

Cryptocurrency Market Trend Analysis and Prediction

Saranya S.[#]

[#]*Financial Engineering, Department of Mechanical Engineering, College of Engineering, Trivandrum, Kerala, India*

Abstract— From a long time, there have been predictions happening in the financial market, such as stock market and bond market. Cryptocurrency market is on rise since the past few years but predictions about the market trend followed by these currencies is still in its nascent state. This study is based on the relationship between current sentiments of public opinion and price variation in this market, and is concerned with predicting whether cryptocurrency market will trend upward or downward using sentiment analysis, which is one among the various applications of natural-language processing. Bitcoin is the selected cryptocurrency. A platform to predict the future market trend of cryptocurrency market is built using Naïve Bayes classifier, in Python language. The validity of the model is tested by finding correlation between sentiment polarity of tweets with price values of respective time period. An accuracy of 89.86% is achieved by this model.

Keywords—Cryptocurrency, Sentiment Analysis, Natural-language processing, Naïve Bayes classifier, Bitcoin, Python.

I. INTRODUCTION

The method of making an attempt to determine the future value of any financial instrument or an asset that is used for trading is known as a financial market prediction. The successful prediction of the future price or trend of any financial instrument could yield significant benefit as in profit and returns. As technology is developing, there are many improvements in the prediction of future prices, market trends and market returns. Financial market prediction is not an entirely new discovery or event. Prediction of mature financial markets have been researched on and studied at length, where stock market is the long standing choice to continue predictions on. . The cryptocurrency market is on an up rise since the past few years, whereas predictions about trends of this market is still in its nascent stage. There are no strategies that are established in this market in order to carry out the trading. 2017 has been a year where cryptocurrency markets dominated in the public making news throughout. Many new applications, platforms and technologies were being launched that year. Also, 2017 was the year when Bitcoin reached its highest with the record price of \$19,850 on 16th of December. It was one of the greatest bull markets in the recent times that showed sudden growth but rapidly fell down below \$12,000 within days. Over the coming days, price of Bitcoin recovered, climbing back beyond \$16,000 and higher in all other cryptocurrency exchanges. In this study, the goal is to build a platform that will predict the future market trend, whether high or low of the cryptocurrency market by selecting Bitcoin as

the preferred cryptocurrency. The prediction will be based on the texts that are tweeted by the Twitter users about Bitcoin, using the method of sentiment analysis. Traditional time series prediction models depend upon data that can be separated into elements such as trend, seasonal and noise, to produce efficiency. These types of models are more suitable for a task such as prediction of variables that depend on seasonal effects. Since Bitcoin market lacks seasonality and is highly volatile, these methods are not effective in its prediction. Machine learning methods are more suitable because of the complexity involved in Bitcoin markets.

The purpose of this study is to find out whether future market trend in a cryptocurrency market can be predicted by using sentiment analysis. This can be categorized as fundamental analysis since it relies on the public opinion and reviews. Future trend can be studied by first collecting tweets about Bitcoin as the primary information and then sentiments are classified positive and negative. Here, supervised machine learning is used for classification and other text mining techniques are used to check text polarity. Naïve Bayes classification algorithm is implemented to check and improve classification accuracy. Data collected is Twitter data where tweets regarding Bitcoin from the year 2014 to 2017 is collected.

The objectives of the study are “To analyse the sentiments of cryptocurrency market trend based on Twitter data” and “To build a model that predicts the future market trend based on sentiments”.

Sentiment Analysis - Sentiment analysis is the process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing [1]. The process of sentiment analysis allows perception and anticipation of a new product, helps organizations to track the reception of new brands and popularity, company reputation and flame or rant detection.

Cryptocurrency – Cryptocurrencies are digital or virtual medium of exchange which is equivalent to other monetary currencies and uses cryptography to ensure security. Cryptocurrency market has currencies other than Bitcoin, like Ethereum, Litecoin, Bitcoin Cash, Ripple, and Stellar which are traded in cryptocurrency exchanges worldwide.

