Problem Statement and Hypothesis

Thinking of utilizing social media in forecasting stock shares and cryptocurrency prices has been started with the boom in using social media as a platform to express opinions, Twitter specifically has become known as a location where news is quickly disseminated in a concise format, According to (Colianni,2015) in 2010 Bollen et al. utilized the Profile of Mood States (POMS) to predict the movement of the Dow Jones Industrial Average with 87.6% accuracy, and in 2009, Go et. al. focused on classifying tweets and used several approaches to achieve an accuracy of 84.2% with Multinomial Naïve Bayes, 79.2% with maximum entropy, and 82.9% using a support vector machine. (Colianni,2015), but social media is still doubtful when it comes to research and statistical analysis because (for several reasons) Twitter samples are not necessarily representative of the population, and this should be taken into consideration from a statistical analysis point of view. (White, 2018)

The problem for this research is measuring the significance of the relationship between the estimated sentiments of Tweets (from Twitter) and the change in Bitcoin prices, through testing the following null and alternate hypotheses:

Null hypothesis: There is no correlation between the estimated sentiments from social media and the change in Bitcoin price.

Alternate Hypothesis: There is a significant relationship between the estimated sentiments from social media and the change in Bitcoin price.

In this study linear regression has been used to infer the significance of the relationships between the independent variable (Bitcoin price changes or ‘diff’) and dependent variable/s, by estimating the strength of the relationship between two variables related to the proportion of variance in the dependent variable that can be explained by the independent variable/variables.

(Freedman, 2009), the null hypothesis that has been tested by the t-test is that each coefficient vanishes (Freedman, 2009), using the corresponding p-value which represents the probability of obtaining test results while the null hypothesis is correct.

Summary of Data Analysis Process

Two public datasets have been used, text tweets with time stamps and time series of Bitcoin prices through the same period of tweets, both have been obtained through Kaggle.com

1. Bitcoin Twitter Data (https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets)
2. Bitcoin Price Data (https://www.kaggle.com/maxwells/btcusd)

The first raw data contained 13 columns and 423,111 records cover the period from February 5th, 2021 to July 5th,2021 with an irregular timestamp, for this analysis only three variables have been used, (Date or timestamp, Text tweets, and number of User_followers)

The text for the tweets was prepared by dropping duplicates and cleaned from mentions (ex. @SomeOne), hashtags (ex. #SomeThing), URL links, and converted to lower case, a user-defined function has been created with the use of regular expression (Naik, 2020) and (Gulsen, 2021), also the date column needed minor cleaning from some misformatted rows, the tweets data frame (df_Bitcoin_tweets) was converted into a time series.

Sentiment analysis of tweets has been achieved using TextBlob, the sentiment property returns a named tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0] (TextBlob, n.d.).Textblob users suggest converting the fraction polarity score into three categories (positive, neutral, and negative) or (+1, 0, and -1) as a best practice for the sentiment polarity score (James, 2018) and (Gulsen, 2021).

```
def getsubjectivity(text):
    return TextBlob(text).sentiment.subjectivity
def getpolarity(text):
    return TextBlob(text).sentiment.polarity
```

```
df_Bitcoin_tweets['subj']=df_Bitcoin_tweets['text'].apply(getsubjectivity)
df_Bitcoin_tweets['polarity']=df_Bitcoin_tweets['text'].apply(getpolarity)
df_Bitcoin_tweets['count']=1
```

```
df_Bitcoin_tweets['polarity'] = df_Bitcoin_tweets['polarity'].apply(lambda x: 1 if x > 0 else (-1 if x < 0 else x))
df_Bitcoin_tweets.head(10)
```

