

Crypto Tweets Sentiment Analysis and Price Prediction

“Social Media and Business Analytics Research Project

Ihtisham Ahmad
ahmad@uni-potsdam.de
810569

Vishal Kumar Lohana
lohana@uni-potsdam.de
804983

Muhammad Zeeshan Azad
azad@uni-potsdam.de
805466

ABSTRACT

Cryptocurrencies are an exponentially growing financial market and highly volatile in price. Its volatile nature makes it very difficult to do predictions by using traditional methods. Twitter is one of the most used social media platforms by crypto investors and users. In this report, we have presented a method to use Twitter to predict cryptocurrency prices using tweets sentiment analysis and historical prices data. We scraped and labeled tweets, preprocessed them, trained an LSTM classifier for sentiment analysis, performed statistical tests and used sentiment score along with volume of tweets to successfully predict direction and magnitude of Ethereum and Near Protocol prices. These predictions can be very beneficial for crypto traders and investors as they can analyze the public opinion and make better profitable decisions.

Keywords

Sentiment Analysis, Cryptocurrency Price Prediction, LSTM, Ethereum, Near Protocol, NLP.

INTRODUCTION

Bitcoin's whitepaper was published in 2009 by an unknown person named 'Satoshi Nakamoto' (Nakamoto, 2009). From that day, the world became aware of blockchain, decentralized finance and a new way of storing and transferring digital asserts. In the last 7 years, cryptocurrency trading has become more and more popular, at one point in 2021 total crypto market cap touched \$ 3 Trillion.

All the cryptocurrencies other than Bitcoin are called altcoins and their number has been increasing rapidly to 20, 000 + in recent years. There are two kinds of platforms where crypto coins can be traded, one are centralized crypto exchanges such as Binance, Coinbase etc and another method is by using decentralized exchanges (DEXs) where coins are transferred from seller to buyer on-chain.

Twitter is one of the major sources of crypto news and informational content. Millions of people read, share and discuss crypto related investment opportunities on twitter regularly, which influence the decision of investors. Many crypto twitter (CT) influencers make predictions for various altcoins and many traders search for the hashtag of coin on twitter before making their investment decision. It is an important area of study to use twitter sentiment analysis along with state of the art deep learning algorithms to forecast prices which can help investors make better decisions.

In tweets data there is a great amount of noise such as fraudulent/scam tweets, sarcastic tweets, influencers promoting shitcoins etc and it is very important to get rid of this noise to process the data for further modeling and predictions. We used various preprocessing techniques, human labeling of tweets and building deep learning models that can differentiate between noise and useful tweets.

In this study two altcoins are examined, Ethereum (ETH) and Near Protocol (NEAR). Both are layer 1 blockchains providing decentralized platforms to build decentralized applications (Casino, Dasaklis, Patsakis, 2019) on top of them. Ethereum was launched in 2015 and was the first blockchain providing smart contract functionality. It is second in crypto ranking only after bitcoin having a \$200 + Billion market cap. Ethereum native token ETH is available to trade in almost all crypto centralized and decentralized exchanges.

Near protocol was launched in 2020, it is an open source proof-of-stake (POS) blockchain development platform. It is a direct competitor of Ethereum offering better functionality. Near has 25th crypto rank and has a \$4 Billion market cap which is very small as compared to Ethereum. We chose Near as we wanted to study one high market cap coin like ETH and one medium market cap coin. High market cap coins need a huge amount of investment to have a small percentage of change in price whereas less market cap coins need relatively less investment or number of investors to move its price. Likewise, NEAR price is more volatile as compared to ETH.

Many studies have been done about the relationship of sentiments and prices of traditional markets, and there have been systems developed that can predict direction and somewhat of magnitude of prices by using sentiment analysis.

There are not many studies done on the relationship of crypto tweets sentiment and crypto prices. The main focus of this study is to find out correlation between twitter sentiment, crypto prices and volume of tweets. Also, to develop a system that can predict direction as well as magnitude of prices.

