Cryptocurrency and the Wisdom of the Crowds

Predicting changes in Cryptocurrency Prices using Public Sentiment

Aashka Trivedi   
*Computer Science Department*  
*New York University*  
aht323@nyu.edu

Minji Kim  
*Computer Science Department*  
*New York University*mk7773@nyu.edu  
  
Omkar Darekar  
*Computer Science Department*  
*New York University*  
oad245@nyu.edu

***Abstract*—Cryptocurrency prices are volatile, to say the least. But, most of this volatility may be attributed to the high dependency of cryptocurrency prices on public opinion. This project aims to study the influence of public sentiment and interest on the price of the most popular cryptocurrency, Bitcoin. Here, we aim to study the specific correlation between changes in bitcoin prices and three social media interest measures- Twitter Sentiment, Reddit Networks and Google Trends. This project analyzes the correlation on two granularities- daily and hourly. We first conduct a Granger Causality Analysis to identify if the three interest measures granger cause a change in bitcoin price. After this, we train a classification model that predicts the difference in bitcoin price per day, and per hour, given a set daily interest ”scores” derived from the aforementioned interest measures. We find that there is a high causation between google trend scores and twitter sentiment scores with bitcoin price. Moreover, we are able to train a classification model that achieves almost 60% accuracy while predicting changes in daily and hourly bitcoin price changes.**

***Index Terms*—sentiment analysis, bitcoin prices, predictive analytics, twitter sentiment, reddit networks, google trends**

1. Introduction

Cryptocurrency is by-and-large unregulated by any government institution, but it is still one of the most volatile forms of currency in the world. This volatility may be attributed to the fact that cryptocurrency prices remain highly dependent on public opinion and investor sentiment. An overall positive outlook towards a specific type of cryptocurrency, or the endorsement of certain types of public influencers (most notably, Elon Musk), may drive mass investments, thus bumping up the value of that form of currency.

In this project, we build on this motivation, and aim to empirically study whether there is a correlation between public sentiment and the changes in cryptocurrency prices, and if there is, then how strong and what type of correlation. The scope of this project will be limited to one form of cryptocurrency, that is, Bitcoin. Bitcoin began being used in 2009, and 12 years later, it remains one of the most popular and valuable forms of digital currency. The intuition here is that if we can build a pipeline to study the effects of public sentiment on one form of cryptocurrency, it can be easily expanded to include different types of digital currency- this remains an important future scope of the project. Public sentiment in this project is measured through three popular public interest measures:

1) Twitter Sentiment: Twitter is a social media platform that allows discussions and interactions in the forms of tweets. It is a highly popular means of micro-blogging, and a powerful indicator of public sentiment, even for bitcoin. In this project we will analyze the sentiment of tweets regarding bitcoin for a specific period of time, and see if this sentiment is an indicator of the direction in which Bitcoin prices will change.

2) Reddit Networks: Reddit allows forum-like discussions, arranged by topics as “subreddits”. Reddit is a popular forum for cryptocurrency and stock-related discussion, and boasts of millions of users. This project analyses the network of discussion centered around bitcoin, and analyses the number of discussions (popularity), along with possible sentiment of those discussions. We will first explore whether popularity itself is a good indicator of change in bitcoin price, and then try to explore sentiment, motivated by prior works [1].

3) Google Trends: Google is one of the most powerful search engines, and its daily search volume for a particular term may be indicative of public interest in that topic. We study the daily-aggregated search numbers for the term “bitcoin”, along with a constellation of other keywords related to the term. While google trends are only indicative of *some* interest, we believe it may be a powerful supportive influencer of bitcoin prices [2], [3].

Through this project, we will discover whether public interest, measured through Twitter Sentiment, Reddit Networks, and Google Search Volume, is able to predict the change in daily and hourly prices of Bitcoin, measured categorically. Daily, we define five categories of changes to bitcoin price- High Negative Change, Moderate Negative Change, No Significant Change, Moderate Positive Change, and High Positive Change. Hourly, we define 3 categories of changes to bitcoin price- Negative Change, No Significant Change, and Positive Change. We aim to study whether a classification model can be trained to predict the aforementioned categories of change in Bitcoin prices.

1. RELATED WORK

This project aims to study the correlation between the price of Bitcoin and public sentiment- specifically, Google Trends, Twitter Sentiment and Reddit Networks. This section briefly discusses the prior work done to study the effect that each of these platforms individually has on cryptocurrency prices.