date	user_followers	date	text	subj	polarity	count
2021-02-05 10:52:04	301.0	2021-02-05 10:52:04	2□ debunking 9 bitcoin myths by _lowry_ #cryptocurrency bitcoin #crypto #blockchain btc...	0.000000	0.0	1
2021-02-05 10:52:04	301.0	2021-02-05 10:52:04	weekend read keen to learn about #crypto assets? check out our reading list! 1□ 2021 cryptomarket outlook...	0.000000	0.0	1
2021-02-05 10:52:06	301.0	2021-02-05 10:52:06	4□ bloomberg lp #cryptooutlook 2021 with #cryptocurrency bitcoin #crypto #blockchain btc...	0.000000	0.0	1
2021-02-05 10:52:07	301.0	2021-02-05 10:52:07	5□ #blockchain 50 2021 by . . #cryptocurrency bitcoin #crypto #blockchain...	0.000000	0.0	1
2021-02-05 10:52:26	37.0	2021-02-05 10:52:26	#reddcoin #rdd to the moon #altcoin #turnreddcoinin1dollar #eth btc bitcoin	0.000000	0.0	1
2021-02-05 10:53:49	100.0	2021-02-05 10:53:49	bitcoin and #eth both have bullish setups for a move higher ... btc it would just be great if daily close (in abo...	0.416667	1.0	1
2021-02-05 10:54:52	69.0	2021-02-05 10:54:52	Sperl 0.06.i have insisted that since 0.02 it will be 0.071. it increased 300% in about 2 months. bitcoin btc...	0.000000	0.0	1
2021-02-05		2021-02-05	#amazing #monopoly #crypto #cryptocurrency #cryptocurrencies #cryptonews			

Figure (1) – Bitcoin Twitter sentiment analysis.

```
df_Bitcoin_tweets['polarity'].hist(bins=5)
```

```
<AxesSubplot:>
```

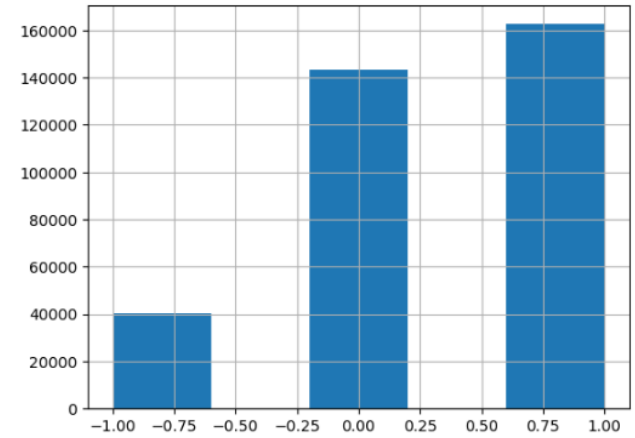

Figure (2) – Bitcoin Twitter sentiment polarity histogram.

Another column or variable has been added to the data and named “count”, which represents the count of tweets and was given a constant value ‘1’, this variable will represent the total number of tweets where Bitcoin was mentioned within one minute period after resampling and summation, for this analysis, the change in Bitcoin price will be represented by the “diff” column, which is a one-step difference was calculated using the diff( ) pandas method applied to Bitcoin close price extracted from the second raw data set.

Both data frames, df_Bitcoin_tweets and df_Bitcoin_Price have been joined together in one clean data frame starts at '2021-02-05 10:52:00' and ends at '2021-05-29 16:59:00' with a one-minute sample rate, the next step was resetting the data frame index and removing the timestamp as it's not necessary for the analysis, outliers have been removing from the variables.

