Correlation Between Bitcoin Price And Sentiment of Public Discourse

Chris Messer, Erin Abbott, Devyn Byrd

Abstract. In this report, we analyzed the sentiment of discussion around Bitcoin in two public forums, and their relationship with the price of Bitcoin. We used a combination of regression models, exploratory data analysis, and principal component analysis and determined that our best performing model was the principal component analysis. We hypothesized that the sentiment around Bitcoin drives the price of Bitcoin. Our exploratory data analysis did not reveal a strong relationship between sentiment and price, so our hypothesis was rejected. However, we discovered interesting correlations between Bitcoin price and the total conversation activity around Bitcoin.

1 Introduction

Bitcoin is a digital currency created in 2009. It is a decentralized form of currency, meaning it is not regulated by a government or central bank. Bitcoin can be used to purchase goods and services, or it can be held as an investment.

The price of Bitcoin is volatile and can change drastically over short periods of time. This is due to several factors, including speculation, market forces, news about the currency, and the availability of exchanges. Bitcoin's price is also influenced by supply and demand. When demand for Bitcoin increases, the price typically rises, and when demand decreases, the price usually drops.

Online forum Reddit.com and social media platform Twitter.com host communities and discourse both in-favor and against a cryptocurrency revolution. The sentiment of this discussion is often a reflection of the price of Bitcoin. This is intuitive; people are more likely to speak negatively of an asset when it is underperforming, and positively when it is over-performing. Using machine learning and data processing tools, we can measure and quantify the sentiment of this discussion.

The use of sentiment analysis to extract insights from customer feedback has been present since the 1950s, and has continually evolved. Social media such as Twitter and Reddit provide diverse exposure to businesses, allowing them to connect to customers, receive feedback, and use sentiment analysis to improve or evolve their products and services. As many people are invested in the cryptocurrency markets and regularly post technical analyses and thoughts, these posts can affect the market.

In this paper, we explore the notion that the relationship between these two values is one of causation rather than correlation. E.g. the sentiment around Bitcoin actually drives the price of Bitcoin. We hypothesize that by looking backwards a period of time n, measured in either days d or hours h, from time t, we can predict the price of Bitcoin at time t by observing the discourse around Bitcoin at time $t\text{-}d|h$ to time t.

2 Methodology

This study has been divided into three parts. First, we discuss the extraction of the data we will analyze using sentiment analysis, and how we approached “tagging” the qualitative discussion around Bitcoin with a quantitative measure of sentiment. Second, we perform some exploratory analysis on the now prepared datasets to explore the shape of the data, and determine any correlations between our two data

populations, from Reddit and from Twitter. Third, we perform a series of regression analyses to attempt to identify the correlation between Bitcoin price and sentiment of public discourse, and analyze each model's appropriateness.

3 Data Preparation

Our analysis will focus on three sets of data:

- Tweets about Bitcoin
- Bitcoin Pricing information
- Reddit comments that contain the word Bitcoin

3.1 Twitter Data

Our Twitter data contains two sets of data we will use for the analysis: the sentiment tags for tweets about Bitcoin, and the Bitcoin pricing information. On first inspection, it appears the data is incomplete. We will need to handle this before running our analysis. First, we separated the Twitter data columns from the Bitcoin pricing data. Then, since we want to analyze the data at different time intervals, we aggregated the data at the day level rather than just the hour interval and stored it as a new dataframe.

3.2 Bitcoin Pricing information

As mentioned above, the Twitter data also contained information on Bitcoin Open/High/Low/Close (OHLC) at each interval. However, as noted, we are missing rows of data in that set, which includes the Bitcoin data at those hours. As such, we determined it appropriate to seek a secondary source for Bitcoin pricing data.

One challenge we faced was locating an OHLC Bitcoin dataset that was at the hour level, as most datasets were aggregated at the day level. For the hour interval, we determined it was appropriate to exclude the missing hours from the analysis. Looking at the data summary above, only 578 records out of 12k were missing. Further, even if we brought in the Bitcoin pricing data for the missing hours, we would still be missing the independent variables for those hours. As the Twitter data was aggregated at the daily level, we determined it was appropriate to pull in Bitcoin prices at the daily interval as well. For this, we used the R package "crypto2". This allowed us to align the aggregated daily Twitter data with daily Bitcoin OHLC data.

