

Sentiment analysis of web data to predict price fluctuation of cryptocurrencies



ETHEREUM



Group 4:

Jainam Jhaveri

Neha Sharma

Rushabh Vakharia

Sumesh Vijayan

Overview:

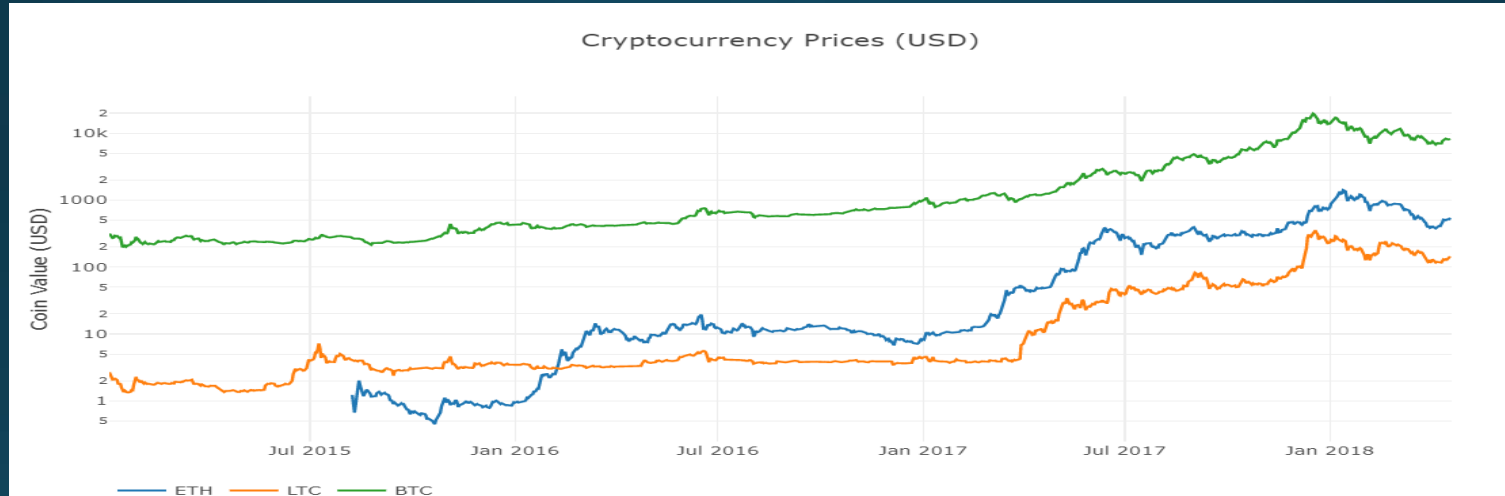
- ❖ As the world's first completely decentralized digital payment system, Bitcoin represents a revolutionary phenomenon in financial markets.
- ❖ Following the path of Bitcoin, other cryptocurrencies like Litecoin and Ethereum also gave customers the option to divulge in the digital currency market.
- ❖ Though existing only in digital form, the rise of these currencies is unstoppable.
- ❖ Amidst this entire hype, there is growing concern focused on cryptocurrencies' considerable price volatility and associated risks.
- ❖ There exists a significant impact of social media on future cryptocurrency returns.
- ❖ The web offers substantial information about bitcoin's acceptance among the general public, as well as daily fluctuations in its market sentiments.
- ❖ This helps investors gain insights into cryptocurrencies' value from this information-rich environment.

Motivation:

- ❖ The cryptocurrency market is pretty unregulated which is driven by demand and supply.
- ❖ It is highly based on the sentiments of the people.
- ❖ In order to help people make calculated and profitable decisions, we are planning to perform analysis on tweets and news headlines from the web for Bitcoin, Litecoin and Ethereum.
- ❖ People usually say, write or tweet things which they are interested or disinterested in.
- ❖ This is the central idea of our analysis and will help us to make the correct decisions.

Price Trends:

- ❖ For the three cryptocurrencies mentioned previously, the price trend from 2014 to 2018 is as seen here:



- ❖ These prices are fetched from exchanges like COINBASE, KRAKEN, OKCOIN and a few more.
- ❖ As can be seen, there is a lot of fluctuation in the prices and its reasons are what we are trying to understand.

