Influence of Social Media on Cryptocurrencies

by Nick Sawicki, Kyran Flynn,
Iris Huang, Jacqueline Jia

Introduction

- Overarching goal: analyze trends between social media and cryptocurrency prices
- Cryptocurrencies are famously volatile as their prices are determined by a relatively small market and without any government backing. Thus, the attitude of the market is likely inferable through social media.
- We narrowed our search for a correlation to reddit data, a common platform for in depth cryptocurrency discussions
    - 3 established cryptocurrencies
    - 3 ‘meme’ cryptocurrencies

Data

Reddit Data

- Collection: Scraped from r/CryptoMoonShots via PRAW
- Cleaning: organized into categories, removed duplicates and incompletes
  - Established coin posts
  - Meme coin posts
  - Established coin comments
  - Meme coin comments

Cryptocurrency Data

- Collection: historical hourly cryptocurrency data obtained via CryptoCompare API
- Cleaning: organized into categories
  - **Established coin data** (Bitcoin, Ethereum, Solana)
  - **Meme coin data** (Dogecoin, Shiba Inu, Sushi)

Hypotheses: Meme vs. Established Coins

1. **How does price volatility compare?**
   a. The mean daily price percentage change for meme coins is significantly greater than that of the established coins (p=5.4331e-12).

2. **Is one discussed more frequently in r/CryptoMoonShots?**
   a. Established cryptocurrencies have a higher Reddit post (p=0.0053) and comment frequency (p=0.00013) within a 24 hour time frame compared to the post and comment frequencies for meme coins.

3. **Is one more associated with traditional financial terminology (such as the federal reserve, interest rates, and GDP)?**
   a. Established coins are more associated with traditional financial terminology as compared to meme coins (p=0.00376).

ML: The Concept

→ Predicting the direction of a coin's fluctuation from reddit posts/comments made within three hours of the fluctuation.

- Prediction made using a deep learning model to process the natural language in posts/comments.
- Natural language cleaning and encoding for numerical representations of posts/comments.

Ultimately: check whether a correlation between input and output is learnable by our model assuming a correlation exists.

ML: Reddit Data Pre-Processing

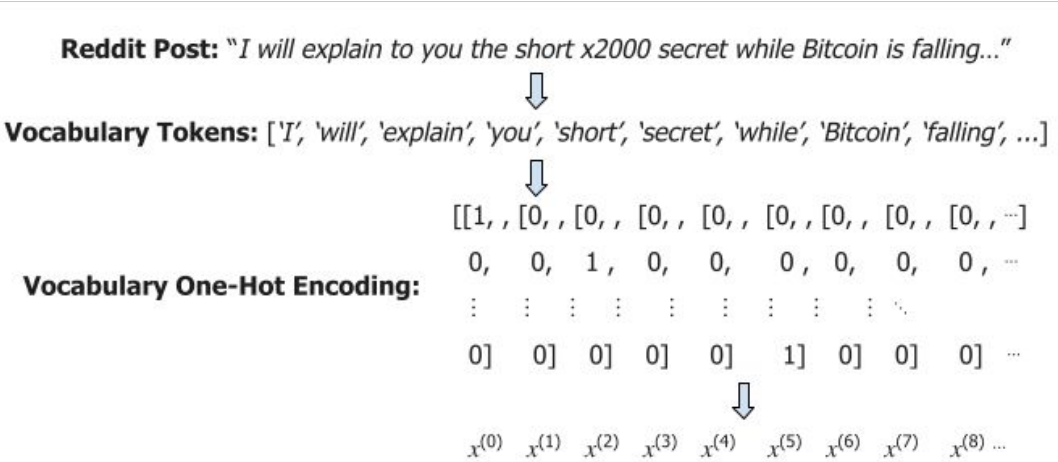
Cleaning Process:

- Tokenization
- common word removal
  - The, I, is , are, etc…
- Emoji separation
  - Separated with a space
- Repeats removed

Then, vocabulary creation for encoding.

→ Minimum threshold frequency in the data to be considered.

Then:



ML: Coin Data Pre-Processing

- For each coin, found the direction of the fluctuation.
- Direction label 1 is up, label 0 is down.
- Here is an example:

Coin: [Open = 15, Close = 20,...]

