COSC 2671 Social Media and Network Analytics

Analysing and Tracking the Sentiment and Topics on Social Media

Kanru Wang 07/04/2018

1. Data Pre-processing and Exploration

1.1 The search word of this project is “dogecoin”, a type of crypto currency.

1.2 The REST API method is used to retrieve past tweets. As of 2018-04-07, if setting `since="2018-03-20", until="2018-04-08"`, tweets from 2018-03-28 07:16:50 to 2018-04-07 08:34:17 is retrieved. Tweets earlier than 2018-03-28 cannot be retrieved. We can infer that the REST API method can retrieve up to 10 days’ past tweets, not any longer or further.

1.3 A total of 7564 tweets are retrieved and used for analysis.

1.4 The stopword list is a combination of (1) English stopword list from nltk.corpus, (2) string.punctuation, (3) some specific stopwords for this analysis, (4) some regular expressions to match certain pattern. Additionally, for CountVectorizer function, it will apply its own stopword list on the tokenized words after our pre-processing.

1.5 The top 20 most frequent hashtags and their frequencies are

```
dogecoin: 2468, bitcoin: 851, doge: 764, litecoin: 426, thug: 369, crypto:
331, cryptocurrency: 307, giveaway: 226, ethereum: 179, free: 176, btc:
174, faucet: 171, ltc: 141, blockchain: 119, reddcoin: 114, freedoge: 110,
acceptslitecoin: 107, altcoin: 103, dashcoin: 93, etherium: 91
```

1.6 The top 20 most frequent userMentions and their frequencies are

```
dogecoin: 436, RedditDogecoin: 412, SatoshiLite: 362, coinok: 261,
dogecointicker: 244, JuanMoreno1771: 194, THUGCOIN: 188, killmefam: 182,
WildchildSings: 112, applancer_crypt: 90, DougPolkPoker: 89,
Applancer_pro: 89, CoinQuality: 86, KingOfBitcoin1: 80, dogecoin_devs: 78,
jetcoins: 75, REGA_fintech: 75, bitfinex: 75, moxy_one: 75, eBoostCoin: 71
```

1.7 The top 20 most frequent words (stemmed) in tweeter text and their frequencies are

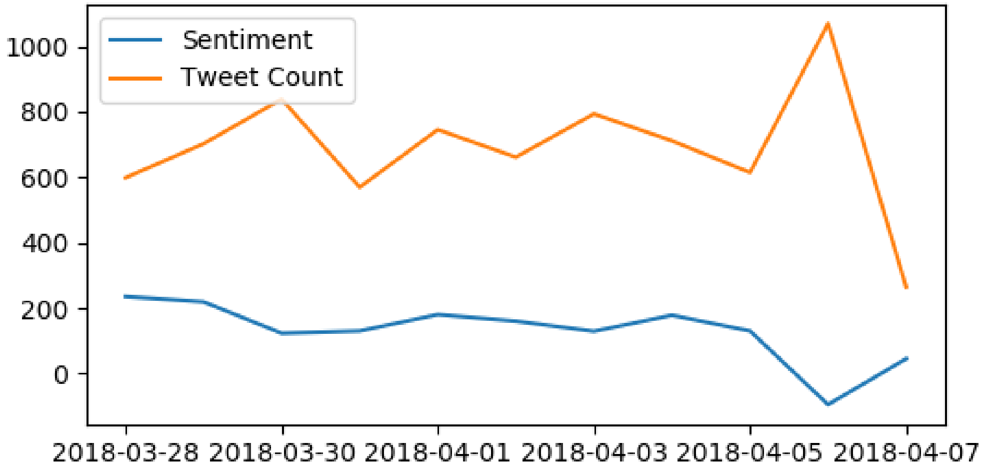
```
dogecoin: 3544, doge: 1709, r: 814, price: 583, btc: 581, free: 549, usd:
544, current: 531, thi: 514, valu: 509, coinbas: 495, it': 492, approx:
483, 1m: 483, vircurex: 483, cryptocurr: 456, fork: 452, @dogecoin: 433,
@redditdogecoin: 412, bitcoin: 408
```

