Dogecoin price prediction using Sentiment Analysis

By:

Group 8

BYGB – 7977 – 002

05/06/2022

Arpita Choudhury

Shayna Lue

Siangling Hsu

Zhiyi Tan

Executive Summary

Software engineers Billy Marcus and Jackson Palmer created dogecoin in late 2013. In early 2021, dogecoin gained cult status on Reddit's WallStreetBets message board where enthusiasts had promised to propel its value "to the moon". This study aims to build a model to find the trend of this cryptocurrency price.

For our study we decided to go with dogecoin specifically because it has shown to be growing in popularity and has gone through major growth periods in 2021. Although there are a few significant studies that are available yet, the cryptocurrency has had huge growth in terms of volume and value. Many people are becoming interested in investing in cryptocurrency. This is what persuaded us to focus our study on predicting the price of dogecoin.

The study mainly focused on sentiment score on twitter posts related with dogecoin. It has two parts. Firstly, the study tries to find the best LSTM model to predict the price of dogecoin. The best model is using sentiment score from Vader and close price. Even though it has the lowest LMSE, it is hard to turn LSTM model into an investment strategy. As a result, the second part of the study wants to predict the general trend of the dogecoin price. It uses the positive component of Vader to predict the rebound point in the trend of dogecoin price.

Problem Statement

As technology continues to advance, cryptocurrency is becoming more popular. Similar to the stock market, the price of cryptocurrency fluctuates. This can cause those that will invest in a cryptocurrency like dogecoin to become skeptical about making an investment. Digital currency also has decentralized system settings, which might affect the accuracy of predictions.

Every year social media continues to grow and become more influential. A single social media post can have a positive or negative effect on the dogecoin price. A popular social media platform, 'Twitter' has shown to have a strong influence on dogecoin prices and trading volume. Especially tweets made by public figures such as Elon Musk. For our project, we decided to look at social signals and sentiment analysis for the prediction of prices of dogecoin to find the best strategy for investing.

Business Goal Analysis

To address the goal of predicting the price of dogecoin, most investors used a form of statistics. They are looking at dogecoin's historical data, such as their patterns, movements, and price charts. Others are choosing to look at political and economic news, whether domestically or globally. Some may even take a quantitative approach by looking at the overall market's

performance history. Text analytics is needed for our study because it allows an investor to get a better understanding. Since dogecoin is a digital currency whose price changes by the second, text analytics can provide a different approach to making a data-driven decision. Using text analytics will give investors the ability to convert unstructured text data and organize it to see insights and trends of dogecoin. Using text analytics can help investors achieve a positive return in their investment. Plus, more companies are using dogecoin as a viable form of payment. This includes companies such as Twitch, Tesla, and AMC Theaters. As dogecoin continues to grow and become more popular it will not only help them, but the companies that use them as payment as well. Also, text analytics can help give a better condensed visual of what other dogecoin investor are doing or what direction their opinions are taking on Twitter

Dataset Description

1. Tweet Data

The Twitter data is scraped using the Twitter Search Scraper. The tweets are scrapped in the English language with the keyword 'dogecoin' and the date range is set from 12/01/2021 to 03/31/2022 (4 months). The dataset included columns like User, Date, Retweet Num, Reply Num, and Like Num. The Retweet Num is taken into account to know if there are any important tweets that had a significant number of retweets or a higher number of likes. The post is used to calculate the sentiment score using different approaches. Overall, a total of 531,967 tweets are collected.

	Date	User	Language	Tweet	ReplyCount	RetweetCount	LikeCount
7	2021-12-23 23:56:50+00:00	CryptoPricing	en	Bitcoin Price (USD): 50793.5 \nEthereum Price ...	1	1	1
8	2021-12-23 23:56:42+00:00	arccaz	en	#Dogecoin now watch it 🚀 #doge1 @elonmusk	6	38	137
9	2021-12-23 23:56:18+00:00	CryptoJolly98	en	@arccaz @Cupidoge #dogecoin 💯💯💯💯 much wooooow ...	0	2	4
10	2021-12-23 23:56:07+00:00	bmurphypointman	en	#surprise #bitcoin #tumblr #twitter #facebook ...	0	0	2
11	2021-12-23 23:55:55+00:00	smile_sessions	en	the year is 2027 I sit down in my amazon gamin...	2	0	3
12	2021-12-23 23:55:44+00:00	oscarfrazier	en	Chart Wars: Why Dogecoin, Ethereum Classic's B...	0	0	0

