A picture containing graphical user interface

Description automatically generated

ISSS 610- Applied Machine Learning

Master of IT in Business

IS ELON MUSK MANIPULATING BITCOIN?

A Time-series Forecasting of Bitcoin Price based on Sentiment Analysis of Elon Musk’s Tweets

**Group 6**

Cheng Di

Joel Yeo Cher Boon

Karishma Yadav

Sean Samual Prajs

Zheng Wanyu

# **Abstract**

This report presents a study on the impact of Elon Musk’s tweets on bitcoin (BTC) prices through a two-fold approach. First, predicting the tweets’ sentiments towards BTC and second, incorporating these sentiments into a downstream time series price forecast together with other parameters as input features. The study encountered numerous constraints and limitations related to data availability and quality. Due to limited data points and high bitcoin price volatility in recent months, the price predictions were sub-optimal. However, the results from t-tests indicate that the Bitcoin Prices are indeed impacted by Elon Musk’s tweet sentiment.

1. **Introduction**

Bitcoin was created in 2009, following the 2008 global financial crisis. A time where many had lost faith in governments and banks. New bitcoins are created through “mining,” a process powered by computers worldwide. Bitcoin transactions are recorded using a distributed ledger technology with no single point of failure called blockchain. The cryptocurrency prices, which are unregulated, and part of a decentralized market meant to be impervious to any single party’s influence, seem to soar and plunge based on the actions of just one e, Elon Musk.

The following two Figures show the impact of Elon Musk’s actions on Twitter on the price of BTC and the number of trades in BTC:

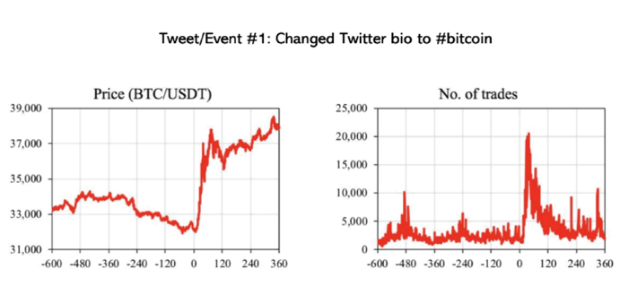


Figure 1: Tweet/Event- Changed Twitter Bio to #bitcoin

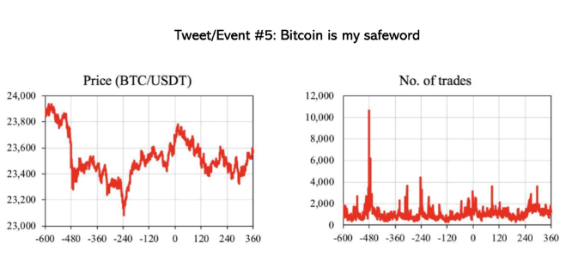


Figure 2: Tweet/Event- Bitcoin is my Safeword

From Figure 1, it looks like Elon Musk changing his twitter bio had a significant positive impact on the price and the number of Bitcoins traded. In Figure 2, however, his tweets did not impact the price. In this study, we aim to investigate if Elon Musk's tweet sentiment can determine the direction in which the Bitcoin price will move and if investors should trade BTC based on Elon Musk's tweets (sentiment).

# **Exploratory Data Analysis**

|  |  |
| --- | --- |
| Graphical user interface, chart, application  Description automatically generated  Figure 3: Elon Musk Monthly Twitter Activity  When looking into Elon Musk’s twitter activity, it is evident that he had made more (and better) use of the platform in the time leading up to this paper. The entrepreneur’s twitter traffic peaked in early 2021 and remained at high levels thereafter.  Judging by the monthly number of retweets and likes, as seen above, Musk’s twitter account was popular in nominal numbers, reaching 100,000 likes and 10,000 retweets at the beginning of 2021. |  |

Graphical user interface, chart, histogram

Description automatically generated

*Figure 4: Bitcoin Prices*

Bitcoin has gained popularity over time, which is reflected in the recent increase in its trading volume. The cryptocurrency has become a lot more volatile as its price started to hike (the difference is calculated as the opening price on any day, d, minus the opening price the next day, d+1).

Chart, histogram

Description automatically generated

*Figure 5: Bitcoin Closing Price with Bollinger Bands*

The coin’s price increase and volatility hike are further reflected in Figure 5. In the graph, the closing price with Bollinger bands is visualised. As shown, the distance between the upper band and lower band increased with time.

