Influence of Social Media on Cryptocurrency Fluctuations

The Flintstones: jjia6, ihuang8, nsawick1, kflynn6

Introduction

With the rise in the popularity and media discussion of cryptocurrencies, we thought it would be interesting to dig deeper into the relationship between coin fluctuations and social media content about coins. Our work focuses on finding distinctions in the markets of more established cryptocurrencies and the meme coins, both in the consumer's behavior and their financial measures. Furthermore, we sought to find a correlation between social media and price fluctuations through a prediction algorithm.

Data

For our analysis, we decided to focus on Reddit data, 3 established coins (Bitcoin, Ethereum, Solana), and 3 meme coins (Dogecoin, Sushi, and Shiba Inu). For each coin, we used the CryptoCompare API to get its historical data hour to hour for the past year. We also used the Python Reddit API Wrapper to scrape the r/CryptoMoonShots subreddit for posts and comments that mention each of the 6 coins.

Hypotheses

Hypothesis 1: Meme coins have a significantly higher mean daily price percentage change than established coins.

Support for Hypothesis 1: We calculated and pooled together the daily price percentage change for the meme coins and the established coins. With these two independent groups, we used a two sample t-test at the 99% significance level and found that the meme coins have a significantly higher mean daily price percentage change (p=5.4331e-12). Figure 1 shows side-by-side histograms of the distribution of mean daily price percentage changes for both categories of coins.

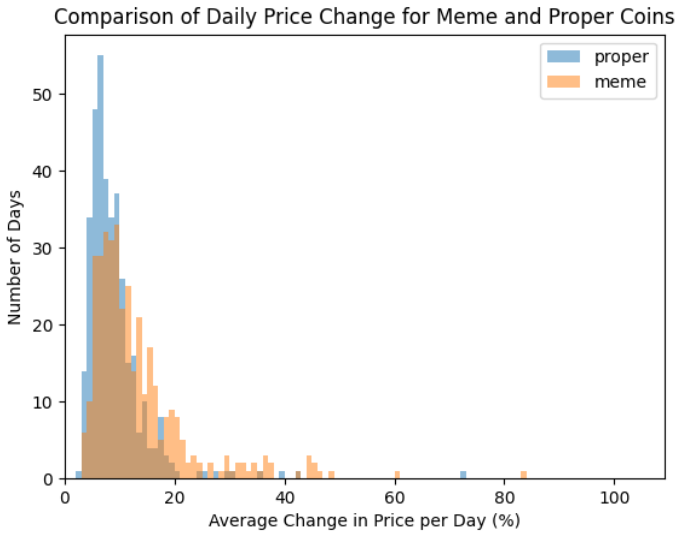


Figure 1

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.

Support for Hypothesis 2: We pooled together the posts for the meme coins and for the established coins respectively to obtain 2 post datasets and 2 comment datasets. For the pair of post datasets, we used a two sample t-test to determine that at the 99% significance level, there is a statistically significant difference between the proper post frequency and the meme post frequency ($p=0.0053$). The same test was conducted for comments ($p=0.00013$). The proper coins had higher post and comment frequencies, meaning they had greater amounts of discussion surrounding them on r/CryptoMoonshots.

Hypothesis 3: Established coins are more associated with traditional financial terminology as compared to meme coins.

Support for Hypothesis 3: Data to validate this claim was scraped from the r/CryptoMoonshots subreddit and subsequently cleaned such that all posts having to deal with meme coins – Dogecoin, Shiba Inu, and Sushi in this case – were separated from posts associated with established coins – Bitcoin, Ethereum, and Solana. 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 in both data sets to determine if any of the posts contained any of the financial terms. Using a two-sample t-test at the 99% significance level, it was discovered that traditional financial terms are more prevalent in established coin reddit posts than in meme coin posts ($p=0.00376$). This aligns with our original hypothesis that established coins such as Bitcoin and Ethereum were more entrenched in traditional financial institutions relative to meme coins like Dogecoin.

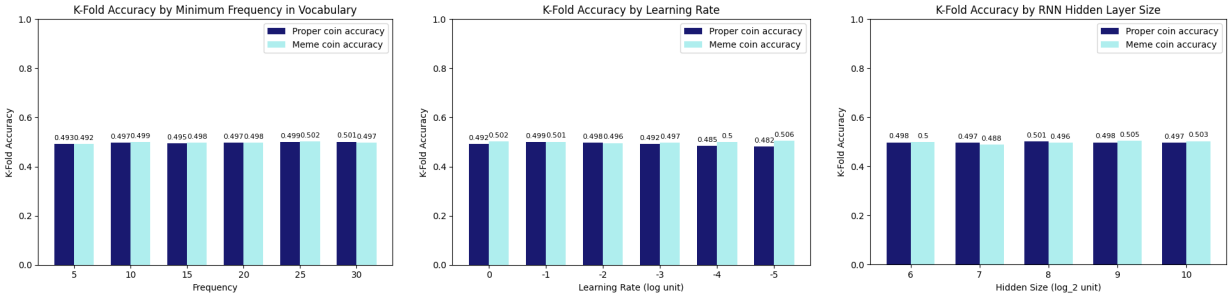
Machine Learning:

Testing whether a correlation exists between a cryptocurrency's fluctuation and social media posts about that cryptocurrency made within three hours of that fluctuation using a prediction algorithm.

Method: We pre-processed the reddit data into a numerical form by finding a vocabulary of words that have an above threshold frequency in our data. After cleaning and tokenizing, we then one-hotted each word by its index in that vocabulary. We constructed a binary prediction task by labeling this data with the direction of the cryptocurrency's subsequent fluctuation. We then utilized a Recurrent Neural Network (RNN) model, a supervised learning algorithm, to input a sequence of words of variable length into the model. Once an entire post/comment has been processed, a prediction is made on the direction of the fluctuation.

Model Design: Our RNN uses a sigmoid activation function between each word input for numerical stability (large posts cause output blowups). For our final ML redesign, we evaluated our model for a range of values for three hyperparameters: the vocabulary threshold frequency, the learning rate of our model, and the RNN hidden layer size (linear layer matrix dimensions).

Results: We used K-Fold cross validation as our performance metric. The results were as follows.



As observed above, our model remained approximately as good as random on a binary prediction task across all hyperparameter values. This can have a multitude of reasons:

1. The RNN model was not appropriate for the learning task.
2. There is no underlying correlation between the reddit data three hours prior to a coin fluctuation and the coin fluctuation's direction.
3. The data was insufficient for our model to find a correlation between reddit posts/comments and coin fluctuations.

These reasons remain within the realm of the data we collected, and do not necessarily generalize to trends between social media and cryptocurrencies.