Instruction and flow of the project:

Unzip the file Project/Paper_code_Files and follow the steps below:

The Following code files were used in the same other, in order to replicate the work, the following files with the necessary resources and data in the same order are needed to be followed.

1. reddit_scraper

In order to get the Reddit comments, we make use of the PRAW API and get the Redditt data, the python notebook reddit_scraper get the data for us.

2. clean_reddit_comments

In this section, we use R in order to clean the comment data. I also use it to filter out bots using a list that I created, which was available here. I also filter out non-English subreddits from a list that I created, which was available here.

3. modify_vader_lexicon

In order to derive sentiment from the comment data, I use VADER (valence Aware Dictionary and sentiment Reasoner) which is a "lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media" that is available in Python). The GitHub repository for VADER is located here. In this section, we modify VADER's lexicon using R by adding domain specific sentiment terms that are relevant to my analysis. we used the Loughran-McDonald financial sentiment corpus which is available here

4. sentiment_analysis

In this section, we use VADER and its modified lexicon (which we did in section 3) in order to classify the sentiment of the comment data, in conjunction with parallel processing.

5. format_data_for_analysis

In this section, I group by date all the sentiment data that I have collected and merged it with bitcoin price data, which I downloaded from yahoo finance which you can find here.

6. stationary_transformation

Here, I transform my time series data so that it is stationary. We included a guideline as well as examples of how to do so. we made functions that asses the stationarity of the data as well as a resource telling you how to interpret the results and what you should do in response to those results.

7. var_model

The model that I chose to forecast my data was vector autoregression, which are used for multivariate time series. I compared the results from my fitted model to test data as well as against a persistent model.