Graphical user interface, application, histogram

Description automatically generated

*Figure 6: Elon Musk Twitter Activity and Bitcoin Prices*

Figure 6 compares changes in Bitcoin price with Musk’s twitter popularity. The two seem to be correlated. The more retweets and likes Musk received the higher the Bitcoin price rose. This relationship drives the hypothesis of this paper and begs the question: is there a causal link between Elon Musk’s tweets and Bitcoin price developments?

# **Approach**

The overall approach for this study is segregated into the sentiment analysis of Elon Musk’s tweets and a time series prediction of BTC prices. First, sentiment models have been trained on tweet datasets with sentiment labels. These models were then applied to Elon's tweets to generate sentiment states. These were fed into time series price forecast models with prices of other crypto currencies and BTC itself.

The metric of evaluation of sentiment models chosen was **micro\_F1 score** due to dataset imbalances. The final evaluation of the price forecasting models was done using **RMSE**.

# **Datasets**

Datasets used for training the sentiment models included: BTC\_tweets.csv (with sentiment labels), Bitcoin Tweets dataset (without sentiment labels) and Vader\_btc.csv.

*BTC\_tweets.csv* contains about 28138 tweet entries and 10 features after null values and duplicates were removed. Among the features only ‘Tweet’ and ‘New\_Sentiment\_State’ were used for model training.

*Vader\_btc.csv* is a tweet dataset with sentiments calculated based on modified VADER analyzer. It has total of 265,771 entries and 6 features. Due to computational limitation, only first 60000 entries are used, before dropping null and duplicates. ‘tweet\_text’ and ‘overall\_sentiment’ are used for model training.

Sentiment prediction is applied on *ElonMuskTweetCombined.csv*, a dataset including all Elon Musk’s tweet from 01/01/2018 to 30/06/2021. The tweets from 01/01/2018 to 11/04/2021 are directly downloaded from Kaggle.com, and the rest (11/04/2021 to 30/06/2021) are scrapped from Twitter. There are totally 9808 entries with 14 features, ‘created\_at’, ‘tweet’, ‘retweets\_count’ and ‘likes\_count’, used in this study.

Price datasets include bitcoin prices, dogecoin prices and Ethereum prices, all downloaded from yahoofinance.com. Closed prices for all the cryptocurrencies were extracted for the time frame 1/1/2018 until 30/6/2021.

# **Part 1: Sentiment Classifiers**

# **Related work**

# Models used for sentiment analysis includes Valence Aware Dictionary for Sentiment Reasoning, support vector machine (SVM) and BERT.

# SVM is a simple, efficient, and robust classification algorithm and provides various kernels to implicitly map the inputs to high-dimensional feature spaces.

# BERT on the other hand, makes use of language and sequential understanding from the masked language modelling and next sentence prediction tasks on large Wikipedia corpora in its pre-trained stage. Model fine-tuning of BERT for countless task adaptations has reported many State-of-the-Art results in the field of NLP, and a task adaptation for Sentiment Classifier was adopted for the project.

**6.1 Valence Aware Dictionary for Sentiment Reasoning (VADER)**

The BTC Tweets sentiment labelled dataset, the only bitcoin tweets sentiment labelled dataset openly available was manually checked and it had some wrong labels. Therefore, we used Valence Aware Dictionary for Sentiment Reasoning (VADER), to train the bitcoin tweets. It is a lexicon-based sentiment model that considers Emoticons, social media slangs, exclamation marks, booster words and differentiates uppercase from lowercase words. The outputs from this model were used to train the subsequent sentiment models to further enhance sentiment accuracy by considering context of the tweet with respect to Bitcoin.

Graphical user interface, text, application

Description automatically generated

Figure 7: VADER Model Flow

# **Results for Models – SVM & BERT**

|  |  |  |
| --- | --- | --- |
| Micro\_F1 score | Trained on BTC\_Tweets | Trained on VADER\_btc |
| SVM | 0.955 | 0.947 |
| BERT | 0.938 | 0.925 |

*Table 1: Sentiment result*

**6.2 Support Vector Machine (SVM)**

A comprehensive data pre-processing is done for the text inputs, including demojize (convert emoji to corresponding text), lowercase, @username removal, http:// link removal, deconstruction, special characters and punctuation removal, stop words removal, single character removal and extra whitespace removal. Specifically, stopwords list downloaded from NLTK.corpus library is modified to remove words with potential sentiment indication, eg, can’t, most, etc.

10% of data is split out as test data, with the rest being training set. TfidfVectorizer is used to convert the raw text to a matrix of TF-IDF features to feed to the SVM model. The model is trained with 5-fold cross-validation and uses ‘f1\_mirco’ as training criteria. Grid search is performed to find optimal regularization parameter ‘C’ and classification kernel ‘kernel’. Among C = {0.01, 0.1, 1} and kernel = {‘linear’, ‘rbf’}, the search result suggests 1 and linear as best parameters. With the same set of settings, the SVM model is trained on BTC\_tweets and Vader\_BTC datasets and can obtain over 0.955 and 0.957 micro\_f1 score on the test sets respectively (Table 1).

