Twitter Sentiment

vs

Bitcoin Price

Ricardo Clemente

https://github.com/ric-clemente

Bitcoin

It is a descentralized digital currency without a bank and can be sent from user to user(peer to peer) without intermediares.

Twitter

Social networking servisse on which users post and interact with messages known as “tweets”

Sentiment analysis

Use the natural language processing to mine and identify the sentiment polarity and subjectivity in sentences.

What is a subjective sentence?

A subjective statement relies on assumptions, beliefs , opinions and influenced by emotions and personal feelings.

What is a objective or neutral sentence?

A objective or neutral sentences relies on facts where the information is provable measurable and observable.

Problem:

Analyze the sentiment extrated from tweets and check if there is any correlation with the Bitcoin price change.

Strategy used :

For our problem we going to use two classifiers.

The first one is going to determine if the sentence is subjective or not.

Then if the sentence it’s subjective we are going to use the polarity classificator to find out if the sentence is positive or negative.

Method Step by Step :

1) Tweet Scraping

Tweets were collected with the tool twitterscrapter that contains the words “BTC” or “Bitcoin”.

Number tweets : 40000 tweets

Date Range: 2017-11-5 to 2018-02-28.

2) Pre Processing

For each dataset and tweet it was necessary to clean the text and convert into vectors.

- **Clean text includes:**

  - Convert text to lowercase

  - Remove numbers

  - Remove punctuation

  - Remove urls

  - Remove hashtags

  - Remove twitter usernames

  - Remove Stop Words(the”, “a”, “on”, “is”, “all”)

  - Use Steeming - Is a process of reducing words to their root form

  - Use Lemmazation- Uses lexical knowledge bases to get the correct base forms of words.

- **Text into vector :**
  After we clean our sentences we are going to use the bag of words concept which creates a dicionary for every existent words in our datasets . One word will correspond a position in the dicionary list which will contain the number of times the corresponding word has occured in the document.

3) Training and Test

For both classifiers it was used Naive Bayes

Subjectivity Classifier:

- **Datasets:**

  - quote.tok.gt9.5000 - contains the subjective sentences.
  - plot.tok.gt9.5000 - contains the non subjective sentences(objective sentences)

- **Train and test performance:**

```
Accuracy Train NB: 0.9119474733262986
Accuracy Test NB: 0.7805907172995781
```

Polarity Classifier:

- **Datasets:**

  - rt-polarity.pos – contains the positive sentences
  - rt-polarity.neg – contains the negative sentences

- **Train and test performance:**

```
Accuracy Train NB: 0.95575
Accuracy Test NB: 0.914
```

4) Classify Tweets

Some tweet classifications:

	likes	text	dates	classification
0	0	Please share a bit or an entire @bitcoin 1EApfAbpN6y4A77dj8W8NbNCgWRUVGvtf4 #blockchain #btc #Crypto #BitcoinCash #bitcoin thanks #280caracteres #ParadisePaperspic.twitter.com/vkiLbCagFR	2017-11-09	Positive
1	1	Guess we should have let the corporate dbags commandeer btc to save $ on transactions? Or how about we just get @coinbase & the others to adopt #segwit? Then we dont have to listen to you rich bab...	2017-11-09	Neutral
2	0	do you guys think this would make /r/investing angry or happy? ;): http://ift.tt/2hghirL #bitcoin #btc	2017-11-09	Positive
3	0	OK, so something weird has been going on lately with #bitcoin. Here is my take on what is going on with #BTC prices after #SegWit2x was canceled, or was it? https://www.financemagnates.com/cryptoc...	2017-11-09	Positive
4	1	The current price of Bitcoin is $7131.75.\nThe current price of BCash is $663.077, or 0.0932815 BTC	2017-11-09	Positive
5	0	Just some basics for noobs to understand that $bcc #bcash is a scam ruled by some people for short term profit. The only #bitcoin is $BTC Always DYOR or cry later\nhttp://www.bestbitcoinexchange.n...	2017-11-09	Neutral
6	1	What would be perfect is if all the big hodlers like Roger Ver sold off all $btc (assuming they already have) and then #BitcoinCore CRASHES. I will still buy core, cuz at the end of the day no one...	2017-11-09	Negative
7	2	I can see $BTC perhaps correcting now and bouncing around $5700 on or near the 16th. IMHO\n#Bitcoin	2017-11-09	Positive
8	1	Are Bitcoin $BTC price % changes explained by the % change in transactions OR hash rate. Please discuss. Both seem to have a close fit, causation always key. @VladZamfir @twobitidiot @TuurDemeeste...	2017-11-09	Neutral
9	1	Tried to isolate #bitcoin's opacity premium, and ended up w/ variable correlated w/ offshore #yuan (2015-present). Q2'17 divergence aside. Not totally surprised. Alt capital flow metric, #yuan pre...	2017-11-09	Negative
10	0	Wait.. So was Segwit2x just the biggest pump and dump of all time or...? https://www.reddit.com/r/ethtrader/comments/7bu4g6/wait_so_was_segwit2x_just_the_biggest_pump_and/?ref=share&ref_source=twi...	2017-11-09	Neutral

