Feature Analysis and Forecasting Price Movement in Crypto Market

Kaavish Report
presented to the academic faculty
by

Muhammad Faizan Kazi	mk04339
Muhammad Ammar	ma05168
Muhammad Jazib Bhatti	mb05082

In partial fulfillment of the requirements for
Bachelor of Science
Computer Science

Dhanani School of Science and Engineering

Habib University

Spring 2022

Feature Analysis and Forecasting Price Movement in Crypto Market

This Kaavish project was supervised by:

My Internal Supervisor
Faculty of Computer Science
Habib University

Approved by the Faculty of Computer Science on _____.

Dedication

This is for our parents for their constant support which kept us going. Thank you for being the continuous source of support and encouragement during the challenges of graduate school and life.

Acknowledgements

I would like to thank our internal supervisors Professor Sohaib Ali and Professor Qasim Pasta for Kaavish I and Kaavish II, respectively, for their guidance throughout this project. Without which this project would not have been possible.
We also appreciate our external supervisor Amer Haider for being a constant source of counsel
Thanks to our friends and colleagues for their enlightening advises which were a great help along the way
Finally we are grateful for Krish Naik (https://www.youtube.com/user/krishnaik06) and the website https://towardsdatascience.com/ for the information and assistance which helped us build the basis of this project.

Abstract

The project aims to capture the price movement of Bitcoin and forecast short term percent changes. It uses ARIMAX model using exogenous independent features which include technical indicators and sentiment of the tweets on Bitcoin. This will help day traders and business entities doing short term trading to determine the right time to enter and exit this highly volatile market.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Problem Statement

The cryptocurrency, a decentralised digital asset built using blockchain technology, has seen an extraordinary growth in popularity and interest since its inception. This digital currency has attracted a lot of interest owing to its extreme volatility, which allows for profitable digital trading. In the last decade, the entire market value of cryptocurrencies has risen from 1 billion to over a trillion dollars[4], with the number continuing to rise. The cryptocurrency market thrives on speculation. To profit, investors gamble on whether prices would rise or fall.These speculative bets result in a large inflow of cash or a large outflow of cash, resulting in significant volatility. While investing in the crypto-market can offer significant profit margins, the risk element is very high owing to the high volatility, which makes small investors wary of risking their modest capital.

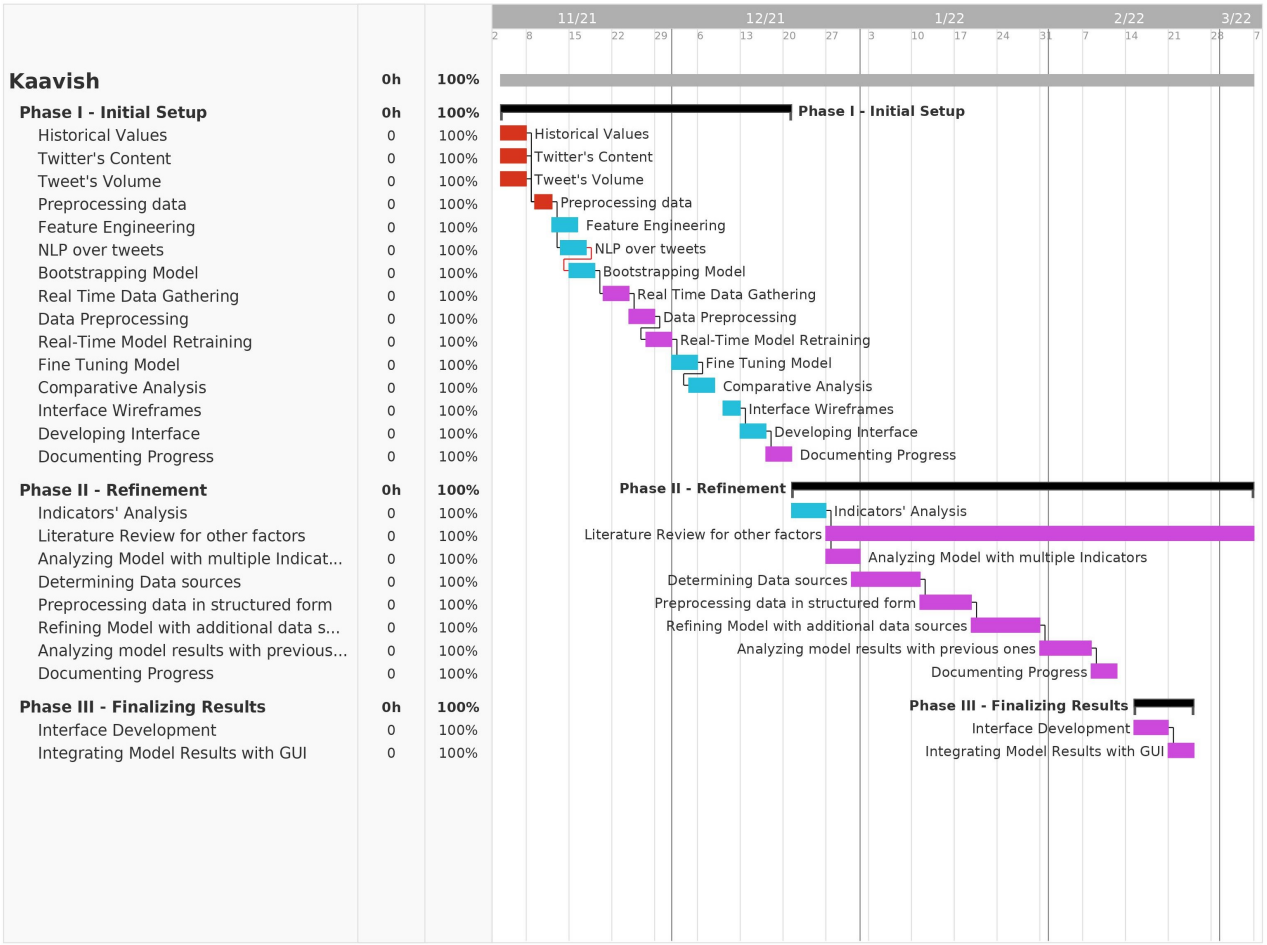
1.2 Proposed Solution

We plan to forecast the price movement of the Bitcoin through machine learning models based on the historical data; the OHLCV table, technical indicators and sentiment analysis of social media regarding Bitcoin. The emergence of social media such as Twitter makes the latest news and social media posts about financial markets widely accessible. Investors have therefore been utilizing such a variety of digital resources to make trading decisions and hence the motivation behind the sentiment analysis module of our project.