3.3 Bitcoin Data

The dataset from Reddit was not as clean as the Twitter dataset. As such, several steps were performed to extract, transform, and load the dataset into model-ready format. This portion of the analysis was performed in Python to take advantage of the natural language processing packages.

The data was sourced from Kaggle, and contained comments on Reddit.com from 2012-2019, with each row representing the comments. To prepare the Reddit data from the kaggle dataset, there were a few transformations that needed to be done. First, we aligned the Reddit comments time frame with the kaggle dataset for the tweets about Bitcoin. The Twitter dataset contains only tweets from August 2017 to January 2019, and the Reddit comments reached all the way back to 2012. Therefore, we trimmed the Reddit comment dataset to match the Twitter dataset.

After filtering on date, 4% of the records had blank values. These records did not appear to be corrupted records on import, but instead were comments attached to other comments. Because these comments had

null values for the datetime they were posted, we were unable to include them in our analyses.

Next, we inspected the data for other formatting concerns. Specifically, we should be sure to handle:

- urls
- special characters
- new lines
- foreign languages
- numbers (typically do not add context to the sentiment)

Urls, special characters, new lines, and numbers were removed from the data. Next, we used a machine learning package for language detection. Because the sentiment model we selected only works on English language, we removed non-english comments (~2% of total comments.)

3.3.1 Sentiment Tagging

To decide on a model, we considered two factors, accuracy of model and consistency with other data sets.

Regarding accuracy of model, we consulted a whitepaper, Social media sentiment analysis for cryptocurrency market that covers accuracy of different sentiment analysis models when trained on cryptocurrency content across Twitter and Reddit. The paper found that of 21 models tested, VaderSentement was one of the strong performing models. Because this is the same model used in the Twitter data kaggle set, we used it for the Reddit sentiment analysis as well. Using this package in Python, we are able to input a sentence, and the sentiment score is output indicating whether the phrase is positive/negative/neutral.

Finally, we aggregated the data. The original data was on a per comment basis, and the Twitter data was aggregated at hourly intervals. So we summed the total positive, negative, and neutral comments at each hour, in aggregate, and across different subReddits.

3.3.2 Final Transformations in R

After preprocessing and tagging of the data was complete, final transformations in R occurred. First, we aggregated the hourly data at a daily interval, to match the sets we created for the Bitcoin and Twitter datasets.

The result of all of these operations is six data frames: an hourly and daily interval aggregation file for the Twitter data, the Reddit data, and the Bitcoin OHCL data. These six data frames were then joined on day (for the daily interval set) and day and hour (for the hourly interval set) to make two dataframes, one for hourly interval and one for daily interval. These two final data frames were saved into csv files for reference, and will be then reloaded into a new workbook for further analysis.

3.3.3 Final Transformations

Lastly, we added columns to the data frame to show the movement of the close price from period *t-n* to period *t*. Predicting the price of Bitcoin at the top of the hour wouldn't be useful if we have to wait until the top of the hour to know the sentiment of all tweets and Reddit comments during that period.

Additionally, we determined it was appropriate to use a rolling sum of total/positive/negative/neutral comments/tweets as the model features. We noted during exploration this produced better model results, specifically with handling and accounting for outliers. The feature names such as 'total_tweets' will refer to the rolling sum of the last *d (days)* or *h (hours)* / periods.

We wrapped these transformations into a function, so that we can pass either a daily or hourly dataframe and a value of *n* and *d* to make transformations on the fly. We will use this to try many different values of

n and d/*h* to see which produces the best model. For initial model evaluation, we chose to use the values *d=7* for daily aggregated models and *h* = 6 hour hourly aggregated models.

4 Data Exploration

The main goal of the data exploration was to visualize the variables we have, see if any initial outliers stand out, and look at some initial correlations to verify if any variables seem to have the relationship we are expecting.

