

Dataset description

What is in the files

In the files provided you will find data quantifying the volume of online discussion mentioning multiple topics, and expressing different emotions and sentiments, around a number of popular cryptocurrencies. You can use it to find out that e.g. during day X, users of Bitcointalk posted N posts about crypto regulations, or that there was two times more bullish than bearish posts published on that day. Both absolute (number of posts/tweets) and relative (percentage of total discussion) numbers are provided. You can also find the amount of “total buzz”, which includes all the entries related to a given cryptocurrency, even if it was not assigned any topic or sentiment.

Naming convention

Every data file is named with the following convention:

`COIN_INTERVAL_[rel].csv`

COIN - Cryptocurrency that this data relates to. Possible values: see Cryptocurrencies section below.

INTERVAL - How the numbers are aggregated (by taking sum over the given time interval). Possible value: 1H, 24H.

rel - Files with names ending with “rel” contain relative number. Files without “rel” contain absolute number. Keep in mind that columns in “rel” files are also named with “rel” suffix, e.g. “Positive_rel” - it can be useful if you want to concatenate features from multiple files.

Data sources

Data included in this repository comes from two data sources: Twitter and Bitcointalk. Bitcointalk dataset contains entire history of relevant Bitcointalk discussions. Twitter contains the full timeline of over 5000 Twitter users identified as most influential in the crypto space.

Cryptocurrencies

There is data provided for each of the following coins:

- Bitcoin
- Ethereum
- Ripple

Keep in mind that for some of them the volume of discussion is smaller, which may result in less informative data.

Time span

Bitcointalk datasets include the entire historical data from Bitcointalk, up until November 2018.

Twitter datasets include data from November 2016 up until November 2018.

Topics and sentiments

In each file (except price and total buzz data sets) there are numbers for 114 topics and sentiments tracked for each data source and cryptocurrency mentioned above. You can find the full list of available indicators in a separate file.