*A. Google Trends*

Previous works [2] have shown that weekly Google Search trends are correlated to Bitcoin prices, however, recent works [3] challenge the notion that Google Trends alone are a strong predictor of Bitcoin price changes. Smuts [3] challenge the notion that such a positive correlation exists, arguing that in 2018 there was a strong negative correlation between Google Search volume and Bitcoin prices. This change may be attributed to the fact that Google Searches may not be an indicator of *positive* interest, but merely interest as a whole; and alone may not be able to measure investor sentiment. This leads us to believe that using Google Trends as a secondary factor of public interest may be more beneficial than using it as a primary factor.

*B. Twitter Sentiment*

After having read multiple research, we found that tweets from a selected group of users who influence the public sentiment more than others with the right classifiers can be an excellent predictor of positive and negative sentiment of a cryptocurrency. In the following work [4] the author goes on to validate his hypothesis of twitter feeds being an efficient predictor of a massive conglomerate like Dow Jones and its stock index. Another similar work [5] proved the efficacy of taking into consideration the tweet volumes of users along with the sentiment to lead the way to building a good predictor. Finally, to validate the work of the above authors the paper [6] goes on to prove a high degree of correlation between the twitter sentiments of verified base of interested users does help in predicting the value of an extremely volatile asset like Bitcoin.

*C. Reddit Network*

In the related work [1] defines that there is a medium- term positive correlation between price and online activity and argues that such a relationship supports the validity of cryptocurrencies as speculative assets. Also this paper has shown that the website Reddit has been successfully used as a data source used to model user behavior. The idea that we focused on from this article is when they combine the engineered features from Reddit communities along with features based on past price fluctuations, the model gives better forecasting. From this point, we intend to use 3 different datasets and try to make the best prediction of the price fluctuations.

1. DATA SOURCES

This section describes in detail the process of collecting, understanding, and preparing our four main data sources for an 8 month period from January to August 2021. The data for each source has been collected and prepared on two granularity levels- daily numbers and hourly numbers.

*A. Google Trend Data*

For this project, we use the daily google search numbers for a constellation of keywords to analyze the interest in certain bitcoin-related topics over time. We use the Pytrends API [7] to scrape this data, and find the daily search numbers for 16 bitcoin-related keywords for the 8 month period from January 2021 to the end of August 2021.

We obtain the keywords by using the ”Related Queries” API call of PyTrends, and choose the top-16 queries as our constellation. The keywords used for analysis are:

*bitcoin, btc, bitcoin price, bitcoin kurs, bitcoin usd,*

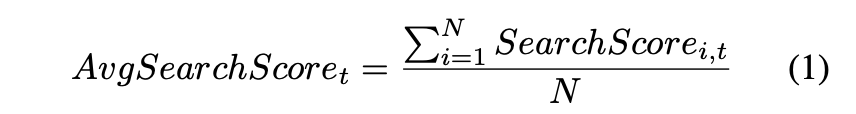
*bitcoin stock,bitcoin dollar, bitcoin euro, buy bitcoin,*

*buy btc, btc usd, btc inr, price btc, btc stock, btc coin,*

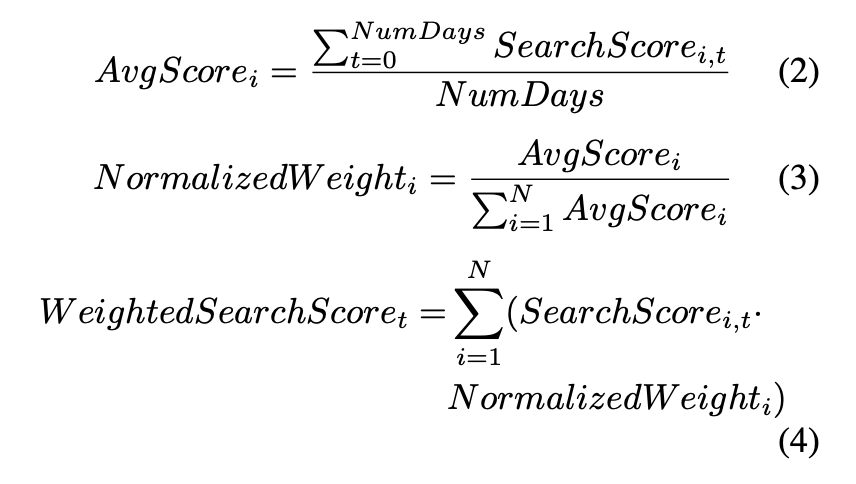
*btc euro*

*1) Daily Google Search Scores: A goal of this process is to perform Granger Causality [8] to identify whether a change in google search numbers would cause a change in bitcoin prices. In order to use a single time series for Google Search Numbers, we propose the use of a Daily Search Score. We experiment with two methods to combine the search numbers for 16 keywords into a single daily score:*