The SVM models have achieved surprisingly satisfactory scores on the labelled datasets. The model takes in raw text and vectorizes the tokens based on TF-IDF features. When the test and training data come from splits of the same dataset, they may establish similar vocabulary, and thus similar TF-IDF features. When SMV model maps the data points to hyperplane for polarity classification, it is then able to achieve a high score. Having said that, the classification is essentially based on statistics of vocabulary rather contextual sequence understanding. When the same model is applied to the Elon’s tweets dataset, it may not be able to transfer the learnings and the sentiment prediction results may not be as accurate as shown on the labelled datasets. However, since the ground truth of Elon’s tweet sentiment is not available, the prediction outcome from SVM models cannot be verified. Therefore, the next BERT model is introduced attempting to predict sentiment state with contextual understandings.

**6.3 BERT**

A different pre-processing approach was taken for BERT than for VADER and SVM. As BERT’s architecture comprises of a Transformers Encoder and makes use of a WordPiece tokenizer along with Positional Embeddings to process text inputs. The positional embeddings are used to model how a token at one position attends to another token at a different position, and sequential data is better captured. As position of words in a sentence will affect the sentiment of the sentence, pre-processing steps such as Lemmatization and stop-word removals are not done. Stop-words can potentially provide context to the intent of the sentence and will be dangerous to remove. As BERT base cased model was implemented, lower casing of the text is not necessary as well. The rationale behind this is that sentiments might be affected by the casing of the words (stop vs STOP). Punctuations are not removed as well because BERT is able to handle punctuation tokens. The final pre-processing steps for BERT includes removing Retweet targets, removal of mentions, removal of URL links, removal of #, removal of control characters and conversion of Emojis into their respective name aliases (😊 emoji -> :smiling\_face:). Emojis are critical in analysing sentiments for tweets, and a proposed solution is to convert the emojis to a text form.

The micro-f1 score resulting from BERT was 93.8% and 92.5% respectively for the labelled BTC tweet dataset that was erroneous, and the VADER labelled dataset. These results are worse than the results from SVM but could ultimately be more accurate in classifying sentiments in Elon Musk’s tweets. This is so as both BTC dataset and VADER dataset had slight uncertainties with the accuracy of the Labels. The rationale for using 2 BERT models trained on BTC dataset and VADER dataset to predict a sentiment output for Elon Musk Tweet dataset is explained in section 9 under Limitations. Ideally, the hypothesis is that the better understanding of sequential text data and natural language understanding from BERT’s pre-train stage on much larger Wikipedia corpora than corpora for tweets related to cryptocurrency would produce more accurate downstream sentiment predictions. Fine-tuning was also done with cryptocurrency related datasets as the hypothesis was that the fine-tuned BERT model is able to better model keywords for Cryptocurrency tweets. As Elon Musk Tweet dataset was unlabelled, and unsupervised sentiment prediction directly on that dataset without fine-tuning creates the impression of more uncertainty, the above process was used instead. This is also explained in Section 9 under Limitations.

1. **Part 2: Timeseries Models**

**7.1 Data Preparation** **for the Price Prediction Models:**

Data Splitting:

Bitcoin prices, Ethereum price and Doge Coin price datasets were downloaded from yahoo finance. The sentiment scores were taken from the datasets generated from the sentiment analysis of Elon Musk's tweets in the first part of this paper. Root Mean Squared Error (RMSE) was used as the evaluation criteria as it gives an absolute number on how much the predicted results deviate from the actual number and could easily be interpreted for model evaluation and comparison. It was observed that the trend in bitcoin price changed significantly over the last couple of months of tha available data. Given that timeseries models are split into train, validation and test sets based on time, there was a drastic difference among the data that was used for training vs the one used for validation vs the one used for testing new data. To eliminate this discrepancy, we chose to slice the data where the price trend was not as varied and ran the same models on the sliced data to check for reduction in the evaluation metric, Root Mean Squared Error (RMSE).

