Using Reddit Sentiment to Predict Bitcoin Price Fluctuations

Sourav Sarkar, Tong Thomas Zhang

ss5645@columbia.edu, ttz2104@columbia.edu

Abstract

Recent trends in Bitcoin and Cryptocurrency price fluctuations have indicated buyer sentiment as a strong driver for the market price. Reddit, a social news aggregation platform, strongly reflects sentiments of specific topics through "subreddits" and we predict that Bitcoin related subreddit sentiment could be used as a measure for predicting Bitcoin price. The goal of this study would be to show significant improvement when using Reddit sentiment to predict Bitcoin price fluctuations as opposed to traditional pricing methods through buy and sell decisions on a portfolio. In particular we are considering data from January to March, a volatile period in terms of Bitcoin price. Overall, this study found no significant difference between using Reddit Sentiment to predict price fluctuations when compared to other strategies such as momentum but can generate valid buy/sell signals

1 Introduction and Motivation

Bitcoin (Nakamoto, 2009) is a cryptocurrency, a form of electronic cash. It is a decentralized digital currency without a central bank or single administrator that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries. In 2018 it made headlines (Wikipedia, 2019) as prices surged then crashed, affecting both experienced and inexperienced investors.

During the course of these price runs, it was recognized by some investors that there were similar price runs in history in which much of the valuation of the security was based on public sentiment. With this, several trading strategies were implemented. Perhaps the most well known strategy was momentum (Chan et al., 2000) - a strategy in which buy and sell decisions are made simply by how much the price is going up or down, as if its going up, it's likely to continue going up as more investors will put money in the security/stock, driving the price up further. This was a popular strategy as it is relatively simple to create and has been proven outside the Bitcoin market, such as with stocks. Another alternative strategy commonly used but harder to implement was the analysis of Twitter Sentiment as a method to gauge public opinion about Bitcoin. Buy and sell decisions were made based off the wisdom of crowds and those posting about Bitcoin on the social media platform. Though harder to implement, there is a lot more flexibility with this sentiment trading strategy since it moves beyond just price data and can be adjusted more flexibly based on certain parameters such as Tweet volume and additional dimensions of sentiment. These trading strategies worked generally well during the Bitcoin price run and the motivation behind this study was to ascertain an even better trading strategy based on sentiment and to determine if it could beat commonly used strategies with any significance.

Though no study had directly compared the two, we hypothesized that a well formed sentiment strategy could possibly outperform the momentum strategy because of the additional flexibility analyzing sentiment comes with, the ability to adjust more parameters could prove useful in fine tuning the strategy. However, rather than using Twitter as done in previous studies, we additionally hypothesized that Reddit may be a better indicator of public sentiment. The reasoning behind this comes from the "subreddit" design of Reddit, in which many sub forums of different topics are created. Among these are "r/Bitcoin", "r/Cryptocurrency" and more. We reason that those posting on these subreddits have more domain expertise than those posting on twitter and analyzing the sentiment of posts and popularity of posts would yield a better indicator for buy and sell decisions for Bitcoin.

2 Related Work

There have been several studies on testing trading strategies in certain markets. The effectiveness of Momentum has been addressed in numerous studies, specifically by AQR (Berger et al., 2009), and has generally been accepted as a good trading strategy. Since analyzing sentiment is both a newer trading strategy, and a trading strategy with more flexibility, there have been many attempts to apply sentiment trading strategies to stocks, particularly sentiment from Twitter, which have yielded success. In the Bitcoin market, it is hypothesized that sentiment strategies may perform even better and the same Twitter sentiment strategies have been applied to Bitcoin markets.