Table 1. Example of Twitter Data

2. Dogecoin Market Data

The historical market price of dogecoin was obtained for the same date range as that of the Twitter data by requesting the Binance resource URI. Next, a JSON object was created to fetch the results. This response object is used to access certain features like the content and the column headers. The dataset included columns like open, close, high and low price, volume and time. For the analysis, only the close price is taken into account.

open_time	open	high	low	close	volume	close_time	quote_volume	trades
2021-12-01 21:37:00	0.2073	0.2075	0.2072	0.2075	176315.0	1638412679999	36572.6341	75
2021-12-01 21:38:00	0.2074	0.2075	0.2074	0.2075	91701.0	1638412739999	19027.4998	60
2021-12-01 21:39:00	0.2076	0.2076	0.2072	0.2073	216619.0	1638412799999	44922.5271	65
2021-12-01 21:40:00	0.2072	0.2073	0.2071	0.2072	221448.0	1638412859999	45882.8145	94

Table 2. Example of Dogecoin Price Data

An obvious observation from the above table is that the price is changing every minute. For the ease of visualization as well as analysis, the time in minutes is converted to hours by taking the average hourly price.

The figure 1 shows the trend of the dogecoin price on an hourly basis. The price showed a significant rise in the month of November and later started to decline till March 2022.

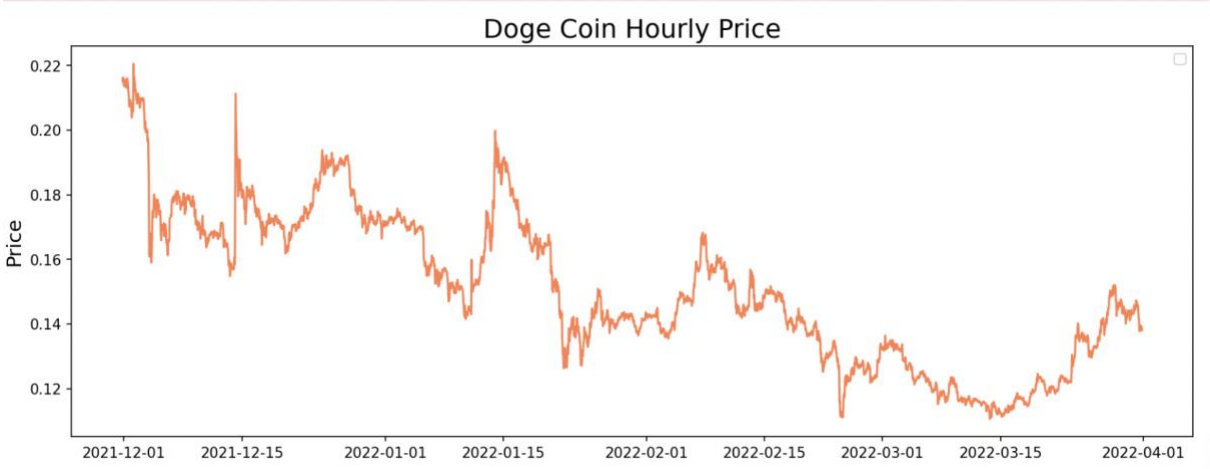

Figure 1. Dogecoin hourly price from 12/1/2021-03/31/2022

As for second part study, we use daily price rather than hourly price. We download it from Yahoo Finance in the same data range.

System Design

Our analysis includes two parts. The first part is mainly for predicting the price of dogecoin. The second part is to analyze the relationship of rebound point and tweets sentiment score.

1. First Part - Predicting the Price of dogecoin:

    a. Preprocessing: The date of tweets posted are four hours late compared to the twitter website's posts. We adjust the date back to same as what we saw on twitter. To perform sentence-level sentiment analysis, we only turn all the terms into lower case.

    b. Grouping into hourly data: After having each sentence's sentiment score, by grouping with an hour, we use average score on every tweets. As for market data, the close price for an hour is the last one close price in minutes dataset.

    c. Selecting Variable: We try different variables as input of LSTM model. The detail description will be depicted in the System Implementation part.