Chart, histogram

Description automatically generated

A picture containing chart

Description automatically generated

Figure 8: Bitcoin Price Charts

The datasets were split into train, validation, and test sets as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data set | Total Time Frame | Training Set | Validation Set | Test Set |
| 1 | 01/01/2018-30/06/2021 | 01/01/2018-08/10/2020 | 09/10/2020-15/06/2021 | 16/06/2021-30/06/2021 |
| 2 | 01/01/2018- 04/10/2020 | 01/01/2018-27/05/2020 | 28/05/2020-04/09/2020 | 05/09/2020- 04/10/2020 |

*Table 2: Data set and Splitting Rules*

Data Pre-Processing:

Elon Musk’s tweets’ sentiments were obtained from the four different sentiment models mentioned above in Table 1 (SVM-BTC, SVM-VADER, BERT-BTC, BERT-VADER). The Arithmetic Mean was taken to transform the tweet sentiments, number of retweets and likes to their respective daily aggregates.

The Bitcoin, Doge Coin and Ethereum price data sets were combined with the outputs generated from the sentiment models built in part 1 of the project. Moreover, time zones for all the datasets were aligned and set to British Summer Time (BST). It was found that the price datasets had 4 days of missing data, which was replaced with the price of the cryptocurrency on the previous day. The tweet sentiment datasets generated above also lacked data for several days corresponding to cryptocurrency prices. On these the sentiment score was set as 0 because no tweet means no sentiment score. The features were scaled to the range (-1,1). Timeseries neural network models expect the input shape to be three-dimensional with dimensions [number of samples, timesteps, features]. The data was reshaped accordingly, with timesteps as a hyperparameter.

For evaluation, Root Mean Squared Error (RMSE) was used as the evaluation criteria as it gives an absolute number on how much the predicted results deviate from the actual number and can be interpreted easily for model evaluation and comparison.

**7.2 ARIMA - BASE STATISTICS MODEL**

ARIMA is a forecast algorithm frequently used in statistics and economics to predict future values based on past information. Although it is not a machine learning model, this project takes it as a statistical baseline for price forecast performance.

At first, the model was fitted with the BTC price only. The best result, RMSE = 2132.23, was achieved with order (p, d, q) = (2, 2, 1), where p is the order of the autoregressive model, d is the degree of differencing, and q is the order of the moving average model. Afterwards, the prices of ETH and DOGE were included as regressors, but the RMSE increased to 5064.40 (refer to Appendix).

**7.3 LONG SHORT-TERM MEMORY**

Long Short-Term Memory (LSTM) is a type of recurrent neural network that can learn the order dependence between items in a sequence. Given the gated architecture of LSTM, which has the ability to manipulate the model’s memory state (memorizing historical data), it is ideal for time series problems.

Models:

LSTM model with 1 LSTM layer, 1 dense layer (50), 1 dropout layer (10%) and 1 more dense layer (1). All the hyperparameters in each layer were decided by hyperparameter tunning. The input features used are all the cryptocurrencies (BTC, dogecoin and Ethereum), the sentiment scores from the BERT-VADER model, Ethereum Prices, Doge Coin Prices, Retweet Counts, Likes Count (timesteps = 7).

# **Input & Results**

|  |  |  |
| --- | --- | --- |
| Input Features | RMSE (Dataset1) | RMSE  (Dataset2) |
| Cryptos, Sentiment Score from SVM-BTC model, Retweet Counts, Likes Count based LSTM Model | 6676.832 | 262.585 |
| Cryptos, Sentiment Score from SVM- VADER model, Retweet Counts, Likes Count based LSTM Model | 3196.585 | 298.624 |
| Cryptos, Sentiment Score from BERT-BTC model, Retweet Counts, Likes Count based LSTM Model | 3131.265 | 253.834 |
| Cryptos, Sentiment Score from BERT-VADER model, Retweet Counts, Likes Count based LSTM Model | 3753.673 | 255.460 |

*Table 3: LSTM Models and Results*

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Figure 9: Model with Dataset1 Figure10: Model with Dataset2

# **Analysis**

With dataset 1, the smallest test RMSE was 3100 with around 40,000 bitcoin prices in test data, and both bias error and variance error are very large. The possible reason is there was a drastic difference among the data that was used for training vs the one used for validation vs the one used for testing new data which we have mentioned in section 6.3.1.

To prove the assumption, dataset 2 is used with the same input columns but a different data split (details can be found in Table 2: Data set and Splitting Rules). RMSE is smallest 253 with around 10,000 bitcoin prices in test data. The result improves a lot from RMSE 3100 with 40,000 value (7.5%) to 250 with 10,000 value (2.5%). So, the LSTM model performance much better when with the dataset2, which proves that the assumption is correct.

**7.4 GATED RECURRENT UNITS:**

Stacked Gated Recurrent Units were used in the models for deep learning for the datasets mentioned above. Sentiment scores from the 4 sentiment models described in part 1 were used to identify the one that gives the lowest RMSE for Bitcoin price prediction and further analysis was carried out to confirm the impact of Elon Musk’s tweet sentiment on Bitcoin price.