2.1 Using Twitter Sentiment to Predict Stock Prices

Traditionally, sentiment analysis has been used as a method for predicting buy and sell signals on stock prices. Studies have found connections between Twitter sentiment and signals in the market. (Abraham et al., 2018) presented significant evidence of dependence between stock price returns and Twitter sentiment in tweets about the companies using an extensive sentiment classification model trained on over 100,000 labeled tweets, rated as ”negative”, ”neutral”, or ”positive”. This result is to be expected as the nature of the stock market is that when a stock is shown to be doing well, more investors will put money into the stock which will drive the price even higher, which if the basis behind the popular momentum trading strategy. The caveat behind this study was the sentiment data that was looked at had to be properly selected, as not all tweets mentioning a certain stock provides particularly relevant sentiment information. We believe subreddits will lessen the tediousness and improve the accuracy of this task as everyone posting on the cryptocurrency subreddits provide more relevant information and will be more familiar with the cryptocurrency domain.

2.2 Using Twitter Sentiment to Predict Bitcoin Prices

More specific studies between Twitter and Bitcoin have been conducted. A study found that Bitcoin has an especially strong relationship with sentiment when compared to stocks. (Pagolu et al., 2016) This strengthens the claim that sentiment may be an especially good method for predicting Bitcoin prices. A main concern in the discussion of this study was that in the model it trained to analyze sentiment, it could more reliably make predictions with a certain subset of Twitter users.

Particularly on Twitter there are many bots which post about Bitcoin, this can skew sentiment in a certain direction. With Reddit, by design bots are more easily detected through both moderators and the up-vote, down-vote feature. Again, this adds to the hypothesis that the use of subreddits may be able to circumvent this.

2.3 Twitter Sentiment and Volume to Predict Bitcoin Prices

Expanding on just sentiment, additional parameters have been added to studies done on Twitter and Bitcoin price. (Pagolu et al., 2016) found a strong correlation between Tweet volume and Bitcoin price. It tested three main models of predicting Bitcoin price in which the parameters of Tweet Sentiment, Tweet Volume, and Google Trends were mixed. Moreover the study found that through its sentiment analysis algorithm, thought it was good at predicting Bitcoin price when it was trending upwards, it proved more unreliable when Bitcoin price was trending downward.

In general this is to show that with sentiment an extra dimension of analysis is added. With the momentum trading strategy, the dimensions are simply the price going up or the price going down. Sentiment has more facets and therefore more refined choices on buy and sell decisions can be made. Thus while there are few ways to implement momentum trading strategies, there are a plethora of ways to implement sentiment trading strategies. As sentiment can reflect more than the public just feeling ”good” or ”bad” about a certain stock. How much a post is upvoted/retweeted, more caveats to the tone of voice in posts, how many posts are being made, and who is posting can all be considered in a sentiment trading strategy.

3 Research Questions

The approach to establishing the method was centered around the following research questions:

- Does Reddit post sentiment give proper buy and sell signals? - Many people discuss cryptocurrencies in famous Reddit sub forums such as r/CryptoCurrency/, r/Bitcoin/, r/ether which generates huge volume of daily data.

Can we consume the data real time and generate reliable signals on price movement?

- Crypto users use various vocabulary like HODL, MOON, FUD, FOMO. Is using a traditional sentiment analyzer effective at analyzing these types of posts?
- Apart from sentiment analysis, with data on upvotes, reddit scores, reputations, post volume etc. Can it be reliably used and how will their performance compare to using sentiment?