⇩

fluctuation = $\frac{20 - 15}{20} = 0.25 > 0$

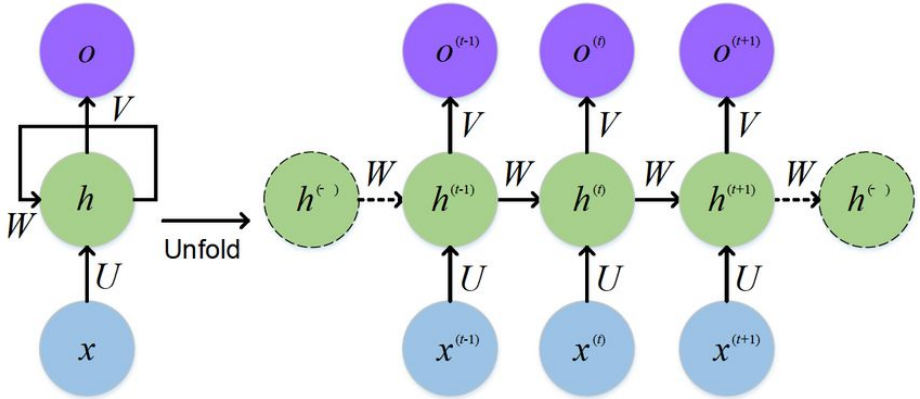
⇩

Direction: 1

- This becomes the prediction label for all reddit posts/comments that mention a coin made 3h before a coin's fluctuation.

ML: RNN Model

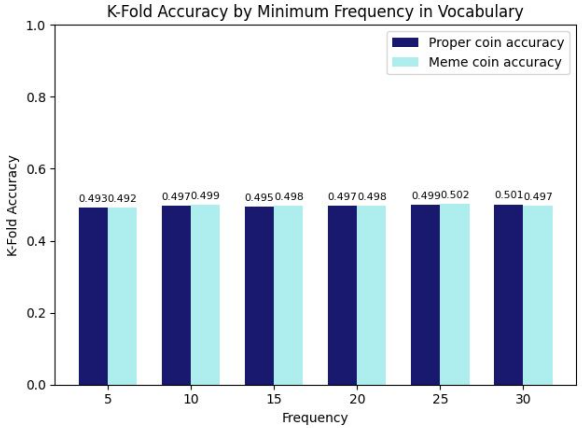
- Using a recurrent neural network, we can pass as inputs sequences of words.
- Each one-hot encoded word of a post/comment is an input $x^{(t)}$.
- Added sigmoid function after each forward pass through W for numerical stability.

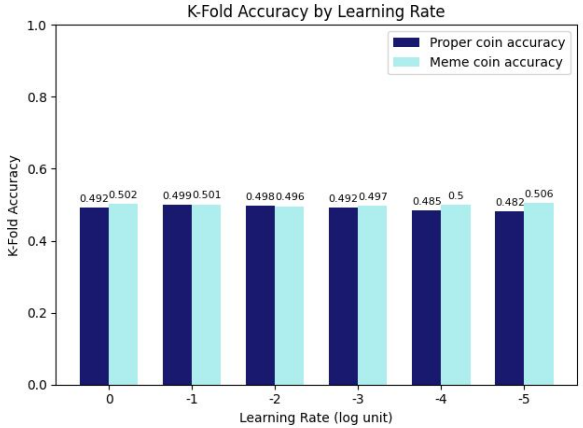


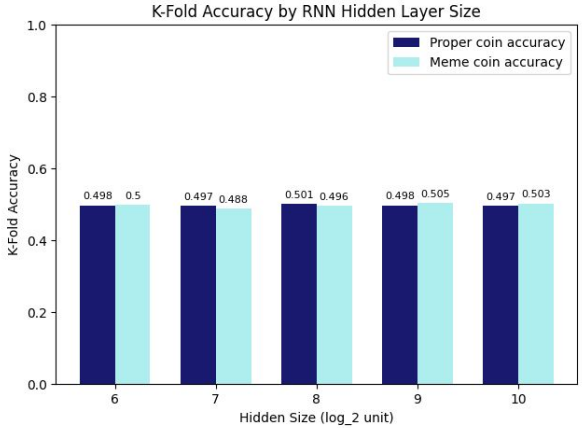
- Loss defined through the prediction of the label of the direction of the coin's fluctuation.

ML: Results

- Redesign included hyperparameter testing for three parameters:







Note: the timeframe parameter and the number of splits in the K-Fold cross validation remained constant.

ML: Implications of Results

Since the model remained as good as random, it implies one of the following three possibilities:

1. There is not enough data for a correlation to be found.
2. The model is not appropriate for the correlation we looked to find in our data.
3. There is no correlation between reddit posts/comments made three hours before and a coin's fluctuation and the direction of that fluctuation in our data.

Note: *in our data*, not in general.

Limitations & Future Research

Limitations:

- Our data is limited to the past year of data for both reddit and coins.
- Our models use only one reddit page, and only 6 cryptocurrencies.
- The RNN's layers are neural networks of relatively low complexity.

Potential Future Work:

- Acquiring more data to confirm hypotheses and better the ML training, including more social media platforms and more coins.
- Improve natural language processing through:
  - Premade word embeddings for encoding.
  - More advanced recurrent models such as LSTMs or GRUs instead of an RNN.

Acknowledgements

- The reddit dataset was sourced through PRAW.
- The cryptocurrency data was obtained through the CryptoCompare API.
- The RNN architecture was inspired by a PyTorch tutorial on character level classification of names.
- The RNN figure was made by Weijiang Feng from ResearchGate.

Thanks for Listening!