1) **Average Search Score**: This score takes the average of the scaled search score provided by the PyTrends API for each keyword in our constellation. This is a simple metric that helps gauge the overall interest, while giving the same importance to every keyword. Here, say we have N keywords in our constellation. Then, for each keyword Keyword, we have a daily search score SeacrhScorei,t for day t. The Averaged Search Score for each day t is calculated as:

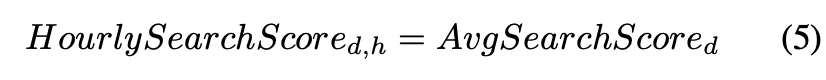
**

2) **Weighted Search Score**: Here, we give each keyword a weight proportional to how much it is usually searched for (on average). We find the average search score *AvgSearchi* for each individual term *keywordi* for the 8 month period that we are analyzing. The weighted average is then computed for all keywords, giving weight proportional to AvgSearchi to the term *keywordi.* The Weighted Search Score for each day t is calculated as:



We then analyze each of these scores to see which has a stronger Granger Causality to Bitcoin Prices.

*2) Hourly Google Search Scores:* We also reduce the granularity of analysis, and determine the relationship between google search prices and *hourly* bitcoin prices. However, Google Search Scores are only provided on a per-day basis, so in order to obtain an *Hourly Search Score*, we simply replicate the Daily Search Score, as described in the previous subsection, to each hour of that day. Here, we only use the Average Search Score. Thus, for each hour h of day d, we get the hourly search score as:



*B. Twitter Sentiment Data*

Twitter data was an important data set for this project, as it was the largest database of public opinions that could very well influence the prices of bitcoin on a given day. Granger Causality on twitter data also gives us the highest score, indicating the relation between the bitcoin price volatility could be highly influenced by tweets about bitcoin, or well-known figures in the world of bitcoin. We could collect about 8 months of tweet data starting from the first month of 2021, until August. To extract relevant tweets, we have used set of 4 most widely used hashtags/tokens to identify bitcoin tweets ie., btc, Bitcoin, BTCUS, BTCTN.

*C. Daily Twitter Sentiment*

We started off this project by aggregating daily sentiment scores for each of our public opinion data sets, which meant that for every day for the 8 months of our data set we had a compound sentiment value drawn from the negative, positive and neutral scores calculated for each tweet. We used this compound score to then feed into our best model to categorize the bitcoin prizes as a result of the classification problem.

*D. Hourly Twitter Sentiment*

After evaluating our model performance on daily sentiment scores from our twitter data set, we came to the conclusion that we could not reach our model’s peak performance due to our data set being rather small in terms of mined data. Hence, we tried to increase the size of the data set by accumulating the hourly sentiment for our tweets, which had a two fold advantage. One, it improves the granularity of the dataset and helps us look at the shifts in public opinion in smaller time ranges. Second, since bitcoin’s asset price is considered to be extremely volatile, by increasing the granularity of the dataset, we are in a better position to track these sudden shifts in public opinion with the hourly sentiment.

*E. Reddit Sentiment Data*

Reddit Network is an online forum that consists of different subReddits which are separated by their specific topics. Users are gathered in place of their interest to discuss and exchange ideas or information. This implies all the information on Reddit is grouped by its main subject, and also the majority of the data being text data.

Therefore, we could simply target the Bitcoin subreddit (r/bitcoin) and scrap all the user comments in the time range of January 2021 to the end of August 2021. We assume all the comments under the subreddit are talking about Bitcoin, so no keyword extraction is executed. Instead, heavy text cleaning is needed such as removing emoji or stopwords. The Reddit API which is provided to users from Reddit to scrape their data only allows to pull a very limited amount of recent comments or submissions from even a few different streams for subreddits, such as hot, new, top, etc. Because of this reason, we employed a third-party API, Pushshift. It enables you to get a large amount of data from any subreddits. The scraped data includes date and time for each comment

*1) Daily Reddit Scores:* To score public mood on Reddit data, VADER was applied. VADER is a package used for sentiment analysis, it provides outstanding sentiment analysis results and also it guarantees relatively faster running time, so it is appropriate to work with large data.