1.3 Intended User

Since our project's aim is to predict the price movement of the Bitcoin, our intended user would be investors wanting to maximize their profit by making informed decisions and avoiding price fluctuations. This may come handy to those investors having little or no experience in the market but still wanting to invest in this volatile market and make profits.

1.4 Project gantt chart and deliverables



1.5 Key Challenges

Since, it has been just 10 odd years when Bitcoin was officially listed, the maximum amount of data we can have is 10 years of historical data. This may throw us a challenge since limited data may affect the accuracy of models, so we may tackle this problem by decreasing the resolution(time-period of the OHLCV) from day-wise data to hourly data if not minute data. This will increase computation required for prepossessing and running models but we will be using university's GPU to avoid that problem.

Although there are many technical indicators which are used to study the trend of the market but they are not generalized. If an indicator performs gives good insight on the market of a specific coin, it may not do that well for another coin. To tackle this issue, we plan to make use of different denominations of indicators (such as SMA and EMA crossovers) and use them as our own custom indicators. So, Feature engineering will be performed on the custom indicators and analysis will be done to analyze the insight they provide.

2. Literature Review

This chapter presents the current state of the art in the domain and talks about other similar work that has been done in this area. It also establishes the novelty of our work by highlighting the differences between the existing work and our work.

We will keep updating this chapter (especially if our project is research-intensive) as our research proceeds and we come across more work related to our problem.

Of course, we take inspiration from [1] but wish the work was typeset in LaTeX[3], e.g. by taking help from [2].

3. Software Requirement Specification (SRS)

3.1 Problem Formulation

The problem that we are targeting for our Kaavish Project is to forecast and analyze the price movement of cryptocurrencies specifically Bitcoin. Many researchers have worked on similar lines, which will be shared in detail in Literature Review. We plan to approach this problem by first reviewing the work done by various researchers, which will provide us with a relevant insight of the patterns and trends being followed in order to solve this problem. Our main objective would be to understand those methodologies and provide our own innovation by using multiple(relevant) data sources through which we can forecast the price movement more effectively. We have divided our approach in the following three modules:

3.1.1 Module I

Literature Review

In order to start a research on any field, one must know the recent studies and researches done in that area. Therefore, our first step is to study recent research papers and articles which have been published in this domain. That will help us identify the different factors and parameters upon which the volatility of the bitcoin market depends. We will be focusing on the state of the art architectures and the datasets along with the models that are being used in them in recent studies.

Identifying Novelty

After having a detailed overview of the work and researches done by researchers in this field, we will have an understanding of the patterns that are currently being followed in order to predict the market or price movement. This will surely help us in getting a start, but our main objective would be to incorporate a unique methodology by using not just a single data source, but multiple data streams using which our model can forecast the price movement more effectively.

3.1.2 Module II

Dataset Preparation

This will be the most important part of our Kaavish journey. Since our research will mostly be revolve around identifying the data streams which manipulate BTC market. These data streams can be as straightforward as OHLCV tables or as complicated and unstructured as the social media sentiments which can heavily impact the price movement of cryptocurrencies. The objective for this sub-module would be to utilize all these data streams and to transform the unstructured data in a more structured format i.e. in a learn-able format for the model such that it can identify whether the price shift would be positive or negative.

Model Training And Refinement

After identifying the data streams, transforming them into structured form, we will be all set to bootstrap our model i.e. training the model. We plan to feed multiple identified data streams and perform feature engineering to be able to achieve the best possible set of features that can yield high accuracies for the forecasting. Feature engineering may involve discovering the correlation of the feature set with the close prices, dimensionality reduction such as the Principle Component Analysis of the technical indicators. Then we plan to fine tune our trained model such it retrains itself in real-time by using the error in previous predictions.

Comparative Analysis and Model Improvement

Once our model is successfully trained, we plan to test it's accuracy with regards to the hit rate, and will be analysing it in real-time over the unseen data. Our aim would be to compare the results of our model with the actual percentage change in the close prices. Since there is always a room for improvement so along the way we will be researching and looking for other relevant parameters which may help us to achieve better results.

3.1.3 Module III

Interface Designing

Once our research is completed, we will move towards designing a user-friendly interface through which users can view the live real-time predictions by our model as well as the history of model's predictions and bitcoin's actual value.

Interface Development Integration

This sub-module as the name suggests, will be dedicated to the development of the interface and integrating our model's results with the interface in a user-friendly manner. Following are the libraries and tools we will be using :

- ReactJS - A frontend library to develop user interface
- Flask - Developing server which will fetch model results and serve it for our application
- React Bootstrap - A frontend styling library for the design and styling of the application

3.2 Datasets Preparation

This section describes the specific dataset(s) used to build our system. An appropriate snapshot of the dataset(s) is also included. Further details, when needed, are presented in the appendix.

3.2.1 OHLCV Historical Data

The dataset which will be the backbone of our project is the historical data. It consists of open, high,low,close and the time frame of coin we are taking into consideration. Open indicates the price which the stock opened with respect to the day and close is it's opposite. Meanwhile the high and low indicates the all-time high and all-time low price of the stock with respect to each day. And finally volume indicates the total number of shares traded in a security over a period.

Currently, we will be using the OHLCV with a resolution of day-wise data from the year 2014 to current data. It contains more than 2500 days of trading data which we will be feeding to our respected models.

The motivation behind considering the historical data is by looking at the range of the assets within the day, it can help us identify the amount of volatility in the market.

Apart from that, we will be using the close prices from the OHLCV and using them for the technical indicators.