A. Literature Review

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity, where the entity can represent events, individuals or topics. These topics will mostly be covered by reviews that express a mutual meaning. However, some researchers stated that OM and SA have slightly different notions. Opinion Mining extracts and also analyses people's opinion about an entity whereas Sentiment Analysis will identify the sentiment expressed in a text then analyses it. Therefore, objective of sentiment analysis is to find views and opinions, identify which sentiments they express, and then categorise the polarity of these sentiments [3]. [4] pointed out that sentiment expressions may not necessarily be always subjective in nature. Anyway, there is no basic fundamental difference between document and sentence level classifications because sentences can be just short documents [5]. Two long and detailed surveys which were presented in [6] and [5] focused mainly on the applications and challenges in sentiment analysis. The techniques used to solve every problem in the field of sentiment analysis were mentioned by them. [7], [8] and [9] have given short surveys that depicted the new trends in sentiment analysis. Various tools and open source packages were used to build the news collection, aggregation engine and also sentiment evaluation engine. It is also stated that the time varied values of News Sentiment will reflect a strong correlation with original stock price variation [10]. When applying machine learning to cryptocurrency was a new field with limited research happening, by using Bayesian regression, an 89% return on investment over fifty days of trading Bitcoin was achieved [11]. Yet another approach predicted the price fluctuation of Bitcoin using random forests algorithm with an accuracy of 98.7% [12]. Another study that focused only on classifying tweets was done and used several approaches to find the best model. It achieved an accuracy of 79.2% with maximum entropy, 84.2% with Multinomial Naïve Bayes and 82.9% using a support vector machine [13]. A text mining based framework is demonstrated to determine the sentiment of news articles and to depict its impact on energy demand. News sentiment was quantified and shown as a time series and was compared with variation in energy demand and prices [14]. The relationship between views for Bitcoin, Bitcoin price and tweets was studied on Google Trends. A weak to moderate correlation was found by the author between Bitcoin price and both positive tweets on Twitter and Google Trends views. The author found this to be proof that they can be used as predictors. Unfortunately, one limitation of this study is that the sample size is only 60 days [15]. From the papers studied, it can be noticed that there have been no model built for the future market trend prediction in cryptocurrency market using Twitter sentiments. So, this can be considered as a research gap, and a model can be built in order to predict sentiments regarding the cryptocurrency market with taking the data as twitter messages known as "tweets" There have been studies where different classifier models are tested for their efficiency, therefore Naïve Bayes model have been selected for performing classification in this work.

B. Methodology and Data Analysis

The methodology adapted to build the prediction model is as shown in Figure 1, in which design of the system is depicted.

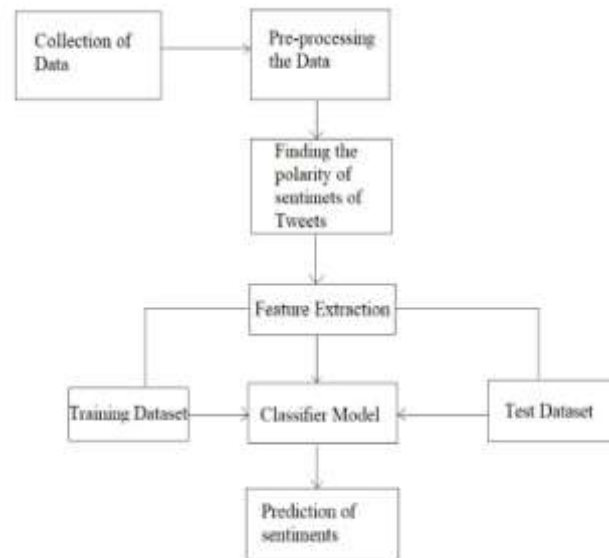


Figure 1. Work Flow Diagram

1. Data collection

Sentiment analysis needs adequate amount of data in the form of a corpus in order to train the model successfully. Data collection can be done from Twitter using different corpuses which are available in the internet or streamed directly into the domain software by using different APIs (Application Programming Interface). There are two types, search API and stream API. The system continuously send queries to the Search API in order to collect tweets, where a small delay is given to correct the rate limit. Here, tweets are collected from different databases and repositories in the internet. Twitter data about Bitcoin from the year 2014 to 2017 are collected and used to train the model. There are around 1.2 million filtered tweets in the dataset used to train this model. A labelled dataset containing 39,580 tweets is created in order to test the model and check its accuracy.