4.1 Variable Visualization

The first thing we did was visualize the distribution of our variables from the Twitter data using histograms to locate outliers. We can see that while the overall sentiment score (Compound_Score) appears to be normally distributed, all of our count variables are skewed, indicating there are some high outlier values that may need to be imputed or removed prior to using the data in our models.

Figure 1. Sentiment Score

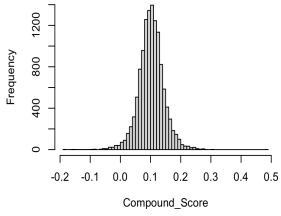


Figure 2. Negative Tweets

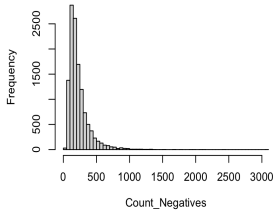


Figure 3. Positive Tweets

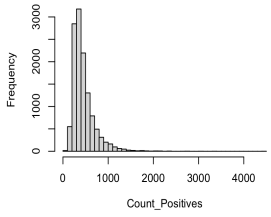
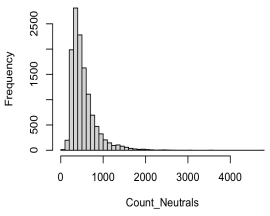


Figure 4. Neutral Tweets



Figures 1, 2, 3, 4 - Histograms illustrating the frequency of sentiment scores, negative tweets, positive tweets, and neutral tweets, respectively.

We repeated that exercise for the count variables from the Reddit data where we saw the same trend.

Figure 5. Positive Reddit Comments

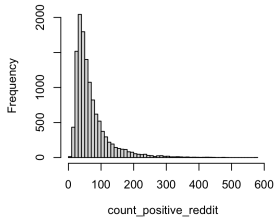


Figure 6. Negative Reddit Comments

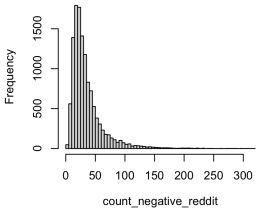
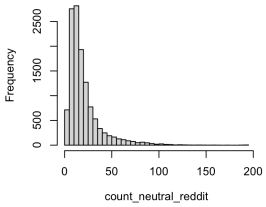


Figure 7. Neutral Reddit Comments



Figures 5, 6, 7 - Histograms illustrating the frequency of positive Reddit comments, negative Reddit comments, and neutral Reddit comments, respectively.

4.2 Variable Correlation

We first merged the hourly Twitter dataset with the hourly Reddit dataset. We then analyzed various variable combinations in order to verify if they have the relationship we are anticipating. We first compared the positive, negative, and neutral counts from each dataset to see if they are positively correlated. Our hypothesis was that if there are a higher number of positive tweets, for example, there would also be a higher number of positive Reddit comments.

Figure 8. Positive Reddit comments vs. Positive Tweets

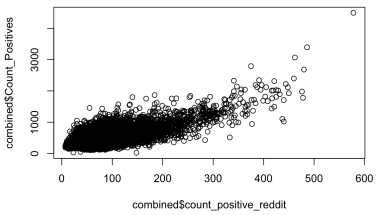


Figure 9. Negative Reddit Comments vs. Negative Tweets

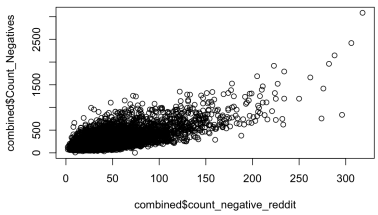
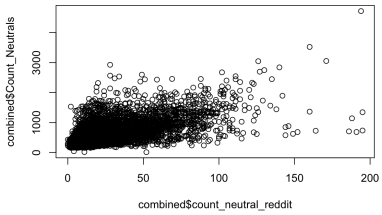


Figure 10. Neutral Reddit Comments vs. Neutral Tweets



Figures 8, 9, 10 - Scatterplots illustrating the relationship between the count of tweets and Reddit comments for each of the three sentiment categories; positive, negative, and neutral.

