Cryptocurrency Price Prediction: Utilizing Twitter Sentiment Analysis and COVID-19 Prediction Model

Samuel Kim
Computer Science
Georgia State University
Atlanta, USA

***Abstract*— The use of natural language processing to extract sentiment from social media has been widely explored to predict future value of stock or cryptocurrency value. Price prediction models developed using sentiment scores extracted by mining textual data from social media platforms have shown social media sentiment can be a reliable indicator of market trends. Our approach to building a prediction model maps sentiment scores extracted from mining Reddit comments to a repurposed SIR-model, a time series forecasting model previously used to predict the number of COVID-19 daily cases. Though our results were not entirely accurate, we can conclude that the prediction model can be a reliable tool to predict price movement for the next day.**

Keywords*— *Cryptocurrency, Social Media, Sentiment Analysis, Time series, Prediction, COVID-19

I. INTRODUCTION

The rise of Bitcoin brought forth much speculation about the future of decentralized currency. Utilizing decentralized currency eradicates the need to rely on a third party such as a bank to transfer funds and leaves no room for government regulation to set its price. This gives the people full control over their own wealth, free of national monetary policies and allowing seamless transactions across the globe at any given time.

The release of Bitcoin presented blockchain technology for the first time in 2009, allowing secure financial transactions without the need for a third party. Since then, the market capitalization of Bitcoin has reached $1.1 trillion dollars as of April 2021. The introduction of blockchain technology sparked world-wide innovation and gave birth to a plethora of cryptocurrencies that show great promise of assimilating into modern society and the global economy. Cryptocurrencies such as Ethereum and Vechain aim to provide solutions to real world issues using their own blockchain implementation. Ethereum allows users to interact with their "smart contracts", which are software programs that contain rules to constitute terms of a contract. These programs can be written in code by developers using Ethereum's native language called Solidity [1]. Smart contracts allow distributed data vending by exchanging data securely on the blockchain as well as enabling credible transactions without authorized third parties [2]. These new implementations of blockchain technology are slowly transforming the world of big data, making blockchain a promising tool in many applications such as health informatics and pharmaceutical supply chain [2].

However, over the course of 12 years we have seen that the cryptocurrency market does not behave like the stock market and experiences high volatility. In 2017 the value of Bitcoin increased approximately 2000% from $863 to $17,000, then 8 weeks later the price dropped by more than half down to $6200 [3]. Today we see cryptocurrencies such as Ethereum and Bitcoin skyrocketing in value, seeing an increase of 1454% and 508% respectively from May 2020 to May 2021. As of May 4th, 2021, Bitcoin's current value sits at $54,348, a considerable increase in value compared to where it sat in October 2020 at just around $11,000. This high volatility provides the need for prediction models that aim to predict market trend and price movement for the cryptocurrency market.

II. BACKGROUND AND RELATED WORK

Public sentiment portrayed through the media and various news outlets throughout the course of history has shown its impact on financial markets, especially the stock market. Negative sentiment linked to a particular company tends to drop the price of the stock, while positive sentiment tends to raise the price of the stock. This section will aim to highlight the effect of social media on the cryptocurrency market as well as current methods being used to build price prediction models.

A. Social Media and Sentiment Analysis

Sentiment analysis uses text mining and natural language processing to classify the emotional mood of a given text. Extracting sentiment from textual data is commonly used by organizations and researchers to analyze financial markets or public opinion of a commodity/product. A company developing a new product or service may want to gauge the opinions of their consumers. The new age of social media has shown that an investment idea can spread like wildfire to the masses.

Tweets from influential people have shown significant impact on financial markets. Throughout the month of February 2021, Elon Musk posted a series of tweets mentioning the cryptocurrency known as Dogecoin. On February 4th, 2021, Musk posted a tweet claiming

"Dogecoin is the people's crypto". The following day the value of Dogecoin shot up 43.67%. Another tweet was posted by Musk that same day and the value of Dogecoin went up 22.81% the following day after that. It seems after every tweet posted by Musk, positive sentiment regarding Dogecoin spread like a disease through word of mouth. The value of Dogecoin hovered around $0.03 cents per coin in the beginning of February 2021. As of May 4th, 2021, the current value sits at around $0.56 cents per coin.

