Predicting Ethereum Price Change Using News Articles

Ashutosh Bansal
19323385

Ming Jun Lim
21337483

Markus Scully
16321114

I. Introduction

THE cryptocurrency market have witnessed an explosion in the recent years with Bitcoin[1] hitting lots of new time high and many investors moving to this trading space [1]. With the internet constantly evolving, social networking sites such as Facebook, Twitter, Instagram and etc have become a means of communication between people. There have been researches in text sentiment analysis to predict price changes for Bitcoin [1][2] as Bitcoin is considered a dominant cryptocurrency. However many other cryptocurrency have been developed and have been in demand such as Ethereum [2]. The project 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 and forums. The problem is converted into a classfication problem by comparing the future market price with 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?

The input to our algorithm is text sourced web scrapers on social networking sites. We then compare several machine learning methods such as k-Nearest Neighbours (kNN), logistic regression and Support Vector Machine (SVM) to output a predicted label, +1 if price increased or -1 if price decreased.

II. Dataset and Features

The dataset consists of Ethereum related articles and class labels gathered from Twitter, Reddit and Binance[3] over the last four years. Data from Twitter and Reddit are merged with a similar column names: *date content* and *popularity*.

A. Data Source

Binance provides a set of API[4], WebSocket endpoints available to collect live data and old historical data[5]. Monthly trade data of ETHUSDT[6] symbol is collected from August, 2017 to October, 2021 with a total of 657,446,190 data points. These data points are processed by calculating the average price by days and discretizing the continuous values into +1 and -1 class labels by comparing it with the previous day resulting into 1,537 data points.

[1] Bitcoin is the first well known cryptocurrency started in the late 2000s

[2] Ethereum is programmable blockchain and cryptocurrency described as Ether

[3] Binance is one of the largest cryptocurrency exchange

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

[5] https://github.com/binance/binance-public-data

[6] ETHUSDT is a symbol for Ethereum and USDT pair, 1 USDT is 1 USD.

B. Feature Extraction

Several pre-processing steps are applied. The first step is tokenization, splitting input character sequences into tokens. The endings of the tokens are chopped of by stemming. The tokens are pruned using a set of stop words as they contribute little to no information such as words "is", "the", "a", etc. Tokens appearing frequently and rarely were also pruned through the `min_df` and `max_df` parameters in tokenization, this step was very important to reduce the feature size. The algorithm uses Natural Language Toolkit `nltk` library providing stemming functions such as `PorterStemmer` and a list of stopwords `nltk.corpus.stopwords.words("english")`.
These sequence of tokens are mapped to a feature vector using the bag of words model, the model consists of hyper-parameter n-grams and cross validation tested in sections below. Some general pre-processing on text data from [2] were applied onto our dataset consisting of:

- Converting all words to lower case.
- Removing words containing numbers.
- Removing symbols from words ("#","@").
- Removing words containing nonenglish characters.

III. Methods

A.

IV. Experiments

V. Results

VI. Discussion

VII. Summary

References

[1] Nguyen Huynh Huy, Bo Dao, Thanh-Tan Mai, Khuong Nguyen-An, et al. Predicting cryptocurrency price movements based on social media. In *2019 International Conference on Advanced Computing and Applications (ACOMP)*, pages 57–64. IEEE, 2019.

[2] Otabek Sattarov, Heung Seok Jeon, Ryumduck Oh, and Jun Dong Lee. Forecasting bitcoin price fluctuation by twitter sentiment analysis. In *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, pages 1–4. IEEE, 2020.