1) **Get sentiment score on each comment**: First apply VADER to every comment in the Reddit data. Once executed with the polarity scores() method the score of four properties will be given.

*1. neg : the negative emotional index*

*2. neu : the neutral emotional index*

*3. pos : the positive emotional index*

*4. compound: sentiment index between -1 and 1 by appropriately combining neg, neu, and pos scores*

We evaluate either the sentiment is positive or negative, based on the compound score. If the score is 0.1 or higher than that then it implies positive mood, and if the score is lower than 0.1 then it implies negative mood.

2) Averaging the scores as daily value: As we are predicting the daily price fluctuation, each score of comments should be unified as a score daily. For this process we choose averaging all comments per daily.

*2) Hourly Reddit Scores:* As the Reddit data scraped is already involving both data and time for each comment, from the step that we applied VADER to each data, we simply separate the scores by hourly and averaging them as we did to get the daily score.

*F. Bitcoin Prices*

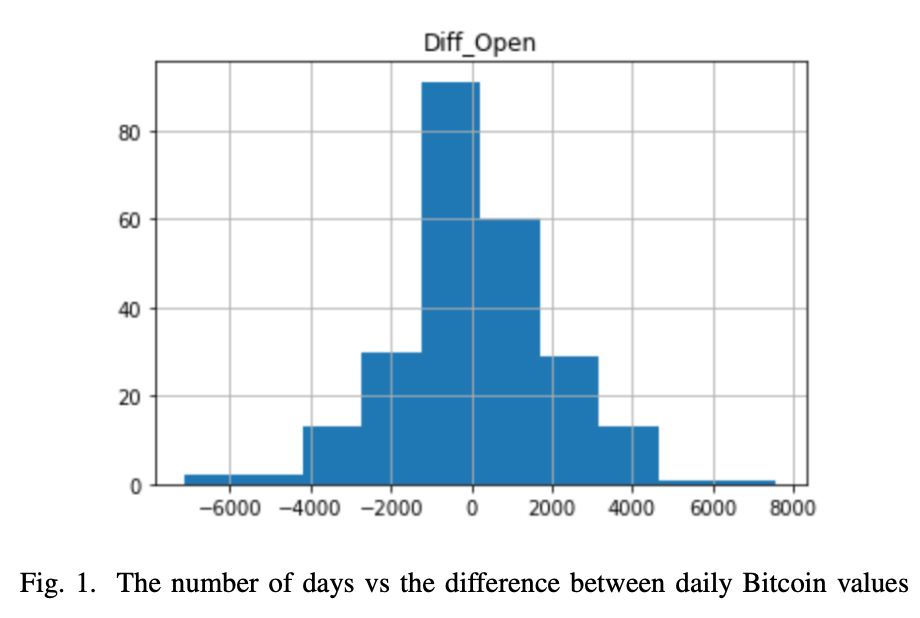
Bitcoin Prices on both an hourly and a daily level are widely available in the form of public datasets and open source APIs created by investment websites and tools. However, it will be very difficult to predict the exact price of bitcoin, which is highly volatile and has a lot of variability. Instead, this project concentrates on *classifying* the change in daily or hourly bitcoin price. This categorization is part of the data preparation step, and is discussed below for both granularities of the time series.

3

*1) Daily Prices:* We use Alpha Vantage’s API [9] to get the daily bitcoin prices for the 8-month period from January 2021 to the end of August 2021. In our analysis, we always use the “Open” Prices for the bitcoin price.

In this project, we want to predict the changes to bitcoin prices using the social media interest measures of Twitter Sentiment, Reddit Sentiment and Google Search Trends. Since it is out of the scope of this project to predict the exact price or change in price of bitcoin, we instead convert the change in bitcoin price to a *categorical* metric. We divide the change of the daily open bitcoin prices into 5 categories:

* 1) Highly Negative Change: when the difference in the prices of Bitcoin at the open of Dayt and Dayt+1 is less than -$4000.
* 2) Moderately Negative Change: when the difference in the prices of Bitcoin at the open of Dayt and Dayt+1 is between -$4000 and -$500.
* 3) No Considerable Change: when the difference in the prices of Bitcoin at the open of Dayt and Dayt+1 is between -$500 and $500.
* 4) Moderately Positive Change: when the difference in the prices of Bitcoin at the open of Dayt and Dayt+1 is between $500 and -$4000.
* 5) Highly Positive Change: when the difference in the prices of Bitcoin at the open of Dayt and Dayt+1 is more than $4000.



