VADER Sentiment Analysis on Utility and Speculative Cryptocurrencies

Srinivasan Sathiamurthy and Rebecca Derham

December 12, 2022

1 Introduction

In our second Coffee Break Experiment, we investigated the relationship between prices and sentiment data gathered from Twitter for two major cryptocurrencies, Bitcoin and Ethereum. We found that the sentiment data did have a modicum of predictive power over prices, and were able to obtain a reasonably well-fitting predictive model for each coin. However, our overall results were relatively weak, which we determined to be likely due to the interference of external factors such as overall market conditions.

In this paper, we extend our previous work by conducting a more thorough analysis of the relationship between Twitter sentiment and cryptocurrency prices. This time, we decided to investigate this relationship for six different cryptocurrencies: Uni Token (UNI), Sol (SOL), Aave (AAVE), Dogecoin (DOGE), Shiba Inu (SHIB), and Apecoin (APE) . The first three coins are so-called "utility coins," in that they arguably derive some amount of fundamental value from usability in their respective underlying ecosystems, while we believe that the latter three are purely speculative and do not have any inherent value. We were interested in including both types of coins in our dataset as we speculated that speculative coins may be more affected by sentiment scores than the utility coins, as investor speculations, which we can likely capture using Twitter, are the primary cause for changes in the prices of speculative coins.

We again have two sets of data: sentiment scores obtained from Twitter, and daily prices. However, this time we built a more comprehensive and accurate set of sentiment data by obtaining a number of tweets each day for each of the cryptocurrencies under consideration, and then using the Valence Aware Dictionary for Sentiment Reasoning (VADER)'s rule-based model to obtain sentiment scores. For prices, we used the difference between the daily log returns of each token and the log returns of Bitcoin over the same time period.

We endeavored to complete three tasks in this project. First, and perhaps most importantly, we determined the existence and strength of any causal relationships between the coins' sentiment features and their market-normalized returns. From those sentiment features with casual relationships with prices, we determined which set of sentiment features best predict market-normalized returns for each coin. Finally, we generated predictive models for each coin and compared the predictability of these models across our two groups of coins.

2 Data

We chose UNI, SOL, and AAVE as our utility coins as each has a use in some environment which arguably is important to crypto as a whole. UNI is the governance token for Uniswap, the largest and most-used decentralized cryptocurrency exchange platform. SOL is used to pay the transaction fees for trading on the second largest smart blockchain (after Ethereum). AAVE is the native token for the reputable AAVE protocol, which allows users to borrow and lend cryptocurrencies safely.

We chose DOGE, SHIB, and APECOIN as our purely speculative coins, as none of them are linked to any such protocols – therefore their prices are not tied to anything relevant to the crypto market as a whole. By contrast, they are usually only talked about in the context of internet meme culture . However, it is worth noting that Apecoin is the underlying token of the BoredApe ecosystem, which is a purely entertainment-driven protocol. BoredApe specializes in selling Bored-Ape NFTs, and building an artificial ecosystem around the public's excitement for related products. However, this ecosystem is relatively self-contained and has no financial implications to the market as a whole, and so this coin's "worth" is artificial and its fundamental value is likely nonexistent.

Our dataset consists of two distinct parts: a set of daily tweets about each coin, as well as price data for each coin.

2.1 Twitter Data

We used SNScrape's Twitter API to search "uni crypto", "aave crypto", "sol crypto", "dogecoin", "shib crypto" and 'apecoin", from November 1st, 2021 through October 31st, 2022. Note that we added "crypto" to some of the search terms as, for example, searching only "uni" or "sol" would clearly yield results unrelated to the cryptocurrencies. We started searching from 6 AM EST on each day since most news outlets publish media around that time, and the bulk of daily trading volume follows between 6 am and noon EST. We wished to obtain at least 1000 tweets per day, as we did for our Coffee Break Experiment. However, the coins under investigation here are significantly smaller and less well-known than the ones we investigated previously, and so for several of our coins, there were less than 1000 search results per day. As a result, it was necessary to loosen this data collection process to obtaining up to 1000 tweets per day. However, each coin still had at least 100 tweets per day with most close to 1000, so we believe this amount of data is still sufficient.