OHLCV Dataset Snippet

	timestamp	open	high	low	volume	close
0	2015-10-08 13:00:00	0.0	245.0	0.0	0.60665438	245.0
1	2015-10-08 14:00:00	245.0	245.0	244.5	4.453648931	245.0
2	2015-10-08 15:00:00	245.0	245.0	244.92	3.016925828	244.92
3	2015-10-08 16:00:00	244.92	244.92	244.25	3.89525246	244.25
4	2015-10-08 17:00:00	244.25	244.99	244.02	3.920632003	244.99
...	...	...	...	...	...	...
57652	2022-05-06 20:00:00	35996.68	36100.0	35865.0	31.55234367	36083.75
57653	2022-05-06 21:00:00	36083.75	36100.0	35882.81	36.33917961	35961.32
57654	2022-05-06 22:00:00	35961.32	36055.6	35884.76	30.0996847	35990.88
57655	2022-05-06 23:00:00	35990.88	36135.0	35978.91	79.57487521	36005.22
57656	2022-05-07 00:00:00	36005.22	36055.08	35944.61	12.34896688	35986.42

57657 rows × 6 columns

3.3 Literature Review

In the area of forecasting the price movement of crypto market, many researchers, analysts and AI/ML enthusiasts have done a lot of research and designed models specifically for this purpose. We tried to study the recent study being done in this field. And have summarized the best findings in multiple researches in this section below. We found that LSTM was the most used model in the recent years. Detailed can be found in this excel sheet.

A paper published by Helder Sebastião & Pedro Godinho [1] in January 2021, in which after training their model from bullish period, they validated their model on the bearish historical data of crypto market. Upon testing over multiple models, two of them produced a annualized Sharpe ratios of 80.17% and 91.35% which had an annualized returns of 9.62% and 5.73%.

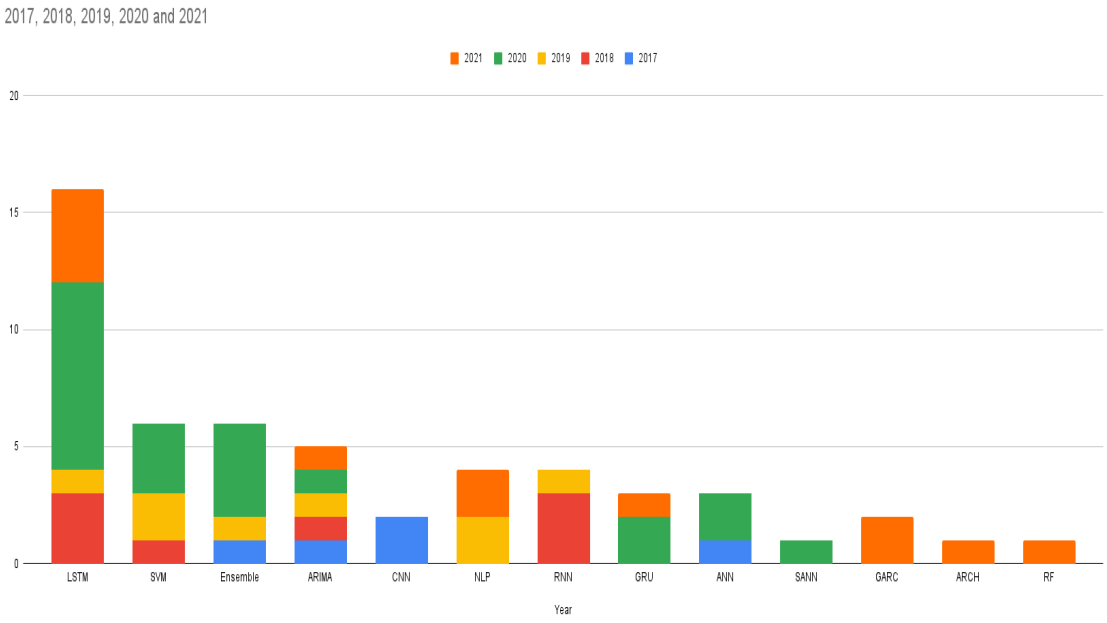


Figure 3.1: Graph representing models used in different studies

Another research and comparative analysis done by Ukranian researchers[2], in which they used Machine Learning Ensemble Algorithms for Forecasting Cryptocurrency prices. The ensemble algorithms included Random Forests (RF) and Stochastic Gradient Boosting Machine (SGBM), their datasets included daily close prices and after training their ensembled models, the out of sample accuracy of short-term prediction daily close prices obtained by the SGBM and RF in terms of Mean Absolute Percentage Error (MAPE) for the three most capitalized cryptocurrencies (BTC, ETH, and XRP) were within 0.92-2.61%.

A recent study by Caux, Bernardini and Viterbo [5], used the infamous model LSTM and GRU in this study. Their objective was to forecaste bitcoin values in a minute-granulated time for the entire next day. Although they used the time-series forecasting methodology but proved that GRU performs better in worst case scenario than LSTM, but in a better scenario both performs same.

Many other researches shows that by implementing different models such as multiple-input deep neural network model[2], neuro-fuzzy controller and neural network mod-

els[3], and machine learning-based classification and regression models we can forecast the price movement of cryptocurrencies namely Bitcoin, Ethereum and Ripple.

Another research done by Mohapatra, Ahmed and Alencar[6], in which they used the tweets for the public sentiment analysis along with OHLCV tables as their data streams. Tweets were refactored by performing VADER analysis, an NLP methodology, which provides the polarity of a document i.e. positivity or negativity on a scale of -1 to +1. Thus, their final dataset was transformed in a tabular form, for which, as they state, decision-tree based models outperforms other. Therefore, they used XGBoost model for their training. Furthermore, after bootstrapping the model, they developed a system where, after a specific time frame, both the data streams provide the dataset collected in that time to the model along with the error of model in predicting the price in that time-frame. Upon which, model retrains itself in real-time and provide with more precise predictions in next round. This was one other reason mentioned for picking XGBoost as their base model because it takes a lot less time to train itself than many others.

Since it is a known fact that, unlike stocks or gold, cryptocurrencies do not have any intrinsic value attached to them. Instead, their price is purely determined by the supply and depend of that specific currency. Heavy demand from buyers will push the value of a digital coin upwards. Therefore, there are multiple ways affecting the price movement of a digital coin which involves, but not restricted to, their team's involvement with the public on social platforms, the use case behind that digital coin, any news appearing regarding the digital coin. All of them can both positively and negatively affect the price movement of the digital coins. All of the said data is not something readily available on any platform. We plan to use this unstructured data and preprocess it in order to feed it to our model along with the historical and real-time price charts including daily highs, lows, open and close prices.