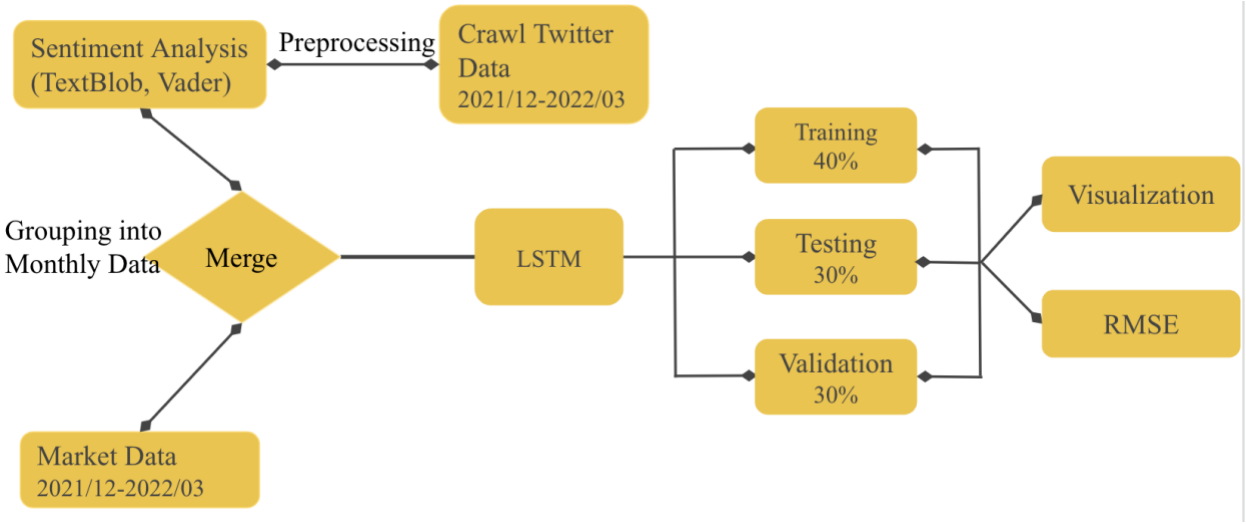

Figure 2. The Workflow For the First Part - LSTM Model Prediction

2. Second Part - Predict the General Trend of the Price :

    Preprocessing and Grouping: We remove the stopwords in each tweets and turn all the letter into lower cases. Before performing the Vader, we combine same day tweets into a document. To be more specifically, we want to perform a document (all the tweet happened in a day) analysis rather than a sentence (each tweet) analysis.

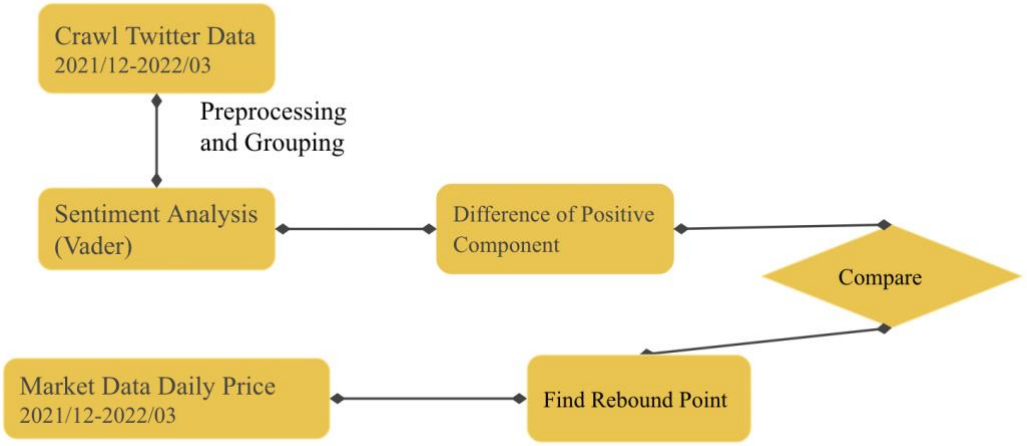

Figure 3. The workflow for the Second Part - Predict the General Trend of the Price

System Implemention:

1. Input of LSTM Model

   After merging, we got table 3. We choose each of variable and close price as LSTM models input.