5) Calculations

For each day it was calculated the % of positive and negative tweets with at least 10 likes. The neutral tweets were ignored.

.

6) Bitcoin historical data

- **It was used the cryptocompare api to get the following historical data:**

  - Bitcoin closing prices

  - Trading volume

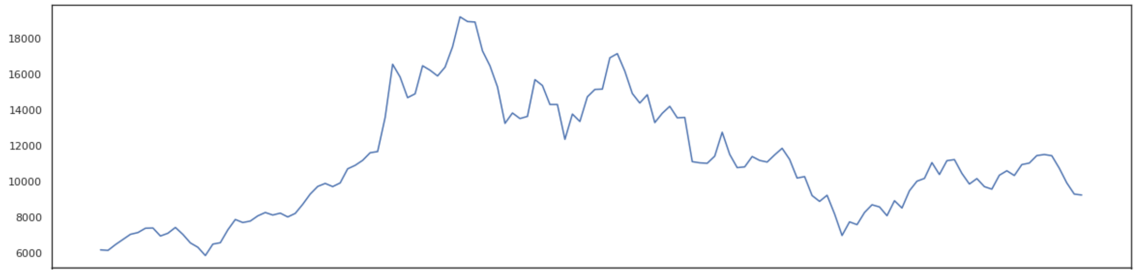
  - Timestamp

The % positive and negative tweets were added to the dataset:

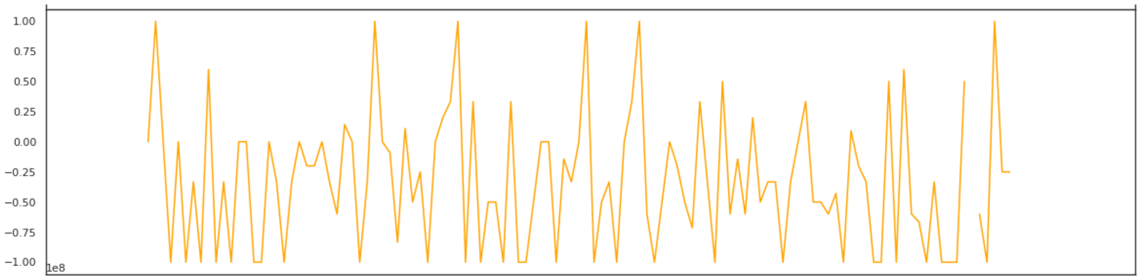
	close	volumeto	timestamp	positive	negative	samples
0	6155.00	3.317674e+07	2017-10-29	NaN	NaN	0
1	6125.00	1.754572e+07	2017-10-30	NaN	NaN	0
2	6450.02	2.411846e+07	2017-10-31	NaN	NaN	0
3	6739.79	3.082127e+07	2017-11-01	NaN	NaN	0
4	7025.00	7.440179e+07	2017-11-02	NaN	NaN	0
5	7124.36	4.260343e+07	2017-11-03	NaN	NaN	0
6	7366.00	3.676271e+07	2017-11-04	NaN	NaN	0
7	7379.17	3.223164e+07	2017-11-05	0.000000	0.000000	0
8	6929.97	4.709258e+07	2017-11-06	1.000000	0.000000	1
9	7084.87	2.911407e+07	2017-11-07	0.000000	0.000000	0
10	7410.00	6.844777e+07	2017-11-08	0.000000	1.000000	1
11	7020.00	4.456109e+07	2017-11-09	0.000000	0.000000	0
12	6543.20	9.082739e+07	2017-11-10	0.000000	1.000000	1
13	6300.00	8.234972e+07	2017-11-11	0.333333	0.666667	6
14	5835.00	1.700409e+08	2017-11-12	0.000000	1.000000	3
15	6475.78	1.136491e+08	2017-11-13	0.800000	0.200000	5
16	6554.92	4.431548e+07	2017-11-14	0.000000	1.000000	2
17	7275.12	6.712821e+07	2017-11-15	0.333333	0.666667	3
18	7858.00	8.362667e+07	2017-11-16	0.000000	1.000000	3
19	7687.51	6.982431e+07	2017-11-17	0.000000	0.000000	0
20	7770.00	4.481716e+07	2017-11-18	0.000000	0.000000	0

