Cryptocurrency Price Prediction using Sentiment Analysis

Abdul Rehman Khurshid
akhurs18@earlham.edu
Computer Science Department at Earlham College
Richmond, Indiana, USA

KEYWORDS

Sentiment Analysis, Cryptocurrency, Price Prediction

ABSTRACT

Predicting cryptocurrency price movements is a well-known problem of interest. In this modern age, social media represents the public sentiment about current events. Twitter especially has attracted a lot of attention from researchers who are studying the public sentiments. Recent studies in natural language processing develop automatic techniques in analyzing sentiment in social media information. This research is directed towards predicting volatile price movement of cryptocurrency by analyzing the sentiment on social media and finding the correlation between them. Machine learning algorithms including support vector machine and linear regression will be used to predict the prices.

1 INTRODUCTION

As the economic and social impact of cryptocurrencies continues to grow rapidly, the importance of related news articles and social media posts also increase. This is particularly true of Tweets, as Twitter is growing popular in the financial world as fresh news and advice are from the top players in the world of finance appear on Twitter [17]. As seen previously with traditional financial markets, there appears to be a relationship between media sentiment and the prices of cryptocurrency coins. While many causes of cryptocurrency price fluctuation can be seen, it is important to explore whether sentiment analysis on available online media can inform predictions on the prices of coins [12]. This research will help people who are planning on investing in cryptocurrency as if price movements are predicted ahead of time, buy and sell points can be determined before large movements.

Bitcoin started 2021 hitting an all-time high of almost $65,000. It closed out the first half of the year down 47 percent from its all time high price [1]. This shows how volatile cryptocurrencies are and why some people are still skeptical to buy them. The volatility in the value of cryptocurrencies means uncertainty for both people investing in them, and those who would want to use it as a currency. Given that cryptocurrency prices do not behave like traditional currencies, such as the U.S. Dollar, the prices are extremely difficult to predict [12].

In addition to its growing popularity, the low barrier of entry and high data availability of the cryptocurrency market makes it an important subject of study [19]. What causes the price change of cryptocurrency is an area of debate [12]. This project analyzes the influence of news and Twitter data to predict price fluctuations for two cryptocurrencies: Bitcoin and Cardano. For this project, methods which utilize sentiment analysis of Tweets are reviewed. This involves utilizing Twitter's API and a Python library called Tweepy to collect and store Tweets which mention Bitcoin or Cardano. The Tweets are then analyzed to create a sentiment score for each day and compared to the price changes to that day to establish if there is a relationship between Twitter sentiment and cryptocurrency price changes. [19]. In addition to this data, Tweet volume and Google Trends relating to crytocurrencies will be utilized to see how the volume affects the price movement. Support vector machine and linear regression are machine learning algorithms which will be used to predict the prices, after the data is gathered from Twitter's API and Google Trends. The results using both of the datasets will be compared and analysed to see which would produce higher efficiency. This would than be used to determine if the sentiment of the public affects the prices of the coins more than the volume of post do. This research would help investors in determining their trading strategies in the future.

As plenty of research is done on this topic with considerably good outcomes which are being used for trading in real time everyday. This research would focus on the how the sentiment of the public and the volume of the Tweet's affects the price movements. The correlation between the results of both of the datasets would be analysed.

2 RELATED WORK

This paper builds on a wide range of research and topics. This section introduces existing research on sentiment analysis on cryptocurrency price pridiction.

Abraham [12] conducted a Twitter sentiment analysis on Tweet's which had hashtags relating to cryptocurrency to predict cryptocurrency prices. The data was analyzed to determine if it would be a valuable input to the final model. VADER sentiment analysis [10] determined Tweets to be more neutral than positive or negative, and if the public sentiment is neutral, this usually does not indicate a pattern for buying or selling. Both Google Trends and tweet volume were highly correlated with price. A linear regression algorithm was used to predict Bitcoin closing daily price. A more complex model than linear regression could be used in future work to improve the results as the results from this paper were taken when the prices were only going up.

Lamon et al [17] used news and social media sentiment on posts relating to cryptocurrencies to predict cryptocurrency prices. The model uses a classifier to learn feature weights that are used for labelling data. Linear support vector classification, multinomial Naive Bayes, and Bernoulli Naive Bayes were tried. However, logistic regression produced the best results. This model was able to predict the largest price increases and decreases correctly. As this research paper analyzed news and Twitter data separately, a more efficient result can be obtained if the model can work with a combination of news and social media data.

Valencia et al [19] proposed a model that used machine learning tools and social media data to predict the prices of cryptocurrencies. The market data for the top 65 crytocurrencies was taken from Cryptocompare public API, and the data needed for the sentiment analysis was taken from Twitter's API. VADER was used to calculate the the sentiment score. This model utilized neural networks (NN), SVM and random forest (RF). The results for this model show that predicting cryptocurrency is possible through sentiment analysis and by using machine learning tools.