   Variable Explanation:

   a. close : The close price in each hour.
   b. score_TextBlob: The sentiment score calculated from TextBlob.
   c. score_TextBlob2: The sentiment score calculated from TextBlob multipled with number of retweet.
   d. compound_NLTK: The sentiment score calculated from NLTK Vader.
   e. compound_NLTK2: The sentiment score calculated from NLTK Vader multiplied with number of retweet.

open_time	close	score_TextBlob	compound_NLTK	score_TextBlob2	compound_NLTK2
12/1/21 0:00	0.2136	0.1573	0.1756	0.4704	0.5253
12/1/21 1:00	0.2151	0.1496	0.2242	2.0514	3.0733
12/1/21 2:00	0.2156	0.1334	0.1922	0.3177	0.4577
12/1/21 3:00	0.2144	0.1656	0.2297	0.0831	0.1153
12/1/21 4:00	0.2145	0.0337	0.1875	0.0793	0.4422
12/1/21 5:00	0.2136	0.0922	0.1750	0.1176	0.2231
12/1/21 6:00	0.2134	0.0995	0.1618	0.1363	0.2218
12/1/21 7:00	0.2131	0.1076	0.1943	0.1000	0.1807
12/1/21 8:00	0.2142	0.0881	0.1979	0.1342	0.3015
12/1/21 9:00	0.2160	0.1126	0.2026	0.1755	0.3158

Table 3. Example of input data for LSTM model

2. Parameters of LSTM Model

We use Keras package to implement LSTM and set input layer as (3,2) , using 3 historical price (t-3,t-2,t-1) to predict price at time t. Another variable 2 means that we are input 2 parameters, like close price and score of TextBlob. For training model, we use mean squared error as loss function , Adam optimizer , learning rate with 0.01 and validate with validation data. To avoid the baseline becomes too perfect, we add one dropout with 0.5 value in the LSTM model and only use 10 epochs. We run 5 models for LSTM, the first one only use close price as input and this is our baseline. Next, we start to add sentiment score, like TextBlob or Vader to the model. Moreover, we believe that a tweet is more influential when more people retweet it. Therefore, we multiple the sentiment score with the number of retweet.

```
Model: "sequential_7"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_7 (LSTM)               (None, 64)                17152

 dropout_4 (Dropout)         (None, 64)                0

 dense_12 (Dense)            (None, 8)                 520

 dense_13 (Dense)            (None, 1)                 9

=================================================================
Total params: 17,681
Trainable params: 17,681
Non-trainable params: 0
_________________________________________________________________
```

```
cp2 = ModelCheckpoint('model2/', save_best_only=True)
model2.compile(loss=MeanSquaredError(), optimizer=Adam(learning_rate=0.01), metrics=[RootMeanSquaredError()])
model2.fit(X2_train, y2_train, validation_data=(X2_val, y2_val), epochs=10, callbacks=[cp2])
```

Figure 4. Example for LSTM model parameters

3. Data Use in Second Part - Predict the General Trend of the Price

Date	Pos_Componen	Diff_Pos_Comp	Close Price	Definition
12/19/21	0.0718	-0.0100	0.1673	
12/20/21	0.0750	0.0032	0.1712	
12/21/21	0.0815	0.0065	0.1732	
12/22/21	0.0739	-0.0076	0.1845	
12/23/21	0.0579	-0.0159	0.1866	
12/24/21	**0.0106**	**-0.0473**	**0.1907**	**REBOUND**
12/25/21	0.0773	0.0667	0.1900	
12/26/21	0.0719	-0.0055	0.1877	
12/27/21	0.0768	0.0049	0.1741	
12/28/21	0.0739	-0.0029	0.1678	

Table 4. Example of Input Data for Second Part - Predict the General Trend of the Price

Pos_Component: The positive component from Vader result.

Diff_Pos_Comp: After applying Vader, we got the positive, negative, neutral components and compound score. Since the document is too large, the compound score is always 1. Thus, we analyze the difference of positive component rather than compound score. We subtract the time t positive component with time t-1 positive component.