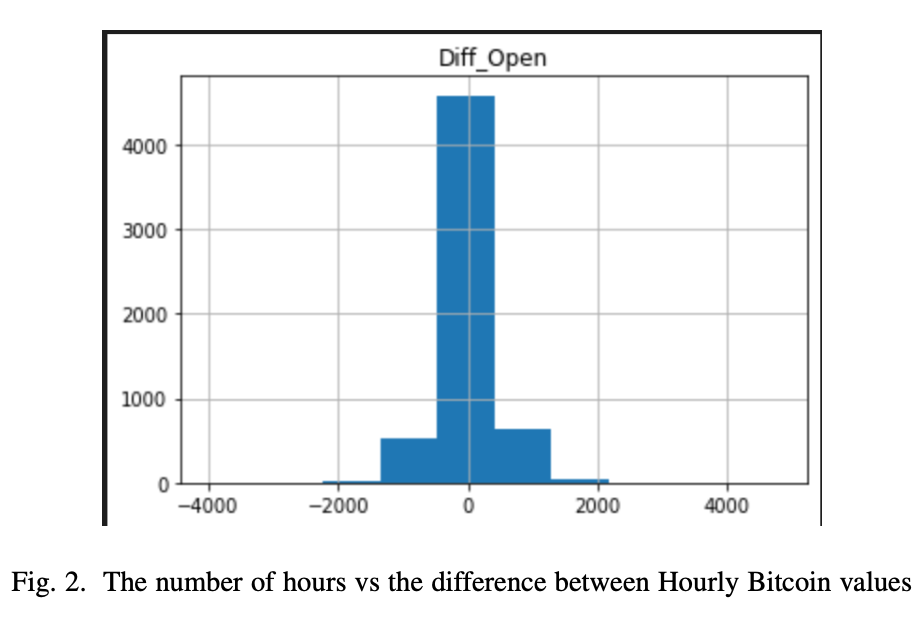
These categories were formed after analysing the number of days that fall into each category’s difference range. This histogram is shown in Figure 1.

*2) Hourly Prices:* We obtain the hourly prices for Bitcoin from Bitstamp through CryptoDataDownload [10]. Similar to daily prices, we convert the change of hourly bitcoin prices to a categorical metric, by dividing it into 3 categories:

* 1) Negative Change: when the difference in the prices of Bitcoin at the open of Hourt and H ourt+1 is less than -$150.
* 2) No Considerable Change: when the difference in the prices of Bitcoin at the open of Hourt and H ourt+1
* is between -$150 and $150.

3) Positive Change: when the difference in the prices of

Bitcoin at the open of than $150.



These categories were formed after analysing the number of hours that fall into each category’s difference range. This histogram is shown in Figure 2.

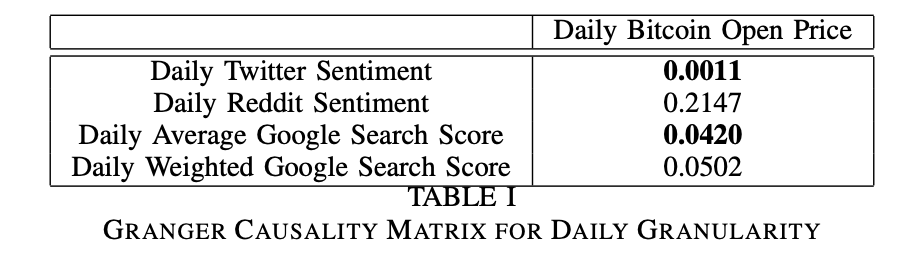
IV. METHODS

For the modelling aspect of our project, we first perform a Granger Causality Analysis in order to determine which of the three interest measures (if any) have a causal relationship with Bitcoin Prices, and then attempt to build a classification model that can predict the type of change in Bitcoin Price. Both methods are described here.