THEORETICAL BACKGROUND / RELATED LITERATURE

Cryptocurrency and Blockchain Technology:

A cryptocurrency is a digital currency in which transactions are stored and verified without any third party. It uses a storage technology known as blockchain which was proposed in 2008 and implemented in 2009. Blockchain is the core technology behind all the cryptocurrencies including bitcoin. With the launch of bitcoin, a paper was also published by 'Satoshi Nakamoto' titled "Bitcoin: A peer-to-peer electronic cash system". In that paper for the first time blockchain was introduced as a backbone of a system which can be a replacement of current banking systems.

As blockchain provides decentralization, security and privacy, it has proved very useful in various other industries such as healthcare, IOT applications, supply chain management etc (Arjunwadkar, Ramageri, 2020). This enormous ability of blockchain to be used for multiple applications resulted in thousands of new cryptocurrencies. Every startup who uses blockchain also launched a cryptocurrency token of their product so they can grab quick investments from investors all over the world. Sometimes these tokens have utility and sometimes only used as investment tokens. There are insane profits for the earlier investors as they buy tokens at very cheap rates before the project gets famous. Looking at these profits, many new investors are opting for crypto related investments instead of traditional markets.

ETH and NEAR both are blockchain platforms where developers or coders can create decentralized applications (Dapps). To understand the concept, we can relate it with the Apple App Store and apps inside it. ETH and NEAR acts like an App Store and Dapps are apps inside them, the only difference is that these applications and their data is stored on blockchain rather than traditional servers.

Sentiment Analysis:

Sentiment Analysis is the part of Natural Language Processing (NLP) domain in which a piece of text is categorized into positive, negative or neutral according to the sentiment, emotions or opinions of the writer. NLP is the collection of algorithms and methods for computers to analyze and understand text data. Data is collected from twitter tweets, facebook posts, customer reviews and from all over the internet. This collected data is preprocessed and then different approaches (Ramachndran, Parvethi, 2019) are used to classify each piece of text.

In general, Lexicon and Machine Learning based approaches are used to classify text in sentiment analysis (Verma, Thakur, 2018). Particularly, in Machine Learning several methodologies like Naive Bayes, Maximum Entropy, K Nearest Neighbor and Neural Networks (LSTM) are used. In this study we have used a type of Neural Networks called Long Short Term Memory (LSTM), which is a famous Neural Network architecture used in the NLP domain. It has ability to

Background on LSTM:

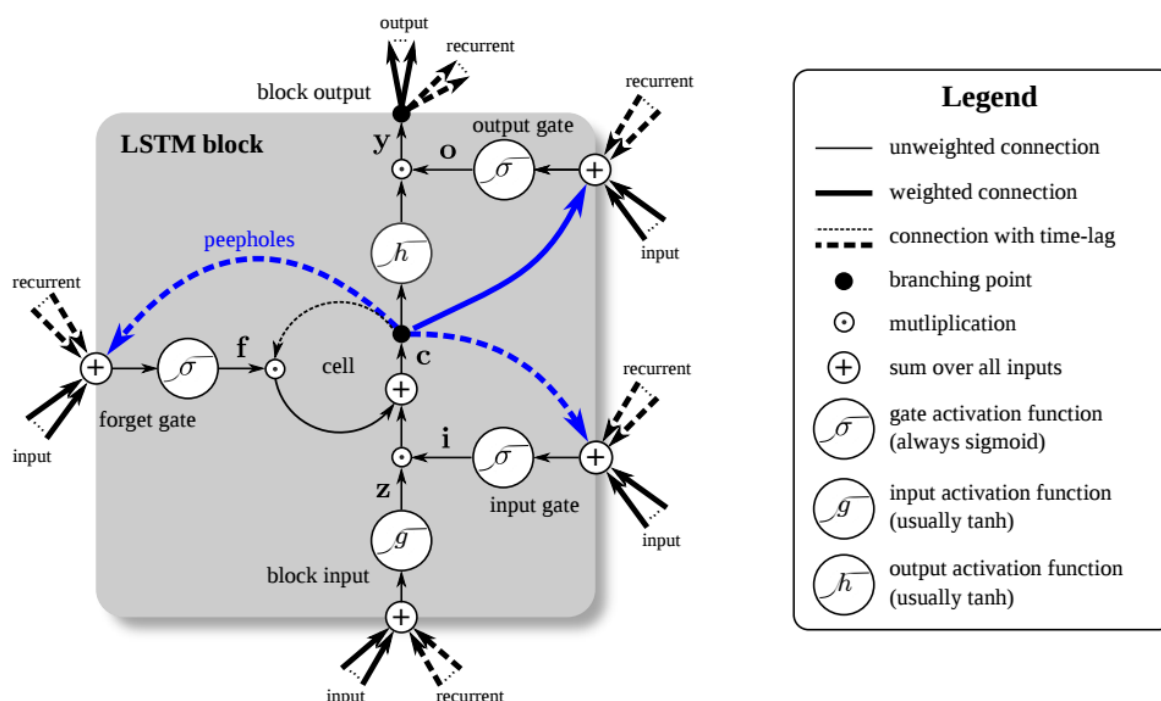


Figure 1: A Long Short-Term Memory (LSTM) unit. Image by Klaus Greff and colleagues as published in *LSTM: A Search Space Odyssey*. Image by [Klaus Greff and colleagues as published in *LSTM: A Search Space Odyssey*](#).

The LSTM model is a derivative of the RNN, a sequential feedforward neural network that utilizes internal states to process variable length sequences. The underlying differences lie in the handling of the vanishing gradient problem during backpropagation. There are three main components to the LSTM cell, namely, the forget gate, the input gate and the output gate. The input gate informs what new information will be stored in the cell state. The forget gate indicates what information will be discarded away from the cell state. The output gate is used to provide the activation of the final output of the LSTM model (Wong, 2021). The equations of gates are as follows:

$$i_t = \sigma(w_i [h_t - 1, x_t] + b_i) \quad (1)$$

$$f_t = \sigma(w_f [h_t - 1, x_t] + b_f) \quad (2)$$

$$o_t = \sigma(w_o [h_t - 1, x_t] + b_o) \quad (3)$$

$$\tilde{c}_t = \tanh(w_c [h_t - 1, x_t] + b_c) \quad (4)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (5)$$

$$h_t = o_t + \tanh c_t \quad (6)$$

where i_t is the input gate, o_t is the output gate and f_t is the forget gate. w_x refers to the weight at the respective gate(x) and $h_t - 1$ is the hidden state from the previous timestamp. The cell state which is also known as the long term memory, is represented by c_t and h_t is the hidden state, representing the short term memory. Note that t is the current timestamp(t). In figure 1 the working of each individual gate and also the entire LSTM block is shown.

Related Work

Authors of “Text based sentiment analysis using LSTM” in 2020, have proposed that long term memory of a recurrent neural network is important for successful and better analysis of the text and predicting sentiments. They further emphasized that choosing a better option for word embeddings also improves the results; as this is the common approach and depends on the dataset and problem, refactoring and modifying data according to the approach in NLP is the best way forward for sentiment analysis and the authors have clearly identified this problem and made the approach to use LSTM based on the dataset. (Allu, Andhavarapu, Bagadi, Murthy, Belusonti, 2020)

In 2014, Researchers at University of Siegen, discussed word association for analyzing stock market news reports and three methods described in Efficient Market Hypothesis, weak, semi-strong, and strong-form. Because the stock market does not fit in any of the above, we can say only a mixture of efficiency and inefficiency can be the right point of view. The authors in this paper indicate that results can be improved when we combine lexicon based word association analysis with real-time data (Fathi, Uhr, Zenkert, 2014).

In “Twitter sentiment analysis approaches: a survey”, the authors have focused on how people these days are expressing their emotions on twitter and mentioned different methods that can be used for sentiment analysis. The machine learning methods discussed in this paper vary from SVM, TF-IDF, LSTM etc and have different results depending on the dataset and problem question. The use of LSTM model for classification combined with spark streaming for real time data processing and sql-based data analytical processing provides a scalable solution for streaming text analytics and have achieved accuracy of 82.1% for positive and 79.9% for negative sentiments (Adwan, Al-Tawil, Dibsi, Huneiti, Shahin, Zayed, 2020).