Data Sources:

- ❖ For our analysis, we will be scraping and extracting data from Twitter and Kaggle along with following websites.
- ❖ Below is the list of data sources for each of the three cryptocurrencies:
- ❖ Bitcoin:
 - ❖ <https://news.bitcoin.com>
 - ❖ <https://www.bloomberg.com>
- ❖ Litecoin:
 - ❖ <https://www.litecoinnews.io>
- ❖ Ethereum:
 - ❖ <https://www.coindesk.com>

Working:

- ❖ For each of the currencies, we will be scraping news headlines from the sites mentioned along with the date of its publication.
- ❖ Similarly, tweets regarding the currencies too will be fetched using the Twitter API.
- ❖ Once the data is scraped and extracted, we will filter out irrelevant data and use VADER to classify the tweets or headlines as positive, negative or neutral.
- ❖ Once this is done, we will be cleaning and clustering the data using the LDA (Latent Dirichlet Allocation) algorithm.
- ❖ Finally, we will be predicting the prices using logistic and linear regression.

Process Diagram:

Scrape news headlines and tweets for Bitcoin, Litecoin and Ethereum

- Sources: Twitter, Kaggle, litecoinnews.io, news.bitcoin.com, cryptocurrencynews.com, bloomberg.com

Use VADER's in-built lexicon file to classify the headlines or tweets as positive, negative or neutral.

Cleaning and preprocessing of data for clustering purpose

Cluster the headlines and tweets using LDA (Latent Dirichlet Allocation)

Perform linear/logistic regression on the clustered data to predict the prices

Twitter:

- To extract tweets from Twitter, we will be using the Tweepy package.
- In the class created for the analysis, we will first be fetching tweets based on the specific keywords like 'bitcoin', 'litecoin', 'ethereum', 'cryptocurrency', 'blockchain'.
- Once these tweets are fetched, we will be cleaning them by removing the unnecessary hyperlinks, words and characters using regular expressions.
- Finally, the sentiment for each of these tweets will be calculated using the Textblob package in Python and stored in a dictionary.

Twitter Sentiment Results:

- ❖ The screenshots attached herewith show how tweets contribute to price fluctuations in cryptocurrencies.
- ❖ Positive and negative tweets can be seen.
- ❖ Also, percentage of the three types of tweets is shown.

```
Positive tweets percentage: 42.857142857142854 %  
Negative tweets percentage: 9.523809523809524 %  
Neutral tweets percentage: 47.61904761904762 %
```

Positive tweets:

```
RT @Peurtoken: Did you know that there is an additional 10% bonus token in our Pre-Sale now? Best Price Ever. Check on https://t.co/o5pyFZP...
```

```
*****
```

```
RT @SocialWalletInc: #Nasdaq CEO say's they'll consider becoming an exchange for #Crypto https://t.co/SgC9jQRxb7 #Blockchain #Bitcoin #btc...
```

```
*****
```

```
RT @RealCryptoGuide: These Small Cap Cryptocurrency Projects Are Worth Researching - https://t.co/yyHmSAuPB7 #Investing #Cryptocurrency #Cr...
```

```
*****
```

```
RT @truechaingroup: With all the new blockchain platforms being developed around the world, TrueChain's leadership on that front is becomin...
```


```
*****
```

```
RT @bethereumteam: Another great #networking session in #WorldBlockchainForum #Dubai filled with valuable feedback! It was good to finally...
```




```
*****
```

```
*****
```

Negative tweets:

```
RT @bethereumteam: Looking for a #TokenSale with a working MVP?   
Less than a week left to get your 30% BONUS: https://t.co/VupC9N9Dcm  
#Cr...
```

```
*****
```

```
Mass Adoption of #Bitcoin   
```

```
1 - 100x Leverage 
```

```
 https://t.co/XokkANT2um 
```

```
Register Now & go Long  or Short... https://t.co/3KVCowtv87
```

```
*****
```

Twitter Drawbacks:

- ❖ The data from twitter is extremely unorganized and unstructured and using it for analysis won't give us accurate results.
- ❖ This makes analyzing twitter data very difficult.
- ❖ Having said that, the results still won't help investors make profitable investing decisions.

Sentiment Analysis:

- ❖ Used VADER as our sentiment analyzer using its in-built lexicon file for the analysis.
- ❖ The analyzer yielded a pretty informative output classifying headlines and tweets into following categories:
 - ❖ Very positive
 - ❖ Somewhat positive
 - ❖ Very negative
 - ❖ Somewhat negative
 - ❖ Neutral
- ❖ VADER classifies sentiment intensity of sentences from -1 to 1.