*A. Granger Causality Analysis*

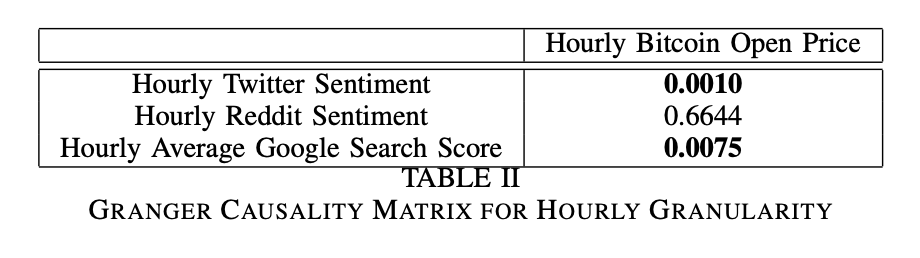
To determine whether Twitter Sentiment, Reddit Sentiment or Google Search Trends have a causal relationship with Bitcoin prices, we first conduct a Granger Causality Analysis [8], [11], [12] amongst these measures. Granger Causality helps to test whether one time series can be used to forecast another. In our analysis, we perform a Chi-square test and use the p-value to determine causality, with a maximum lag of 15. Here, we consider a significance threshold of 0.05, so, a result of less than 0.05 would indicate that the variable X granger causes the variable Y. We conduct the granger causality analysis on two granularities- daily and hourly.

*1) Daily Granularity:* Granger causality can only be measured between *stationary* timeseries, i.e, processes whose unconditional probability and statistical descriptors would not change in time. Our analysis showed that Daily Bitcoin Prices and Daily Twitter Sentiment are non-sationary, thus, to make them stationary, we calculate the daily difference between the values.



The Granger Causality Matrix for Daily values is shown in Table I. As discussed above, if the entry M[i,j] is lesser than the significant value (0.05), then it indicates that *i granger causes j*. Thus, twitter sentiment and averaged google score granger cause bitcoin prices.

*2) Hourly Granularity:* Granger causality can only be measured between *stationary* timeseries, i.e, processes whose unconditional probability and statistical descriptors would not change in time. Our analysis showed that Hourly Bitcoin Prices and Hourly Reddit Sentiment are non-sationary, thus, to make them stationary, we calculate the daily difference between the values.

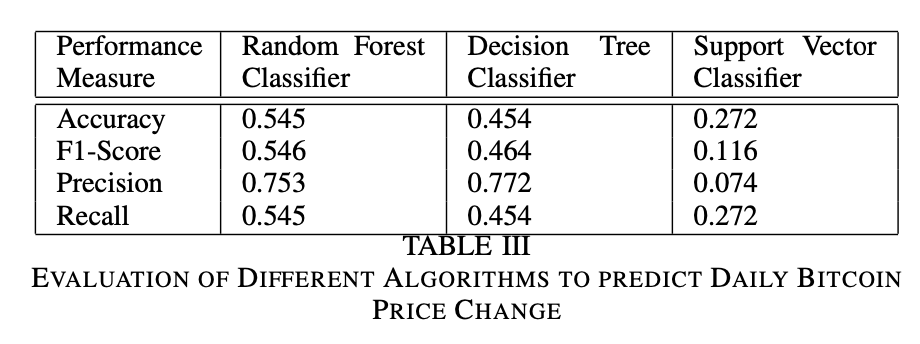


The Granger Causality Matrix for Hourly values is shown in Table II. As discussed above, if the entry M[i,j] is lesser than the significant value (0.05), then it indicates that *i granger causes j*. Thus, twitter sentiment and averaged google score granger cause bitcoin prices.

*B. Predictive Modelling*

One goal of this work is to create a model that can predict the direction of change Bitcoin prices given the twitter sentiment scores, google trend scores, and reddit sentiment scores of the day or hour before. Now, collection of the datasets was tricky, because there was a lot of missing values, especially on the hourly granularity (for example, it is not necessary that a new comment is left on reddit every hour). We handled the missing values as follows:

* We first remove any columns for which both the twitter sentiment score and the reddit sentiment score is unavailable.
* In the remaining data points, we fill the missing values with the *average* of the sentiment value. We found this to be the best way of filling missing values, as compared to simply dropping the rows (which drastically reduced the size of our dataset), or filling it with 0’s (which reduced accuracy of the model).

