CS1951A Data Science Final Project:

Influence of Social Media on Cryptocurrency Fluctuations

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

Introduction

- Our overarching goal was to 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, and six cryptocurrencies, three of which are relatively established and three of which are considered 'meme' coins but are nonetheless traded.

Datasets

Data Collection:

Reddit Dataset:
- Used Python Reddit API Wrapper (PRAW) to scrape the r/CryptoMoonShots subreddit (chosen for its popularity and use).
- Collected posts and comments from the past year that mentioned our six coins of interest.

Cryptocurrency Dataset:
- Used the CryptoCompare API to get the hourly historical data of coins for the past year.
- Collected data for 6 cryptocurrencies: 3 more established coins (Bitcoin, Ethereum, Solana) and 3 meme coins (Dogecoin, Sushi, Shiba Inu).

Data Cleaning:

Reddit Dataset:
- Organized the data into four dataframes (meme coin comments, meme coin posts, established coin comments, established coin posts), removing duplicates and incompletes.

Cryptocurrency Dataset:
- Merged the data into two dataframes for meme coin data and established coin data.

Resulting Data:

Reddit Comments:
- Attributes: **ID**, associated **post ID**, **created UTC** (time posted), **body** (text), **score** (# of upvotes), which **coin** the comment mentions.

Reddit Posts:
- Attributes: **ID**, **created UTC** (time posted), **title**, **selftext** (text), **# of comments**, **score** (# of upvotes), **upvote ratio** (# of upvotes/total # of votes), which **coin** the post mentions.

Coin Data:
- Attributes for all coins are: time, **high** (highest USD price during the hour), **low** (lowest USD price during the hour), **open** (USD price at the hour), **volume from** (total amount of the coin traded into USD during the hour following time), **volume to** (total amount of USD traded into coin during hour following time), **close** (USD price at the end of the hour following time).

Hypotheses

General Methodology:
- We used a two sample t-test with a 99% significance level to test our hypotheses and performed the appropriate data extraction for each hypothesis.

Hypothesis 1: The mean daily price percentage change (daily PPC) for meme coins is significantly greater than that of the established coins.
- Goal: Determine whether meme coins experience greater daily price fluctuations than established coins and are thus more volatile.
- Data extraction and processing:
  - For each coin, the daily PPC was calculated using the equation: (maximum daily price - minimum daily price) / minimum daily price*100.
  - We then combined the percentages for the meme coins and established coins respectively to form two independent samples.
- Result:
  - Meme cryptocurrencies have a significantly higher mean daily price percentage change (p=5.4331e-12).
  - Figure 3 shows a side-by-side comparison of the distribution of daily price percentage changes for meme coins and established (proper) coins.

Data Schema

Reddit Comment Data

ID	Post ID	Created UTC	Body	Score (# upvotes)	Coin

Reddit Post Data

ID	Created UTC	Title	Selftext	# comments	Score (# upvotes)	Upvote ratio	Coin

Coin Data (For Each Coin)