Additionally, we decided to shift from using R's quanteda package and the Loughran and McDonald financial sentiment dictionary to simply using Python's VADER package. After some exploratory data analysis, we discovered that the financial sentiment dictionary had only 3700 words associated with relevant sentiment features, and moreover this dictionary does not include terms such as "gain" which are common in the crypto social media. By contrast, VADER is specifically designed to capture sentiment in social media while the financial dictionary is trained on formal 10-K statements which, while specific to finance, will likely not be as successful at analyzing Twitter data. Additionally, quanteda is too slow to run sentiment analysis on each tweet individually, which eliminates the ability to obtain information regarding variance in sentiment on each day.

2.2 Price Data

We believe that the results from our previous work were likely made less significant due to the impact of external factors such as the general conditions of the cryptocurrency market. In an attempt to remedy this, we used a comparison between each coin and Bitcoin (BTC) prices instead of each coin's raw prices. As there are no highly liquid or popular ETFs for cryptocurrencies which could be used as an indication of current market conditions, we had no choice but to choose a coin ourselves. BTC is widely considered to be a good indicator of overall market conditions in the crypto market, as it is the largest and most liquid cryptocurrency by volume. Moreover, it is linked to a stable and reputable protocol which has not had any ground breaking changes in the past year and hence has not had fluctuations purely due to its development over the past year.

We obtained price data for the six currencies under investigation, as well as for BTC, from CoinMarketCap, using end of day closing prices. Since crypto markets support a decentralized network of computers which run 24/7, markets are deemed to close at 00:00 UTC. This observation permits the convenient use of sentiments from tweets gathered earlier in the day to predict end-of-day prices.

After obtaining raw price data, we calculated market-normalized returns by subtracting the log returns of Bitcoin from the log returns of the coins' prices. We did so for three main reasons. First, log returns are symmetric while simple returns are not; logarithmic returns of equal magnitude but opposite signs cancel each other out and so this property helps reduce directional bias in our models. Secondly, subtracting Bitcoin's log returns effectively considers only the changes in price that are independent of overall market conditions. Notably, this does not account for the possibility of the target token being a scalar multiple of BTC (in which case our resulting time series would still be correlated with market returns), but fortunately we did not run into that scenario for these six coins. Finally, this process rendered our price data to be stationary, which is a property required by the statistical methods we use later in the paper.

Figure 1 displays the raw closing prices over the full year for 5 coins of interest plus BTC. APE is not included as there was a large spike which would render the graph unreadable, but in general follows a similar downwards trend. In addition, the prices are not to scale but have been divided by some constant in order to fit in one graph.

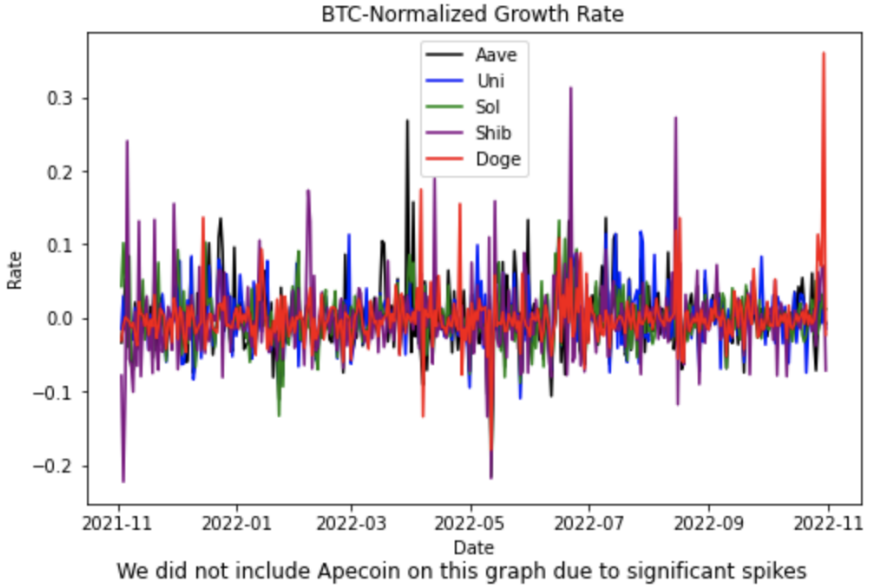
Figure 1: Closing Prices

3 Methods

To obtain our sentiment features, we ran VADER on each tweet and obtained the set of scores "%positive", "%neutral", "%negative", and "compound". The first three scores are percentages that add up to 100% and report which percent of each tweet's tokens are deemed to be positive, negative, or neutral. The "compound" feature is a normalized, weighted function of the first three scores which determines how positive or negative a tweet is overall. Since a tweet is generally considered to be positive if its compound score is at least 0.05, negative if its compound score is at most -0.05, and neutral otherwise, we created another feature called "Compound Indicator", which was 1 if the tweet is scored as positive, 0 if neutral, and -1 if negative. Using these five scores (positive, negative, neutral, compound, and compound indicator), we obtained a total of ten features by calculating the daily mean and variance of each score.