Rebound Point : We define the points manually. For the rebound points, it needs to fit two qualifications. First, the price trend must be V-shaped or A-shaped. We find the V or A shaped with figure 5. In the figure 5 , point A and B do not match the qualification, because they lack the future consecutive trend. Second, price fluctuation must be more than 10%. Since the change of price exists big V/A or small V/A, to focus on big V/A shaped, we pick the fluctuation larger than 10%.That is, it requires consecutive days of price increases followed by consecutive days of price decreases, or consecutive days of price decreases followed by consecutive days of increases, as shown in the figure 5. The turning point in the price trend in the picture, we call it the rebound point, as the name suggests, this point represents the price from a big rise to a big fall, or from a big fall to a big rise. In the end, we only have 3 rebound point between 12/1/2021 to 03/31/2022.

During the process of finding rebound point, we define the sector for consecutive decrease (green background in table 4), the sector for consecutive decrease (red background in table 4) and rebound point (orange background in table 4).

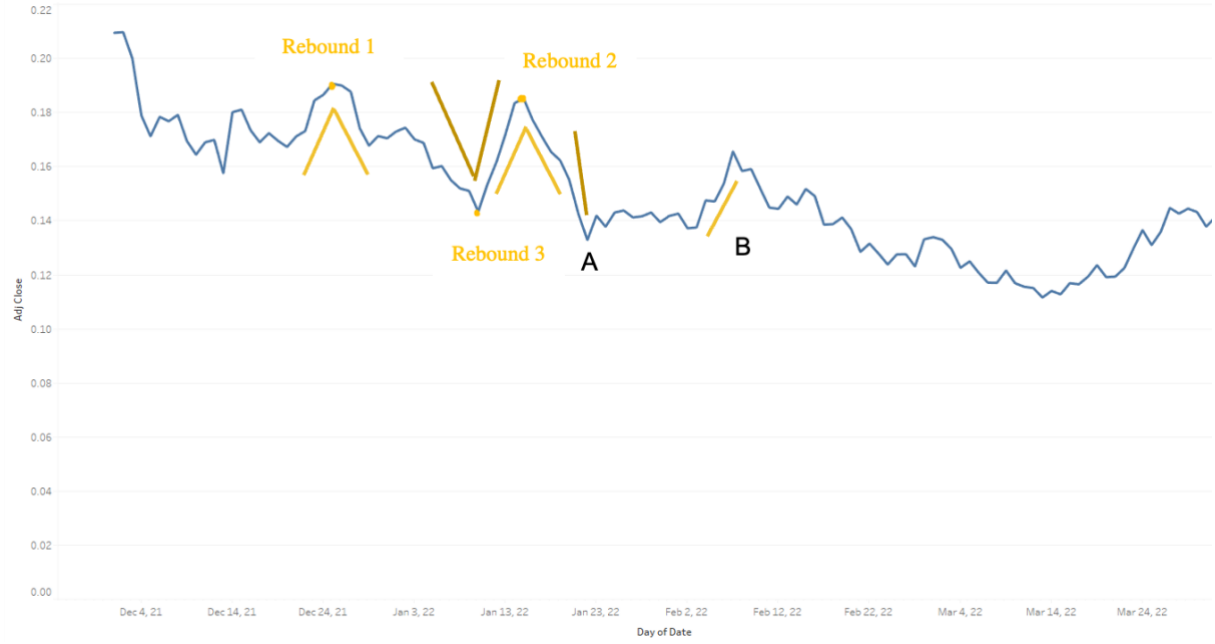

Figure 5. Price of dogecoin with Rebound Point and example for V/A shape

Evaluation of First Part - Predict Dogecoin Price with LSTM :

1. RMSE :

According to the table 5, all models perform better than the baseline. The best model is using close price and Vader score (t-2,t-1,t) to predict price at time t with the lowest RMSE 0.0046. Looking at its figure 8, the predict price is really close to the real price.

	Variable input in LSTM	RMSE of Test Data
1	Close Price (Baseline)	0.0058
2	Close Price + Score of TextBlob	0.0233
3	**Close Price + Vader Score**	**0.0046**
4	Close Price + Score of TextBlob x Retweet Count	0.0163
5	Close Price + Vader Score x Retweet Count	0.0195

Table 5. RMSE Results of LSTM Models

2. Visualization:

Plotting time series is important, because it is easy to be deceived with RMSE score. Take the fourth model for example, its RMSE is only 0.0163, it already looks like a great model. However, looking at its figure 9, the prediction line (blue line) did not react the trend of the real price in the beginning.