In general, the positive and negative comments do seem to have a positive correlation, while the neutral appears to have a weaker correlation. For this reason, multicollinearity could be present in models that use combinations of these variables, so this is something we will have to keep in mind.

The last thing we did was plot a few of these variables ($d = 7$ *and* $d = 6$ for hour and day, respectively.) against the change in closing price of Bitcoin from one hour to the next (lag n-1) in order to see if any had a clear linear relationship. None of the combinations we tried had an obvious linear relationship. Below we see these plots for the count of positive tweets as well as the count of positive Reddit comments.

Figure 11. Comments/Tweets vs. Price Change, Daily

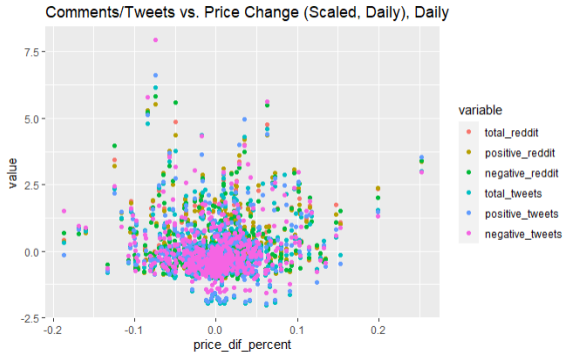


Figure 11 - Scatterplot illustrating the relationship between the count of different sentiment Reddit comments and tweets versus the change in daily Bitcoin price.

However, this does not mean that they will not prove to be good predictors. They may need to be used in conjunction with other variables or we could need to use a different lag in order to calculate the change in Bitcoin price. We may also need to perform non-linear transformations on some of the variables. All of these options will be explored in upcoming stages of the project.

We then looked at the relationship between the price of Bitcoin against the closing price of Bitcoin on each day. Here, there seemed to be an apparent visual correlation here.

Figure 12. Reddit/Twitter Activity vs. Price, Daily

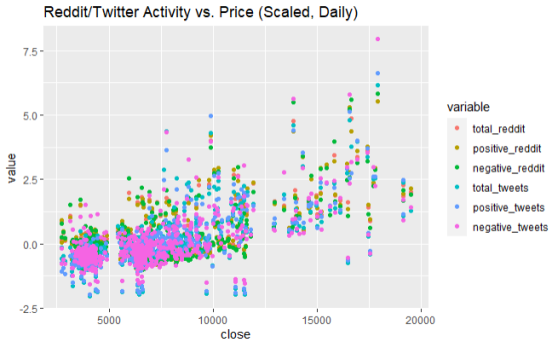


Figure 12 - Scatterplot illustrating the relationship between the closing price of Bitcoin and the value of Bitcoin.

4.3 Sentiment & Price Correlation

Lastly, we observed the relationship between the percent positive/negative tweets/Reddit Comments and the price of Bitcoin. Here, we noted no correlation between the mix of positive and negative talk and the price. If our hypothesis were true (that the sentiment of the discussion impacts the price) then we would have seen positive percent of tweets/comments spike when the price spikes, and negative percent spike when price dropped. As such, the remainder of our analysis will focus on where we did see a correlation - total tweet/comment activity and price.

Figure 13. Sentiment Mix vs. Price, Daily

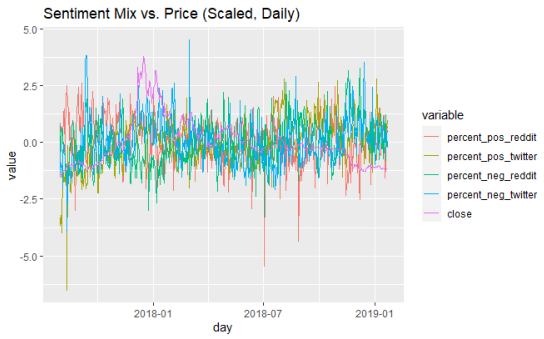


Figure 13 - illustrating the relationship between the percentage of positive and negative tweets and Reddit comments, versus the daily closing price of Bitcoin

5 Data Analysis

The next phase of this project will revolve around building models to see which combination of variables will produce the best model. For each model, we will evaluate it to ensure it does not violate any of the assumptions of a linear regression model. Those assumptions are:

- Collinearity
- Non-linearity of the response-predictor relationships
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points

For each model, if an assumption has been violated, we will move on to analyzing another model and discard the model from final consideration.