**MODELS & RESULTS**

Stacked GRUs with 3 GRU layers and 1 dense layer were used for all the models. The input features used are Bitcoin Price (timesteps = 7), Sentiment Score from the BERT-VADER model, Ethereum Prices, Doge Coin Prices, Retweet Counts, Likes Count.

Several combinations of number of GRU units, batch size, timesteps were tried to arrive at the optimum values of 64 batch size and 7 timesteps that gave the least RMSE. Early stopping with a patience value of 10 was used to run 1000 epochs to arrive the least train & validation losses.

|  |  |  |
| --- | --- | --- |
| Sentiment Model Output used | RMSE (Dataset1) | Result Charts: Predicted vs Actual Bitcoin Prices |
| SVM-BTC Sentiment score based GRU Model with Retweet counts, Likes Counts, ETH and Doge Prices | 2871.87 | Chart, line chart  Description automatically generated |
| SVM-VADER Sentiment score based GRU Model with Retweet counts, Likes Counts, ETH and Doge Prices | 3227.28 | Chart, line chart  Description automatically generated |
| BERT-BTC Sentiment score based GRU Model with Retweet counts, Likes Counts, ETH and Doge Prices | 2630.22 | Chart, line chart  Description automatically generated |
| BERT-VADER Sentiment score based GRU Model with Retweet counts, Likes Counts, ETH and Doge Prices | 1979.14 | Chart, line chart  Description automatically generated |

*Table 4: GRU Models and Results for different sentiment inputs*

**Analysis**

The results show that for GRU, the BERT-VADER Sentiment score based GRU Model has the least RMSE, therefore BERT-VADER sentiment score is used as input for further experiments. The RMSE for BERT-VADER sentiment score input could be due to better sentiment prediction by the sentiment model. This result enables a weak but useful verification of the hypothesis in section 6.3. BERT Sentiment outputs appear to capture Elon Musk’s tweets sentiment in the context of Bitcoin better than the other models.

**7.5 Bi-LSTM**

Building on the GRU results where the best performance was given by running on the BERT-VADER dataset, a last model experiment was conducted with a stacked BI-LSTM architecture on only the BERT-VADER dataset. The purpose of this section is to serve as an add-on to compare if the more complex architecture of an LSTM, coupled with a bi-directional input sequence capture would allow for better forecasting than a unidirectional GRU. The following model architecture was implemented for the experiment. A bi-directional LSTM layer was connected to a dropout layer, followed by another bi-LSTM and dropout layer. This was then connected to a final bi-LSTM layer and dropout layer, before entering a fully connected layer.

# **Experiments**

We were interested to find out if the Bi-directional nature of Bi-LSTM could help to improve the forecasting ability of the models. The same evaluation metric of RMSE was used for the model, and the model was subsequently trained with the 2 different timeseries as explained previously in section 4.3.1. Furthermore, the same configurations were kept from the GRU experiments (GRU models 1 & 3), and the same testing range was used as well for a direct comparison.

Hyperparameter tuning was performed using random search in keras tuner, and the following hyperparameters were decided upon. Dropout rate for layer 1 = 0.4, Dropout rate for layer 2 = 0.30000000000000004, Dropout rate for layer 3 = 0.2, batch size = 32 and learning rate = 0.0001. Likewise for the other models, Adam optimizer was also used. The experimental BI-LSTM model was only done on the BERT-VADER dataset as it was found to have the best results from the GRU model experiments.

# **Results**

|  |  |  |
| --- | --- | --- |
| Dataset | RMSE | Results Charts: Predicted vs Actual Bitcoin Prices |
| 1 | 2190.75 | Chart, line chart  Description automatically generated |
| 2 | 265.13 | Chart, line chart  Description automatically generated |

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Model | Input Features | Test Set RMSE |
| 1 | Bi-LSTM | Bitcoin Price (timesteps = 5), Sentiment Score from the BERT-VADER model, Likes, Rt, ETH, Doge | 2190.75 |
| 2 | Bi-LSTM | Bitcoin Price (timesteps = 7), Sentiment Score from the BERT-VADER model, Likes, Rt, ETH, Doge | 265.13 |

*Table 4: Bi-LSTM Models and Results*

# **Analysis**

The results obtained for Bi-LSTM are worse than the results obtained from GRU. A possible explanation for this is that the dataset is extremely small. GRUs perform better than LSTM architectures in scenarios where the dataset size is small, and the results shown confirm this finding. A potential improvement would be to consider using Bi-directional GRUs instead of LSTMs.