Even for the best model (Method 3. Close Price + Vader Score), after visualizing, the prediction line does not look like always consistent with the orange line. Not to mentioned the other models with worse RMSE score.

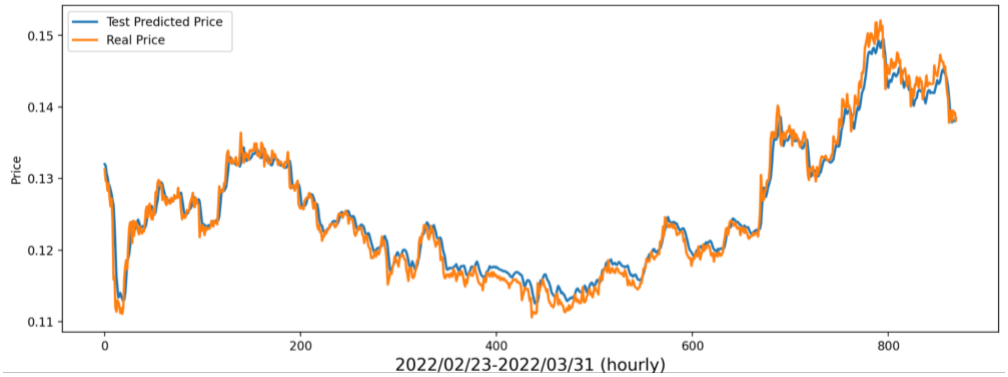

Figure 6. Result of Test Data along with Close Price in LSTM Model (Baseline)

