account. Also, we have taken a list of 3 emojis for each of the Happy and Sad emojis but there are many emojis which can be used for classifying a positive emotion and Negative emotion such as

Happy/Joyful Emoji's list:

Sad/Angry Emoji's list:

If we have taken these emojis into account, our emoji contained tweets count would have been increased significantly from mere 11,000 (out of 2.5 Million tweets).

8. ACKNOWLEDGMENTS

The authors wish to thank Mr. Vincent Malic for numerous helpful discussions and comments and for his guidance and valuable suggestions. We also want to extend thanks to the Twitter and Quandl for providing the data related to tweets and Bitcoin price respectively.

9.1. CITATIONS

[1] Kim YB, Lee J, Park N, Choo J, Kim J-H, Kim CH (2017) When Bitcoin encounters information in an online forum: Using text mining to analyze user opinions and predict value fluctuation. PLoS ONE 12(5): e0177630. Retrieved from https://doi.org/10.1371/journal.pone.0177630

[2] Ladislav Kristoufek, Scientific Reports 3, Article Number:3415 (2013) doi:.10.1038./rsep03415 BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Retrieved from https://www.nature.com/articles/srep03415

[3] Jermain C. Kaminski, Ph.D., Media Lab (2016), Nowcasting the Bitcoin Market with Twitter Signals, Cornell University, USA. Retrieved from https://arxiv.org/abs/1406.7577v3

[4] Bo Pang, Lillian Lee (2008), Opinion Mining and Sentiment Analysis, Computer Science Department, Cornell University, Ithaca, NY. Retrieved from http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[5] Martina Matta, Ilaria Lunesu, Michele Marchesi, Università degli Studi di Cagliari Piazza d'Armi, 09123 Cagliari, Italy Bitcoin Spread Prediction Using Social And Web Search Media. Retrieved from http://ceur-ws.org/Vol-1388/DeCat2015-paper3.pdf

[6] Evita Stenqvist, Jacob Lonno (2017), Predicting Bitcoin price fluctuation with Twitter sentiment analysis. Retrieved from http://www.diva-portal.org/smash/get/diva2:1110776/FULLTEXT01.pdf

9.2. REFERENCES

[a]. Steven Loria, sloria1@gmail.com. TextBlob: Simplified Text Processing, Retrieved from: http://textblob.readthedocs.io/en/dev/

[b]. C.J. Hutto, cjhutto@gatech.edu, Georgia Institute of Technology, Atlanta. Retrieved from: https://github.com/cjhutto/vaderSentiment

[c]. fneilson, Simplest sentiment analysis in Python with AFINN. Retrieved from: https://gist.github.com/fnielsen/4183541