3.4 External Interfaces

We expect every project to have at least of the following subsections. This section must be aligned with your project deliverables. Please consult with your project supervisor regarding which of the following section(s) you should include in your report

3.4.1 Visualization

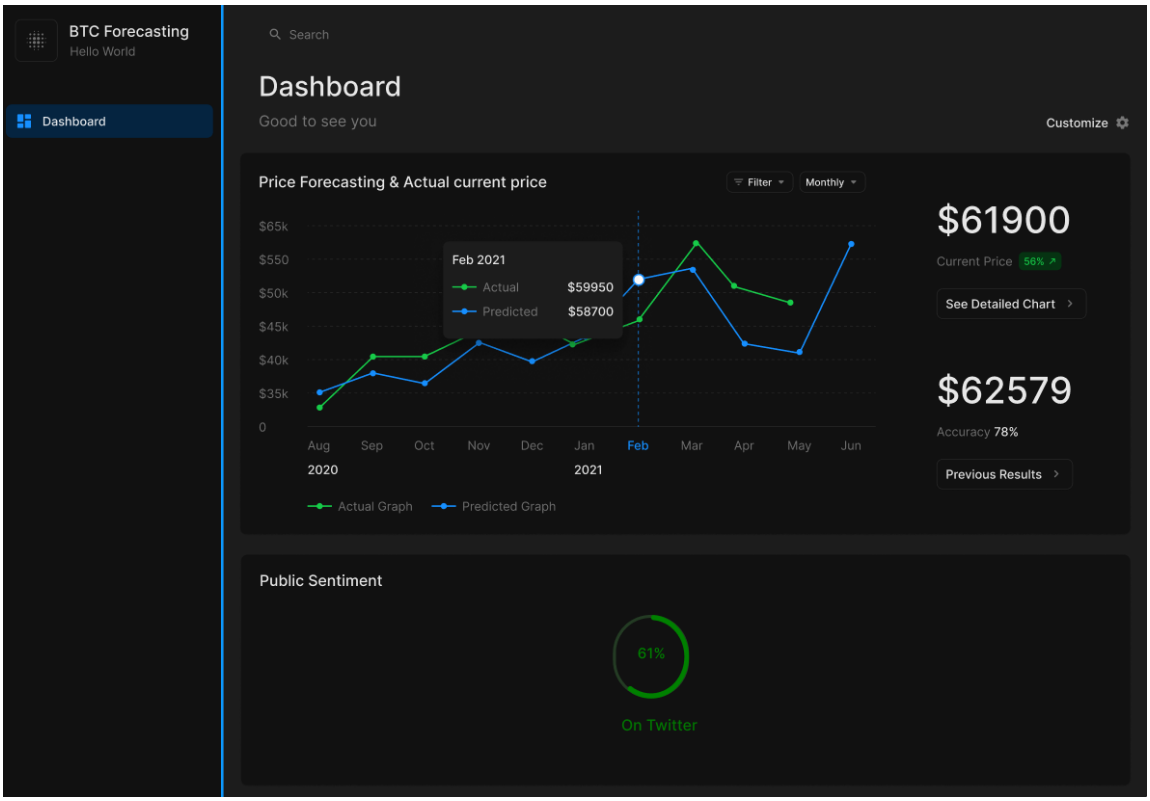


Figure 3.2: Representing our model results

Graph Section

There will be 2 graphs,one will be the realtime close price and the other will the close price predicted by the model. A drop-down bar will filter the chart for whatever resolution the user selects i.e yearly, monthly or daily. A legend at the bottom will be indicating will be indicating which line represents the actual close and which line the predicted close.

Public Sentiment

The polarity of the sentiments of the public from different social media platforms will be shown in the circles with the polarity of the sentiments as percentages.

4. Software Design Specification (SDS)

This chapter provides important artifacts related to design of our project.

4.1 Proposed System Model

Our pipeline consists of multiple data sources i.e The historical prices of Bitcoin with a resolution of 1 day. and 18 Million tweets regarding bitcoin from the year 2017 to 20. Preprocessing includes cleaning those tweets through various NLP techniques such as removing special characters, stop words, stemming and lemmatization. Then to extract the sentiments of those tweets, they are processed by Vader Analysis which assigns scores to every tweet taking into account the retweets, follower count and likes of that specific tweet. The historical prices along with the scores of the tweets are fed to the machine learning models to generate prediction for the next day. The model will be fed on daily basis via daily tweets and OHLCV values to generate buy/sell signals on a daily basis.

4.1.1 XGBoost Regressor

Since our dataset was in a tabular format, an XGBoost regressor would produce better forecasting results as proposed by this paper[6]. The model will be provided with the final score from the Vader analysis on daily basis, previous day's closing price and volume of the bitcoin traded on that day. Once the XGBoost regressor is trained on a period of 5 years, it will be tested on a period of one year. The evaluation metrics will be based on Percent chnage of actual prices from the previous day and predicted percentage change.

4.1.2 Stacked LSTM

In parallel with the XGBoost model, a stacked LSTM due to its ability to capture volatility will also be trained. However, since we will be using closing price, tweet scores and volume of bitcoin, LSTM will be trained for a multivariate data feed. Different lookbacks will tried and tested based on the evaluation metrics. Three three layers of LSTM with 50 neurons will be trained initially with a dropout of 0.2 on around 100 epochs.