Opinion mining can be seen as a subset of sentiment analysis, described as the analysis of textual data found on social media platforms such as Twitter, Facebook, Instagram, etc. Sentiment analysis is commonly performed on Twitter tweets to extract sentiment scores from each tweet. Textual data from tweets are classified as positive (+1), negative (-1), and neutral (0). The two main approaches used to perform sentiment analysis are either machine learning-based or lexicon-based [3]. Machine learning-based approaches are based on classification models while lexicon-based methods use a sentiment dictionary to match opinion words with the mined text data to determine sentiment [3].

B. Current Techniques

Sentiment scores mapped to time-series models built using machine-learning based approaches have proven to show remarkable accuracy. The paper cited in [4] makes use of sentiment scores extracted from roughly 1500 tweets as features to a time-series prediction model built using Long Short-Term Memory Networks. Predictions on the test set for Bitcoin, Litecoin, and Ethereum obtained a mean absolute error of 0.045, with a difference of only a few US dollars between predicted and actual price.

Long Short-Term Networks are essentially recurrent neural networks with LSTM units and can also be used to perform sentiment analysis as shown by the paper cited in [5]. LSTM has the capability to remember long sequences of data which makes it suitable for textual data found in Twitter and news data [5]. The paper in [5] also uses multi-class classification to extend the range of twitter sentiment score labels from -2 to +2 (very negative to very positive) [5]. The paper uses Random Forest Regression Algorithm with bootstrapping, where the data is trained on multiple decision trees and the final predicted output is computed by the average prediction from these multiple trees [5]. However, concluded results show that sentiment scores have minimum impact on prices possibly due to the scores being neutrally-skewed.

The paper "Predicting cryptocurrency price bubbles using social media data and epidemic modeling" [6] repurposes a Hidden Markov Model, previously utilized to detect influenza epidemic outbreaks to predict cryptocurrency price bubbles. A Hidden Markov Model predicts whether the state of a population is in an epidemic or non-epidemic state. The paper applies this model to Twitter data to categorize as 'trending' vs 'non-trending' to implement a trading strategy that outperforms a buy and hold strategy [6]. This methodology is based on the idea that financial price bubbles are linked with the epidemic-like spread of an investment idea [6]. The author of [6] quotes from another source, "Shiller in fact favors a more epidemic-like definition, describing a bubble as occurring by psychological contagion, where the news of price increases spurs investors' enthusiasm which spreads contagiously and brings in a larger group of investors, drawn in by envy and excitement about the previous price rises [7]".

III. PROPOSED APPROACH

The ideology presented by the paper cited in [6] describing investment ideas as an epidemic-like spread through word of mouth motivated our hypothesis of being able to repurpose a time-series model previously used to predict daily COVID-19 cases. We will be repurposing an SIR-model to allow for mapping of sentiment scores in order to predict the next day closing price for the currencies Bitcoin and Dogecoin. The Mean Absolute Percentage Error calculated between actual and predicted price will be used to determine accuracy of our predictions.

A. *SIR model*

Variations of the SIR model such as the SEIR, SEIQR, and SIRD have been used by epidemiologists to predict the number of daily infected COVID-19 cases. Given a constant population, the model divides the population into 3 categories: susceptible (S), infected (I), and recovered/removed (R) [6]. Members of the population transition from one category to another over time, based on a set of differential equations which tracks the change in the number of members in each category with respect to time (t):

$$\frac{ds}{dt} = -b\, s(t)\, i(t)$$

$$\frac{di}{dt} = b\, s(t)\, i(t) - k\, i(t)$$

$$\frac{dr}{dt} = k\, i(t)$$

Fig 1. SIR derivative equations [7]

B. *Repurposing the SIR model*

Based on the set of differential equations, the number of susceptible members increase or decrease according to the rate of contact between infective and susceptible members. This rate of contact is denoted as *-b* or negative beta. Similarly, the number of infective members increase with a positive rate of contact *b* or positive beta, but decrease given a fixed rate of recovery denoted as *k*.