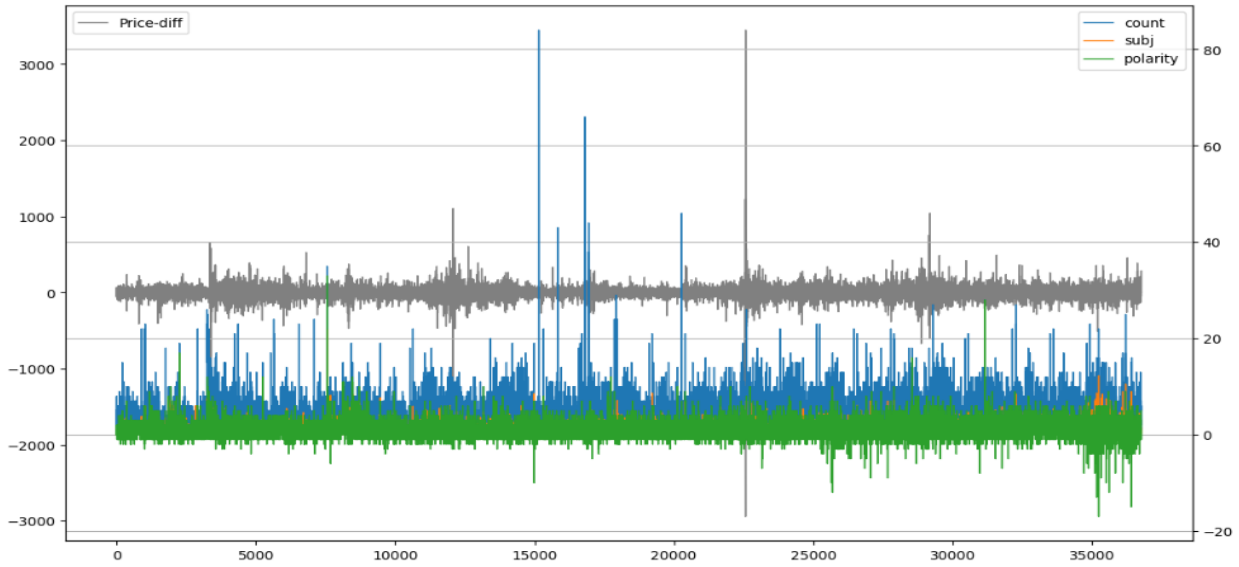

Figure (3) – Bitcoin Twitter Data, timestamp removed, data points stacked (before removing outliers).

The final clean and prepared data frame consists of 29437 rows each of them represents one minute period and five columns contain both the independent variables ['count', 'polarity', 'subj' and' user_followers'] for this one minute period and the corresponding dependent variable ['diff'] which represents the change of Bitcoin price.

Scipy.stats.OLS (ordinary least square) has been used for linear regression implementation, the modeling process started with (1) separating the dependent variable or the target ('diff') and the independent variable ('polarity'), (2) adding a constant to the independent variable array, (3) splitting the inputs into train and test data sets (70% of the data used for training the model and 30% left out for testing), (4) initialized then fitted the OLS model to the

train inputs (dependent and independent variables), (5) used the fitted model to estimate the parameters (model summary, t-statistic, p-value and input coefficient), the test data points have been used to evaluate the model.

```
X = sm.add_constant(X) #adding constant that will represent the intercept of the MLR model

# Splitting the data into train and test portions
X_train, X_test, y_train, y_test = sklearn.model_selection.train_test_split(X, y, test_size = 0.30)

# LR implementation
linear_regression = sm.OLS(y_train,X_train)
fitted_model = linear_regression.fit()
residuals =fitted_model.predict(X_test)-y_test
itercept= fitted_model.params[0]
coefficient =fitted_model.params[1]
```

Figure (4) – Linear regression model.

Outline of Findings

The null hypothesis will be that each coefficient vanishes, except for the intercept. the degree of significance of the regression t-statistics can be evaluated using the corresponding p-value which represents the probability of obtaining test results at least as extreme as the results actually observed, and the null hypothesis is correct. (Freedman, 2009)

```
# MLR summary:
print(fitted_model.summary())
```

OLS Regression Results

Dep. Variable:	diff	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	2.645
Date:	Sun, 25 Jul 2021	Prob (F-statistic):	0.104
Time:	19:59:40	Log-Likelihood:	-1.0983e+05
No. Observations:	20605	AIC:	2.197e+05
Df Residuals:	20603	BIC:	2.197e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0607	0.434	-0.140	0.889	-0.911	0.790
polarity	0.4741	0.291	1.626	0.104	-0.097	1.045

Omnibus:	2.676	Durbin-Watson:	2.006
Prob(Omnibus):	0.262	Jarque-Bera (JB):	2.725
Skew:	-0.008	Prob(JB):	0.256
Kurtosis:	3.054	Cond. No.	2.25

Figure (5) – Linear regression model summary (notice p-value 0.104).

```
# Linear regression plot
fig, ax = plt.subplots(figsize=(8, 6))
#plt.yscale('Log')
plt.scatter((X[inputb]),y,s=5,label='Data points')
line= range(int(X[inputb].min()),int(X[inputb].max()+1))
line2 = sm.add_constant(line)
plt.plot(line,fitted_model.predict(line2),'r--',label='Regression Line')
plt.legend(loc='upper right')
plt.ylabel(target)
plt.xlabel(inputb)
plt.grid()
plt.show()
```