1. **T-Tests on GRU Models**

For the 2 datasets, T-Tests were conducted on the test set predictions for input features with and without the Elon Musk’s tweet sentiments, to confirm that the difference in predictions (lower RMSE) for both was statistically significant.

The following models were built for predicting the Bitcoin Price:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data set | Model | Input Features | Test Set RMSE | T-Test Results |
| 1 | GRU Model 1 | Bitcoin Price (timesteps = 5), Sentiment Score from the BERT-VADER model | 2093.69 | p-value: 0.0230 |
| 1 | GRU Model 2 | Bitcoin Price (timesteps = 5) | 2564.75 |
| 2 | GRU Model 3 | Bitcoin Price (timesteps = 7), Sentiment Score from the BERT-VADER model | 198.10 | p-value: 0.0085 |
| 2 | GRU Model 4 | Bitcoin Price (timesteps = 7) | 321.44 |

*Table 5: GRU Models and Results for different features and time frames*

Plots showing the results, Predicted Bitcoin Prices vs Actual Bitcoin Prices for the GRU Models

GRU Models Results for Dataset 1:

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 11,12: Stacked GRU Models 1 & 2

GRU Models Results for Dataset 2:

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 13,14 : Predicted vs Actual BTC price for GRU Models 3 & 4

# **Analysis**

Table 5 and Figures 11,12,13 & 14 show that Dataset 1 models in general, have significantly higher RMSE than that of Dataset 2 models, the bitcoin price trendlines are also better for Dataset 2 models when compared with those of Dataset 1 models. Therefore, it can be inferred that the timeseries models predict better when the training, validation and set data have the similar trend and not very stark differences which are present in dataset 1.

The results for the GRU models show that the RMSE for GRU models for both the datasets/timeframes is lower for the models where Elon Musk’s tweet sentiment is included as the input feature vs the ones where Bitcoin price alone is used as the input feature. Additionally, even the trend lines are better fitted for the model where sentiment scores of Elon Musk’s tweets are included, overall suggesting that the sentiment of Elon Musk’s tweets does have an impact on the price of Bitcoin.

To confirm the hypothesis, T-Tests for sample independence were conducted and the p-values were found to be 0.0230 < 0.05 (significance level) and 0.0085 < 0.05 (significance level) for dataset 1 (GRU Model 1 and GRU Model 2) & dataset 2 (GRU Model 3 and GRU Model 4) respectively. Therefore, the Null Hypothesis that the RMSEs for dataset 1 and dataset 2 are the same was rejected at 95% confidence level and it was concluded that the difference between the RMSEs was statistically significant at this confidence level. Hence, Elon Musk’s tweet sentiment does affect Bitcoin price quite significantly.

1. **Limitations of Project**

There were several limitations that we encountered while proceeding with the study. The limitations will be explained in 2 parts, Sentiment Limitations and Forecasting Limitations.

Firstly, for Sentiment Limitations, there was a lack of available and trusted datasets with correctly annotated labels for Bitcoin Tweet Sentiments. While we initially sourced a dataset with labels, we quickly realised that this dataset was actually erroneous in its labelling. We crawled various data sites and bases and was unable to find a good, labelled dataset regarding sentiment towards Bitcoin, or Cryptocurrency. A consultation via email was sent to the NLP professor, to check if there were ways that we could circumvent this problem. He verified that one of the possible solutions we proposed - the use of a rule-based model (VADER) with custom tweaked weightages of words to output a base sentiment score on a dataset, followed by using a machine learning model to train and output a final sentiment score was a viable option. Primarily, the benefits of using a pre-trained model like BERT to train again on the VADER output was to hopefully exploit the latent understanding of sequence and language from the masked language modelling on huge corpora in the BERT pre-training stage. As we were not able to verify confidently that our sentiment outputs were as accurate as it could be, we introduced the use of SVM as well, and adopted various experiments downstream in the forecasting models to attempt to see the best sentiment model output to use for future tweets.

Secondly for forecasting limitation, the dataset is extremely small, as Elon Musk’s tweets are widely reported to only affect Bitcoin and Cryptocurrency prices from 2018 onwards. We also had to perform a test, validation, and test split, that further reduced the available training data.

Chart, histogram

Description automatically generated

*Figure 15 Historical Bitcoin Price*

Furthermore, as can be seen from the Bitcoin price graph, an extremely huge spike was recorded after October 2020. This characteristic regarding the volatility of the price data is problematic as our training data would have only captured part of the spike, validation data would have captured most of only the dip, while the test data would have captured a relative plateau. Forecasting on a plateau after training on the spike and validating on the dip would cause predictions to become extremely erroneous.