Sentiment Analysis Results:

❖ Bitcoin:

- ✓ 1166 neutral headlines
- ✓ 252 somewhat negative headlines
- ✓ 50 very negative headlines
- ✓ 328 somewhat positive headlines
- ✓ 52 very positive headlines
- ✓ 1848 total headlines

❖ Ethereum:

- ✓ 235 neutral headlines
- ✓ 46 somewhat negative headlines
- ✓ 7 very negative headlines
- ✓ 71 somewhat positive headlines
- ✓ 10 very positive headlines
- ✓ 369 total headlines

❖ Litecoin:

- ✓ 34 neutral headlines
- ✓ 9 somewhat negative headlines
- ✓ 0 very negative headlines
- ✓ 11 somewhat positive headlines
- ✓ 2 very positive headlines
- ✓ 56 total headlines

Data Cleaning & Preprocessing:

- ❖ Tokenized the headlines and tweets.
- ❖ Removed stop-words.
- ❖ Removed punctuations.
- ❖ Performed lemmatization of unique words.
- ❖ Removed frequent keywords from lemmatized headline tokens to improve clustering performance.

Text Clustering:

- ❖ Implemented the Latent Dirichlet Allocation (LDA) unsupervised learning algorithm.
 - ❖ Created 5-clustered and 4-clustered documents into topics by word similarity.
- ❖ Built a prediction model utilizing the cluster breakdown.
- ❖ Clusters were used as input to predictive regression models
 - ❖ Clustering improved the results of linear and logistic regression model which was insignificant otherwise.

Price Prediction:

- ❖ There was no significant correlation between the overall daily sentiment and the price movement.
- ❖ But a specific cluster will have a greater impact on the daily price movements.
- ❖ Hence, it was necessary to determine specific sentiment scores for the cluster on a given day.
- ❖ We used days 1 and 3 in our case.

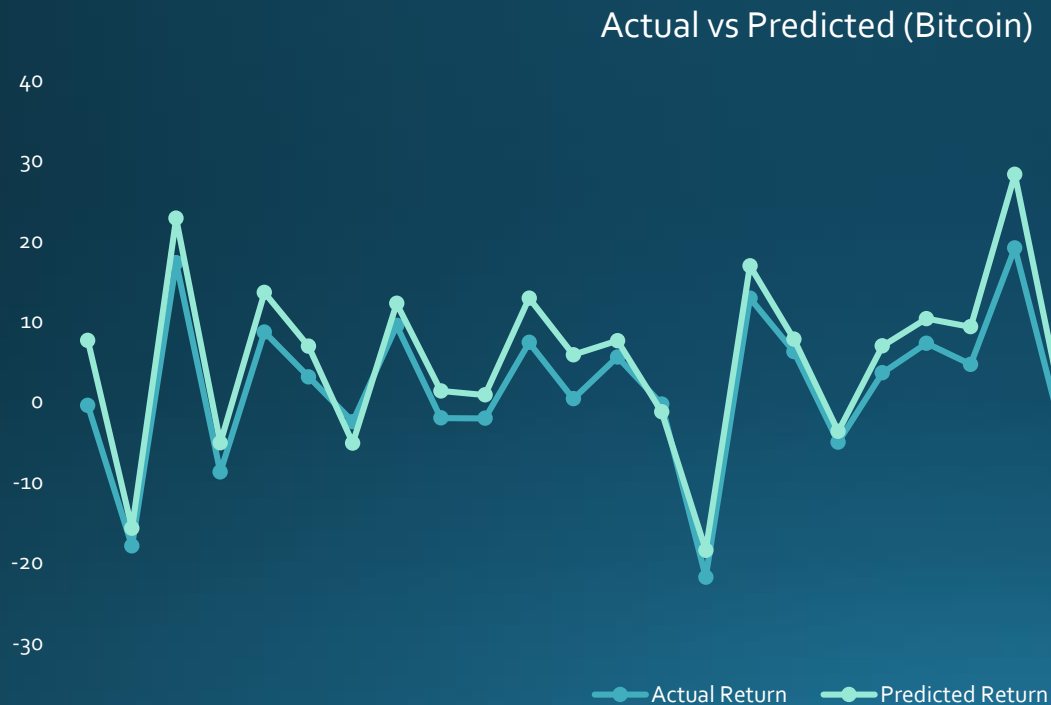
Price prediction using Linear Regression:

- ❖ Test 1: Linear Regression
- ❖ In order to predict the exact price value we used linear regression.
- ❖ Ran 4 multiple linear regression models in total:
 - ❖ 5 clusters 1 day price movement
 - ❖ 4 clusters 1 day price movement
 - ❖ 5 clusters 3 day price movement
 - ❖ 4 clusters 3 day price movement

