INF 550 Project Proposal: Sentiment Ananlysis of Blockchain Companies based on Twitter

Yuan Li, Xiaofan Zhao, Qi Zhong

October 3, 2018

1 Abstract

Since its appearance back in 2008, blockchain technique has been an enduring focus of various fields including computer science, finance, and economy. Numerous companies started their business based on this technology and its applications. Some of these companies have great influence on blockchain ecosystems, while some of them are facing the challenges of trust and reputation. As one of the main platforms of communications and discussion between participants of the blockchain ecosystem, in the levels of individuals, non-profit foundations and companies, Twitter could be an applicable data resource to help us understand more about these companies. In this project, we are going to focus on a portion of these companies, look into the attitude and keywords when Twitter users are talking about them and the location distribution, try to figure out the connection between the attitude trends towards those companies, and the prize trends of five different blockchain-based cryptocurrencies(BTC, BCH, ETH, XRP, EOS). Also, by studying the pattern of these tweets, we are going to gather and analyze more information about their relation, clusters, and existence of paid posters.

List of companies:

Coinbase, Karken, BitPay, Bitmain, bitcoin.com, Binance, Bitstamp, Xapo, Shapeshift, Purse.io

2 Questions and Analysis

2.1 General popularity about blockchain and companies

- The popularity time trend of keyword: ”blockchain”.
- Then we plot the overall popularity time trend of the companies in the list, if the trend is similar to the trend of blockchain then we can say that we have a relatively good sample of blockchain companies.

2.2 Sentiment analysis of blockchain companies

- Sentiment and popularity of companies.
- Term frequency in negative and positive tweets for each company.
- Plotting the companies’ sentiment trends and cryptocurrencies trends plot side by side to see the possible correlation between them.
- By pointing the location of a tweet with a color related to attitude towards the company, we can have the location distribution information of them. Some of the companies could have more supporters on specific locations.

2.3 Connection of companies

By calculating how often are people talking about two companies together, we’re going to try to make a matrix to evaluate how related those companies are and generate the graph to visualize the connection and clusters.

2.4 Posting patterns and the involvement of paid posters

We assume that paid posters might have a clear pattern of posting the time during a day, and the sentiment of companies who are hiring paid posters could have a polarizing feature. So for the companies in our list, we are going to make an accumulated post number trend over 24 hours and the histogram of sentiment for each company, to see if they're hiring paid posters.

3 Datasets description

3.1 Data collection

Tweet scraping Most data will be achieved from Twitter. Tweepy package of Python will be used to create API through which Twitter messages can be fetched with specific keywords. The data collected are finally saved in the form of csv files.

Factors: keywords and time span To only get information that is necessary to this study, a list of factors need to be set as the category of information to achieve from the Twitter messages. For this study, the factors are: the user id, text, time when the message was created, location, hashtag, number of the user followers, number of accounts the user follows, repost or original, is there picture in the Tweet or not, number of comments, number of likes.

The keywords for searching and information scraping are the names of the companies and "blockchain". The total time span of the data is approximately 3 weeks.

3.2 Popularity index and trend

When conducting popularity and trend analysis based on the dataset, the extent of popularity of blockchain and different blockchain companies will be quantified using RiteTag, a tool that can provide real-time Twitter hashtag popularity index and trend. Then the result of this analysis can be illustrated in the form of lines chart using R.

3.3 Price information of cryptocurrencies

As mentioned before, the price variation of cryptocurrencies will be compared with the trend of sentiment change. Here, the information of the price can be achieve from CoinMarketCap via certain API.

3.4 Positive/negative word frequency

The analysis of strong suits or drawbacks of companies with positive or negative sentiments is an important part of sentiment analysis in this study. During this process, positive/negative word frequency counting is a must. This task will be implemented by coding in Python to count the word frequency for the text column in csv files.

3.5 tools

For the data collection, we are going to use python and tweepy; and for analysis, python and R pakages could be used. For visualization, we use R for normal plots and Tableau for geographical information presenting.

Contribution .

Yuan Li: data collection and analyzing;
Xiaofan Zhao: data collection and analyzing, visualization;
Qi Zhong: data collection and analyzing, Presentation preparation.

Reference .

Basic concepts of blockchain: https://en.wikipedia.org/wiki/Blockchain

https://ritetag.com/best-hashtags-for/blockchain

Usage of Tweepy: http://docs.tweepy.org/en/v3.5.0/getting-started.html

Orlando Troisi et al. Big data and sentiment analysis to highlight decision behaviours: a case study for student population https://doi.org/10.1080/0144929X.2018.1502355

Bo Pang er al. Opinion Mining and Sentiment Analysis http://dx.doi.org/10.1561/1500000011