Thirdly, windowed time lagged cross correlations (WTLCC) between bitcoin with dogecoin or Ethereum was not consistent across different periods of the time series especially for bitcoin against dogecoin, the observed correlation is very weak near in 2020. This means that the different time series are not always highly and consistently correlative with each other, and the results of forecast modelling across different time periods of the time series will be affected accordingly.

Background pattern

Description automatically generated Background pattern

Description automatically generated

Figure 16: WTLCC between Figure 17: WTLCC between bitcoin and dogecoin bitcoin and Ethereum

1. **Post study**

The problem with regards to volatility of the Cryptocurrency timeseries is confirmed by the Bollinger Bands technical indicator commonly used in finance to chart volatility of prices. Loosely, the wider the bands (in orange), the more volatile an instrument is during that period. Bitcoin and most other cryptocurrencies have fluctuating Bollinger band ranges at different points in their timeseries, with an extremely wide band occurring largely after October 2020.

Chart, histogram

Description automatically generated

*Figure 18:* *Bitcoin price chart with Bollinger bands*

The group then decided to cut-off the data before October 2020 and test our hypothesis of whether time series forecasting of Cryptocurrency prices on less volatile periods would at least give more favourable results. This is done with ***dataset 2*** in the price forecasting models and much better RMSE scores were reported than on the initial timeframe. Minimally, we can deduce that machine learning models are not suitable for price forecasting of extremely volatile instruments, such as Cryptocurrency. Perhaps when more data is available in the future, and more characteristic spikes and dips of volatile periods are captured by the timeseries, a different conclusion could be arrived at.

Another post-study done by the group was to experiment with traditional statistical models, and other machine learning models for timeseries forecasting. Statistical models are traditionally more popular than machine learning models for instrument-price forecasting as they can better capture lag elements such as moving averages and make use of statistical methods to detrend. Models such as ARIMA, Prophet and Autoregression were performed on Bitcoin prices alone in ***dataset 1***, to check if these traditional methods would have performed better under volatile conditions. A summary of the best results is shown in the following table.

|  |  |
| --- | --- |
| Models Used | RMSE |
| GRU (Best ML model) | 1979.14 |
| ARIMA | 2132.24 |
| PROPHET | 24416.72 |
| AUTOREGRESSION | 1664.04 |

A simple Machine Learning model such as PROPHET gave horrendous results as it was unable to account for the volatility present. ARIMA is relatively poor in predicting series with major turning points and the underlying theoretical statistical relationship are not as distinct as simple forecast models like autoregression. GRUs performed much better than the other LSTM models, as our dataset size is extremely small and limited. This characteristic makes GRUs an ideal choice over LSTM machine learning models if a machine learning model is chosen to be implemented for cryptocurrency cases. Lastly, Autoregression has a stronger underlying statistical relationship for timeseries than a vanilla ARIMA model. Autoregression models are also suitable for timeseries to model using lags of itself, rather than moving averages. This is perhaps one of the reasons why Autoregression model performed the best when predicting the volatile time. Another key factor is that by using lags of itself, autoregression can better model for the volatile hikes and dips than the machine learning models we have implemented.

# **Conclusion**

To conclude, our study explains the many limitations of attempting to forecast Cryptocurrency prices using Sentiment from Elon Musk’s tweets. These limitations firstly include a limited dataset where Elon Musk’s tweets only started to affect Cryptocurrency prices from 2018 onwards. Next, there is a lack of accurately labelled sentiment tweet datasets with regards to Cryptocurrency. A manual annotation of future datasets would be more beneficial than using an unsupervised rule-based method. Also, it is extremely hard to model for lags and accurately measure the duration of the impact of a particular tweet on Cryptocurrency prices as each tweet impact does not follow any pattern.

Furthermore, with the countless Macroeconomic and financial indicators available, it is almost impossible to model for all news and indicators which perhaps might have affected the forecast. There were thoughts of using USD prices, Macroeconomic indicators such as GDP, CPI and Money Supply indexes, but it was found to be extremely difficult to model these indicators into the dataset based on the contradictions from Economic and Financial Theory. This conclusion is not groundless, as a member of the group has an Economic and Finance degree.

Next, the volatility of Cryptocurrency renders most models almost useless in financial and trading use-case. The high RMSE reported across all models, including models that did not integrate tweet sentiment, would simply mean that the trader would lose money most of the time by using statistical or machine learning models. Contrary to results reported on online resources such as Kaggle, the projects done then were all ‘luckily’ modelling for the period where Cryptocurrency did not experience the increased volatility. This can be seen when the time period is matched to the Figure showing the Bollinger bands.