Next, we first needed to test for stationarity of both the prices and the sentiment scores, as stationarity is required to use the tools needed for our analysis — stationarity requires that at any time, the joint distribution of future data is the same as the joint distribution of past data. To do so, we used the Augmented Dickey Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. ADF tests for whether a series contains a unit root, which points towards a general drift in the assumed random-walk process of generating the data, which is a feature of a non-stationary time seires. Therefore, each series must reject this test. Next, the KPSS test is used to determine whether a series is trend-stationary. This tells us whether the series is mean-reverting around some linear function. As this is a feature of stationary time series, each series must fail to reject the null hypothesis. Fortunately, our market-normalized log returns satisfy stationarity for both of these tests already. The below graph illustrates this visually, since each series is centered at 0 and seems to have constant variance (which are conditions for weak stationarity):

Figure 2: Stationarity



We found the sentiment features to reject ADF (as desired), but also reject KPSS (undesired), which suggested that the series is a type of nonstationary process called difference-stationary. This was corrected by computing first differences for each sentiment feature, which then proceeded to reject the ADF and fail to reject the KPSS tests, as desired.

After obtaining stationary data, we applied Granger-Causality tests to each coin to determine whether there is a significant relationship between the coin's sentiment features and its market-normalized log returns. This test also determines if these relationships are unidirectional, with only one feature influencing another, or whether there is a bidirectional feedback relationship.

Finally, we turned to the main part of our analysis. We split our data into a training set, which consists of the first 11 months of data, and a testing set, which is the last month. We then selected a lag to be used in a vector autoregressive model (VAR) for each coin by using AIC. We then compared the log likelihood and the RMSE errors of our models across the two groups of coins to determine how much of a factor sentiment plays in predicting the rate of returns for each type of coin.

