Cryptocurrency Price Prediction Using News and Social Media Sentiment

Connor Lamon, Eric Nielsen, Eric Redondo
{conlamon, nielsene, eredondo}@stanford.edu

Background

As the economic and social impacts of cryptocurrencies continue to grow rapidly, so does the prevalence of related news articles and social media posts. Similar to traditional financial markets, it appears likely that there is a relationship between media sentiment and the prices of cryptocurrency coins.

Our goal was to explore whether sentiment analysis on news headlines, tweets, and Reddit posts can inform predictions on future price changes of the cryptocurrency, Bitcoin.

Method

Data

Time Period

- **Train**: 01/01/2017 to 02/06/2018; **Test**: 02/07/2018 to 03/06/2018

Media Data *(acquired / cleansed via web scraping and API use)*

- **~4,000 news headlines** *(scraped from coindesk [2])*
- **~100,000 tweets** *(~23,000 from prominent people in crypto community [4])*
- **~110,000 Reddit post titles** *(from prominent crypto subreddits [3])*

Coin Data *(acquired via web scraping and API use)*

- Hourly market data for Bitcoin (sample below)

Timestamp	Price	Volume (Bitcoin)	Volume (USD)
2017-01-01 02:00:00	973.37	184.7	179,446.94

Labeling

- In the final implementation, data points are labeled with the **percent change in coin price 1hr, 2hr, 6hr, 12hr, and 24hr in the future**; we make predictions for each of these time periods

Prediction Aggregation

- Final prediction is the percent change for each text piece; these values were then grouped by sign and magnitude based on time

Models

A separate model was constructed for each media source; all use the **Keras** [5] framework, with Tensorflow [6] backend, and the same general design. First, input text passes through an **embedding layer.** Second, the encoded text passes through a **bi-directional LSTM.** Finally, a **linear activation** is applied to perform the regression task in generating price change predictions. Models are trained using **mini-batches** (sizes proportional to input size).

Embedding Layer

- **Attempt 1: Pre-trained word2vec**
  - Used until it was realized that this did not include embeddings for many crypto-related words found in data (e.g., Bitcoin)
- **Attempt 2: Custom word2vec**
  - Trained a custom word2vec model using the skip-gram method and all the data available (headlines, tweets and Reddit posts)

Bi-Directional LSTM Layer

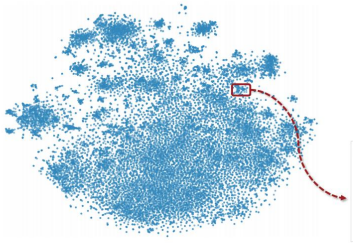
- Initially, two bi-LSTM layers connected by a dropout layer were used; however, this produced similar results to using a single bi-LSTM layer, so the final models use a single layer

Linear Activation Layer

- Since the goal was to perform a regression analysis and output a numerical percent change, a linear output layer was used with MSE loss function

Results

T-SNE Visualization of Custom Crypto Word Vectors



Query Top N Similar Words ('mining'):	
Similar Word	Cosine Distance
rig	0.648
pool	0.635
mine	0.592
gpu	0.582

Interesting Analogies:

'bitcoin' -> 'eth' as 'ethereum' -> ? :: **'btc'**

'litecoin' -> 'eth' as 'ethereum' -> ? :: **'ltc'**

Closest vector to ICO?

Result: Tokensale

Cos Distance = 0.663

Aggregated Price Change Predictions

News Headlines

Pred. Time	Accuracy Predicting Price Change Direction: Overall	For Top 20% Price Increases	For Top 20% Price Decreases
+1 hr	49.25%	42.65%	51.47%
+2 hr	48.36%	42.03%	45.71%
+6 hr	53.13%	46.48%	48.48%
+12 hr	**54.03%**	57.97%	44.93%
+24 hr	54.03%	60.00%	46.48%

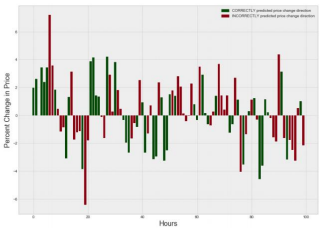
Tweets

Pred. Time	Accuracy Predicting Price Change Direction: Overall	For Top 20% Price Increases	For Top 20% Price Decreases
+1 hr	49.62%	48.28%	48.00%
+2 hr	51.88%	56.92%	54.60%
+6 hr	54.44%	74.59%	41.14%
+12 hr	**57.74%**	81.38%	35.85%
+24 hr	54.89%	85.15%	25.00%

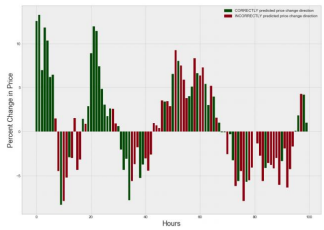
Reddit

Pred. Time	Accuracy Predicting Price Change Direction: Overall	For Top 20% Price Increases	For Top 20% Price Decreases
+1 hr	48.81%	47.97%	51.11%
+2 hr	50.45%	54.76%	47.01%
+6 hr	50.60%	55.74%	50.00%
+12 hr	**54.48%**	65.29%	46.32%
+24 hr	50.15%	62.50%	36.76%

Reddit +2hr Predictions

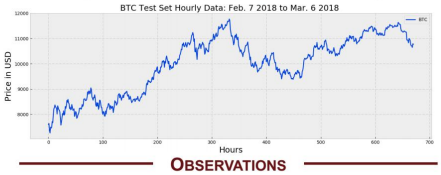


Reddit +12hr Predictions



Discussion

BTC Price during Test Set Time Period



Observations

The model appears to perform best when predicting price increases over longer time periods:

- There is a general increase in Bitcoin price during the test set period; the model appears to be biased towards predicting increases, which could be due to the training set
- The model performs best overall attempting to predict +12 hour price changes using tweets from prominent members of the crypto community

Conclusion

Though the model is still being improved, our two main project objectives were completed.

- ✓ Develop a model that makes cryptocurrency price predictions using non-technical data
- ✓ Consistently achieve greater than 50% prediction accuracy

Future Work

We plan to conduct additional experiments and make updates to the model in the future. Current priorities include:

- Training the cryptocurrency focused word2vec model on more data (millions of data points)
- Trying a doc2vec based model to attempt to understand the relationship between full headlines / tweets / Reddit posts and price changes
- Integrating additional types of media (e.g., from other news sources, Slack channels) and larger volumes of input
- Investigating different strategies for labeling training data (e.g., using a NN to label based on text sentiment)

References

[1] Bitcoincharts; https://bitcoincharts.com/charts/
[2] Coindesk; https://www.coindesk.com/
[3] Reddit API; https://www.reddit.com/dev/api/
[4] Twitter API; https://developer.twitter.com/en/docs/tweets/search/overview
[5] Keras; Chollet, Francois and others, https://github.com/keras-team/keras
[6] Tensorflow; https://www.tensorflow.org/