Later on, with the growth of the cryptocurrency market and people relying heavily on twitter for crypto-related news, the researchers, Şaşmaz and Tek in 2021 started to notice some similar behavior between tweets and the crypto market. In the paper “Twitter sentiment analysis for cryptocurrencies” we can see that the authors have found 91% correlation between BTC and ETH and trained a random forest classifier for sentiment analysis on the tweets of crypto currencies which achieved accuracy of 77%. They have also described how the tweets and prices correlate with each other and found positive correlation (Sasmaz, Tek, 2021).

In 2022, The authors of the paper “Social media sentiment analysis for cryptocurrency market prediction” have solely focused on the impact of sentiments on price movements of bitcoin. When conducting analysis the authors have had multiple obstacles with the datasets including sarcasm,

idioms, negation, non-textual data etc which clearly shows that each model behaves differently based on the provided data. They have also discussed and concluded that interpretable artificial intelligence or interpretable sentiment analysis methods are more valuable for predicting the sentiments from tweets and achieve better results (Raheman, Kolonin, Fridkins, Ansari, Vishwas, 2022).

METHODOLOGY

Data Collection and Preprocessing:

For tweets data we have used an SNS scraper to extract tweets from “01.01.2021” to “25.05.2022” for Ethereum and Near. This scraper does not use an API which makes it easy and fast to collect data from twitter. We have used “ETH”, “Ethereum”, “NEAR” hashtags to extract only tweets that are relevant to these two under-study cryptocurrencies. In total we have gathered 304000 tweets for Ethereum and 50000 tweets for Near Protocol.

Prices Data:

To collect historical prices for Ethereum and Near we have used an API from coinmarketcap and fetch data from “01.01.2021” to “25.05.2022”. We got Date, Low, High, Open, Close, volume and market cap data for each day. Both prices and tweets data is stored in csv format and Pandas DataFrame is used to perform various operations on it. We need an extra column in prices data which can present the percentage of change in price on a particular day. We did it by using following formula :

$$\text{Percentage in change} = (\text{Closing Price} - \text{Opening Price} / \text{Opening Price}) * 100$$




Market Cap: It is the total amount of investment in an asset/coin.

Volume: It is the total sum of transactions that happens in one day for a particular coin.

Tweets Data:

Tweets are in raw form and need to be preprocessed and filtered a lot to make them able to generate meaningful sentiment scores. An overview of the methods that we have used for preprocessing are shown in Table 1. A lot of regular expressions are used to filter unnecessary content.

First of all text is converted into lower case, then by using regular expressions we eliminate all the links in the text and also separate usernames if there are some by space and remove the ‘@’ symbol. We removed ‘#’ and all other unwanted characters that are not letters. Then the text is converted into tokens and NLTK ‘stop words’ collection is used to filter stop words from the text. After filtration text is passed for stemming and lemmatization and then all the tokens are joined to form a text shape again.

Text	Action
 #Ethereum \$ETH Number of Addresses in Profit (7d MA) just reached a 16-month low of 48,313,450.268\n\nView metric:\nhttps://t.co/9t2b8JZ83s https://t.co/pmSwT2a4TM'	Original
 #ethereum \$eth number of addresses in profit (7d ma) just reached a 16-month low of 48,313,450.268 view metric: https://t.co/9t2b8jz83s https://t.co/pmswt2a4tm	To lower case
 #ethereum \$eth number of addresses in profit (7d ma) just reached a 16-month low of 48,313,450.268 view metric:	Used Regular Expression for links: 'https?:\\W[a-zA-Z0-9@:.% \\+~#=?&:-]*'