Stenqvist and Lönn [18] investigated public sentiment from a posts on Twitter to predict the price of Bitcoin. 2.27 million Bitcoin-related Tweets were gathered for sentiment analysis to indicate a price change for the near future. This is done by a method of solely attributing fall or rise based on the severity of aggregated Twitter sentiment change over time periods ranging between 5 minutes and 4 hours, and then shifting these predictions forward in time 1,2, 3 or 4 time periods to indicate the corresponding BTC interval time. This method yielded an 83 percent accuracy. Furthermore, a prediction was only made when the mean of sentiment was limited by a minimum 2.2 percent change. This analysis can be improved in the future by adding a domain-specific lexicon which would yield a more representative sentiment.

Huang et al [16] proposed a long short-term memory (LTSM) sentiment analysis model. The data gathered to identify the sentiment was from the most popular Chinese social media platform, Sina-Weibo. A LSTM-based recurrent neural network was used along with the historical cryptocurrency prices to predict the price trend for the future. The results yielded an 87 per cent accuracy rate. This was 15.4 higher than the traditional autoregression method.

Bharathi and Geetha [14] investigated the sentiment of the public using Twitter data based on cryptocurrency hashtags to predict stock market prices. The correlation between the stock market values and sentiments in the social media data is established using an algorithm for sentiment analysis. The moving average method was used as an indicator to predict prices. The results show that the moving average method, in addition to sentiment analysis, yields a 14.43 per cent higher efficiency than when only the moving average method is used. This study is valuable as it uses both sentiment analysis and machine learning algorithms to predict the prices and examines the efficiency of both methods when used separately.

Salb et al [13] predicted the price movements of cryptocurrencies using a support vector machine with Sine Cosine algorithm (SCA). SVM is used for two thing. The first is to select a appropriate kernel function and fine-tuning their parameters. The second is to search for the best decision plane, which is an optimization challenge. Using a non-linear transformation, the kernel function assists in the creation of linear decision planes. The Sine Cosine algorithm produces multiple initial, random solutions that enables them to swing towards or outwards the optimal solution. Some other algorithms tested in this reseach were Optimized SVM-PSO and SVM. This research indicates that the proposed SVM-eSCA approach obtained the best accuracy of all methods included in the analysis.

3 DATASET

The training and testing datasets are from Twitter's API, Bitinfocharts and Google Trends [4][2][**?**]. In terms of cryptocurrency prices, Coinbase API will be utilized as it is one of the biggest cryptocurrency exchanges in the United States. The data will be stored in a Postgres data management system [3].

4 DESIGN AND IMPLEMENTATION

4.1 Framework

This project can be divided into two components. The first part is to use the text data collected from various tweets so a sentiment analysis model can be created which would extract the average sentiment for Bitcoin and Cardano. We would then use the information from this model to train a price prediction classifier which will predict the direction of the cryptocurrencies prices. Based on the results, the output will be a visual representation of the predicted prices.

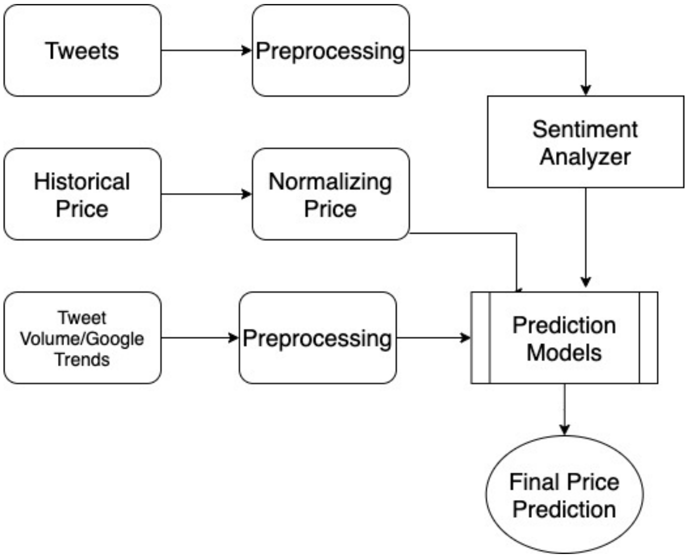

Figure 1: Framework of the project

4.2 Sentiment analysis

Sentiment Analysis is a text classification tool that analyses a piece of writing and computationally determines whether the sentiment is positive, negative or neutral [11].

4.2.1 VADER. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically made for sentiments expressed on social media. VADER calculates a sentiment score which can be used to find out how positive or negative a sentiment is [10]. VADER was selected for multiple reasons: (i) it is an open source tool; (ii) it is human-validated; and (iii) it is specifically attuned for Twitter content [19]. A polarization score will be calculated using the data by applying the geometric mean of the average of the positive and the negative sentiment of all the Tweets with the intention of using the score as a dimension to emotional valence.