We then trained our classification model with about 95% of training data, and tested our model on the remaining 5% of training data. The major drawback was the small data size, so in order to obtain a model that performs decently, we needed to keep the training set as large as possible.

For both the daily and hourly granularity, we train a Random Forest Classifier, a Decision Tree Classifier and a Support Vector Classifier to experiment with the results. We then select the best performing classifier and apply Grid Search in order to find the best hyperparameters for the given model.

V. EXPERIMENTAL RESULTS

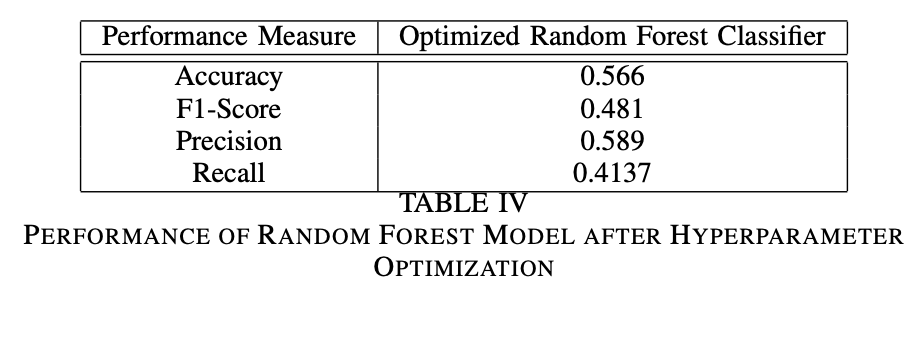
Here, we describe the results of our two models- which predict change in bitcoin price using sentiment data on both a daily and an hourly granularity.

*A. Model to Predict Daily Bitcoin Price Changes*

On a daily granularity, we find the *Random Forest Classifier*

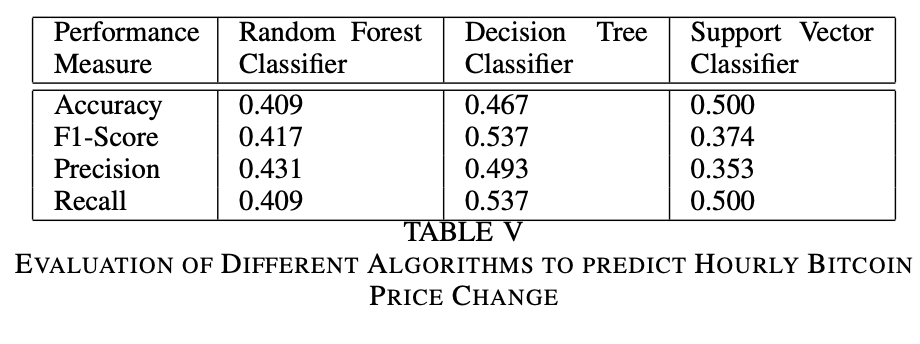
to be the best performing model, as shown in Table III. Please note that here, the F1-Score, Precision and Recall are the weighted average of the 5 possible classes.

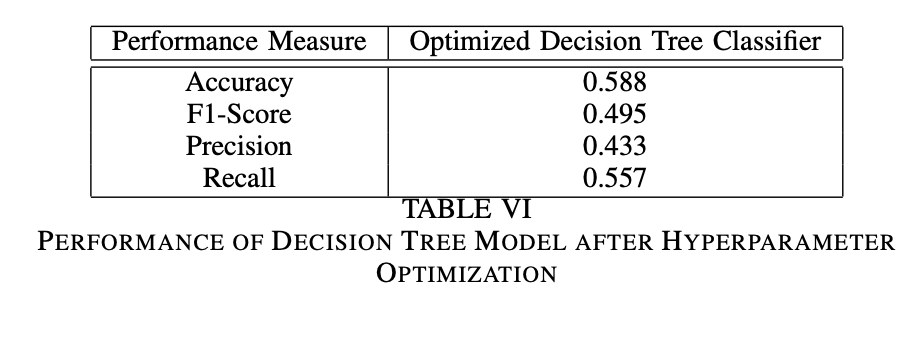
We choose the Random Forest Classifier for further hyperparameter tuning, where we search on candidate values for the max depth, max features, minimum samples for a leaf, minimum samples for a split, number of estimators and whether bootstrapping is needed. We perform a Grid Search, fitting 3 folds for each of 648 candidates, for a total of 1944 fits.