Given a constant population N:

$$N = S + I + R$$

S = S(t) is the number of *susceptible* individuals,

I = I(t) is the number of *infected* individuals, and

R = R(t) is the number of *recovered* individuals.

Fig 2. Distribution of population [7]

We redefine the terms that comprise the constant population N in order to map sentiment scores to the SIR model, where N = total number of sentiment scores:

Susceptibles (S) = number of neutral (0) scores

Infectives (I) = number of positive (+1) scores

Recovered (R) = number of negative scores

Our prediction equation will associate price with positive sentiment, where an increase in positive sentiment will increase the price, while a decrease in positive sentiment will decrease the price.

Our first approach will define the equation for the number of infectives or price in our case:

Where the output of I = predicted price of the following day,

I = current day price + (% change in number of positive sentiment scores between current and previous day * current day price)

Our second approach will multiply the average of price per positive sentiment score based on our initial two weeks of data, by the number of positive sentiment scores during the first week of April 2021.

C. *Data Retrieval*

Our cryptocurrency historical data was downloaded from Kaggle.com, where we were able to find the OHLCV (Open, High, Low, Close, Volume, Market Cap) values for Dogecoin and Bitcoin from February 2021. Python was used to process the OHLCV.csv file into a Panda Data-frame shown in Figure 3.

dates	high	low	open	close	volume	market_cap
2/1/2021 23:59	34638.21349	32384.22811	33114.57724	33537.17682	6.140040e+10	6.24349e+11
2/2/2021 23:59	35896.88214	33489.21867	33533.20067	35510.28904	6.308859e+10	6.611150e+11
2/3/2021 23:59	37480.18789	35443.98273	35510.82121	37472.09010	6.116682e+10	6.976730e+11
2/4/2021 23:59	38592.17638	36317.49881	37475.10403	36926.06447	6.883807e+10	6.875430e+11
2/5/2021 23:59	38225.90595	36658.76354	36931.54565	38144.30686	5.859807e+10	7.102670e+11
2/6/2021 23:59	40846.54690	38138.38834	38138.38834	39266.01073	7.132603e+10	7.311920e+11
2/7/2021 23:59	39621.83549	37446.15388	39250.19051	38903.44148	6.550064e+10	7.244790e+11
2/8/2021 23:59	46203.93144	38076.32281	38886.82729	46196.46372	1.014670e+11	8.603430e+11
2/9/2021 23:59	48003.72396	45166.96004	46184.99147	46481.10424	9.180985e+10	8.656830e+11
2/10/2021 23:59	47145.56820	43881.15260	46469.76120	44918.18449	8.730109e+10	8.366170e+11
2/11/2021 23:59	48463.46713	44187.76235	44898.71161	47909.33119	8.138891e+10	8.923650e+11
2/12/2021 23:59	48745.73380	46424.97782	47877.03437	47504.85118	7.655504e+10	8.848740e+11
2/13/2021 23:59	48047.74459	46392.28233	47491.20256	47105.51747	7.025046e+10	8.774790e+11
2/14/2021 23:59	49487.64087	47114.50959	47114.50959	48717.29021	7.124868e+10	9.075510e+11
2/15/2021 23:59	48875.57161	46347.47709	48696.53666	47945.05683	7.706990e+10	8.932100e+11
2/16/2021 23:59	50341.10325	47201.30375	47944.45801	49199.87134	7.704958e+10	9.166230e+11
2/17/2021 23:59	52533.91431	49072.37714	49207.27643	52149.00754	8.082055e+10	9.716120e+11
2/18/2021 23:59	52474.10725	51015.76455	52140.97272	51679.79669	5.205472e+10	9.629150e+11
2/19/2021 23:59	56113.65055	50937.27572	51675.98129	55888.13368	6.349550e+10	1.041380e+12
2/20/2021 23:59	57505.22819	54626.55978	55887.33571	56099.52051	6.814546e+10	1.045370e+12
2/21/2021 23:59	58330.57214	55672.60951	56068.56825	57539.94367	5.189759e+10	1.072260e+12
2/22/2021 23:59	57533.38933	48967.56519	57532.73886	54207.31907	9.205242e+10	1.010210e+12
2/23/2021 23:59	54204.92976	45290.59027	54204.92976	48824.42687	1.061020e+11	9.099260e+11
2/24/2021 23:59	51290.13669	47213.49816	48835.08766	49705.33332	6.369552e+10	9.263930e+11
2/25/2021 23:59	51948.96698	47093.85302	49709.08242	47093.85302	5.450657e+10	8.777660e+11
2/26/2021 23:59	48370.78526	44454.84211	47180.46405	46339.76008	3.509680e+11	8.637520e+11
2/27/2021 23:59	48253.27010	45269.02577	46344.77224	46188.45128	4.591095e+10	8.609780e+11