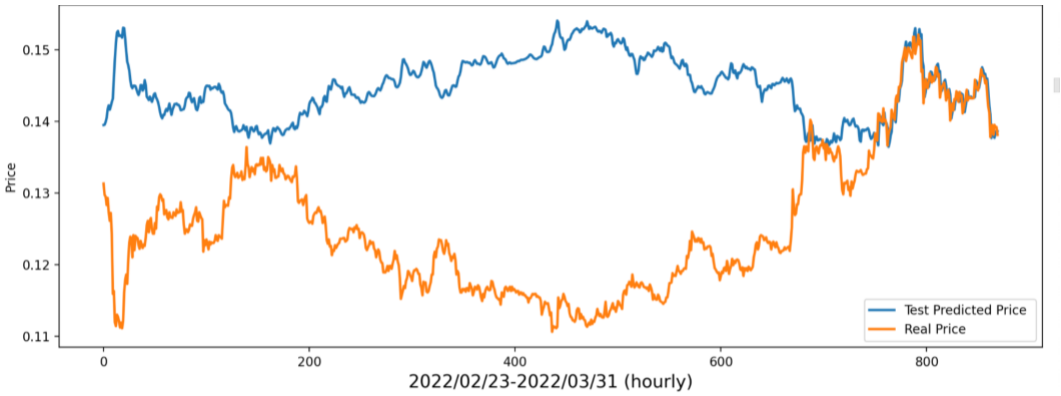

Figure 7. Result of Test Data with Close Price and TextBlob Score in LSTM Model

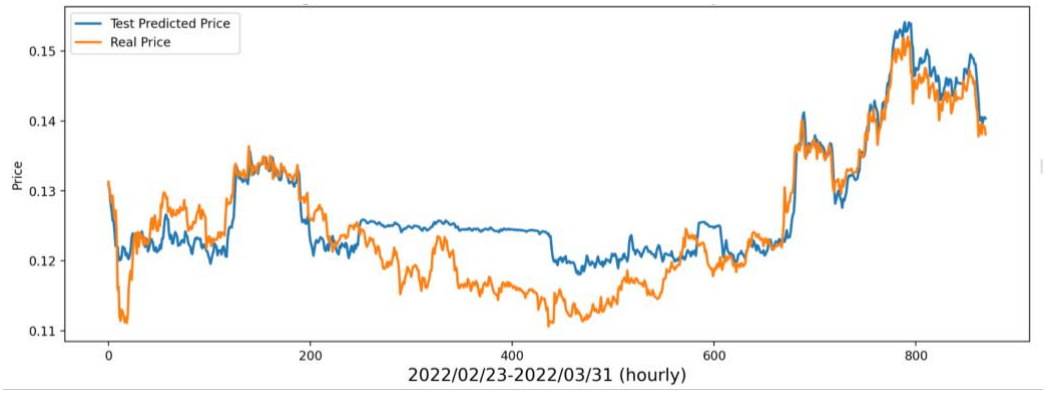

Figure 8. Result of Test Data with Close Price and Vader Score in LSTM Model

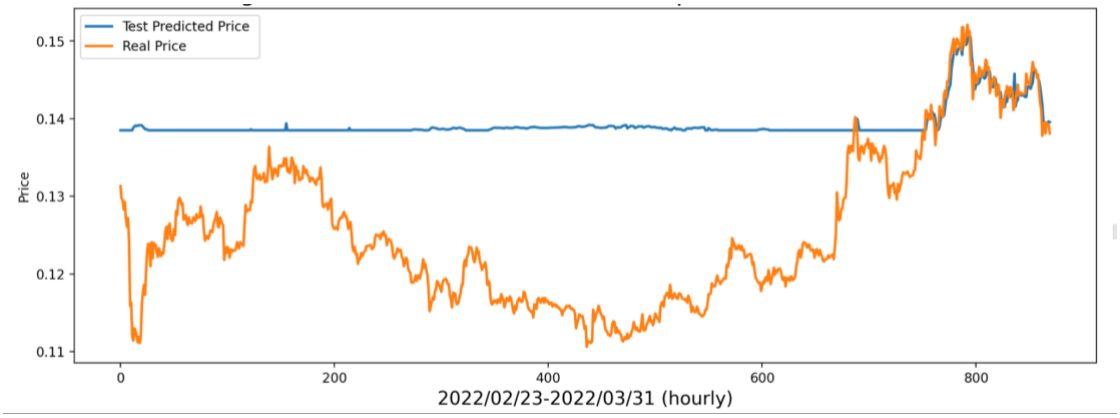

Figure 9. Result of Test Data with Close Price and TextBlob Score Multiplied with Rewteet Number in LSTM Model

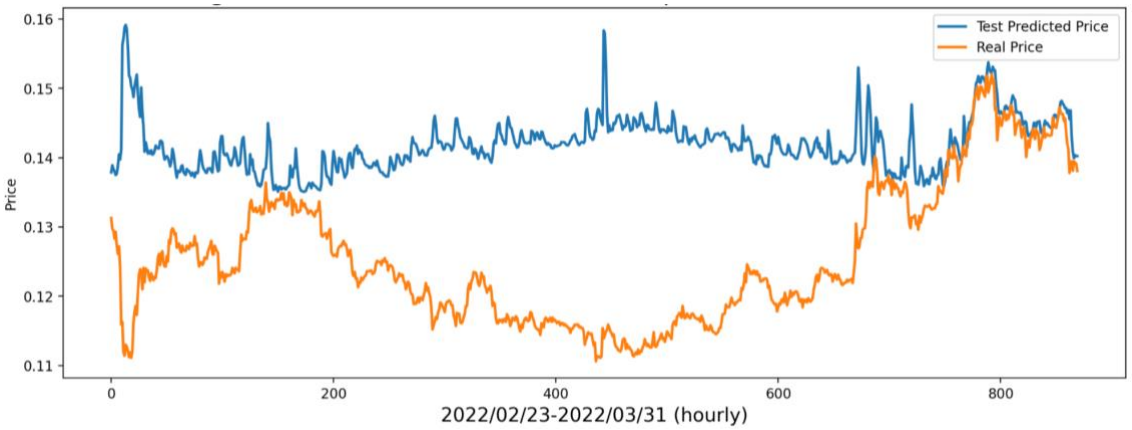

Figure 10. Result of Test Data with Close Price and Vader Score Multiplied with Rewteet Number in LSTM Model