2. Text Pre-processing

Pre-processing of data is executed so as to clean it and to prepare the text for carrying out the process of classification. Almost all the tweets and texts from public interaction platforms will contain many words that will never contribute to the sentiment that the text is trying to convey. Most of the tweets will include tags, symbols, emoticons and characters that are not meaningful at all. In order to accurately obtain a tweet's sentiment, filtering the tweets must be done in order to avoid noise from its original state. Before processing or cleaning, many of the words and characters in the tweets adds noise to the process of sentiment analysis. This procedure will significantly reduce the size of the input text document. Some of the important features of pre-processing are:

- Tokenization
- Stemming
- Elimination of stop-words
- Whitespace removal

3. Sentiment Polarity Detection

A dictionary based approach is used to detect the polarity of the sentiments of the tweets. In order to build the dictionary, there should be two main groups of word collections; positive words and negative words. Now the words in data corpus can be matched against these word lists and the polarity score of that document can be calculated. "Loughran and McDonald" dictionary can be used for the detection of polarity of tweets. This dictionary contains finance related word collection. In this dictionary there are 2006 positive words in the positive words list and 4782 negative words in the negative words list. Polarization algorithm using NLP is used to detect the polarity of the sentiments.

4. Feature Extraction

The most important step in the sentiment classification problem is to extract certain features and select text features. Some of the features are:

- Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting or uses term frequency weights to indicate the relative importance of features.
- Parts of speech: Finding adjectives, as they are important indicators of opinions.
- Opinion words and phrases: These are the words that are commonly used to express opinions including good or bad, like or hate.
- Negations: The appearance of negative words may change the opinion orientation.

5. Training and Testing the Classifier Model

The classifier model algorithm selected to build the prediction platform in this work is Naïve Bayes Classification model. Training the classifier using a dataset is an important aspect of supervised learning techniques. A part of the data collected is set apart as training dataset and the other part is the testing dataset. This training data is given as input to the classifier to make it easier for the prediction of unknown data by this classifier. A portion of dataset kept apart as testing data is given as input to the classifier model. When an unknown dataset is given, the classifier will predict accurately if training and testing of data were done properly. In this study, classification process is done by building a Naïve Bayes classifier that belongs to the family of "probabilistic classifiers". The Naive Bayes classification technique is based on Bayesian theorem and it is particularly suitable in the cases where dimensionality of the inputs is high. Even if the technique is simple, Naive Bayes can sometimes outperform more sophisticated classification methods. Naive Bayes

classification algorithm mainly works for binary or two-class and multi-class classification problems.

C. Results and Discussion

The entire focus of study is to build a model that predicts the market trend with respect to sentiments of tweets given as inputs. The prediction accuracy of classifier will aid in improving quality of analysis. After using inbuilt function to calculate accuracy, the accuracy of model is found to be 89.86%. Correlation was found to be positive between public sentiments of Twitter data and price values of Bitcoin. The value of Pearson Correlation Coefficient is found to be 0.685. Input is given as a group of tweets about Bitcoin on a particular day. The model will predict that market trend is high when more than 70% of input data have positive polarity, and it will predict market trend is low when more than 70% of input data have negative polarity. Tweets are found to have an influence on the price of Bitcoin on next day.