Fig 3. OHLCV data for Bitcoin in February 2021

Twitter was our initial social media platform choice to retrieve textual data from. However, there was an issue where although the code to request data from Twitter API was commented out, running the code still sent the requests and resulted in maxing out the number of requests per month. We hope in the future the Twitter developer team will address this issue. As a result, we decided to use Reddit as our backup platform for textual data. Using Pushshift API (a wrapper API for Reddit's API), we were able to retrieve roughly 1500 Reddit comments from each cryptocurrency's respective subreddits, r/Bitcoin and r/Dogecoin. We encountered another issue where there was a large gap in the text data, missing historical comments dated from Feb 7th to Feb 10th. Due to the large gap, we opted to only scrape comments from Feb 12th to Feb 26th and perform sentiment analysis and predictions based on these two weeks. The Reddit comments were stored into a .json file in preparation for data cleaning.

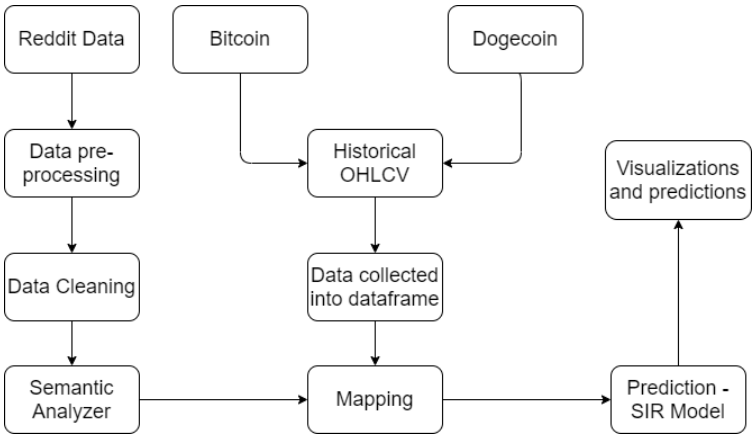

Fig 4. System Architecture

D. *Data Pre-Processing*

Textual data from social media contains a lot of noisy data such as emoticons and other symbols that would interfere with the process of analyzing sentiment. Before we can extract sentiment scores, we must clean the text data of these noisy values by removing stop words, emojis, and any unnecessary characters as mentioned in [4].

Our data cleaning process:

- Remove emojis
- Tokenizing strings: breaking apart string comments word by word into individual tokens
- Convert all tokens to lowercase
- Remove stop words: words that do not add much information to a sentence
- Lemmatizing: trim words down to their root word

We used a Python library, NLTK which has functions that allow us to perform our data cleaning process.

