Cryptocurrency Market Prediction Using Twitter

Final Report

Will Badart
University of Notre Dame
South Bend, Indiana
wbadart@nd.edu

Matthew Fabian
University of Notre Dame
South Bend, Indiana
mfabian1@nd.edu

Shane Ryan
University of Notre Dame
South Bend, Indiana
sryan8@nd.edu

Mara Staines
University of Notre Dame
South Bend, Indiana
mstaines@nd.edu

ABSTRACT

Cryptocurrency as a commodity has proven to be quite volatile, making prediction of market trends highly valuable. Our team leveraged the constant stream of tweets discussing cryptocurrency on Twitter in order to build a model predicting changes in a cryptocurrency's value. This model showed some correlation between tweet sentiment and currency value, suggesting this area should be explored further.

1 INTRODUCTION

Every day, Twitter users produce more than 500 million tweets[1]. This makes Twitter an ideal resource for social sensing — the use of humans as sensors. Users tweet and retweet about events, places, emotions, and more; the aggregation of this data has proven powerful for event detection and prediction. One of the fastest-growing conversations taking place on the platform is over cryptocurrency.

The first decentralized cryptocurrency, Bitcoin, was created in 2009. Currently, over 3500 different cryptocurrencies exist, many simple derivations of Bitcoin. The overall market capitalization of cryptocurrencies has ballooned over the past year to 278 billion dollars. As an example of this rapid growth (and volatility), Bitcoin was being exchanged for $1,200 per coin in April 2017. It reached a high of $19,500 per coin in December 2017, and is at $9,600 per coin as of May 2018.

Because cryptocurrency as a whole is a growing market that is both widely discussed and highly volatile, our team believes that tweets may reflect trends in cryptocurrency value. This belief is rooted in three factors. First, many cryptocurrency investors are speculative. They purchase a cryptocurrency not for its intrinsic value, but because they believe the price will increase and they can later sell it for a profit. Second, these speculative investments are based on news stories, suggestions, and trends in the market. Speculative investors don't purchase cryptocurrencies randomly, rather, they choose to do so because of some sort of input. Finally, we believe that these news stories, suggestions, and market trends can be identified on Twitter before most potential investors have the opportunity to find, analyze, and act on them. In that way, successfully using Twitter data to predict changes in cryptocurrency markets would provide a significant advantage in cryptocurrency trading.

In this project, we will build a model to predict changes in the value of several major cryptocurrencies based on real-time data from Twitter.

2 APPROACH

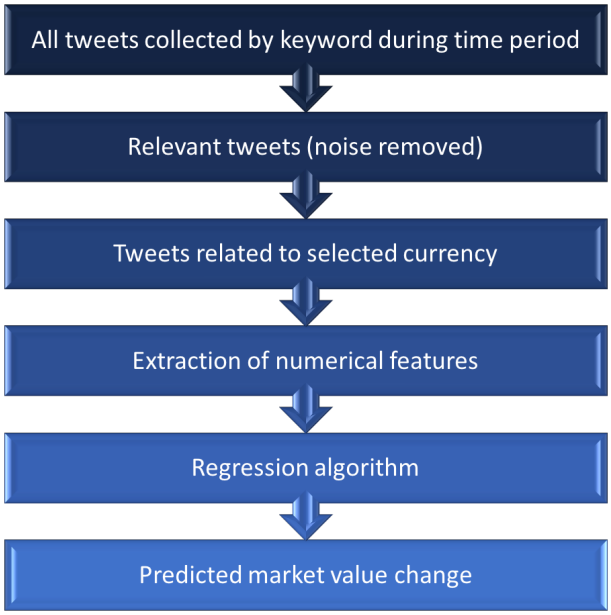
2.1 Data Sources

Data for this project is collected from Twitter. We streamed tweets using the Twitter API, which allows for tweets to be collected in near real-time. The filtering capability of the Twitter API allowed us to only collect tweets that contain certain keywords. We selected both general keywords (e.g. `cryptocurrency`, `altcoin`) and cryptocurrency-specific keywords (e.g. `bitcoin`, `BTC`, `ethereum`). If a tweet contains one or more of these keywords, our program stored the tweet body as well as important metadata (user ID, timestamp, retweets, favorites, etc.) in JSON format and appended it to a data file.

Many sites provide historical data on cryptocurrency value. We utilized the API of CryptoCompare.com to retrieve cryptocurrency values for the time-span covered by the collected tweets. Initial analysis was completed using only Bitcoin values, but in future work could expand to include `Bitcoin`, `Ethereum`, `Bitcoin Cash`, `Litecoin`, and `Rippler`.