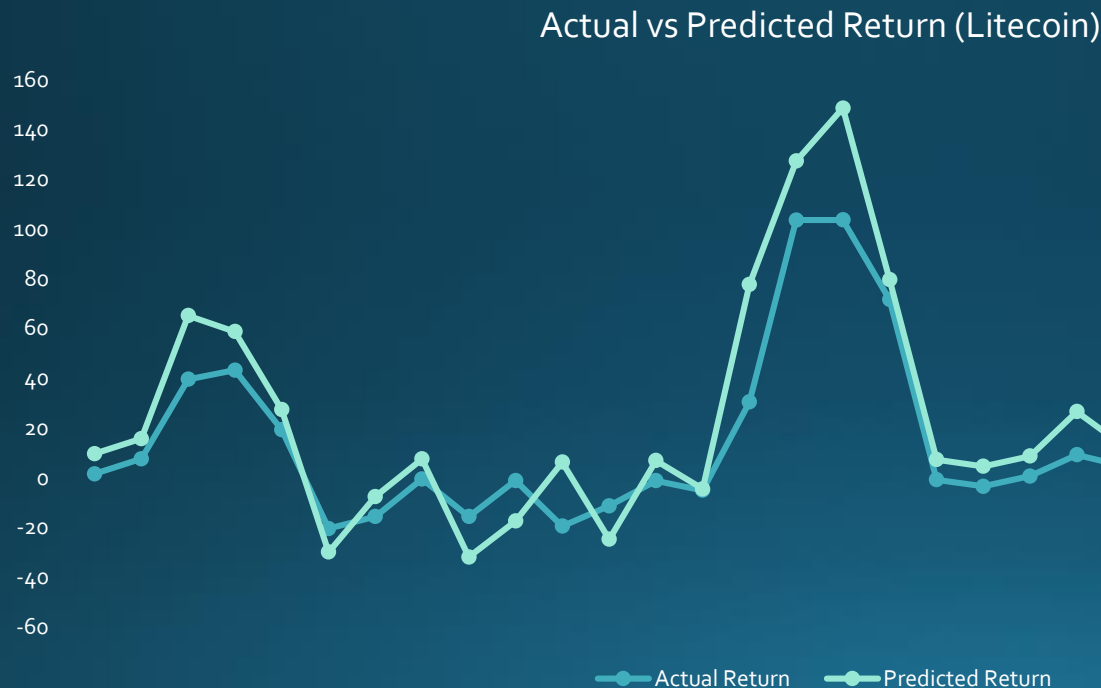
Linear Regression Results:

- ❖ On analyzing the R-squared values for linear regression performed on Bitcoin for all 4 model types, we found that model 3 worked the best with value 0.063
- ❖ On analyzing the R-squared values for linear regression performed on Litecoin for all 4 model types, we found that model 1 worked the best with value 0.151
- ❖ On analyzing the R-squared values for linear regression performed on Ethereum for all 4 model types, we found that model 3 worked the best with value 0.413

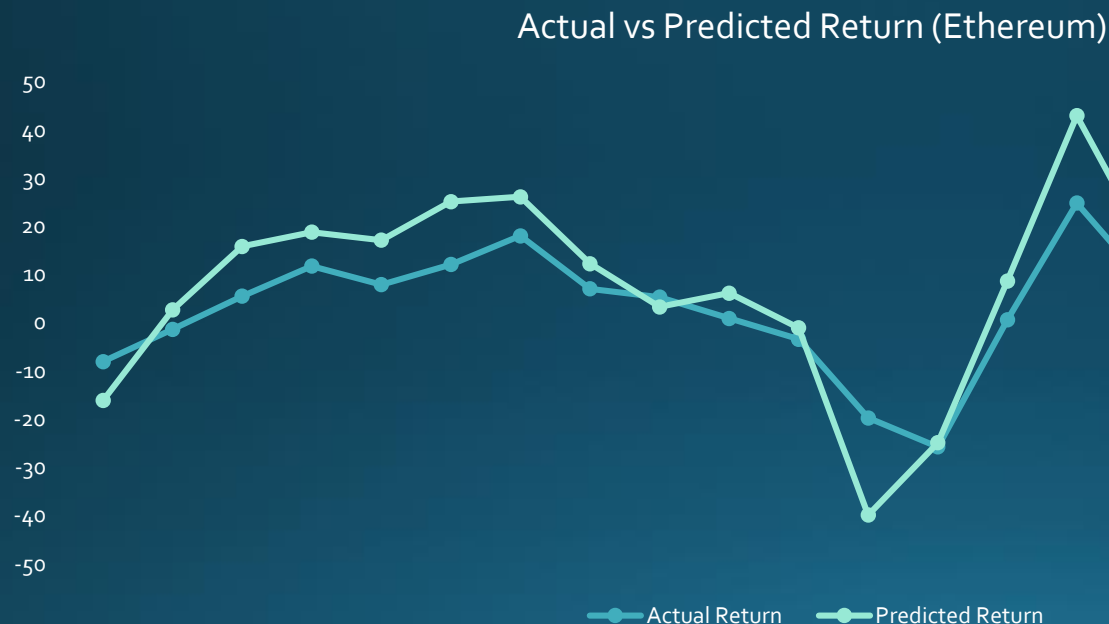
Actual vs Predicted Returns(Bitcoin):