4 Method

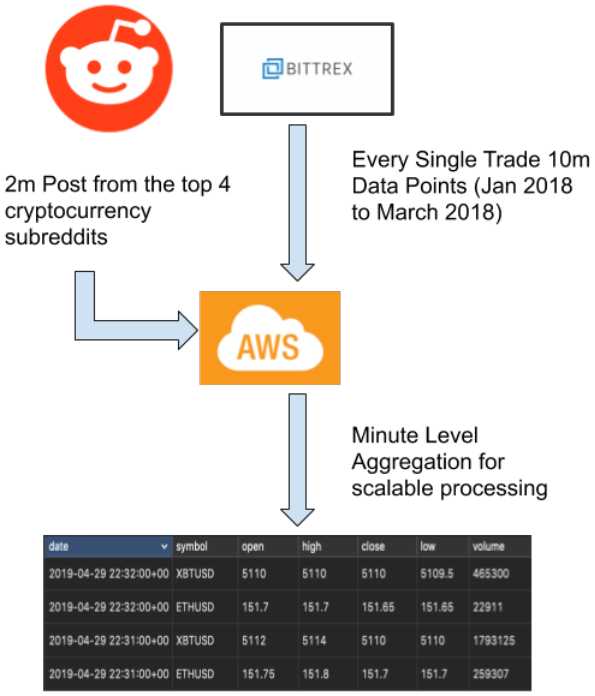
4.1 Assumptions and Simplifications

Since trading strategies generally can be adjusted with a great deal of flexibility, this study will be testing a Reddit sentiment trading strategy and a basic momentum trading strategy following the same rules on buy and sell signals. These strategies will either long the whole portfolio or short the whole portfolio every time there is a decision signal. Although this may not be realistic in terms of maintaining a portfolio risk level, it will still better isolate the two strategies when directly comparing Momentum and Sentiment.

4.2 Data Collection and Cleaning

In Figure 1 we highlight our data collection process. Reddit post data was collected using the Reddit API and Bitcoin Price data was collected using data from Bittrex from Jan 2018 to March 2018 (the main Crypto Bear Run) which comprised of around 10 million trading data points. It should be noted that the Reddit API does not send posts that were deleted. This could affect the study in that when people change their opinions about Bitcoin after the price has changed, then they are more likely to delete their post. We extracted the Reddit Data from the top 4 (Kilkenny, 2019) cryptocurrency subreddits namely - r/Cryptocurrency, r/Bitcoin, r/btc, r/ethtrader which constituted around 2 million reddit posts for Jan 2018 to March 2018 Period. The data was scraped for bots posting spam. Additionally, the data contained a few blank spots in it where there was a post in the data but no associated date and time, for the purpose of this study, all such posts were removed from the data set. Also we used NLTK to for priliminary cleaning of the reddit post (eg: stop word removal). Also if we

Figure 1: Data Pipeline



directly sompare the trading data with the reddit data, there might be inaccuracies in timing data and which might give us the false impression if we look into the future. Hence to avoid that, converted the trading data to minutely OHCLV bars.

4.3 Control Models - Momentum and Monte Carlo

As mentioned previously, this study uses the momentum trading algorithm as a control method for evaluating the success of the sentiment analysis method. Additionally, we will be comparing a monte carlo simulation of making trade decisions randomly to show the effectiveness of both trading strategies. For these strategies we used the same assumptions as our proposed strategy, which was to buy and sell the whole portfolio each time a signal was created. The two momentum models used were a three day momentum model and a five day momentum model. This model makes buy signals when the day price is above the three day moving average and five day moving average respectively.

In Figure 2, we see that the five day momentum trading strategy gave around a 57% return on the portfolio across three months. During a period where the Bitcoin price fluctuated from 13,700 dollars to 7,500 dollars.

In Figure 3, we tune the Momentum Trading

Figure 2: Five Day Momentum Trading Algorithm

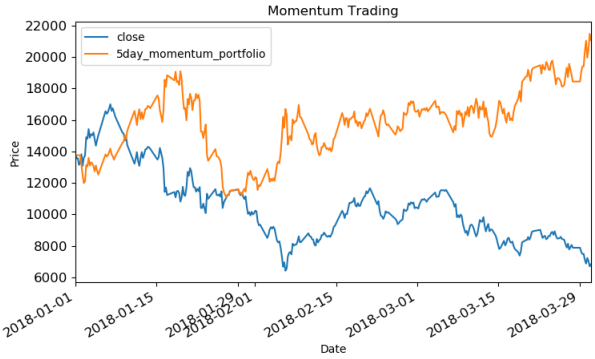
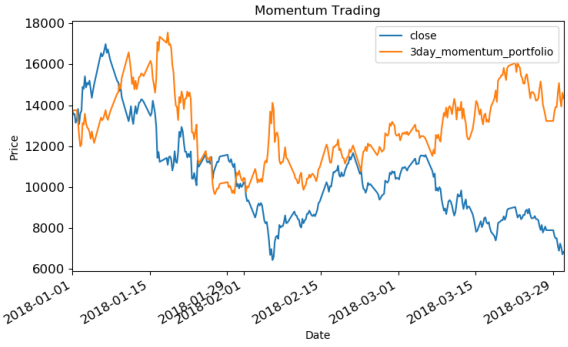