2. Sentiment Analysis

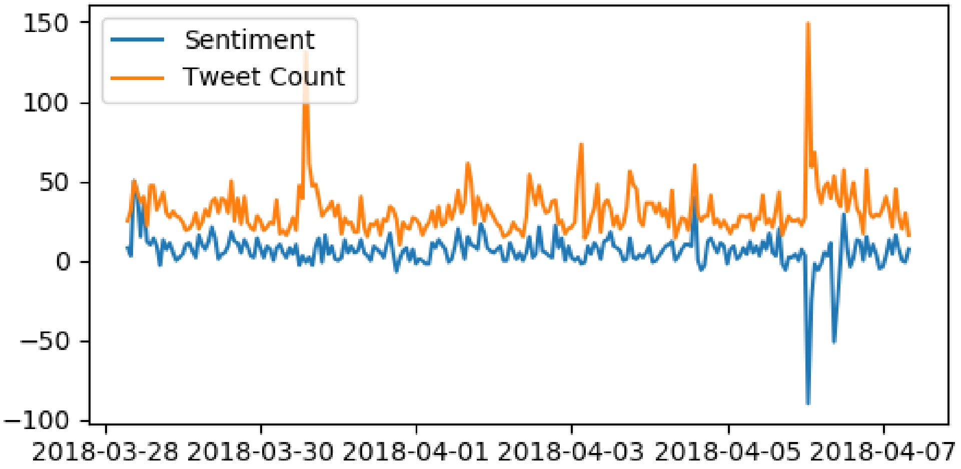
2.1 Using generic positive/negative word collection

We have a collection of generic positive words and a collection of generic negative words. The sentiment value in the plot is the total count of positive words minus the total count of negative words in the text in tweets appeared at each time interval. This calculation take into consideration both the volume of tweets during a certain time interval and the sentiment of each tweet.

2.1.1 By days:



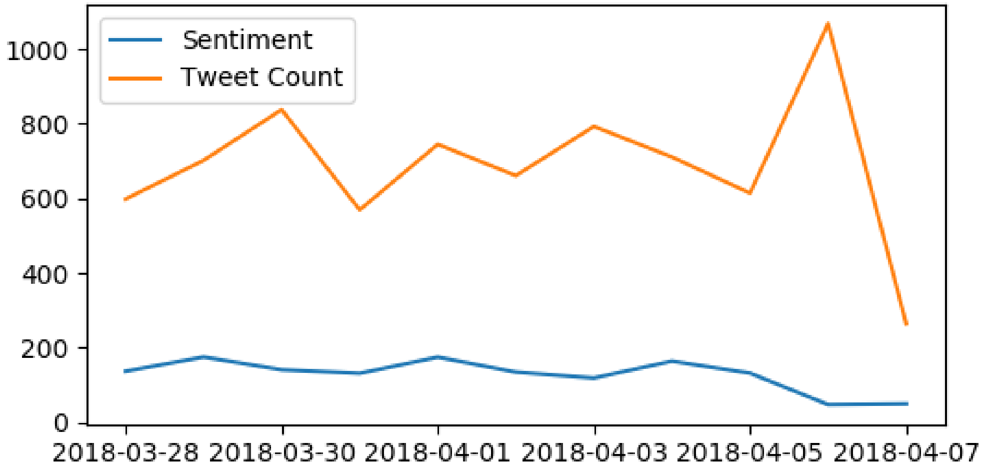
2.1.2 By hours:



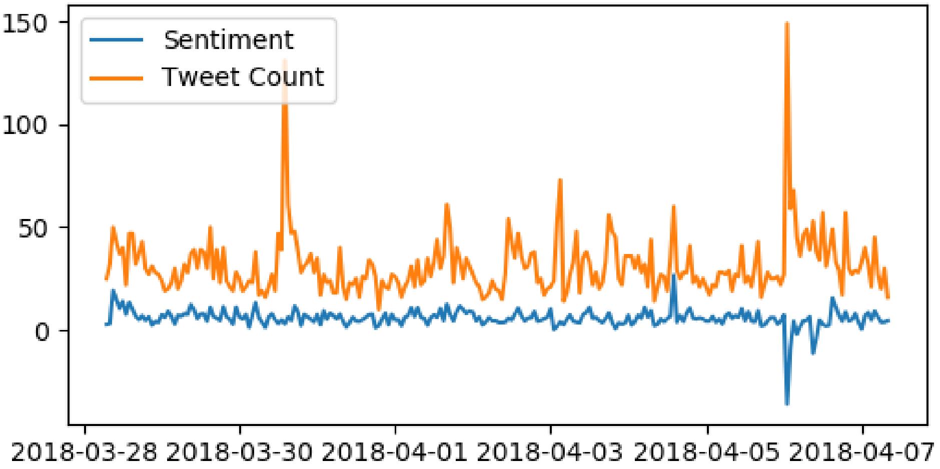
2.2 Using Vader sentiment analysis

Vader method’s lexicon was built from social media, so it is expected to be more relevant to analysing the sentiment of tweets. It also take into consideration punctuation, capitalization, adverb and presence of “but”. Therefore we should not pre-process the text (although it does not make a difference in our case).

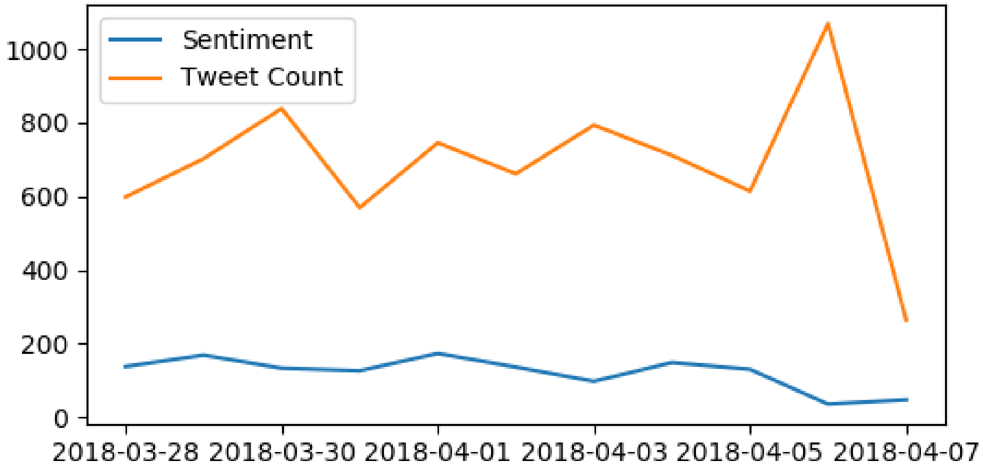
2.2.1 By days, using pre-processed tokens:



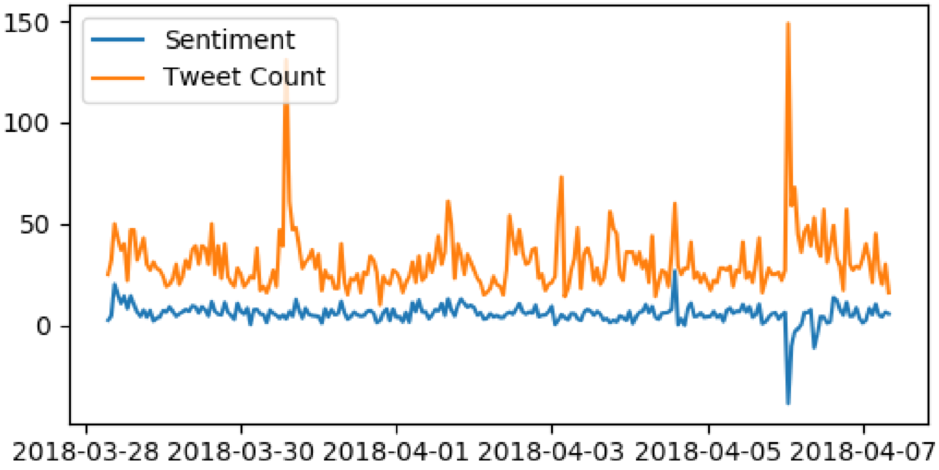
2.2.2 By hours, using pre-processed tokens:



2.2.3 By days, using raw tweet text:



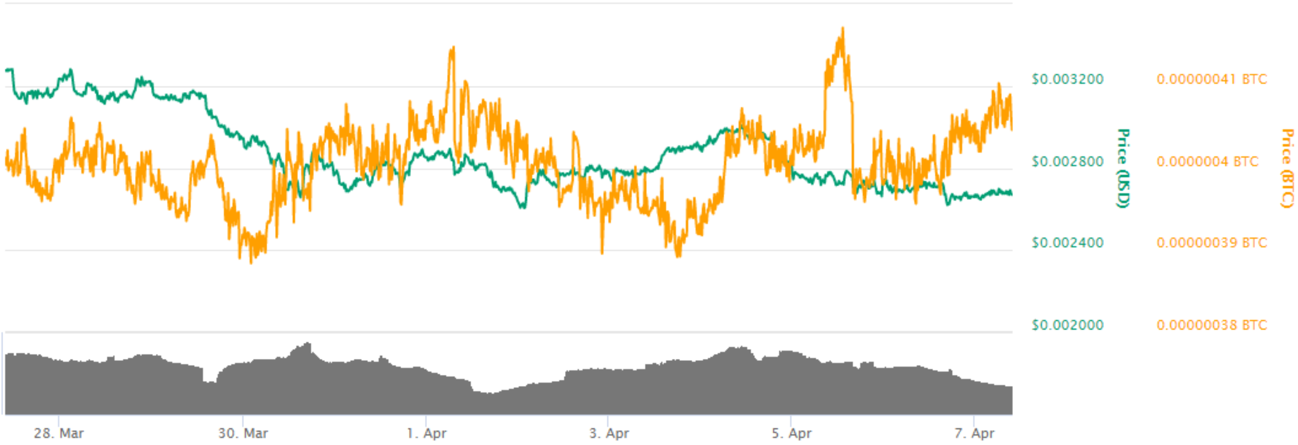
2.2.4 By hours, using raw tweet text:



2.3 Discussion

All plots agree that 2018-04-06 has negative sentiment. The main reason is that one tweet has been retweeted many times during that period of time. Here is the tweet text “`RT @SatoshiLite: So sad. It's time to shut down Dogecoin. https://t.co/WKdjAQjKBt`”. In fact it is a joke and does not have negative sentiment. The fact that it is a joke also partially explains the high volume of tweets during that period of time.

Alternatively, we can explain the negative sentiment by a price drop. The plot in below is the Dogecoin price in USD (green line) and in Bitcoin (orange line) around the same period of time from coindesk.com. The price of Dogecoin had an about 10% drop based on USD and an about 20% drop based on Bitcoin, on around 2018-04-05. However, there are still a lot of rises and drops in the price that are not reflected in the sentiment plot.