E. *Performing Sentiment Analysis*

For our sentiment analyzer, we utilized a lexicon-based approach to analyze our textual data. After storing our cleaned text data into a JSON dictionary, we processed the comments with VADER(Valence Aware Dictionary and Sentiment Reasoner) to extract sentiment scores normalized between -1 (most negative sentiment) and +1 (most positive sentiment). After extracting sentiment scores for each date from Feb 12th to Feb 26th , we noticed the total number of sentiment scores for every day were not the same. In order to maintain a constant population N with regard to our SIR model equation, we scaled the total number of sentiment scores from each day to the average number of total scores

over the course of the two weeks. Scaling each day to maintain a constant number of total scores was necessary in order to compute consistent prediction values. Figure 5 shows the count of sentiment scores before scaling.

02/12/2021	Positive	Neutral	Negative	Total
Bitcoin	286	3063	182	3531
Percent Ratio	8.099%	86.746%	5.154%	100%

Fig 5. Unscaled sentiment scores from 02/12/2021 for Bitcoin

The total number of sentiment scores during the two-week period averaged to a value of 2586 total sentiment scores. The scaling process started with calculating the percent ratio out of the total number of scores for each category and applied the ratio to the average number of total scores, 2586. Figure 6 shows the result of sentiment scores after scaling.

02/12/2021	Positive	Neutral	Negative	Total
Bitcoin	210	2243	133	2586
Percent Ratio	8.099%	86.746%	5.154%	100%

Fig 6. Scaled sentiment scores from 02/12/2021 for Bitcoin

IV. EXPERIMENTAL RESULTS

A. *First approach predictions*

Our first approach to predicting the next day's closing price applied the percent change of positive sentiment scores, to the current day's closing price. However, we conclude this method is not reliable as this heavily relies on accurately predicted price movement in a consistent manner. The shortfall of this approach shows that when a price is predicted in the opposite direction from the actual price, the rest of the predictions will follow and increase the variance in actual vs predicted values. Our first approach predictions values calculated a Mean absolute percentage error of 6.683087%.

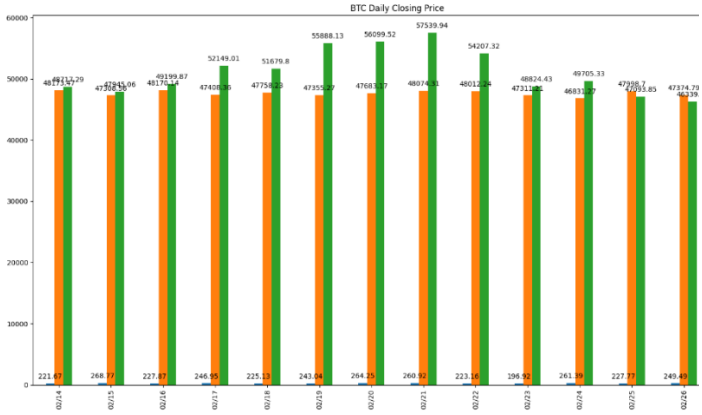

Fig 7. 02/14-02/27 Bitcoin prediction results

Figure 7 shows our prediction results, and you can see the prediction was fairly accurate and correctly followed price movement corresponding to change in positive sentiment the first 3 days. However, you can see although positive sentiment decreased from Feb 15th to Feb 16th, the actual price increased by approximately 6%. We assume this may be due to Reddit's pessimistic outlook for that particular day. Combining text data from Twitter and Reddit may yield better accurate sentiment scores. We have seen very similar results regarding Dogecoin that will be shown towards the end of this section.

B. *Second approach predictions*

Our second approach predictions were not much better than our first approach. Our predication values calculated a Mean absolute percentage error of 6.84193%, which is actually slightly worse than our first approach. We noticed a similar trend to our first approach where although positive sentiment dropped between April 3rd to April 4th, the price increased slightly on April 5th .