Figure 3: Three Day Momentum Trading Algorithm



strategy to a three day trading strategy. The return when using three day momentum strategy is only around 10%. By only moving the momentum window by 2 days, the percent return decreased by around 47%. This is to show that when tuning the trading strategies, the returns can widely vary.

Therefore in order to get a general sense of performance in Bitcoin markets, we performed a monte carlo simulation of 10,000 random trading strategies. To do this we simply feed random buy and sell signals at different frequencies 10,000 times and plot the returns.

Figure 4 represents the 10,000 random portfolios that were made in the simulation. Figure 5 shows their distribution. Only about the top 20 percent of portfolios yield positive returns. So we can see that the standard three day and five day momentum trading strategies place in the top 20 percent of trading with random buy and sell signals. Since testing significance with trading strategies is difficult, we determine relative success by seeing if using Reddit sentiment with very simplified assumptions can yield positive returns.

Figure 4: Monte Carlo Simulation

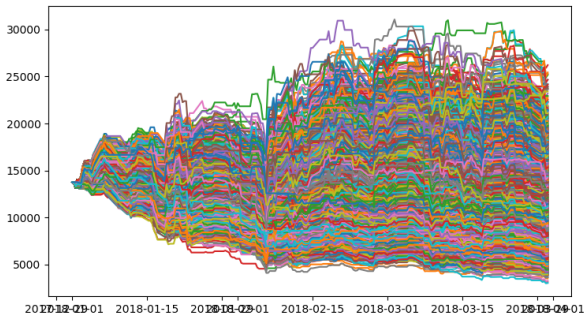
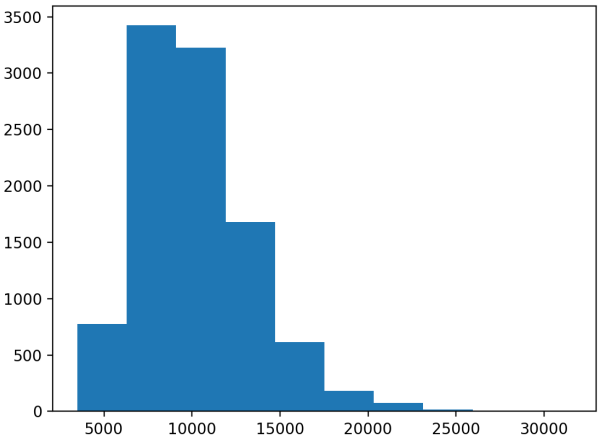