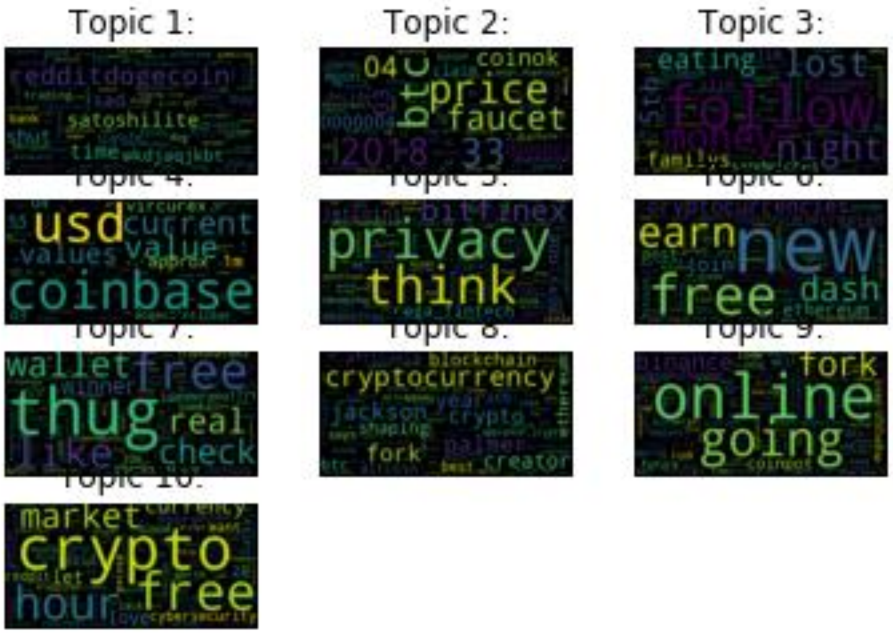
3. Topic Modelling

3.1 Method

All 7564 tweets are used for modelling. When using CountVectorizer function from sklearn, (1) ignore terms that have a document frequency higher than 95%, (2) ignore terms that have a document frequency lower than 2 occurrences, and (3) build a vocabulary that only consider the top 1500 words ordered by term frequency across the corpus. Extra stopwords are used, i.e. 'dogecoin', 'dogecoins', 'doge', 'https', 'coin', 'coins', 'bitcoin', 'litecoin'.

Latent Dirichlet Allocation on 10 topics is applied to pre-processed words from tweet text. The maximum number of iteration is 10. Learning method is set to “online” in order to speed up the process. The LDA model is visualized using word-cloud.

3.2 Result



```
Topic 0:
redditdogecoin time satoshilite shut sad wkdjaqjkbt buy bank trading
transfer
Topic 1:
btc price 2018 33 faucet 04 coinok 03 0000004 0000003
Topic 2:
follow money night lost eating 5th familys giveaway slim jim
Topic 3:
coinbase usd current value values approx vircurex 1m 55 04
Topic 4:
privacy think bitfinex rega_fintech jetcoins moxy_one eboostcoin
ebtcfoundation info wa
Topic 5:
new free earn dash cryptocurrencies join ethereum created post airdrop
Topic 6:
thug free like wallet real check winner juanmoreno1771 20k looks
Topic 7:
cryptocurrency fork palmer jackson year crypto creator blockchain shaping
ethereum
Topic 8:
```

```
online going fork binance coinpot dougpolkpoker future dogecoin_devs day
code
Topic 9:
crypto free hour market currency love let cybersecurity 28 want
```

3.3 Discussion

Topic 0 is all about retweeting this tweet “`RT @SatoshiLite: So sad. It's time to shut down Dogecoin. https://t.co/WKdjAQjKBt`”. As mentioned before, it is a joke and does not have negative sentiment.

Topic 1 is about (1) opportunities to earn faucet reward, and (2) coinok, an important Dogecoin community twitter account that frequently posts Dogecoin exchange rate.

Topic 2 is all about retweeting this tweet “`RT @killmefam: lost all my money to dogecoin this is my familys 5th night eating slim jim sandwiches`”.

Topic 3 is all about a twitter account “dogecointicker” which frequently posts Dogecoin exchange rate.

Topic 4 is about privacy. The most common tweet is “`If you guys think that a dogecoin fork with fake privacy is going to be the privacy coin of the future, you are on the Verge of a huge letdown`”.

Topic 5 is again about earning faucet reward.

Topic 6 is about thug coin.

Topic 7 is about Jackson Palmer, the creator of Dogecoin.

Topic 8 is again about privacy and about the most common tweet in Topic 4.

Topic 9 has an unclear topic.

Since similar topics appeared, it seems that a smaller number of targeted topics will result in more interpretable topics.

4. Summary

Twitter data of longer time horizon is needed to conclude if tweet sentiment is actually responding to Dogecoin price movement. LDA is a good tool to spot important events. LDA generated topics are sometimes overwhelmed by popular tweets that are retweeted for many times, so that LDA may fail to describe more meaningful topics.