Actual vs Predicted Returns(Litecoin):



Actual vs Predicted Returns(Ethereum):



Price prediction using Logistic Regression:

- ❖ Test 2: Logistic Regression
- ❖ Instead of predicting exact values we want to predict if the prices will rise or fall.
- ❖ This is also sometimes useful to investors.
- ❖ Like investors take a long or short positions in stock market, they can decide to buy or sell the cryptocurrencies in the same manner.
- ❖ Ran 4 multiple logistic regression models in total (same as before)

Logistic Regression Results:

- ❖ On performing logistic regression for Bitcoin and Ethereum, we found out that the model type with 5 clusters and 3 day price movement was still better than other models.

❖ Bitcoin

	precision	recall	f1-score	support
0	0.21	0.59	0.31	17
1	0.88	0.58	0.70	90
avg / total	0.77	0.58	0.64	107
	precision	recall	f1-score	support
0	0.31	0.58	0.41	26
1	0.81	0.59	0.69	81
avg / total	0.69	0.59	0.62	107
	precision	recall	f1-score	support
0	0.08	1.00	0.14	1
1	1.00	0.65	0.79	34
avg / total	0.97	0.66	0.77	35
	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	0.63	0.77	35
avg / total	1.00	0.63	0.77	35

❖ 1 day 5 clusters result

❖ 1 day 4 clusters result

❖ 3 day 5 clusters result

❖ 3 day 4 clusters result

Ethereum

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	0.59	0.74	22
avg / total	1.00	0.59	0.74	22
	precision	recall	f1-score	support
0	0.11	1.00	0.20	1
1	1.00	0.62	0.76	21
avg / total	0.96	0.64	0.74	22
	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	0.73	0.84	22
avg / total	1.00	0.73	0.84	22
	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	0.73	0.84	22
avg / total	1.00	0.73	0.84	22

Logistic Regression Results:

- ❖ For Litecoin, model 1 (1 day 5 clusters) gives better results.

	precision	recall	f1-score	support
0	1.00	0.59	0.75	32
1	0.07	1.00	0.13	1
avg / total	0.97	0.61	0.73	33

	precision	recall	f1-score	support
0	1.00	0.59	0.75	32
1	0.07	1.00	0.13	1
avg / total	0.97	0.61	0.73	33

	precision	recall	f1-score	support
0	0.82	0.54	0.65	26
1	0.25	0.57	0.35	7
avg / total	0.70	0.55	0.59	33

	precision	recall	f1-score	support
0	0.94	0.53	0.68	30
1	0.12	0.67	0.21	3
avg / total	0.87	0.55	0.64	33

- ❖ 1 day 5 clusters result

- ❖ 1 day 4 clusters result

- ❖ 3 day 5 clusters result

- ❖ 3 day 4 clusters result

Drawbacks:

- ❖ Some tweets or news headlines are not classified correctly which makes sentiments imperfect.
- ❖ Clustering too is not really ideal and may need lots of more tweets and headlines to form more diverse groupings.
- ❖ Predicting something like cryptocurrency trading using simple modeling techniques is a really complex task.

Conclusion:

- ❖ The tests we ran actually yielded similar results with the five cluster model looking at the price in three days performing the best for Bitcoin and Ethereum.
- ❖ For Litecoin, the five cluster model looking at the price in 1 day performs the best.
- ❖ Proves that the price of cryptocurrencies is tied to the news coverage of the currency itself.

Thank you