Figure 5: Monte Carlo Simulation Return Distribution

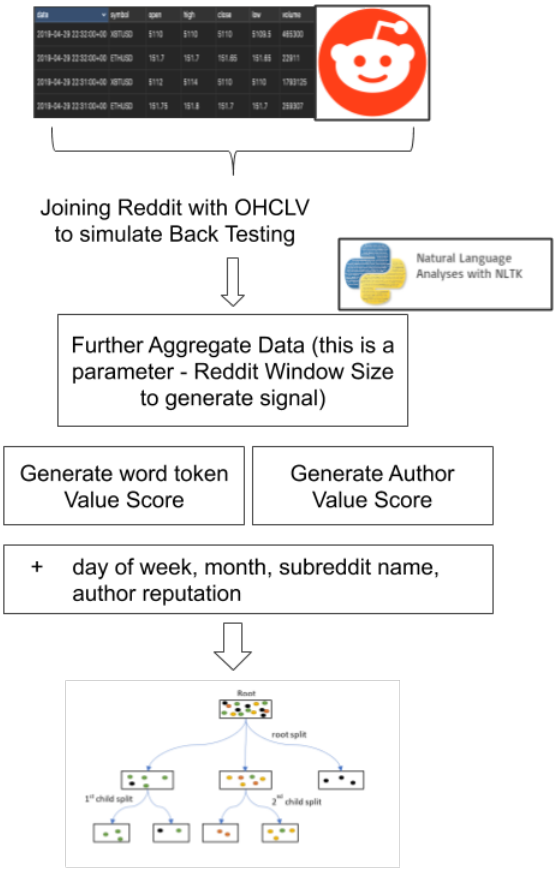


4.4 Reddit Based Methods

We joined our Reddit post (a time series of Reddit posts) with our minutely trading data such that for each post we would have the trading price of BTC at the time of posting. Also as Crypto markets are open 24/7, we just took the average of the Open, High, Close and Low prices. Next we grouped the data in contiguous trading periods. Each group consisted of Reddit posts with associated metadata (author, subredddit, author reputation, upvote, downvote, etc.) with the next available minutely price. As traditional sentiment analyzers might not work that well with Reddit posts, we developed our own sentiment analyzer which learns to find price correlated keywords on its own.

For each author, we defined an author score which is proportional to the fraction of times Bitcoin price increased after the author posted. Similarly we defined a word score. The author score relates user activity to Bitcoin prices. For exam-

Figure 6: Reddit Based System



ple many times in Reddit, users plan Pump-and-Dump Schemes(Buying a significant amount of Bitcoin to drive its price up so other people buy Bitcoin, then selling Bitcoin at an artificially high price where after the price falls back to the norm). If the same users takes part repeatedly we can detect it. Similarly word score helps us to find the correlation among word occurrences and price movements. We augment this with user reputation count. We also had upvotes and downvotes data, but we feel that can be noisy as people can upvote or downvote posts after observing the price as well. So we will not be able to observe the true upvote and downvote data as Reddit only give us total number of upvotes and downvotes rather than a timeline of how much a post was upvoted or downvoted.

Next we take data from the past trading blocks totaling 10000 Reddit posts and their respective pricing data. Next for each post in the current block, we generate the signal through our random forest (Breiman, 2001) classification model which we will describe in a later section. Then we take the average of all the buy/sell signals in the block. If we have more than 60% buy signals then we go long (buy Bitcoin). If we have more than 60% sell signals then we go short. Otherwise we do not do anything.

We retrain the random forest at every trading window. After predicting the signal for the current window, we look at the average Bitcoin price at the next window. If the price was more than 10% we make it a buy data point in the model. If it is less than 10% then we take it as a sell data point. Otherwise we mark it as neutral. Now we look back at the past data blocks and collect as much data such that we have total 10,000 data points (Note : for the initial period when we do not have data we just generate signals using our momentum model).

4.5 Establishing Trading Rules and Control

Trading was controlled by the simplifying rules of using all the money in the portfolio to long or short Bitcoin with immediate methods to do so. Additionally we assume that there are no transaction costs in buying and selling.

4.6 Model Evaluation

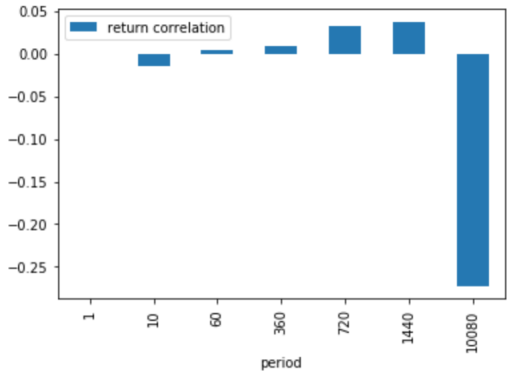
To evaluate our model we compare it to the common momentum trading algorithms as well as a monte carlo simulation of random buy and sell decisions across the trading time period. The purpose of the monte carlo simulation is to show that these strategies have meaning when compared to trading Bitcoin randomly, and momentum should be used as a baseline of comparison as it is a popular, simple to implement strategy. We back test all our models with Bitcoin price data from January to March.