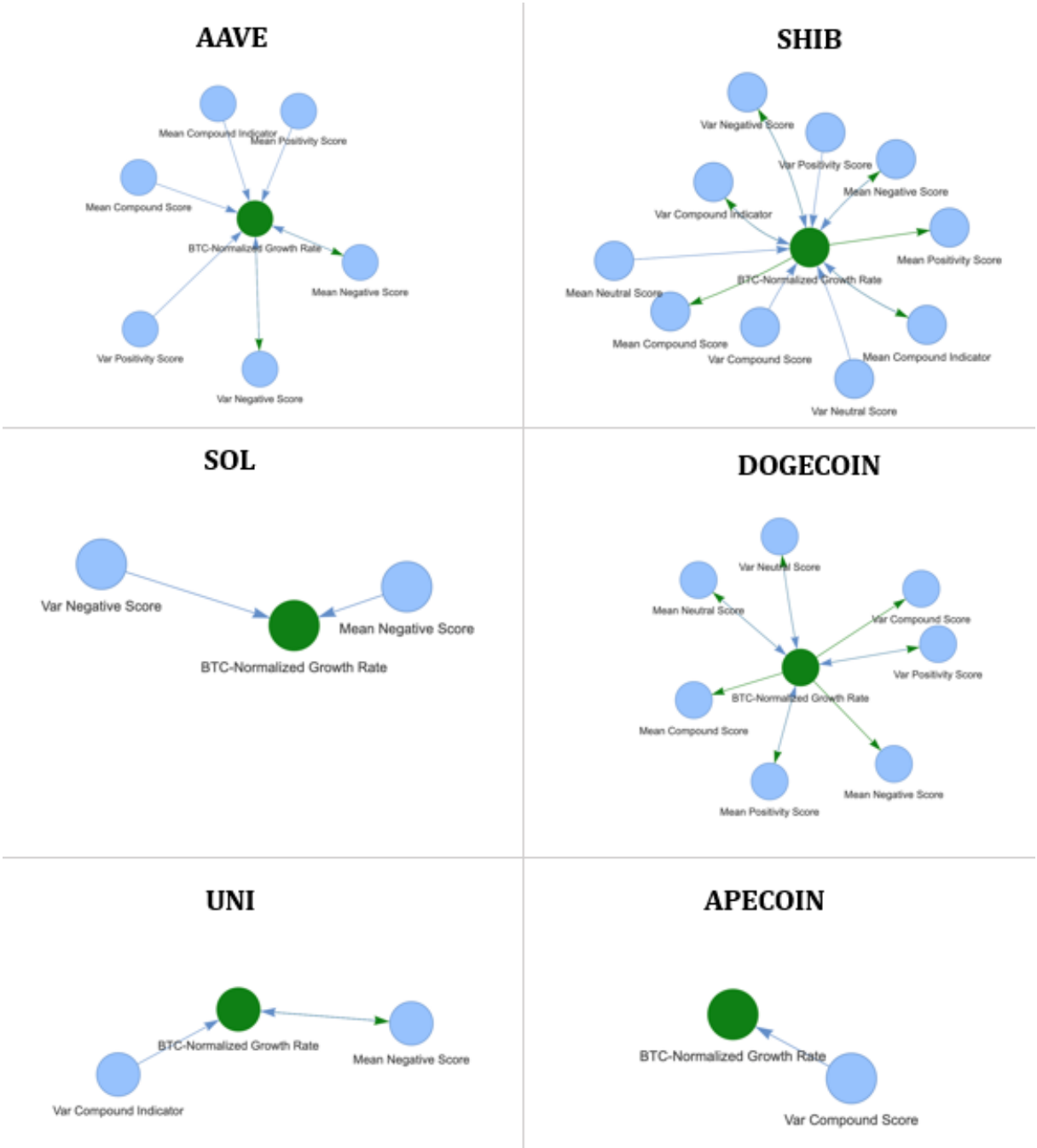
4 Results

Applying the Granger Causality test for the ten features and market normalized returns of each coin, we obtained a matrix of p-values indicating whether there is a causal relationship between each directed pair of variables. We discarded the pairs of variables which did not include the market-normalized returns, as we are only interested in predicting returns from sentiments, not sentiments from other sentiments. The results of parsing through these pairs to find those with a significant causal relationship are shown in Figure 3.

We observe that among our six coins, AAVE, SHIB, and DOGE not only have the most causal relationships between the sentiment features, but they also have the most bidirectional ones. This could suggest that the value of those coins stemmed more from sentiment than underlying protocol changes, or that there were simply more fluctuations of sentiment over time which contributed to investors' opinions of those coins. On the other hand, APE had only one causal relationship, which was with the variance in compound score feature. This is a bit odd since we believed APE to be primarily sentiment driven. However, we did also observe that APE had significantly higher fluctuations in market-normalized log returns than the other five coins. This could suggest a number of explanations – it is possible that price changes of APE were driven by shocks to its underlying protocol (the aforementioned BoredAPE ecosystem). On the other hand, APE may have been more sensitive to underlying market changes than other coins since its ecosystem is not tangibly connected to financial markets. As such, we are inclined to view APE as part of the first group of utility coins for the remainder of our analysis.

As a VAR model depends on a "lag parameter", or what maximum lag to use between when a sentiment is detected and when it influences the price, it was necessary to determine the optimal lag parameter before fitting a final model for each coin. We did so by minimizing the AIC scores of the model across possible lags up to 12 days. The computed optimal lags are displayed in Figure 4. All of the optimal lags are roughly within the span of a week, with 5 of the six coins ranging between 5-7 days, and only SOL having a optimal lag of 9 days.