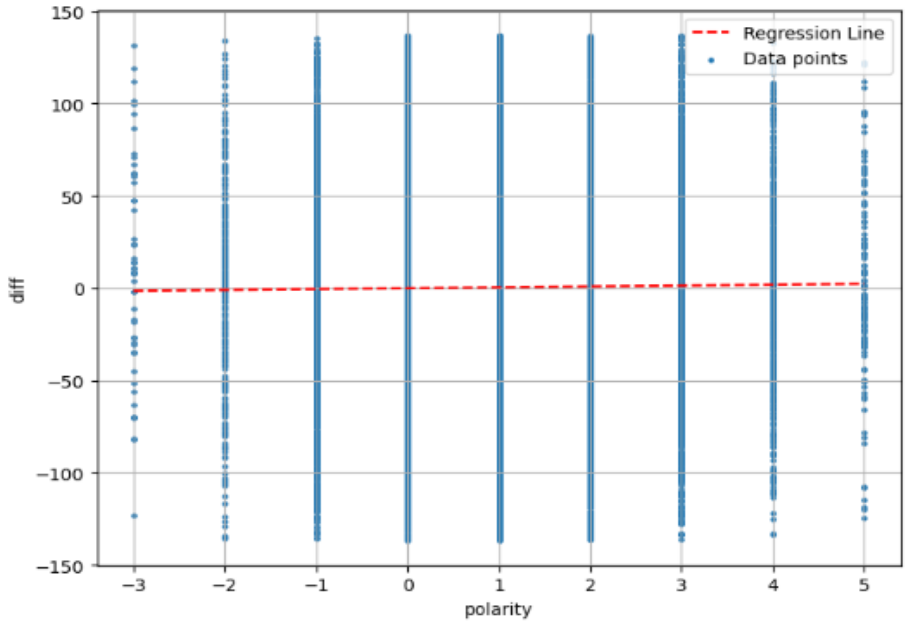

Figure (6) – Correlation between sentiments polarity and Bitcoin price differences (zero correlation), dotted red line represents the linear regression model.

Comprehensively, the linear regression has been implemented between the dependent variable ['diff'] and each independent variable separately ['count', 'user_followers', 'subj', 'polarity'], then every possible pair of input variables, every possible combination of three variables and so on using the explained steps (1 to 5), through all possible combinations of variables and the total number of tests was 15 as summarized in the following table, the maximum R^2 obtained is approximately zero (0.000185)

Test_no. for target	Target	R2	No. of inputs	coefs_list	p-values
1	diff	0.000004	1	count 0.067335,	count 0.742073,
2	diff	0.000049	1	user_followers -0.000065,	user_followers 0.253961,
3	diff	0.000049	1	subj 0.538057,	subj 0.252352,
4	diff	0.000045	1	polarity 0.280032,	polarity 0.275428,
5	diff	0.000048	2	count 0.214765, user_followers -0.000057,	count 0.335329, user_followers 0.360279,
6	diff	0.000097	2	count -0.252177, subj 0.966731,	count 0.336325, subj 0.108665,
7	diff	0.000132	2	count -0.102538, polarity 0.514895,	count 0.643626, polarity 0.064380,
8	diff	0.000147	2	user_followers -0.000094, subj 0.689390,	user_followers 0.106152, subj 0.150652,
9	diff	0.000146	2	user_followers -0.000089, polarity 0.345873,	user_followers 0.120886, polarity 0.179376,
10	diff	0.000082	2	subj 0.460455, polarity 0.177298,	subj 0.411837, polarity 0.561699,
11	diff	0.000156	3	count -0.235845, user_followers -0.000033, subj 1.155264,	count 0.401813, user_followers 0.601976, subj 0.056153,
12	diff	0.000098	3	count 0.014558, user_followers -0.000050, polarity 0.365392,	count 0.951879, user_followers 0.429089, polarity 0.188920,
13	diff	0.000103	3	count -0.212190, subj 0.782586, polarity 0.161240,	count 0.419465, subj 0.240958, polarity 0.599665,
14	diff	0.000185	3	user_followers -0.000067, subj 0.557448, polarity 0.301649,	user_followers 0.248866, subj 0.327439, polarity 0.322805,
15	diff	0.000163	4	count -0.056509, user_followers -0.000069, subj 0.907563, polarity 0.055152,	count 0.841532, user_followers 0.270820, subj 0.175814, polarity 0.857797,

Figure (7) – Summary of 15 linear regression tests.