5 Results

5.1 Using a Traditional Post Sentiment Analyzer

Earlier we hypothesized that traditional sentiment analysis might not work well with Reddit cyrpto data because of the unique language used on Reddit (HODL, MOON, FUD, etc.). This was demonstrated in the development of the model. For each trading window size, first we aggregated all the Reddit posts in that window in a single window. We used the textblob (Textblob, 2019) library to compute the subjectivity and polarity score for each of the post windows. As post window is a proxy of over all Reddit discussion in that period,

Figure 7: Subjectivity Correlation



this was the most appropriate methodology. Next we computed percentage price return of Bitcoin for the current window size and plotted the correlation values and p-values for all trading windows.

Figure 8: Subjectivity P-Values

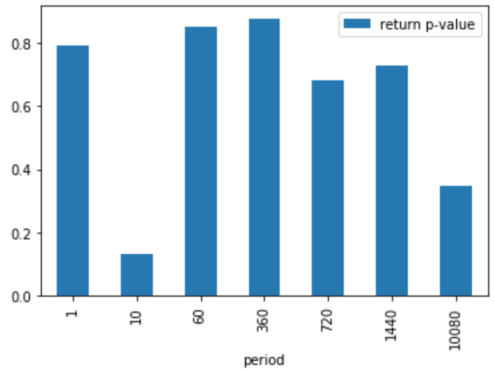
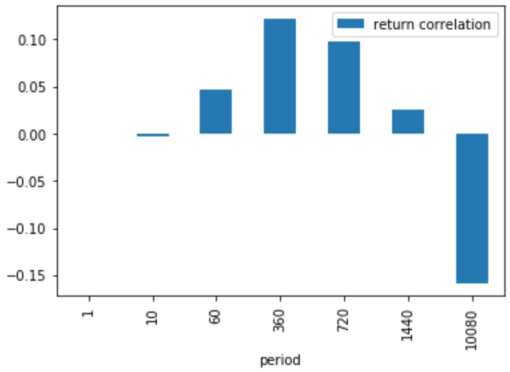
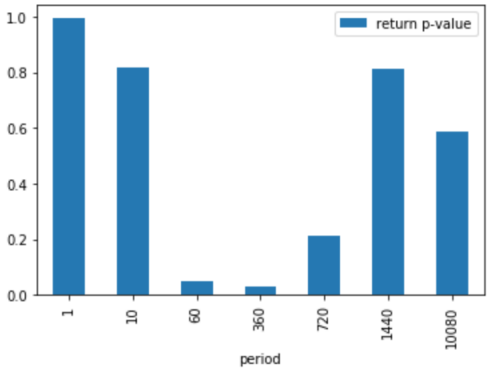


Figure 9: Polarity Correlation



In Figure 7 we plot the correlation of post subjectivity with percentage return values. In Figure 8 we plot the p-values of post subjectivity with percentage return values. In Figure 9 we plot the correlation of post polarity with percentage return values. In Figure 10 we plot the p-values of post polarity with percentage return values.

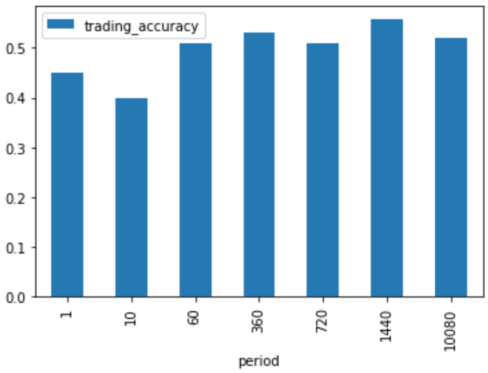
Figure 10: Polarity P-Values



In all cases, higher p-values and lower correlation values suggest that there is not much statistical significance between the output of traditional sentiment analyzers and return values of Bitcoin and this can be attributed to the peculiar and niche nature of Reddit cryptocurrency posts.