4.3 Prediction models

Results generated from the Sentiment Analysis model will be used to train a classifier which will make a price prediction for Bitcoin and Cardano. Two price prediction models, linear regression and SVM will be tested. These models will be compared to determine which model produces the best outcome. The predicted prices would be compared with the actual prices of the coins, which will be taken from the Coinbase API.

4.3.1 SVM. Support vector machines are supervised learning algorithms that can perform nonlinear classifications by mapping data to higher dimensions using the kernel trick [15]. A kernelized SVM can compute complex transformations in terms of similarity calculations between pairs of points in the higher dimensional feature space where the transformed feature representation is implicit [7]. SVM was selected for multiple reasons : (i) SVM works relatively well when there is a clear margin of separation between classes; (ii) SVM is more effective in high dimensional spaces; and (iii) SVM is relatively memory efficient [6].

4.3.2 Linear regression. Linear Regression is a machine learning algorithm that is based on a supervised regression algorithm. Regression models target prediction values based on independent variables. It is used for finding out the relationship between variables and forecasting. Different regression models are unique based on the relationship between the dependent and independent variables, they consider the number of independent variables being used [8]. Linear regression will be used because of these reasons : (i) Regression analysis is more versatile and has wide applicability; and (ii) Learning Regression Analysis will give you a better understanding of statistical inference overall [9].

4.4 Evaluation

To evaluate the reliability of each model accuracy, precision, recall will be used which are defined as follows:

$$Accurracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1}$$

$$Precision = \frac{t_p}{t_p + f_p} \tag{2}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{3}$$

where,
tp = Number of true positive values
tn = Number of true negative values
fp = Number of false positive values
fp = Number of false negative values

Accuracy is defined as a ratio of correctly predicted observation to the total observations, Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, recall is the ratio of relevant classified samples among the total amount of relevant samples. [19][5].

5 BUDGET

This project is primarily a Python project. In addition to Python packages some visualization tools such as Tableau will be used. All the computational resources are available at Earlham. There is no cost expected at this time because the data set and all software is already available online at no cost.

6 TIMELINE

CS488 Deadlines	
Date	Work
Week 1	
Week 2	First draft of paper due
Week 3	First release of software/model due
Week 4	
Week 5	Second draft of paper due
Week 6	Second release of software/model due
Week 7	
Week 8 (finals)	Final paper and software due

7 ACKNOWLEDGEMENT

I would like to thank Dr. David Barbella for their support and feedback on my research proposal.

REFERENCES

[1] [n.d.]. Bitcoin had a wildly volatile first half. Here are 5 of the biggest risks ahead.

[2] [n.d.]. bitinfocharts. www.bitinfocharts.com.

[3] [n.d.]. Coinbase Developer Platform. https://developers.coinbase.com/.

[4] [n.d.]. Google Trends. https://trends.google.com/trends/?geo=US.

[5] 2016. Accuracy, Precision, Recall F1 Score: Interpretation of Performance Measures. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

[6] 2019. Top 4 advantages and disadvantages of Support Vector Machine or SVM. https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107.

[7] 2020. Introduction to Support Vector Machines (SVM). https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/.

[8] 2020. ML | Linear Regression vs Logistic Regression. https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/.

[9] 2021. 3 Reasons Why You Should Use Linear Regression Models Instead of Neural Networks. https://towardsdatascience.com/3-reasons-why-you-should-use-linear-regression-models-instead-of-neural-networks-16820319d644.

[10] 2021. Python | Sentiment Analysis using VADER. https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/.

[11] 2021. Twitter Sentiment Analysis using Python. https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/.

[12] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review* 1, 3 (2018), 1.

[13] Nebojša Bačanin Džakula et al. 2021. Cryptocurrency Forecasting Using Optimized Support Vector Machine with Sine Cosine Metaheuristics Algorithm. In *Sinteza 2021-International Scientific Conference on Information Technology and Data Related Research*. Singidunum University, 315–321.

[14] Shri Bharathi and Angelina Geetha. 2017. Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems* 10, 3 (2017), 146–154.

[15] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. 2015. Algorithmic trading of cryptocurrency based on Twitter sentiment analysis. *CS229 Project* (2015), 1–5.

[16] Xin Huang, Wenbin Zhang, Yiyi Huang, Xuejiao Tang, Mingli Zhang, Jayachander Surbiryala, Vasileios Iosifidis, Zhen Liu, and Ji Zhang. 2021. LSTM Based Sentiment Analysis for Cryptocurrency Prediction. *arXiv preprint arXiv:2103.14804* (2021).

[17] Connor Lamon, Eric Nielsen, and Eric Redondo. 2017. Cryptocurrency price prediction using news and social media sentiment. *SMU Data Sci. Rev* 1, 3 (2017), 1–22.

[18] Evita Stenqvist and Jacob Lönnö. 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis.

[19] Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. 2019. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy* 21, 6 (2019), 589.