Evaluation of Second Part - Predict the General Trend of the Price ：

Date	Pos_Component	Diff_Pos_Comp	Close Price	Definition
12/19/21	0.0718	-0.0100	0.1673	
12/20/21	0.0750	0.0032	0.1712	
12/21/21	0.0815	0.0065	0.1732	
12/22/21	0.0739	-0.0076	0.1845	
12/23/21	0.0579	-0.0159	0.1866	
12/24/21	**0.0106**	**-0.0473**	**0.1907**	**REBOUND**
12/25/21	0.0773	0.0667	0.1900	
12/26/21	0.0719	-0.0055	0.1877	
12/27/21	0.0768	0.0049	0.1741	
12/28/21	0.0739	-0.0029	0.1678	
1/1/22	0.0791	0.0624	0.1744	
1/2/22	0.0671	-0.0120	0.1701	
1/3/22	0.0674	0.0002	0.1688	
1/4/22	0.0788	0.0114	0.1594	
1/5/22	0.0769	-0.0018	0.1602	
1/6/22	0.0854	0.0084	0.1550	
1/7/22	0.0738	-0.0116	0.1520	
1/8/22	0.0164	-0.0574	0.1511	
1/9/22	0.0795	**0.0630**	0.1434	REBOUND
1/10/22	0.0771	-0.0023	0.1534	
1/11/22	0.0817	0.0045	0.1617	
1/12/22	0.0731	-0.0086	0.1720	
1/13/22	0.0819	0.0089	0.1835	
1/14/22	0.0829	0.0010	0.1851	
1/15/22	0.0123	**-0.0706**	0.1772	REBOUND
1/16/22	0.0823	0.0700	0.1711	
1/17/22	0.0856	0.0033	0.1655	
1/18/22	0.0849	-0.0007	0.1624	
1/19/22	0.0928	0.0079	0.1552	

Table 6.Second Part Results - predict the general trend of the price

In each rebound points, the value of Diff_Pos_Comp are all greater than 80% percentile (0.0322) of Diff_Pos_Comp. For example, on December 24, 2021, when the price continued to increase in the previous four days, the daily positive sentiment analysis scores were concentrated in the

range of 0.05-0.08, but on the day of the rebound 1(the 24th), the positive sentiment analysis scores plummeted to 0.01, followed by it started a 4-day sharp decline.The increase and decrease reached 13.7% and 13.6% respectively. In the same way, on January 15, 2022(rebound 2), the positive component value also dropped sharply, and the price also changed from a sharp rise for several consecutive days to a sharp fall. The increase and decrease reached 17% and 28.6% this time. Both times were V-shaped price movements, with prices rising first and then falling.

Another example is the third rebound point on January 9, 2022 is the only V-shaped price trend, with prices falling first and then rising. The Positive sentiment analysis score increased from 0.01 on January 8 to 0.079 on January 9, followed by five consecutive days of price increases.

Conclusion and Future Scope

For the first part, we use LSTM model with different variables to find out the best way to predict dogecoin price. Even though we have a better model - close price + Vader score, it did not looks like closely to the real price when visualizing it. In addition, in LSTM model, we always need the previous 3 parameters to predict, and can only predict the next one hour rather than the next 10 hours of data. We found that it is difficult to predict the specific price of dogecoin with existing data. That's the main reason why we have the second part analysis - predict the general trend of the price.

As for the second part, the rebound point of each large price fluctuation is predicted by analyzing the rebound point of the sentiment analysis score, within our existing data range, the prediction accuracy is as high as 100%. In future applications, the same method can be used to observe the daily positive sentiment score in the days of continuous rise and fall. When the positive sentiment score changes greatly, it can be regarded as a price rebound is coming · so as to make corresponding investment strategies. However, this is only a prediction based on our existing data, and there is no guarantee that 100% accuracy will be maintained in a longer timeline. In future learning, we expect to automate the part of collecting data. Finally, If we want to develop an investment strategy, we need to use the rate of return as evaluation instead of RMSE.

References

https://towardsdatascience.com/how-to-scrape-more-information-from-tweets-on-twitter-44fd540b8a1f

https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/

https://www.analyticsvidhya.com/blog/2020/10/multivariate-multi-step-time-series-forecasting-using-stacked-lstm-sequence-to-sequence-autoencoder-in-tensorflow-2-0-keras/

https://machinelearningmastery.com/reshape-input-data-long-short-term-memory-networks-kera

https://capital.com/how-to-predict-cryptocurrency-prices

https://www.youtube.com/watch?v=kGdbPnMCdOg&t=2589s
https://eudl.eu/pdf/10.4108/eai.29-9-2021.171188