We ran our trading back-testing for the different trading windows. Starting from 1 minute to 7 day windows. For each window, we experimented with the previously mentioned methodology. For each window, we logged the trading decision and we have compared it with whether Bitcoin price increases or decreases in the next window. One thing to note here is that unlike normal classification problems where we need a very high accuracy, here if we consistently hit a percentage of more than 50% we essentially make money. Before the trading simulation, it is important to show how the Reddit scores behave with the return values.

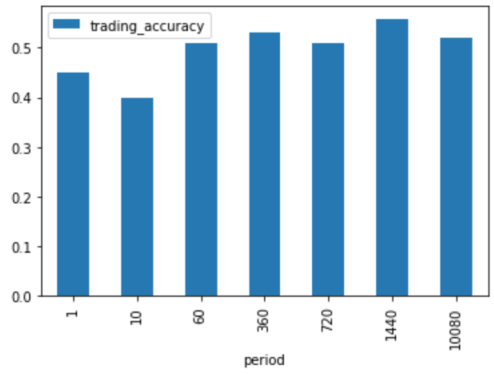
Figure 11: Reddit System with Returns Correlation



5.2 Using a Custom Post Analyzer

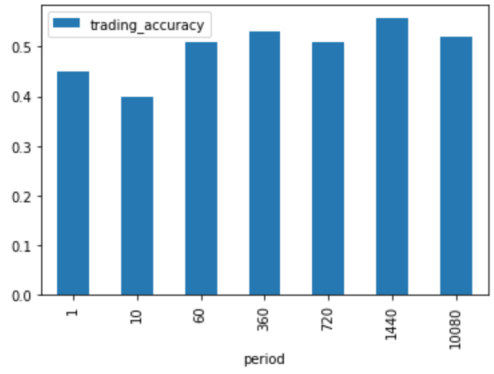
In our method, instead of calculating sentiment score using a tradional sentiment analyzer, we compute post scores by running our random forest classifier. Then we compute its probability values with trading returns. In Figure 11 we show correlation values with its respective trading period and in 12 we show the respective p-values as well. It should be noted that this testing returned relatively high correlation between half day to full day windows.

Figure 12: Reddit System with Returns p-values



With this in mind, we look at our results for the trading back-test with our generated signals

Figure 13: Trading Accuracy with Trading Window



In Figure 13 We note that as we increase the trading window length, the accuracy initially increases and it peaks at around 1 day. We have used the trading signals to long / short Bitcoin and we have calculated the final return for each trading window. In Figure 14 we note that we achieve the highest return again when the trading window is 1 day, which is consistent with our earlier accuracy plot.

In Figure 15 we show one of our back-testing

Figure 14: Return vs Trading Window

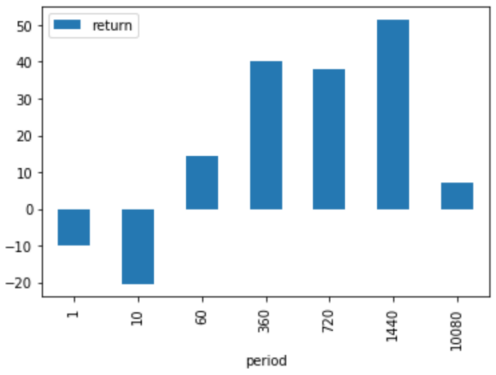
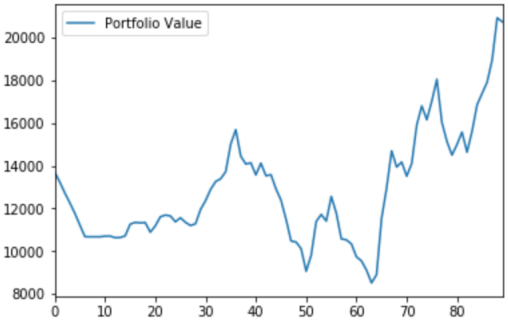


Figure 15: Portfolio Value throughout the process



results. At each window we long / short according to our model and at the end of each window we calculate our portfolio net worth and plot it. As we earlier noted, this trading window returned the highest accuracy in price movement prediction. Using this the portfolio returned around 50% return as well.

