Predicting Ethereum Price Change Using News Articles

Group Project Proposal
Group 35

Ashutosh Bansal (19323385)
Ming Jun Lim (21337483)
Markus Scully (16321114)

CS7CS4 Machine Learning
October 27, 2021

1 Motivation

The market price is determined by the demand and supply of traders in the market. In this group project, the team will be tackling Ethereum market price by estimating if the future price increases or decreases based on text processing sourced from the web such as social networking sites, news sites and forums. The problem will be converted into a classification problem by comparing the future next market price with the current market price instead of a regression problem of estimating the future market price. Can news or posts of Ethereum articles really affect the market price?

2 Dataset

Binance is a cryptocurrency exchange containing many popular cryptocurrencies such as Ethereum (ETH), Bitcoin (BTC), Litecoin (LTC), etc. The market price of ETH will be sourced from the platform and some processing to convert the continuous value to discrete positive (+1) and negative (-1) class labels for the project. Binance provides a set of API [1] and WebSocket endpoints available for the team to collect the data.

Reddit and Twitter are online social networking platforms that allow for public topic-based communication. We intend to scrape these platforms using the open source Python libraries praw for Reddit and snscrape for Twitter. Both websites offer means of narrowing the search space to only posts which are likely to be relevant to Ethereum - subreddits on Reddit, and hashtags on Twitter.

Websites like Coindesk[2], the conversation[3] contain prices, articles and analysis and can be accessed using Python scraper libraries like beautiful-soup.

3 Method

For the analysis of textual data retrieved from social media and news platforms, we intend to extract features by employing the bag of words model. Multiple variations of this will be tested, including the use of bi- and trigrams, word pruning, and n-gram hashing.

The team is planning to use KNN and logistic regression, while working on weekly assignments, it has been observed that the performance of these models is good so that could be a good starting point. The team will look to Improve the models by selecting optimal 'k', 'q', 'c' values and choosing appropriate penalty (L1 or L2) whatever suits best to our models.

[1] `https://github.com/binance/binance-spot-api-docs/blob/master/web-socket-streams.md`

[2] `https://www.coindesk.com/price/ethereum/`

[3] `https://theconversation.com/global/topics/ether-39759`

4 Intended Experiments

Cross-validation will be used to determine the hyperparameters for learning techniques mentioned in the previous section. The model performance of each learning technique will be evaluated by precision, accuracy, F1-score, ROC and AUC. Baseline models such as always predicting the mean and random prediction will be used to compare model performances.