 #ethereum number of addresses in profit (7d ma) just reached a 16-month low of 48,313,450.268 view metric:	Separating usernames: :'\@[a-zA-Z0-9]*', ' ', text)
ethereum number of addresses in profit d ma just reached a month low of view metric ethereum number of addresses in profit ma just reached month low of view metric	Replaced everything not a letter with space: ['^a-zA-Z\']
['ethereum', 'eth', 'number', 'of', 'addresses', 'in', 'profit', 'd', 'ma', 'just', 'reached', 'a', 'month', 'low', 'of', 'view', 'metric']	tokens
['ethereum', 'eth', 'number', 'addresses', 'profit', 'reached', 'month', 'low', 'view', 'metric']	filtered stop words
'ethereum eth number address profit reach month low view metric'	lemmatization, stemming

Table 1

Data Labeling:

To train a model we need labeled training and testing datasets. For this purpose, we labeled around 3000 tweets manually into four following categories, few examples can be seen in Table 2.

- **Positive:** Tweets that are giving a positive emotion or sentiment about the cryptocurrency are labeled as positive. Most of the time such tweets are giving an expression that the price will rise or the project has a really bright future.
- **Negative:** Tweets in which the writer is making a point that the cryptocurrency might go down or overall its a bad project or any bad news are labeled as Negative.
- **Neutral:** Tweets having no negative or positive sentiment about the cryptocurrency are labeled as neutral.
- **Irrelevant:** Tweets that are irrelevant to the under-study cryptocurrency are labeled as irrelevant.

Tweet Text	Sentiment
broke resistance on the hour candle ethereum	Positive
free download zumo we both get some free cryptocurrency when you sign up use my zumo username samstoddart to get our reward download on your appstore terms eth ethereum'	Irrelevant
ethereum is bearish and looking bad	Negative
'binance the stacking page no longer shows how much ethereum we stacked but only how much we have not very practical because we earn interest on stacked amount not on total amount thanks'	Neutral

Table 2

We added a fourth category 'irrelevant' as a label in our dataset because we noticed that there are many tweets which are mentioning Ethereum or Near in their hashtags but in the text or context of the tweet there is nothing about them. Either the person who tweeted is using hashtags to get noticed and promote some other irrelevant coin in his tweet or it's a scam tweet. In both cases we categorized it as irrelevant. Interestingly, neutral and irrelevant are dominant classes in our dataset.

Baseline Sentiment Analysis Models (BERT + Spacy):

We have used two baseline models to compare and evaluate our sentiment analysis model, a pre-trained BERT model from hugging face library and a pre-trained NaiveBayes model from Spacy. Both are trained for sentiment analysis particularly and are used widely. The purpose is to see how much better results our model can produce as compared to the ones already existing.

BERT model works in two simple steps, first is to tokenize text by using pre-trained tokenizer and second is to predict sentiment score of the text. BERT gives us positive, negative and neutral labels with their sentiment score as well.

For Spacy we need spacy.textblob pipeline and then after processing our data through it we predict the polarity score which ranges from -1 to 1. Less than 0 is negative, more than 0 is positive and 0 is neutral.

LSTM Model:

Vocabulary :

We need a vocabulary that includes all the words in our tweets dataset, so that we can convert input of LSTM into tokens or indexes of our vocabulary. We created a vocabulary which consists of unique words that occur in all tweets including both Ethereum and Near dataset. Figure 2 consists of the top 40 most frequent words in our vocabulary and In total we got 97378 unique words.