Time (hr)	High	Low	Open	Volume from	Volume to	Close

Hypothesis Results

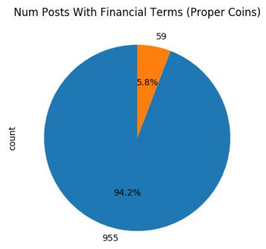

Fig. 1

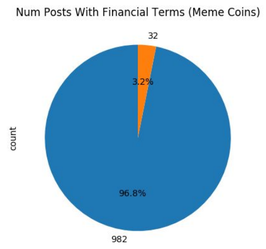

Fig. 2

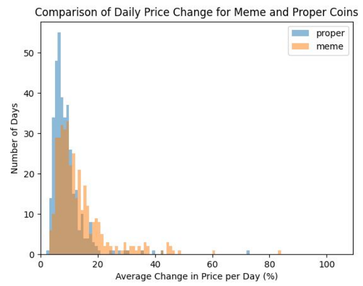

Fig. 3

Machine Learning Model

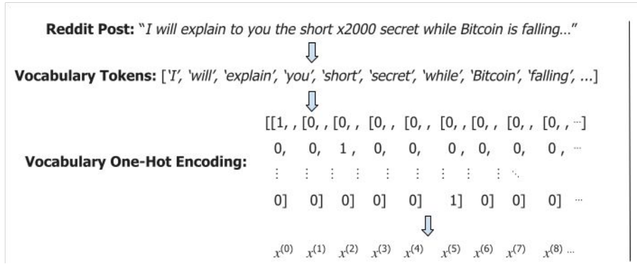

Fig. 4

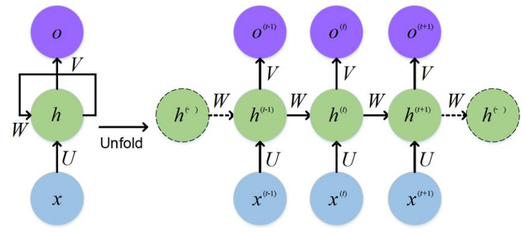

Fig. 5

Machine Learning Results

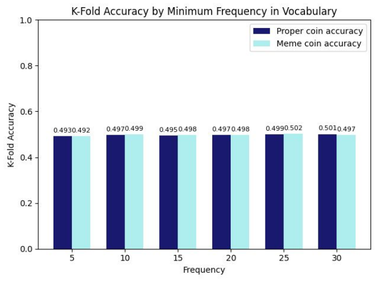

Fig. 6

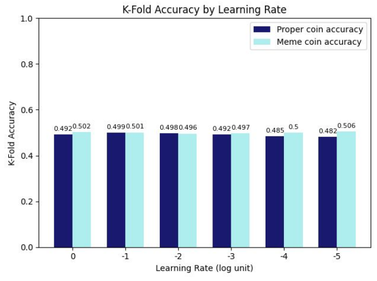

Fig. 7

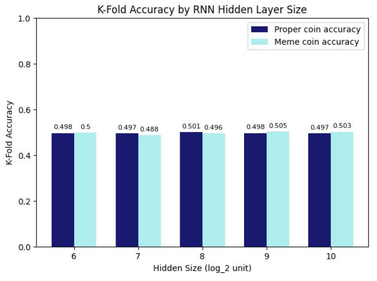

Fig. 8

Hypotheses

Hypothesis 2: Established cryptocurrencies have a higher Reddit post and comment frequency within a 24 hour time frame compared to the post and comment frequencies for meme coins.
- Data extraction and processing:
  - We pooled together the Reddit data (posts and comments separately) for the meme coins and for the established coins respectively to obtain 2 post datasets and 2 comment datasets.
- Results:
  - For the pair of post datasets, there is a statistically significant difference between the established post frequency and the meme post frequency (p=0.0053). The same test was conducted for comments (p=0.00013).
  - Thus, the results show that the established coins are discussed more frequently in r/CryptoMoonShots.

Hypothesis 3: Established coins are more associated with traditional financial terminology as compared to meme coins.
- Data extraction and processing:
  - A list of financial terms such as the federal reserve, interest rates, and gdp which are associated with traditional financial institutions were cross-referenced with all of the posts for the meme and established coins to determine if any of the posts contained any of the financial terms.
- Results:
  - Traditional financial terms are more prevalent in established coin reddit posts than in meme coin posts (p=0.00376)
  - Figures 1 and 2 the breakdown of posts with and without financial terms for established (proper) and meme coins.

Machine Learning

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

Data Cleaning:
- Hourly coin fluctuations were calculated as prediction labels, and with the direction of the fluctuation labeled '0' for down and '1' for up.
- The reddit data was tokenized, and common words were removed. The emojis were kept but separated, and repeated words removed.
- A vocabulary was created for the established and meme coins based on a threshold minimum frequency in the data. The data was then one-hotted by index into this vocabulary, as seen in figure 4.

Model:
- Used a Recurrent Neural Network (RNN) to pass as input the sequences of words in each post/comment, labelled $x^{(t)}$ in figure 4 & 5.
- Sigmoid function between recurrent layers for numerical stability.
- Tested the accuracy of three parameters (vocabulary threshold frequency, model learning rate, hidden layer size W) in figures 6 to 8.

Results: The model performed as good as random for a binary choice on all parameter settings, implying one or more of the following.
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.
3. There is no correlation between reddit posts/comments made three hours before and a coin's fluctuation in our data.

Limitations

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

Future Research

- 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 and more advanced recurrent models such as LSTMs or GRUs instead of an RNN.

Acknowledgments

The reddit dataset was sourced through PRAW, the cryptocurrency data was obtained through the CryptoCompare API, and the RNN architecture was inspired by a PyTorch tutorial on character level classification of names. Figure 5 is by Weijiang Feng from ResearchGate.