Because we saw above that sentiment and pricing movements violated the “Non-linearity of the response-predictor relationships” assumption, we focused our analysis on the relationship between Reddit comment/Twitter activity and the closing price of Bitcoin. Each analysis will be performed at the time lag of $n=1$. If one model performs significantly better than the others, we will explore other time lags.

Regarding validation, because each data point is correlated with the preceding *d/h* data points, we cannot split the data randomly for validation purposes. As such, we used the first 80% of data to validate, and the last 20% for validation.

5.1 Simple Linear Regression

The first model evaluated was a simple linear regression model.We used the counts of each day for each type of activity as the independent variables, and performed one regression at the daily aggregation and one at the hourly aggregation.

Table 1. Model 1 Variable Selection

Independent Variables	Dependent Variable	Aggregation	Lag/Lead
'positive_Reddit', 'neutral_Reddit', 'negative_Reddit', 'positive_tweets', 'neutral_tweets', 'negative_tweets'	lead_close	Daily, Hourly	n = 1 d = 7 h = 6

Table 2. Model 1 Performance

Model (n, d/h)	r^2	Adjusted r^2	V. r^2
Daily (1, 7)	.805	.803	.798
Hourly (1, 6)	.543	.543	.541

The model aggregated at the daily level performed better with a r^2 value of .805. After splitting into train/validation sets, that value dropped to .798. This suggests that 79.8% of the variance in Bitcoins price can be explained by the conversation activity around Bitcoin.

5.1.1 Collinearity

The first thing we want to check here is collinearity. We saw in our exploratory data analysis that these variables appeared to be correlated. To check, we will calculate the variance inflation factor for each variable. If the VIF is above 5 for any of the variables, we should discard it from the model and reevaluate.

```
positive_reddit  neutral_reddit negative_reddit positive_tweets  neutral_tweets negative_tweets
       50.71504        15.28203        46.17748        16.50701        10.59575        11.08261
```

In our case, all variables had a VIF > 5. As such, we discard the model from consideration.

To reduce collinearity in our model, we have several options. We can introduce an interaction term, perform principal component analysis, or remove variables.

5.2 Interaction Term

For this model, we will try using an interaction term. For our interaction term, we will use the volume traded at period t-n. Logically, the amount price moves may be tied to not just how much people are talking about Bitcoin, but also how much Bitcoin is being traded. For example, people may be talking very negatively about Bitcoin at a given moment, but what if no one is trading on this information? As such, we included this as our interaction term.

Table 3. Model 2 Variable Selection

Independent Variables	Dependent Variable	Aggregation	Lag/Lead
'positive_Reddit'*volume, 'neutral_Reddit'*volume, 'negative_Reddit'*volume, 'positive_tweets'*volume, 'neutral_tweets'*volume, 'negative_tweets'*volume, volume	lead_close	Daily, Hourly	n = 1 d = 7 h = 6

Table 4. Model 2 Performance

Model (n, d/h)	r^2	Adjusted r^2	V r^2
Daily (1, 7)	.871	.868	.812

Hourly (1, 6)	.555	.554	.492

Here, we again see a strong r^2 value of .812. This is likely due to volume being a pretty close proxy for how excited people are about Bitcoin, thus driving the price up.

5.2.1 Collinearity

Again, we check the collinearity. Here we see

```
positive_reddit  neutral_reddit negative_reddit positive_tweets  neutral_tweets negative_tweets
       50.71504        15.28203        46.17748        16.50701        10.59575        11.08261
```

In our case, all variables had a VIF > 5. As such, we discard the model from consideration.