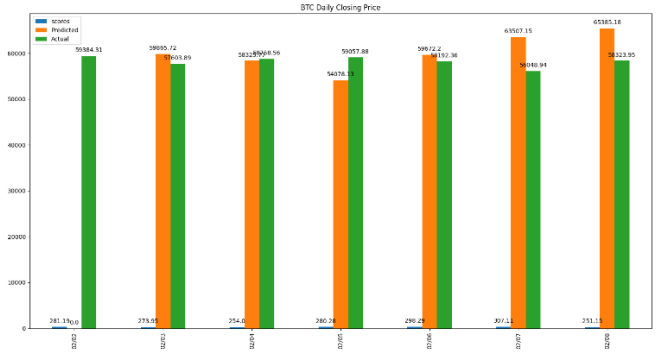

Fig 8. 04/02 – 04/08 Bitcoin prediction results

Figure 9 shows our prediction results for Dogecoin which follow the same pattern where the actual price moved in an opposite direction compared to difference in positive sentiment.

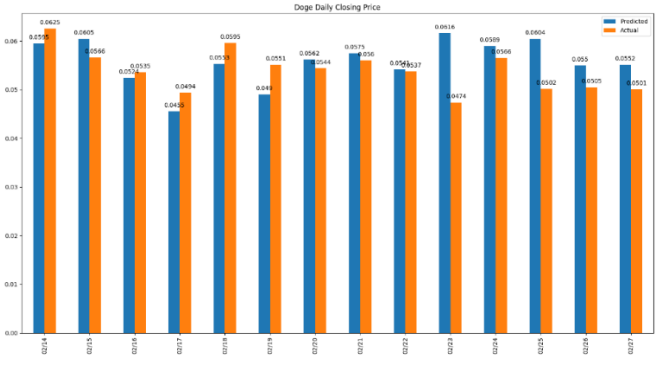

Fig 9. Doge predictions using our first approach

V. CONCLUSIONS AND FUTURE WORK

Although our predictions were not entirely accurate, we conclude that public sentiment from social media can still be used as a reliable indicator to predict market trend. Given the constant and simplistic nature of the SIR model, we can assume there are other factors that must be considered and implemented when predicting price of a cryptocurrency. In the future, we will continue to improve this model and implement additional factors to make this model more dynamic. We believe incorporating additional weight to sentiment scores such as follower count of users, number of retweets and likes can improve the accuracy of this model, as we feel these are factors that can play an important part in extracting accurate sentiment.

References

[1] H. Agrawal, "Coinsutra," 2019. [Online]. Available: https://coinsutra.com/smart-contracts/.

[2] J. Zhou, F. Tang, H. Zhu, N. Nan and Z. Zhou, "Distributed Data Vending on Blockchain," in *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Halifax, NS, Canada, 2018.

[3] S. Mohapatra, N. Ahmed and P. Alencar, "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments," in *IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, 2019.

[4] A. P. Patil, T. Akarsh and A. Parkavi, "A Study of Opinion Mining and Data Mining Techniques to Analyse the Cryptocurrency Market," in *3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 2018.

[5] A. Inamdar, A. Bhagtani, S. Bhatt and P. M. Shetty, "Predicting Cryptocurrency Value using Sentiment Analysis," in *International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019.

[6] R. C. Phillips and D. Gorse, "Predicting cryptocurrency price bubbles using social media data and epidemic modelling," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, HI, 2017.

[7] D. S. a. L. Moore, "Mathematical Association of America," [Online]. Available: https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model. [Accessed 2021].

[8] P. R. Rizun, "Medium.com," 28 March 2019. [Online]. Available: https://medium.com/@peter_r/visualizing-htlcs-and-the-lightning-networks-dirty-little-secret-cb9b5773a0. [Accessed 29th January 01/29/2021].