Figure 3: **Sentiment Features with Causual Relationships**



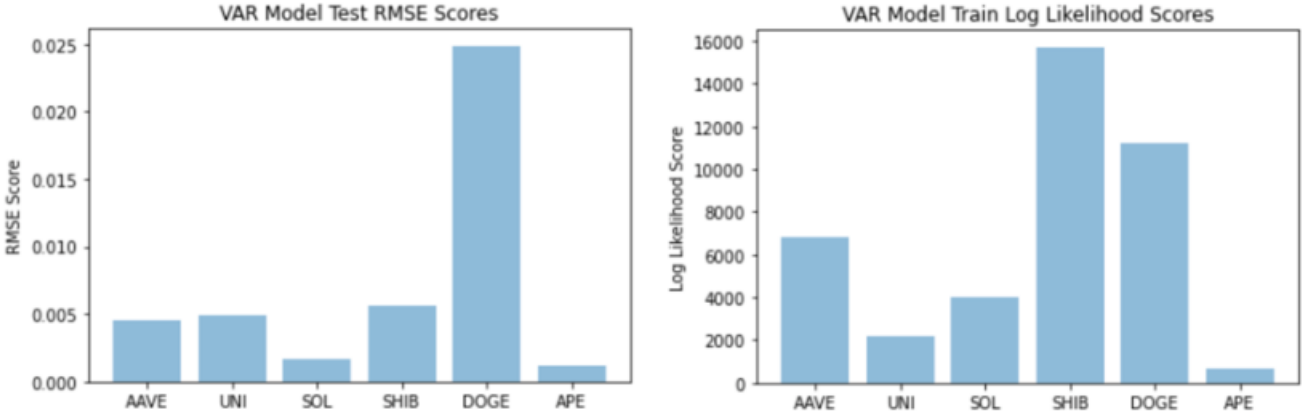
This suggests that it takes approximately a week for sentiment features to be reflected in prices. This makes sense since several news sources publish weekly summaries of the crypto markets, and many less-serious crypto investors trade upon this information.

Figure 4: **Optimal Lags (Days)**

AAVE	UNI	SOL	SHIB	DOGE	APE
5	7	9	6	5	5

After training VAR models on the first 11 months of data using these optimal lag parameters, the resulting log likelihood values for DOGE and SHIB (which we deemed purely speculative) were much higher than the other coins. This indicates that our models did the best job of predicting price from sentiment on the coins for which we hypothesized sentiment would influence price to a larger extent. Moreover, APE's log likelihood for its VAR model was much lower than all of the coins. This is in line with our observation that APE's market-normalized log returns had more significant shocks throughout the year than the other five coins as these shocks are significantly harder to model using VAR. As for the other coins, AAVE, UNI, and SOL have small RMSE values but also small log-likelihoods which suggests a decent amount of predictive power although not as strong as for the speculative coins.

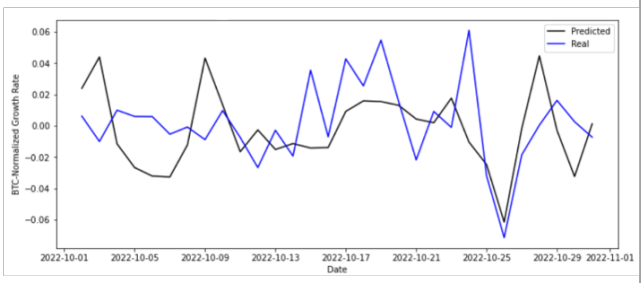
Figure 5: Model Results



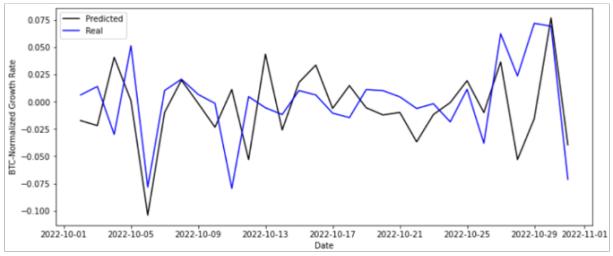
To further evaluate the fits, we calculated the Durbin-Watson test statistic for each model. This score measures the amount of correlation left in the residuals of the model on a scale from 0 to 4, with values close to 2 representing no correlation and values near 0 or 4 indicating positive and negative correlations, respectively. None of our six models had values deviating from 2 by more than 0.07, which are quite good results. This indicates that there are no long-run confounding factors which the models fail to capture. For Apecoin in particular, this indicates that although the coin has experienced more market shocks than the other coins, our VAR model still successfully captures the relationship between the predictor and returns.