5.3 Principal Component Analysis

Now, we will instead try to reduce the number of variables by using principal component analysis. This will take all of our features, make a linear combination of them, and give us a new set of features. While this will not yield any interesting information about the coefficient of the regression model, it should give us independent variables and thus allow us to not violate the collinearity assumption.

Table 5. Model 3 Variable Selection

Independent Variables	**Dependent Variable**	**Aggregation**	**Lag/Lead**
Principal Components of: 'positive_Reddit', 'neutral_Reddit', 'negative_Reddit', 'positive_tweets', 'neutral_tweets', 'negative_tweets'	lead_close	Daily, Hourly	n = 1 d = 7 h = 6

Table 6. Model 3 Performance

Model (n, d/h)	**r^2**	**Adjusted r^2**	**V. r^2**
Daily (1, 7)	.765	.763	.756
Hourly (1, 6)	.454	.454	.452

Figure 14. Principal vs. Variation

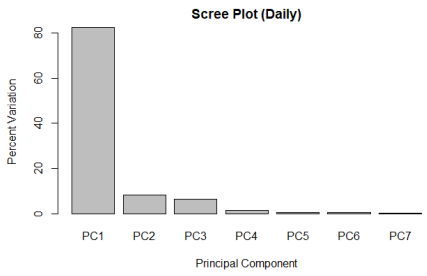


Figure 14 - Percentage distribution of the variation in the independent terms between each principal component

After generating our principal components (PCs), we examined them using a Scree plot to determine how many to include in our regression analysis. In the plot on the right, we can see that the first principal

component comprises 80% of the variation in the independent terms. The first three principal components make up 97.3% of the variation in independent variables. As such, we elected to use the first three principal components in our regression analysis. See model results above.

5.3.1 Collinearity

Using principal component analysis ensures that none of our principal components will have collinearity. We can check this by calculating the VIF for each PC. Using the VIF function in R, we can see that each variable has a VIF of 1. As such, we progress along to evaluating other assumptions of this model.

5.3.2 Correlation of Error Terms

An important assumption is that error terms e1, e2, ...,en are uncorrelated. If they aren't, then we have autocorrelation. To check this, we performed a Durbin Watson Test. Our p value is 0 for both the daily and hourly model, which means we can reject the null hypothesis and conclude that the residuals in this regression model are autocorrelated.

The Durbin-Watson statistic is only suitable for ordered time or spatial series. Because we transformed our data into cross section variables with PCA, we cannot use it to detect autocorrelation.[1] The impact of autocorrelation on PCA and PCA-based SPC is neither well understood nor properly documented. As such, for the purposes of this paper, we will ignore this assumption.[2]

5.3.3 Heteroskedasticity (non-constant variance of error terms)

Figure 15. Residuals vs. Fitted

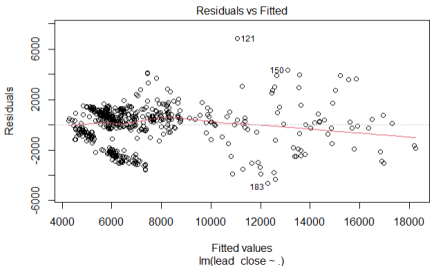


Figure 15 - Scatterplot illustrating heteroskedasticity between the closing price of Bitcoin and the residuals

Above, we can see that there is not a strong pattern emerging. There is some drift towards the end, however, it is not strongly correlated outside of a few observations. As such, we conclude our model does not violate this assumption.

In addition to heteroskedasticity, we also want to check and see if our residuals are normally distributed and if they are spread equally along the range of fitted values.