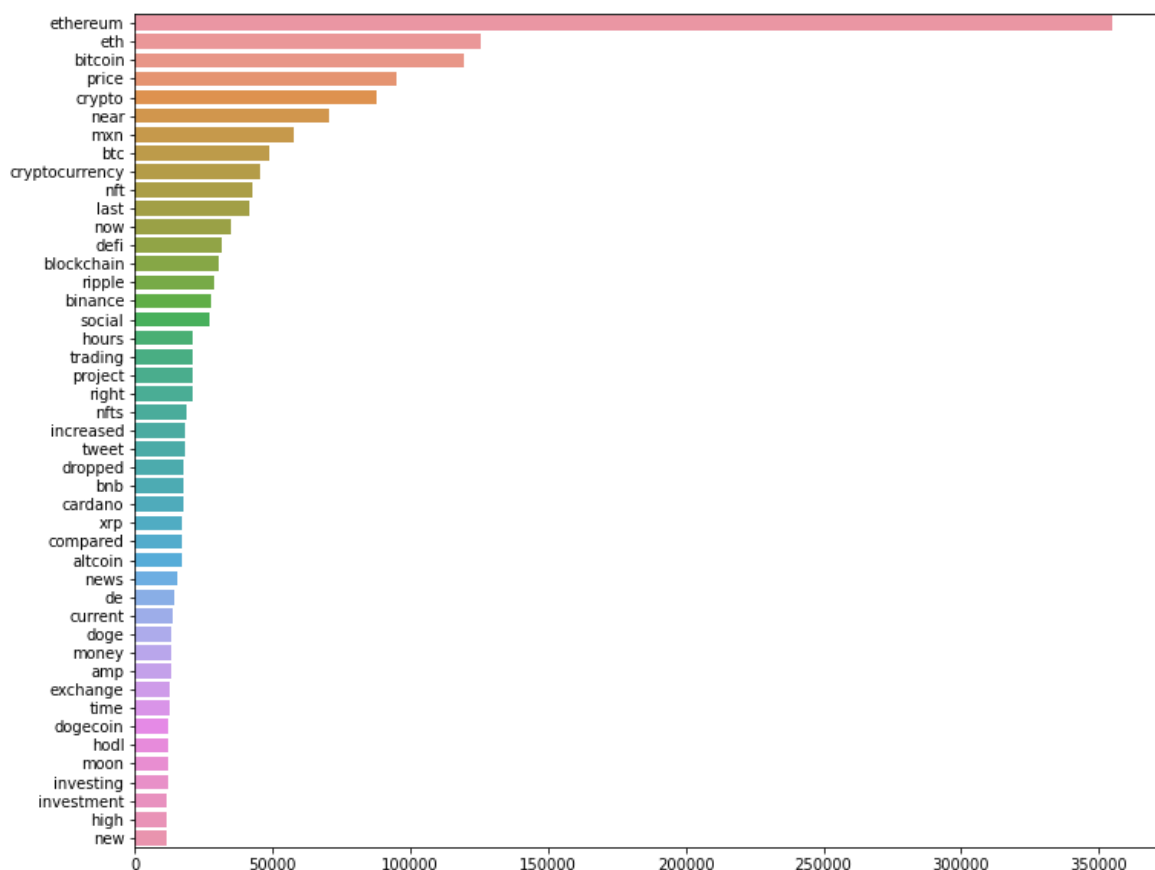
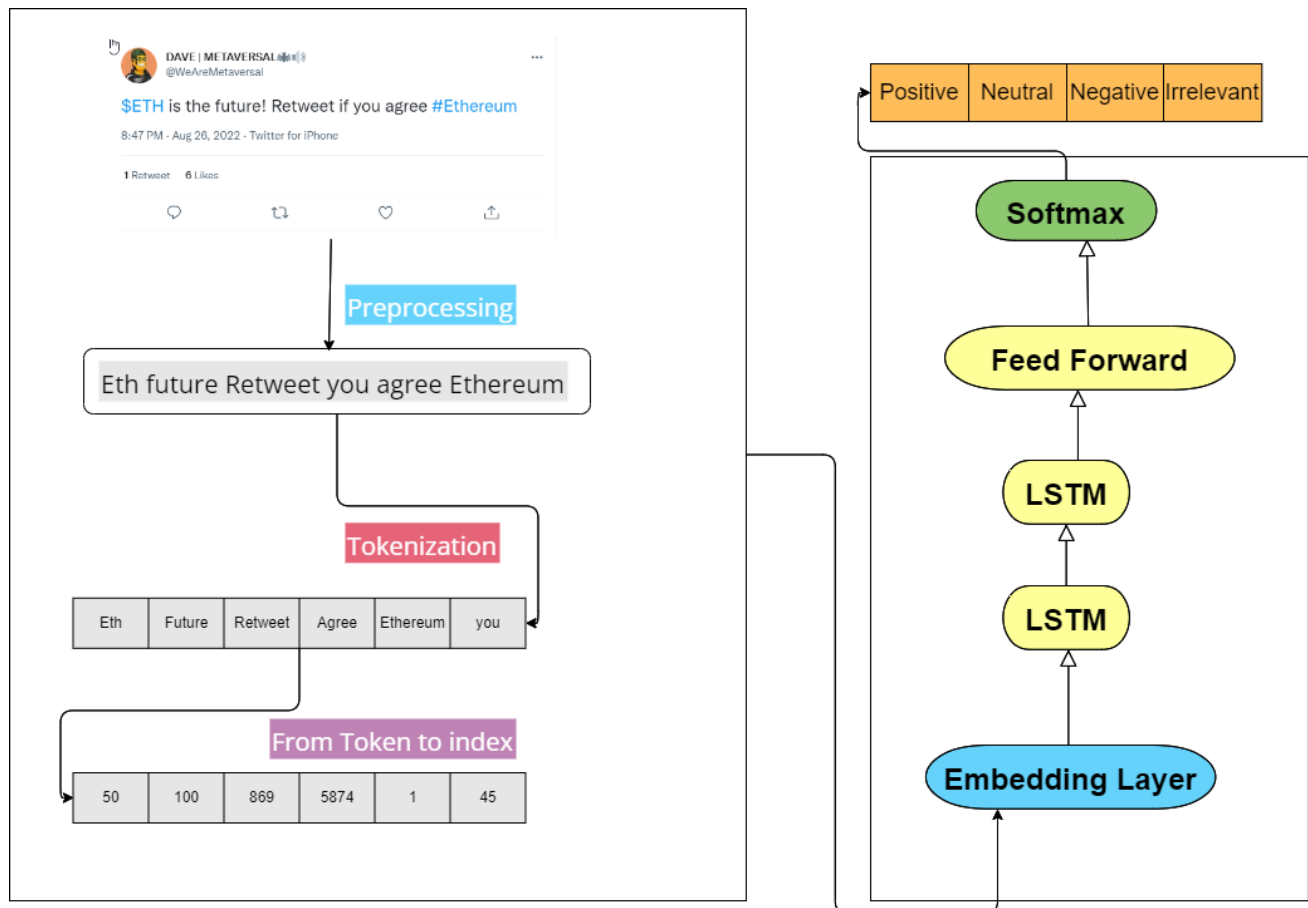