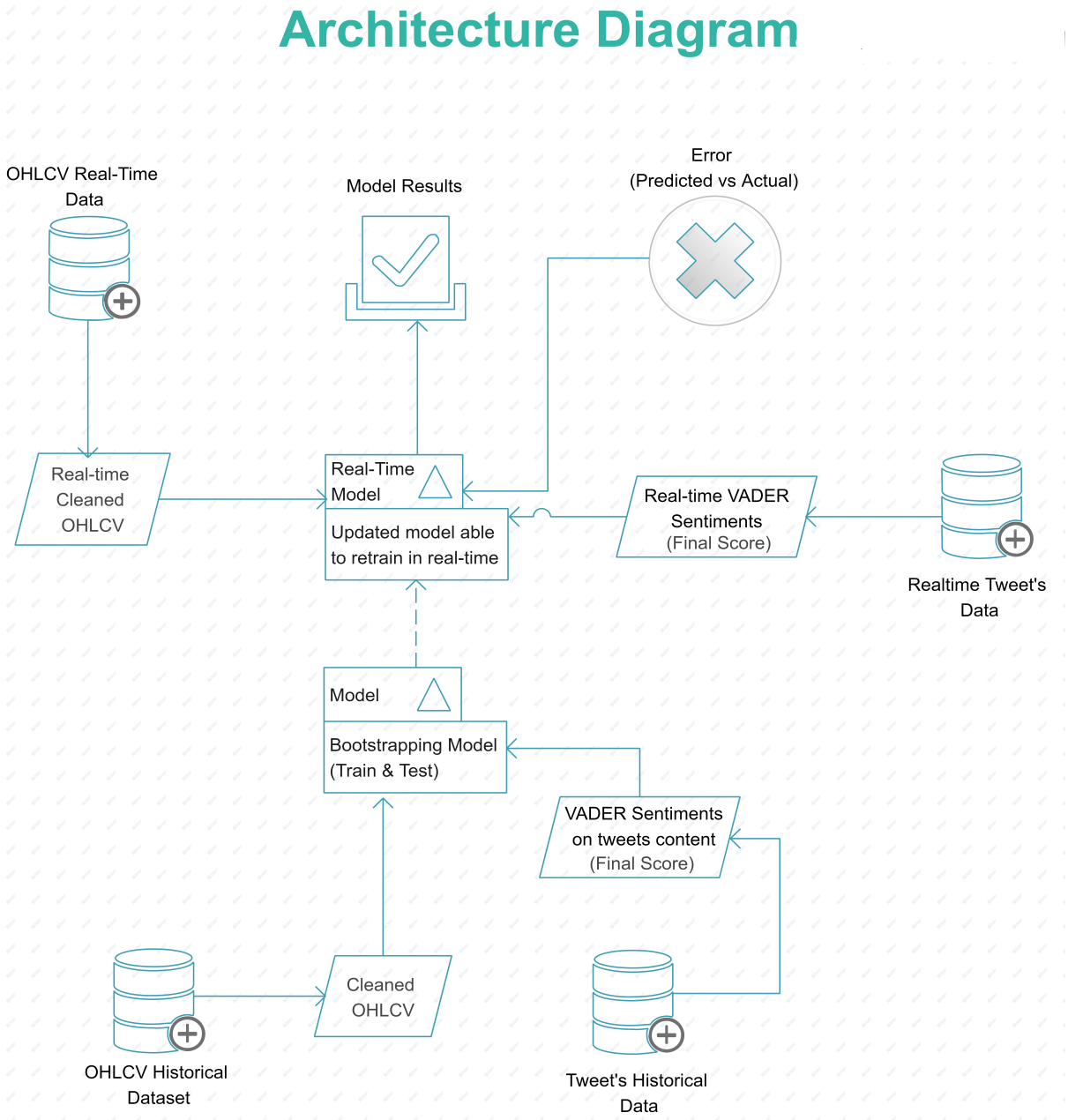
4.1.3 ARIMAX

One of the most prominent statistical models used in time series forecasting is the ARIMA Model. It is a regression based model which computes the lag between the predictions and the actual data and then trains on the the residuals. Extending it further, the ARIMAX is a multivariate model which takes into account exogenous independent variables to predict the next sequence. The most significant aspect about this model is the fact that it differences the data, making it stationary when it is trained. This helps it to generalize the model better, negating the most recurring problem of seasonality.

4.1.4 Other Data Streams

After initializing our model, we will be working on importing other data streams which includes but not restricted to, recent crypto related news and articles on which crypto is directly dependent. Other than that, as suggested by our external supervisor, since big economies in the world like USA and China move the world, their global economic events such as Monthly Inflation Reports, Unemployment, Interest Rates and other economic drivers would have an impact on the price movement of the Bitcoin. Therefore, these are the few data streams that we plan to incorporate in our research application and will help our model predict more efficiently, hopefully.

4.2 Architecture of Diagram



4.3 Experimental Design

Initially, we plan to experiment with multiple designs on our models and data streams, few of which we have implemented and discuss below.

The first approach would be divide all of our dataset in test and train sections, with 70-30 ratio. Then train our model over tweets' sentiment and OHLCV tables and analyze the test results. Comparative analysis of XGBoost Regressor and Stacked LSTM results will help us understand whether a neural network based model or decision-tree based ensemble model would be better for our use-case. Below are the initial results of XGBoost Regressor and Stacked LSTM model.