Figure 16. Normal Q-Q

Figure 17. Scale-Location

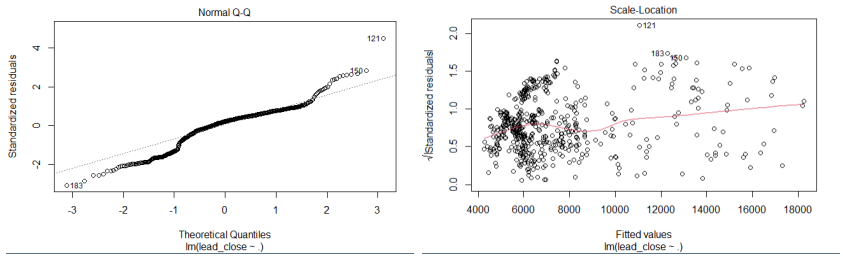


Figure 16, 17 - Scatterplots illustrating distributions and spread of residuals, respectively

Here we can see our residuals are not very skewed, meaning our model is performing well across our entire dataset. We do see some skewness emerge towards the right of the Q-Q plot, however, it only appears to be a few points, and thus may be caused by outliers. Otherwise, the residuals have a pseudo normal distribution.

5.3.4 Outliers and High Leverage Points

Lastly, we look at outliers and leverage. On the first chart, we have two points that seem to be outliers, as their cook's distance is greater than .5. We note that one of the points, 121, is the same point on the Q-Q plot that fell outside of a normally distributed range for residuals. However, looking at the leverage plot, there are no points that appear to be high leverage outliers. This means we could likely remove these points and not have much of an impact on the model. Thus, we conclude that the points are outliers, but are not impacting the overall model and do not need to be removed.

Figure 18. Cook's distance

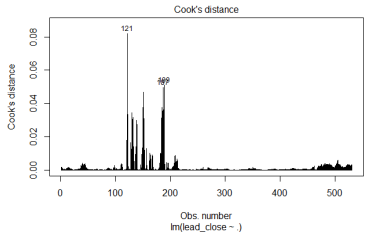


Figure 19. Residuals vs. Leverage

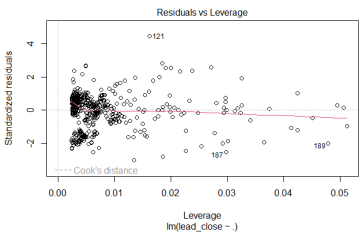


Figure 18, 19 - Bar chart illustrating cook's distance for each observation, and scatterplot illustrating high leverage outliers, respectively.

5.4 Linear Model - Less Variables

Finally, we try a very simple linear regression model using only the total tweets and Reddit comments, and volume traded.

Table 5. Model 4 Variable Selection

Independent Variables	**Dependent Variable**	**Aggregation**	**Lag/Lead**
'total_Reddit' 'total_tweets'	lead_close	Daily, Hourly	n = 1 d = 7 h = 6

Table 6. Model 4 Performance

Model (n, d/h)	**r^2**	**Adjusted r^2**	**V. r^2**
Daily (1, 7)	.774	.772	.768
Hourly (1, 6)	.462	.461	.461

5.4.1 Collinearity

In this model, the VIF factor for all independent variables was < 5 for both the hourly and daily model. As such, the model does not violate this assumption.

5.4.2 Correlation of Error Terms

This model's p-value was below .05 for the Durbin-Watson test, thus we reject the null hypothesis and conclude the error terms are autocorrelated. As such, no further investigation is performed.

6. Conclusion

We have proved there is a correlation between the price of Bitcoin and the amount of discourse surrounding Bitcoin, but we have not proved causation. Regardless, the analysis above shows us there is valuable information that can be extracted from the discourse surrounding Bitcoin. This information could be incorporated into a trading bot that purchases and sells Bitcoin based on indicators calculated using methods described above.

The above models all showed one thing very clearly: there is a definite correlation between the public discourse and the price activity, with just the total number of tweets and Reddit comments accounting for around 75% of the variance in price (simple model - less variables). However, no model showed that the amount of negative or positive sentiment specifically influences the price, rather it's just the total activity that is possibly influencing the price. As such, we reject our hypothesis and conclude that sentiment is not an important indicator– instead the total volume of discourse is.

Reference List

1. Chen Y (2016) Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression. PLoS ONE 11(1): e0146865. doi:10.1371/journal.pone.0146865
2. Vanhatalo E, Kulahci M (2015) Impact of Autocorrelation on Principal Components and Their Use in Statistical Process Control. Qual. Reliab. Engng. Int., 32: 1483– 1500. doi: 10.1002/qre.1858.