As indicated by the tests summary table and the detailed model's summary, It can be concluded that the proportion of variance in the Bitcoin price changes or 'diff' that can be explained by the independent variables and specifically the estimated sentiments from tweets is too small, Figure (6) explains that the relationship has a slope (polarity variable's coefficient) too close to zero, and the corresponding P-value larger than 0.05 (p-value 0.104 in the case of polarity variable) indicates a very high probability of obtaining test results without any influence from the estimated sentiments, the null hypothesis couldn't be rejected, and the statement stated "There is no correlation between the estimated sentiments from social media and the change in Bitcoin price" is still a valid assumption.

Limitations of the Used Techniques and Tools

The analysis was limited due to some factors, for example, the high sensitivity of linear regression to noises and outliers. (Kumar, 2019), using the Textblob python library for sentiment analysis added some uncertainty to the analysis as it tended to label most analyzed texts as

neutral (zero sentiment polarity). Also, the analysis results represent the correlation at a relatively high frequency or sample rate (one minute) while a significant interaction between social media and Bitcoin price changes is expected to take longer to happen.

Proposed Actions

For future studies about the same subject, it is highly recommended to collect more comprehensive data sets, that allow using different time frames (or sample rate) longer than one minute (hours or one day) as this would provide a more well-rounded view, testing different sentiment analysis techniques can be very beneficial, another suggestion for future studies would be to include data from other social media platforms not only from Twitter.

Expected Benefits of the Study

Although the study failed to reject the null hypothesis and couldn't find a significant relationship between the estimated sentiments from social media and the change in Bitcoin price, yet it can provide some benefits and information such as,(1) highlighted the unlikelihood of interaction between social media and high-frequency data, (2) highlighting the weak relationship between estimated sentiments and changes in Bitcoin price, (3) supported the doubtfulness in the statistical significance of social media samples, (4) examined the Textblob library and highlighted some of its limitations, (5) highlighted the limitation of linear regression for prediction in the case of sentiment analysis. (6) the main benefits of the study will be using or following the recommended actions in future studies that are interested in a similar subject, such as, (1) testing different sentiment analysis techniques, (2) including data from other social media platforms and not only from Twitter, (3) adding frequency or samples rate as a variable in the analysis by using lower frequency data, (4) using different machine learning techniques for prediction of changes in Bitcoin prices as explained by (Colianni,2015).

References

Colianni, Stuart et. al. (2015) Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis. Retrieved July 18th, 2021 from http://cs229.stanford.edu/proj2015/029_report.pdf

Freedman, David A. (27 April 2009). Statistical Models: Theory and Practice. Cambridge University Press. ISBN 978-1-139-47731-4.

Gulsen, Furkan (2021) Bitcoin Sentiment Analysis, Retrieved July 18th, 2021 from https://www.kaggle.com/codeblogger/bitcoin-sentiment-analysis

James, Sangeetha. (2018) Stackoverflow question, Retrieved July 18th, 2021 from https://stackoverflow.com/questions/51209514/how-does-textblob-calculate-sentiment-polarity-how-can-i-calculate-a-value-for

Kumar, Naresh. (2019) Advantages and Disadvantages of Linear Regression in Machine Learning, Retrieved July 18th, 2021 from https://theprofessionalspoint.blogspot.com/2019/05/advantages-and-disadvantages-of-linear.html

Naik, Krish. (2020) Natural-Language-Processing, Retrieved July 18th, 2021 from https://github.com/krishnaik06/Natural-Language-Processing/

TextBlob. (n.d.). Simplified Text Processing, Retrieved July 18th, 2021 from https://textblob.readthedocs.io/en/dev/

White, Swede. (2018) Analyzing Correlations between #BTC Tweets and Bitcoin Trading Using Natural Language Processing. (Wolfram Technology Conference), Retrieved July 18th, 2021 from https://www.youtube.com/watch?v=DNU2SW4Lb2Y