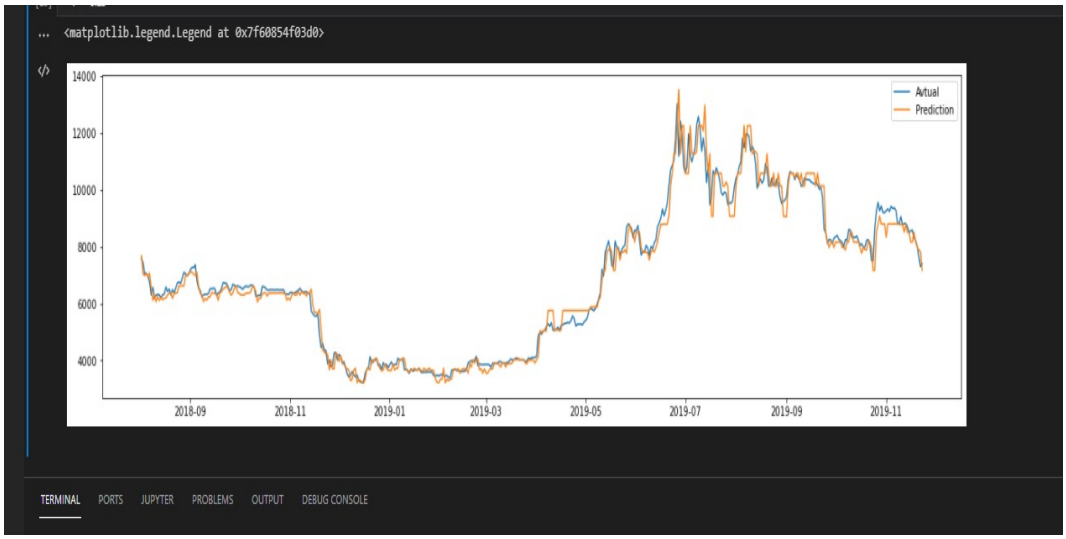


Figure 4.1: XGBoost Regressor Price Prediction vs Actual Price

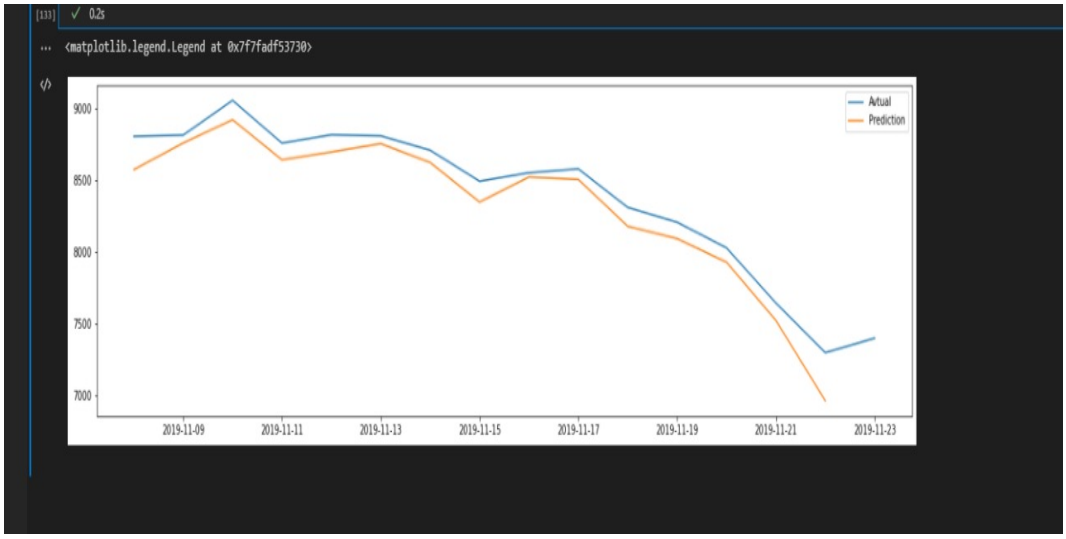


Figure 4.2: Stacked LSTM Price Prediction vs Actual Price

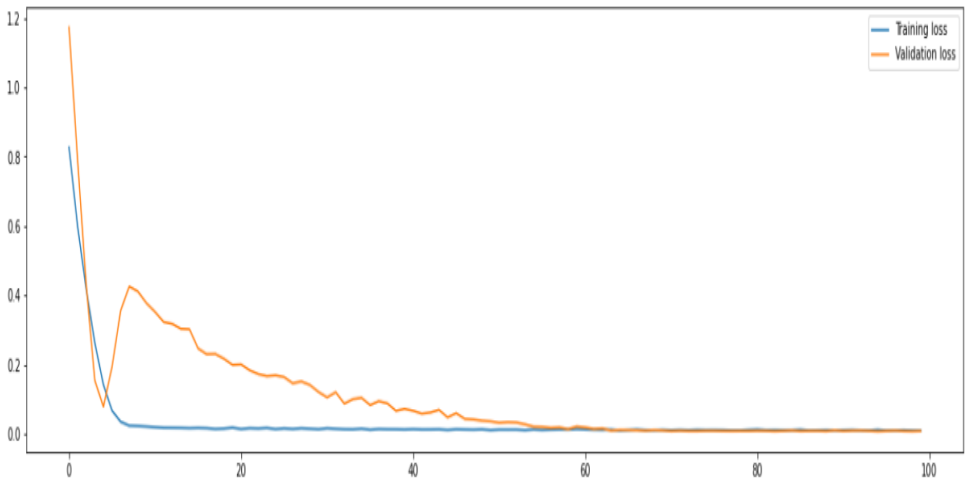


Figure 4.3: Losses of stacked LSTM w.r.t Epochs

In order to determine how our tweet's sentiment will be affecting our model, we will be training and testing our model over the OHLCV tables only. Then, analyzing that with our previous results will help us understanding the importance of tweet's data.

After analyzing all of the above cases, we will be integrating other data streams into our model, and will be analyzing and comparing the results to make the most out of our research.

4.4 Evaluation Metrics

Since our end product will be generating buy/sell signals, then it is vital that we evaluate our machine learning models based on those metrics. We plan to evaluate each of our model based on the hit rates. If the model predicts a certain amount of percentage change on a given day and the actual closing price changes equal to or more than that percentage change, that will be considered that as a hit. Conversely, if the model correctly predicts a certain amount of percentage change and the close changes less than that percentage change, then we consider that as a miss.

So lets say, the model predicts a PCT change of 0.1% for the next day and the actual close price changes positive 0.2% then it will be considered as a hit. If in this case the actual close price changes negative 0.1% percent then it will be labelled as a hit.

The hit rate signifies the proportion of correct predictions of hitting the target with respect to the total number of attempts.

5. Application

The final application consists of a frontend, backend, database and our models' pipelines attached to it. The details of each component is provided below:

5.1 Frontend

frontend details

5.2 Backend

The backend language being used in our application is Flask, a micro-framework for Python. The required APIs we've developed for the project are:

5.2.1 APIs Developed

/predict:

This API will be run as a cron job i.e. will run automatically for every hour. This will predict the percentage change in the next hour, for instance, if it is run 3 PM, it will predict the percentage change for 4 PM. Multiple pipelines developed will be integrated in this API, which will help in predicting the required value.

/get-latest:

This will be GET request, where the returned value will be the predicted close price for bitcoin for next hour.

/get-history:

This will be GET request with body containing the start and end timestamp. The returned values will be a list of actual and predicted close prices at every interval i.e. hour, during the given timestamps.

/get-sentiment:

This will be GET request with body containing the start and end timestamp. The returned values will be a list of public sentiment on twitter regarding bitcoin at every interval i.e. hour, during the given timestamps.

5.3 Database and Pipeline

Pipelines

All of the pipelines which will be integrated in our model are as follows:

- Fetch OHLCV parameters values from Polygon, a provider for such data for previous hour
- Fetch tweets related to bitcoin using twitter full archive developer's API for last hour
- Data cleaning of all the tweets and OHLCV values retrieved
- Sentiment Analysis of the tweets fetched using VADER Analysis, calculating their final scores and averaging it out.