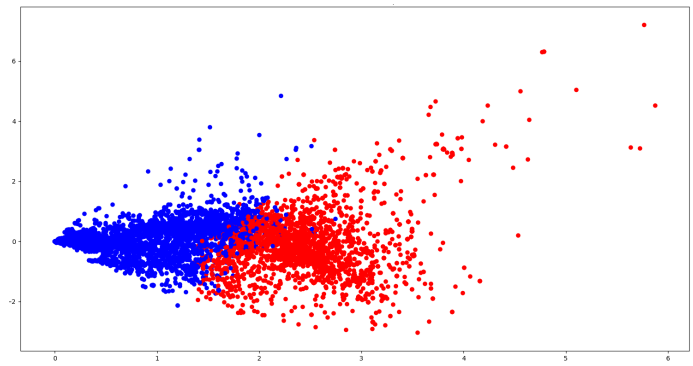
2.2 Methodology

Figure 1: Flow of data processing



Once data is collected from Twitter, it is cleaned, categorized, and used in a regression model to make currency value predictions (Figure 1). Twitter data is noisy, so the first component of data processing eliminates tweets irrelevant to cryptocurrency which made it past the API filter. Our strategy for this task is to use a clustering algorithm to group tweets as relevant or noise. The data is tokenized using the NLTK TweetTokenizer, which splits the tweet text into distinct words. This submodule is superior to simply splitting the data on whitespace, because it has been coded for the type of text encountered on Twitter. For example, it can split text to preserve common emoticons. The tokenized tweets are used to generate a matrix of token counts. The similarity of two tweets' token count vectors is the measurement used as distance in the clustering algorithm. Tweets are clustered using the KMeans algorithm (Scikit Learn MiniBatchKMeans due to data size). This algorithm fits the purpose well, as the number of clusters is known (n=2), and outliers are unlikely due to the character limit imposed on tweets (280 characters). Figure 2 shows a visual representation of the collected tweets after the clustering.

Figure 2: Relevant/irrelevant clustering Projected to 2D space.

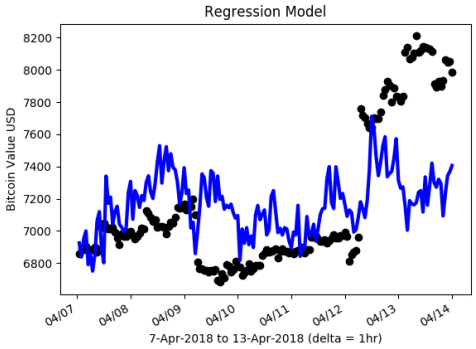


Once noise was removed, data from the tweets was used to calculate several features for use in the regression model. Features that were calculated for each time-span include the number of tweets, average sentiment of tweets, and standard deviation of tweet sentiment. A tweet's sentiment was calculated using TextBlob, which relies on NLTK. The sentiment is a score from -1 to 1 (where -1 is extremely negative and 1 is extremely positive). A completely neutral tweet would have a sentiment score of 0. Standard deviation of sentiment aims to capture volatility in tweet content. Many neutral tweets would have an average sentiment close to 0, as would equal numbers of highly positive and highly negative tweets. However, the standard deviation for neutral tweets would be significantly smaller than that for tweets with opposing sentiment.

Due to constraints on time granularity when collecting historical currency prices, tweets are grouped by the hour to generate an instance and calculate its features. One week's worth of continuous data has been selected for modeling, allowing us to test the limits of our model in predicting micro-trends. This is especially useful for cryptocurrencies (as opposed to the stock market) due to the excessive volatility that can be found in a coin's price.

For the regression model, we use a Ridge Regression. Regression for 1 week with 1-hour data points is displayed below in Figure 3. This model showed some predictive ability (R^2 = 0.18).

Figure 3: Regression model for 1 week of data.



3 RESULTS

Evaluation for our model relies on knowing the actual cryptocurrency values over time. Comparison of our predicted changes to actual changes was measured by examining R^2. Our goal was to maximize the accuracy of our predictions and minimize the time needed to make a prediction. We found that one week of tweets (delta = 1 hour) provides mild predictive ability: $R^2 = 0.18$. Our first attempt had 24 datapoints (1 day), and performed significantly worse ($R^2 = 0.02$)

The accuracy for our model's predictive ability can be measured in two ways: on historical data (on days for which we've collected tweets) and on current data (predicting towards the future). The first method, performed in this research, involves building a model on historical data, then predicting the target value for the time-span immediately following the last data point. This prediction was compared to the actual value from CryptoCompare.com. The second method, using current data, streams live data into the model, predicts the target value at the end of the data, then compares the predicted value to the true value once the time of the prediction is reached. We would like to attempt this type of validation in the future.

To evaluate the efficacy of clustering for noise removal, we perform regression twice over the same time period– once with all tweets, and once with only tweets from the relevant cluster. R^2 was significantly lower for the "less noisy" tweets– 0.18 vs 0.14. This shows that the intuition to remove tweets a human views as irrelevant is not necessarily reflective of the ground truth. In future work, we would like to rework the clustering model in order to see if any sort of tweet sample size reduction could improve accuracy.

Overall, tweet sentiment has a modest correlation with cryptocurrency value. This initial research suggests that Twitter data could prove valuable in predicting crypto market trends with the right data set and features.

4 DISCUSSION

This research has presented several unique challenges. While Twitter data is plentiful, ensuring quality is difficult. First, technical resources must be chosen carefully to handle the extended, continuous up-time of a Twitter streaming program as well as the large files resulting from this stream. Because the data is temporal, missing a continuous chunk of data can hinder the model. In addition, it is difficult to extract quality text from Tweets. When we conceptualize Twitter, we think of our friends, family, and journalists sharing complete thoughts. In reality, a large number of tweets that are picked up by keyword filters are bots, scams, giveaways, or simply incoherent. With our desire to leverage tweet sentiment, removing this noise becomes extremely important. What makes a tweet irrelevant is not clear cut, and we do not have any access to ground truths. As we discovered, the tweets we assumed to be irrelevant actually positively impacted the model. This project has been a great experience because in the real world, data is messy, noisy, and incomplete. Developing strategies to still find value in this data has helped us all grow immensely over the course of the semester.

REFERENCES

[1] David Sayce. [n. d.]. Number of tweets per day? ([n. d.]).