7) Plot the results from dataset

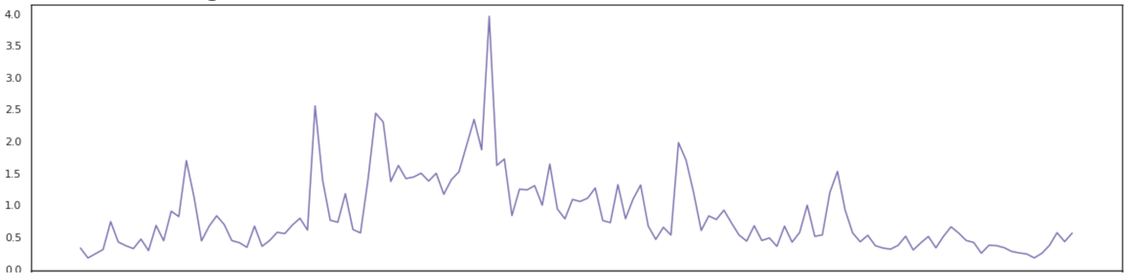
- **Bitcoin Close Prices:**



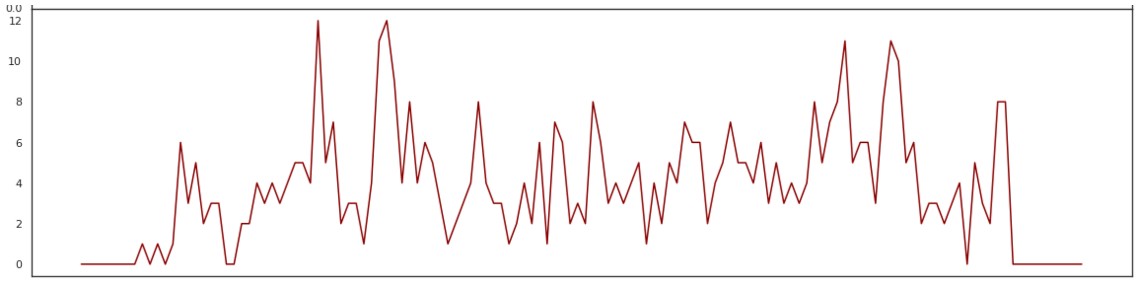
- **% Positive and Negative tweets**



- **Trading Volume**



- **Number of tweets >=10 likes**



8) Plot Heatmap and check for coorelation between features

We can see some correlation between the price,trading volume, and the number of tweets.

References:

Motivation to do it:

https://hackernoon.com/twitter-scraping-text-mining-and-sentiment-analysis-using-python-b95e792a4d64

Theory:

http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf

Pre Processing:

https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908

https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/

Datasets used:

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Code Inspiration:

http://blog.chapagain.com.np/python-nltk-twitter-sentiment-analysis-natural-language-processing-nlp/

https://www.kaggle.com/paul92s/bitcoin-lstm-model-with-tweet-volume-and-sentiment

API for historical data:

https://min-api.cryptocompare.com/