Database

The database used for our application will be MongoDB (NoSQL). The data we're storing is all the OHLCV records collected, tweets' sentiment after calculating final score using VADER Analysis, actual close price, percentage change and our predicted close price of bitcoin. Initially we planned to store the tweets' content as well, but due to the legal obligations as stated in Twitter's regulation document for the APIs, we are required to not store them. The final schema of our database is as follows:

Table	
PK	ID: (int)
	pred_value: (float)
	actual_value: (float)
	pred_perc_change: (float)
	timestamp: (date)
	tweet_sentiment: (float)

Figure 5.1: Database Schema

6. Experiments and Results

6.1 Model Finalized

The model that was performing the best out of all was the ARIMAX model. We used the auto arima module of STATSMODEL API which converged the best on the P, Q, D values of (5, 1, 5) with the best AIC score. Regarding the exogenous features, our finalized features were open and low percent changes with previous days passed though an EMA indicator with a time period of 2. An Exponential Moving Average Indicator (EMA) is able to capture short term movements extremely well and since our model is based on short term predictions, an EMA indicator best represents the short term movements. The two indicators that we finalized were EMA2 and SMA2 which were passed through the close price percent changes of previous hours. And our final feature was the tweet sentiment which was resampled on an interval on 60 minutes. Our prediction variable was percentage change of the close price of the current hour and the next hour.

6.2 Data Effectiveness

The effectiveness of our features can be seen by the plots. The EMA2 and SMA2 indicators which were fed with the close percent changes w.r.t the current and previous hour and the close represents the percent change w.r.t the current and next hour. It can be seen that these indicators were able to catch around ninety percent of the volatility. Apart from that, the open and low pct changes along with the tweet sentiment had a good enough correlation with our prediction variable.

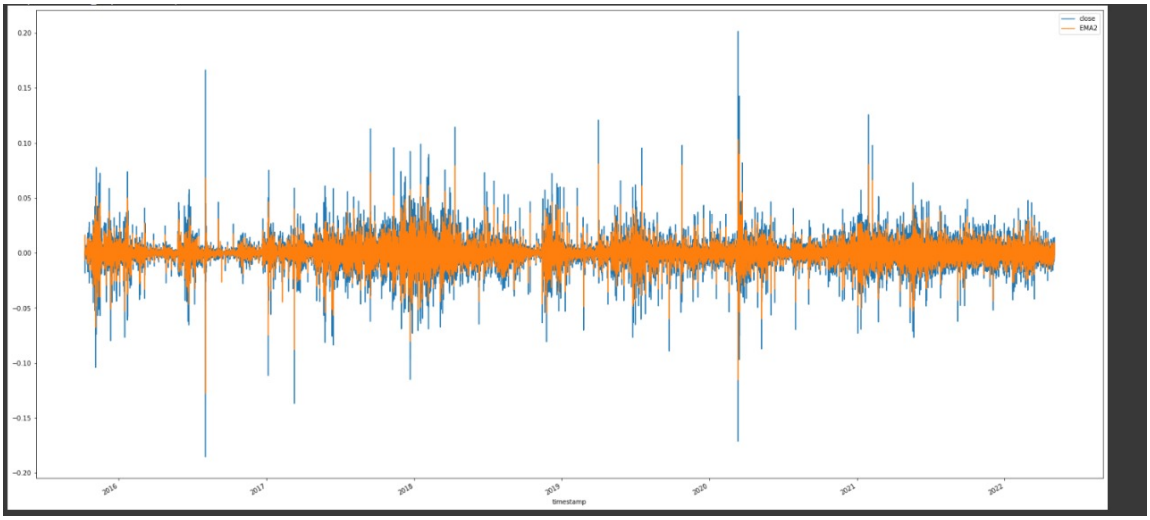


Figure 6.1: EMA 2 and Next Close PCT

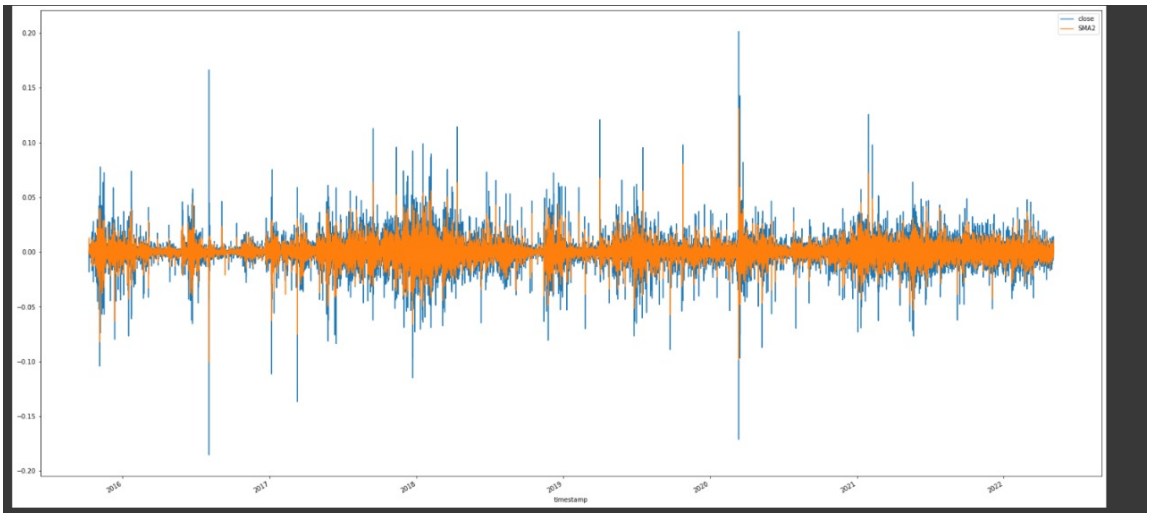


Figure 6.2: SMA 2 and Next Close PCT

6.3 Results

The most impressive feature of our selected model is that it has the ability to predict the percentage change of the most volatile coin in an extremely low resolution of 60 minutes with a long prediction accuracy of over 55% and short prediction accuracy of 65%. This may seem like normal accuracy figures but it is to kept in mind this is a model giving predictions every hour in a day, so to compare it with a daily model wouldn't be logical. It can be seen in the predicted percent changes that model is able to pick up long and short jumps just before the actual percent changes, which is the most difficult task in time series forecasting.