Figure 6: **Predictions**

AAVE

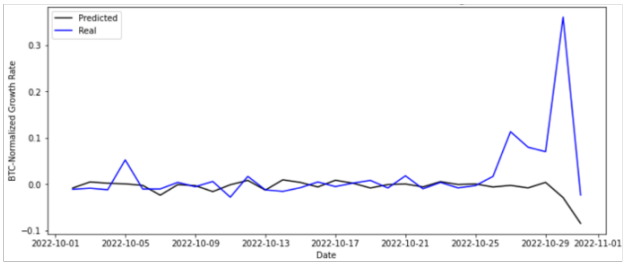


SHIB

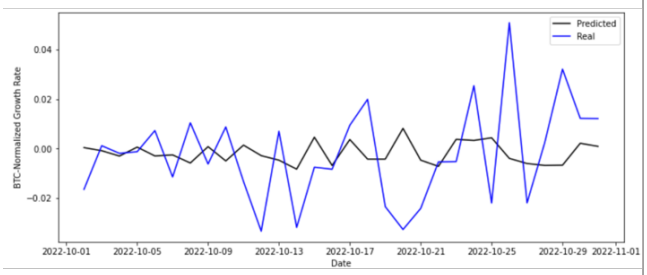


SOL

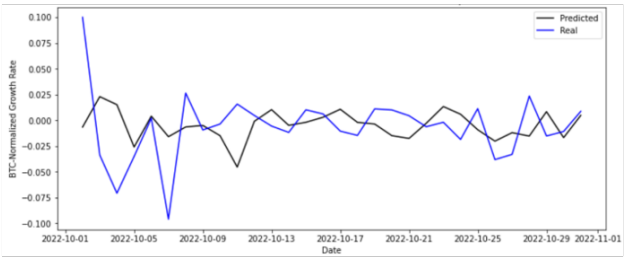
DOGECOIN



UNI



APECOIN



Finally, we fit our models on the testing data (the last month of data), yielding the graphs in Figure 6. These graphs show that, for the most part, these models do a relatively good job at predicting moderate price fluctuations, and therefore may be useful for long-run trading against Bitcoin. For most of the coins, the VAR models do not seem to underestimate the magnitude of larger price fluctuations, which indicates that they could be useful for capturing not only the direction of relative movements against the market, but also the magnitude of such movements. However, this observation does not seem to hold for UNI or DOGE, or, in general, for sharp fluctuations such as those at the beginning of October for APE. In fact, due to these fluctuations, we determine that the RMSE is not a good indicator of how predictive our models are, since it tends to overweight the impact of these sharp spikes.

5 Conclusion

We found that using sentiment data from tweets does a relatively good job of predicting price fluctuations, despite how tenuous of a connection this may be expected to be. Our results show that the predictive power is greater for overall market shifts than for capturing smaller daily price changes, which is to be expected. The strength of each model also varies a lot between different coins – as we predicted, the model seems to fit better to SHIB and DOGE, which we classified as speculative coins, and not as well to the other four which have some amount of inherent value. However, more research is certainly needed before claiming that this trend holds in general.

A major weakness of this project is that much more granular data is needed, since our assumption that sentiment affects each day's growth rate is a bit bold; given the occurrence of a price shock on an arbitrary day, the observed sentiment on that day could have been from after the price shock and thus not predictive. We therefore need a greater amount of trading data to determine whether sentiments affected intraday price shocks or vice versa. However, such granular data is very difficult to obtain for cryptocurrencies and so it was not feasible given the scope of this project.

One direction for further research would be to find an indicator which prioritizes directional movement over capturing magnitude of market shocks. This would better report how predictive our models are than relying on RMSE, which penalizes large price shocks to an undesirable extent. We would also like to apply these results to pairs trading with either Bitcoin or possibly a self-made crypto ETF to see how useful our models are. We could also compare such models against pure pairs trading without the sentiment information and see if our sentiment adds value to the pairs trading model.

6 References

[1] Lucas Louca (2021), *Why Use Logarithmic Returns*, Blog.

[2] (2022), *Historical Price Date*, Yahoo Finance.

[3] Joseph Perktold (2019), Stationarity and Detrending (ADF/KPSS), StatsModels.

[4] Selva Prabhakaran (2019), Vector Autoregression (VAR) - Comphrensive Guide with Examples in Python, MachineLearningPlus.

[5] James D. Hamilton (1994) *Time Series Analysis*, Princeton University Press.

[6] CFI Team (2022) *Durbin Watson Statistic*, Corporate Finance Institute.

[7] Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014.