The performance of the model after tuning the hyperparameters is shown in Table IV. As seen, we achieve a 56.6% accuracy, which is much better than a random choice between the 5 possible bitcoin price change categories.

*B. Model to Predict Hourly Bitcoin Price Changes*

On an hourly granularity, we find the *Decision Tree Classifier* to be the best performing model, as shown in Table V. Here, the F1-Score, Precision and Recall are the weighted average of the 3 possible classes.

We choose the Decision Tree Classifier for further hyper- parameter tuning, where we search on candidate values for the max depth, max features, minimum samples for a leaf, minimum samples for a split, and splitting type. We perform a Grid Search, fitting 3 folds for each of 162 candidates, for a total of 486 fits.

The performance of the model after tuning the hyperpa- rameters is shown in Table VI. As seen, we achieve a 58.8% accuracy, which is much better than a random choice between the 3 possible bitcoin price change categories.

*C. Discussion*

While we can see that our models do significantly better than random guessing, however they still don’t achieve accuracies high enough that they can be deployed. One hint on how to improve accuracies may be noticed in the observation that the hourly model does better than the daily one. This points to the fact that the model may require much more training data to improve the accuracy. That, along with other models may help improve performance.

VI. CONCLUSION

In this work, we study the causatory relationship between public interest measures and bitcoin prices for the period of 8 months. We find that there is a strong granger causality between bitcoin prices and interest measures including twitter sentiment and google search trends. Furthermore, we build models that achieve about 60% accuracy in predicting the direction of bitcoin price change on both an hourly and a daily granularity.

VII. FUTURE WORK

This project has a lot of scope in the future. First, more data from a larger period of time can be collected to create higher performing models. We know that even non optimized models work relatively well when fed large amounts of data as compared to state of the art models. Hence, we believe that collecting at least 2-3 years of relevant twitter, reddit and google trends data would significantly improve the model performance. Moreover, better analytics techniques can be used to improve the data preparation. We used simple averaging to get sentiment scores on a daily and hourly basis, but better means of scoring can also be explored. Finally, an important future scope is to extend this analysis to different types of cryptocurrencies, and study whether the same relation- ships still hold. Since we have only used traditional machine learning models in the course of developing a full fledged classification model, there are certainly better performing models ie neural networks that have a more sophisticated approach when it comes to classification tasks. Therefore using state of art neural networks for classification should give us a much better score, as neural networks are considered universal learners when presented with a large enough data set.

REFERENCES

1. S. Wooley, A. Edmonds, A. Bagavathi, and S. Krishnan, “Extracting cryptocurrency price movements from the reddit network sentiment,” in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 500–505. [Online]. Available: https://webpages.uncc.edu/ skrish21/papers/PID6202775.pdf.
2. L. Kristoufek, “Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era,” *Scientific Reports*, vol. 3, no. 1, 2013.
3. N. Smuts, “What drives cryptocurrency prices? an investigation of google trends and telegram sentiment,” *SIGMETRICS Perform. Eval. Rev.*, vol. 46, no. 3, p. 131–134, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3308897.3308955.
4. S. Colianni and M. Signorotti, “Algorithmic trading of cryptocurrency based on twitter sentiment analysis,” 2015.
5. J. Abraham, “Cryptocurrency price prediction using tweetvolumes and sentiment analysis,” 2018.
6. A. Urquhart and Wang, “Does twitter predict bitcoin?” 2019.
7. J. Hogue and B. DeWilde, “Pytrends.” [Online]. Available:  
   <https://pypi.org/project/pytrends/>
8. C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: <http://www.jstor.org/stable/1912791>
9. Alpha Vantage, “API Documentatio.” [Online]. Available: <https://www.alphavantage.co/documentation/>
10. “Bitstamp exchange data.” [Online]. Available: <https://www.cryptodatadownload.com/data/bitstamp>
11. C. Granger, “Some recent development in a concept of causality,” Journal of Econometrics, vol.39, no.1, pp.199–211, 1988. [Online]. Available:https://www.sciencedirect.com/science/article/pii/0304407688900450
12. C.W.J. Granger, “Testing for causality: A personal view point,” Journal of Economic Dynamics and Control, vol.2, no.1, pp.329–352, 1980. [Online]. Available: https://EconPapers.repec.org/RePEc:eee:dyncon:v:2:y:1980:i:1:p:329