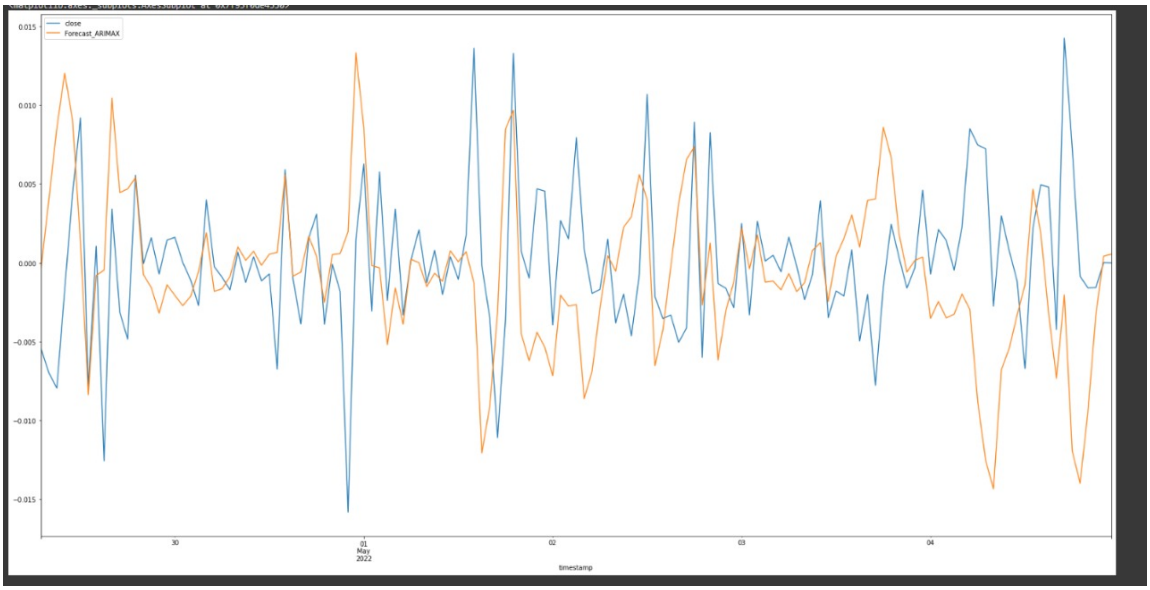


Figure 6.3: ARIMAX Predictions PCT change

7. Conclusion and Future Work

7.1 Conclusion

This project had both research and development aspects to it. The development came later on. Initially, we had to research on the domain by following the related papers published recently. Then researching about the modules that were mostly recommended by the researchers and experts in the field.

After deciding the models to be used and identifying the data streams to be used, we performed the comparative analysis over the results obtained from each. For collecting the dataset, twitters' official API had to be used which delayed our development but was worth it, because after providing all the details, and having discussion regarding the usage, we got access to the premium version of Twitter's API.

Using these APIs and finalizing models, developed an intuitive and simplistic frontend. Then our main focus was to implement all the pipelines which we designed initially as in our architecture diagram. The backend was then integrated with frontend to make the application work as expected.

The distinctive quality of this project that as explained technically in results, all the working can be used for any other coin as well. All we have to do is just to update the data streams as per the coin whose close price needs to be predicted.

7.2 Future Work

The most important aspect when predicting a commodity like bitcoin is to find out what are its driving features. We focused more on short term prediction and relied much of our work on technical indicators and tweets. However, our work can be further extended by figuring out the fundamentals of Bitcoin. Research may be conducted on economical indicators, Inflation of the Super Powers, World politics such as war etc. If a collection of features which drive the price of Bitcoin can be determined, then this model can be converted to make long term models. Access to these features often require subscriptions which was a barrier for us since this was an academic paper.

As a product point of view, we believe that the system developed, can be deployed on a Cloud service along with all the pipelines. After refining the model or maybe adding predictions and services for different digital currencies as well by updating the data streams of the model, services can be provide to authenticated users on subscription basis. We were unable to deploy the service due to the heavy computation required in our application due to which any cloud platform providing free services could not take the load.

Appendix A. More Math

We had sentiment of every tweet in our data using different NLP techniques, but in order to determine the relevance of a certain tweet the following formula was devised. This formula determines how relevant the tweet is, which is based upon the different parameters.

$$FinalScore = Tweet_{Sentiment} * (Tweet_{likes} + 1) * (Tweet_{retweets} + 1) * (Tweet_{followersCount} + 1)$$

Appendix B. Data

Here is a dump of our 2TB data set. Enjoy!

Appendix C. Code

Here is our code.

Our code can be found at `https://github.com/faizan-gc/FYP-BitForc`.

Bibliography

[1] Helder Sebastião Pedro Godinho (2021) *Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Financ Innov 7, 3*

[2] V. Derbentseva and V. Babenko and K. Khrustalevc and H. Obruchd and S. Khrustalovac (2021) *Comparative Performance of Machine Learning Ensemble Algorithms for Forecasting Cryptocurrency Prices*, International Journal of Engineering.

[3] Ioannis E. Livieris and Niki Kiriakidou and Stavros Stavroyiannis and Panagiotis Pintelas *An Advanced CNN-LSTM Model for Cryptocurrency Forecasting*

[4] "Crypto rally: Total market cap hits new all-time high of $2.8T$;1T added in just over a month's time - The Financial Express." https://www.financialexpress.com/market/crypto-rally-total-market-cap-hits-new-all-time-high-of-2-8t-1t-added-in-just-over-a-months-time/2362504/ (accessed Nov. 07, 2021).

[5] "Vista do Short-Term Forecasting in Bitcoin Time Series Using LSTM and GRU RNNs." https://sol.sbc.org.br/index.php/kdmile/article/view/11964/11829 (accessed Nov. 07, 2021).

[6] S. Mohapatra, N. Ahmed, and P. Alencar, "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments," Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019, pp. 5544–5551, Feb. 2020, doi: 10.1109/BigData47090.2019.9006554.