This study happens to model for Cryptocurrency after the volatility spike was experienced, which contributes to a huge reason why the reported RMSE was not ideal. The study concludes that minimally for Cryptocurrency forecasting on a limited and volatile dataset, statistical models that model lags outperform machine learning models. However, these statistical models are also not usable as the results reported would still mean that the traders would have lost monies invested. Both machine learning and statistical models do not outperform technical indicators or fundamental analysis strategies for Cryptocurrency.

There are 2 underlying reasons for this finding. Firstly, Cryptocurrencies are not store of value unlike other instruments due to its volatility. Price hikes and dips can be completely random, or due to sentiment impacts not only directly from Elon Musk’s tweets, but also from general retail investors (think reddit). Modelling for all possible scenarios and factors would be almost impossible, and the dimensions of the features would increase exponentially. Secondly, volatile time series have different characteristics across different time periods. A portion of the full time series could be stationary, yet another portion could be non-stationary. Likewise, models predicting on non-volatile periods would appear to be usable, but this is a false conclusion as seen in the result comparisons between ***dataset 1 and dataset 2*** when the future suddenly appears to become volatile. This underlying reason is the concept of financial random walk theory for instrument and stock prices, where on average in an extended period, prices will follow a random walk movement. This theory is a huge explanation of why models that perform better in certain time periods of the time series will fare badly for another time period and is widely expounded on in Finance and Economics when applying statistical models as well.

Finally, even though the result of the study is not ideal, the conducted T-test between Bitcoin prices and Sentiments from Elon Musk’s tweets confirm that there is indeed a relationship between Bitcoin prices and Sentiments. Going forward, this information will allow users to use Sentiments as a feature in attempting to predict Bitcoin, or Cryptocurrency prices. Implementation of new models such as Self-Attention will also perhaps provide new insights or push new boundaries in usable Forecasting models for volatile instrument price forecasting.

# **References**

Ahmad, M., Aftab, S., Salman, M., & Hameed, N. (2018). Sentiment analysis using Svm: A systematic literature review. *International Journal of Advanced Computer Science and Applications*, *9*(2). <https://doi.org/10.14569/ijacsa.2018.090226>

“Bitcoin USD (BTC-USD) Stock Historical Prices & Data - Yahoo Finance.”” *Finance.yahoo.com*, <https://finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD>

Delobelle, P., & Berendt, B. (2019). Time to take emoji seriously: They vastly improve casual conversational models. *arXiv preprint arXiv:1910.13793.*

“Dogecoin USD (DOGE-USD) Price, News, Quote & History - Yahoo Finance.” *Finance.yahoo.com*, <https://finance.yahoo.com/quote/DOGE-USD/history?p=DOGE-USD>

“Ethereum USD (ETH-USD) Price, News, Quote & History - Yahoo Finance.” *Finance.yahoo.com*, <https://finance.yahoo.com/quote/ETH-USD/history?p=ETH-USD>

Mercado, A.. “BTC Tweets Sentiment - Dataset by Mercal.” *Data.world*, 8 Aug. 2018, <https://data.world/mercal/btc-tweets-sentiment>. Accessed 15 Aug. 2021.

Olteanu, A. “All Elon Musk’s Tweets.” *Kaggle.com*, 17 Apr. 2021, <https://www.kaggle.com/andradaolteanu/all-elon-musks-tweets>

Smith, T. “Random Walk Theory.” *Investopedia, 2019*, <https://www.investopedia.com/terms/r/randomwalktheory.asp>.

Yang, S., Yu,X., & Zhou, Y.(2020). "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI), 2020, pp. 98-101, doi: 10.1109/IWECAI50956.2020.00027.

**Appendix**

**ARIMA Forecast Results**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

**PROPHET Forecast Results**

RMSE of PROPHET model = 24416.72

Chart

Description automatically generated

Chart, line chart

Description automatically generated

**AUTOREGRESSION Forecast Results**

A loop was done to find the best lags to consider for forecasting. Results of various lags are reported in the table below, with the optimal lag being 21 days.

|  |  |  |
| --- | --- | --- |
| N Lags | Prediction Chart | RMSE |
| 14 | Chart, line chart  Description automatically generated | 1664.04 |
| 4 | Chart, line chart  Description automatically generated | 1743.73 |
| 2 | Chart, line chart  Description automatically generated | 1750.24 |

Autoregression pre-processing includes performing an Augmented Dickey-Fuller test to check for stationarity. The null hypothesis of the ADF test is that a unit root is present. In simpler terms, H0 refers to a time series being non-stationary. If we cannot reject H0, the timeseries is non-stationary and we have to perform differencing or log-differencing to make the timeseries stationary for the statistical autoregression to work. A detailed procedure is chartered in the AMLProjectProphetAR py file.