II. CONCLUSIONS

The objectives of this study to analyse the Twitter sentiments about cryptocurrency market and to build a platform that will predict the future market trend is satisfied. The prediction is based on the texts that are tweeted by the Twitter users about Bitcoin, using the method of sentiment analysis. The classifier model was built on Naïve Bayes classification algorithm. The model built will successfully predict if the market trend is going to be high or low when we input the tweets. Finding future trend of a cryptocurrency market which is highly fluctuating in nature is a very crucial task because it depends on many factors. In this study, it is presumed that public tweets and market price are related to each other, and that public opinion on a platform like Twitter definitely have the capacity to fluctuate the market trend. As tweets and public comments will capture the sentiment or emotion about the current market, this sentiment is detected and based on the words in the tweets, an overall polarity can be obtained. If the tweets are positive, then it can be stated that this will impact the market to trend upwards. If the tweets are negative, then it can be stated that this will impact the market to trend downwards. With an accuracy of 89.86 %, it would be possible for the model to predict the market trend given any tweets about any coin in cryptocurrency market.

III. LIMITATIONS AND FUTURE WORK

This study can be extended by using Twitter data about cryptocurrencies other than Bitcoin. Apart from Twitter data, comments and opinions from Reddit, Quora or any other discussion platforms can also be used. News articles can also be considered in order to analyse sentiments of public about any currency. This work can also be expanded as a comparison study of different classifier algorithms. For this, other classification models such as SVM, Regression Classifier, and RandomForest can be built and a performance analysis study can be carried out. An improvement that can be implemented to this work is to stream data directly from Twitter to the

classifier and predict sentiments of these tweets without any time gap. Then, prediction can be done for even price variations that may occur on the same day itself if market is being very active.

ACKNOWLEDGMENT

Author is thankful to the faculty at College of Engineering, Trivandrum and Infinity Lab of UST Global at Kulathoor Campus for guidance and technical support.

REFERENCES

- [1] Vishal A. Kharde and S. S. Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques", *International Journal of Computer Applications (0975-8887) Volume 139- No.11, April '16*.
- [2] Satoshi Nakamoto (2008), "Bitcoin: A peer-to-peer electronic cash system", *academia.edu*
- [3] Tsytssarau Mikalai, Palpanas Themis (2012) "Survey on mining subjective data on the web", *Data Min Knowl Discov* 2012;24: 478–514.
- [4] Wilson T., Wiebe J., and Hoffman P., (2005), "Recognizing contextual polarity in phrase-level sentiment analysis", *Proceedings of HLT/EMNLP*;2005.
- [5] Liu B., (2012), "Sentiment analysis and opinion mining", *Synth Lect Human Lang Technol* 2012.
- [6] Pang B, Lee L., (2008), "Opinion mining and sentiment analysis", *Found Trends Inform Retrieval* 2008;2:1–135.
- [7] Cambria E, Schuller B, Xia Y and Havasi C, (2013), "New avenues in opinion mining and sentiment analysis", *IEEE Intell Syst* 2013;28:15–21.
- [8] Feldman R., (2013), "Techniques and applications for sentiment analysis", *Commun ACM* 2013; 56:82–9.
- [9] Montoyo Andre's, Marti'nez-Barco Patricio, and Balahur Alexandra, (2012), "Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments", *Decis Support Syst* 2012;53:675–9.
- [10] Anurag Nagar, Michael Hashar, (2015), "Using Text and Data Mining Techniques to Extract Stock Market Sentiment from Live News Streams", *IPICSIT vol.xx IACSIT, DOI=10.1.1.462.9734*.
- [11] Shah, Devavrat and Kang Zhang, (2015), "Bayesian Regression and Bitcoin", <http://arxiv.org/pdf/1410.1231v1.pdf>;6Oct2014.
- [12] Isaac Madan, Shaurya Saluja and Aojia Zhao, (2015), "Automated Bitcoin Trading via Machine Learning Algorithms", *Department of Computer Science, Stanford University*.
- [13] Go, Alec, Lei Huang and Richa Bhayani, (2009), "Twitter Sentiment Analysis", *Entropy* 17.
- [14] W. B. Yu, B. R. Lea and B. Guruswamy, (2011), "A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting", *International Journal of Electronic Business Management* ; 5(3): 211-224.
- [15] M. Matta, I. Lunesu, and M. Marchessi, (2015), "Bitcoin Spread Prediction using Social and Web Search Media", *Proceedings of DeCAT* 2015, <https://www.researchgate.net/publication/279917417>.