6 Conclusion and Discussion

In conclusion, we have shown the method that it is possible to create a Reddit sentiment based trading bot on par with traditional strategies such as momentum while using simple assumptions. To refer back to our initial research questions, we found that using certain windows, Reddit can generate proper buy/sell signals. However, as with all trading strategies it is difficult to test significance because of the constantly changing market environment and the relatively new appearance of Bitcoin. We additionally found that the correlation between traditional sentiment analyzers on Reddit Bitcoin posts and the actual price movement was relatively low, due to the different style of language used on Reddit. Therefore, when analyzing Reddit posts, it is imperative to look deeply into how to develop

a model to gauge post sentiment or rating the post in general. Finally, we showed that using other data that Reddit offers can be proven relatively useful in refining post score as we did with user reputation. However, it is important to note that there could be specific instances of noise. Such as the fact that with our data, upvotes and downvotes could not be used as many people could have upvoted the post after the fact of Bitcoin price going up or down.

6.1 Explanation of Results

We believe our our strategy works on a broader basis because momentum strategies and sentiment strategies reflect the same underlying value. If investors see something is doing well, they are more likely to put money into it in hopes of making more and causing the price to go up more. If investors see something is doing poorly they are more likely to take money out, causing the price to fall down more. Momentum strategies reflect this with price data while sentiment strategies reflect this with investor opinions through those who post about the asset.

6.2 Important Notes on Simplification

As there was little prior research done on the topic of Reddit sentiment and Bitcoin price fluctuations, there were simplifications that could be looked more deeply at. When trading the whole portfolio, we had no standard measure of calculating portfolio risk or any other common metrics for any of our models, it may have been more realistic to take these factors into account in comparing how risky each trading method was. We also only covered a bear market in the Cryptocurrency, but this same study could be conducted in bull markets or different bear markets as well to see if the strategies remain consistent.

6.3 Proposed Further Research

Besides the direct improvements/variations of the study mentioned in the previous section, first in future research there should be more studies conducted on directly comparing sentiment based strategies (Twitter vs. Reddit). We have determined from this study and other studies have shown that using sentiment off these social media websites can generate good buy/sell signals. However, is using one social media platform superior to another?

Another possible area of further research is simply tweaking the aspects of Reddit and trading. First of all, experimenting with long/short amounts based on sentiment as in our study we simply traded the whole portfolio value. Moreover, using Reddit and seeing if paying attention to certain data (sentiment vs. upvotes vs. type of subreddit) matters more in a significant way. One particular interesting study that could be done would be collecting live Reddit data of upvotes and downvotes and using those as dimensions to predict Bitcoin price. Additionally, for the purpose of this study we mainly looked at sentiment of the post, and did not pay attention or vary the particular subreddits the posts were coming from. For example, would following r/Bitcoin be more profitable than following r/Cryptocurrency, is there an optimal mix that produces the best result?

References

Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. 2018. Predicting bitcoin price fluctuations combining twitter sentiment and twitter volume. *CoRR*, abs/1610.09225.

Adam L. Berger, Ronen Israel, and Tobias J. Moskowitz. 2009. The case for momentum investing.

Leo Breiman. 2001. Random forests. *Mach. Learn.*, 45(1):5–32.

Kalok Chan, Allaudeen Hameed, and Wilson Tong. 2000. Profitability of momentum stragegies in the international equity markets. *Journal of Financial and Quantitative Analysis*, 35(2):153172.

Dylan Kilkenny. 2019. Cryptosub : Top reddit subreddits.

Satoshi Nakamoto. 2009. Bitcoin: A peer-to-peer electronic cash system.

Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. *CoRR*, abs/1610.09225.

Textblob. 2019. Textblob.

Wikipedia. 2019. Crypto currency bubble.