Figure 2

Architecture:

We have used a simple LSTM architecture (Allu, Andhavarapu, Bagadi, Murthy, 2020) with one embedding layer, a dropout layer, 2 Fully Connected Layers and a softmax at the end to get the probabilities for each class. Before feeding data into our model we did all the preprocessing. In the figure 3 the overall structure of our system is shown starting from a tweet and ending with its sentiment class.



Sentiment Analysis - Architecture

Figure 3

Further implementation details of LSTM model in pytorch:

- Each tweet is tokenized and padded at the end of the sequence to ensure a fixed length. The maximum sequence length was set to 30.
- Spatial dropout drops entire 1D feature maps instead of individual elements; set to 0.25.
- The number of units in the LSTM was set to 1024.
- The dropout was set to 0.5 where it drops out the input/output gate.
- The recurrent dropout was set to 0.5 where it masks connections between the recurrent units.
- The dropout was set to 0.2 just before final output.
- The final output layer was to a sigmoid so as to generate an output.

Training and Evaluation criteria for sentiment analysis:

We have used 80% of the labeled dataset for training and 20% for testing. We trained our model for 30 epochs and chose the best model weights by comparing the validation score. Precision-Recall scores are used as evaluation metrics.

Evaluation score can be seen in Figure 5. Left graph is showing loss values along with the validation cycle on x-axis and the right graph is showing accuracy in training, accuracy in validation, F1 score in training and F1 score in validation. Where formula for F1 score is mentioned below:

$$F_1 \text{ Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

The model was first trained on Ethereum and Near Dataset separately. We have tried different approaches, first we merged both Ethereum and Near dataset and trained a model on it. The performance for Ethereum tweets was good but Near tweets were not showing the same results. It was due to the different distribution of labels in both datasets.

As Near is less popular, has less tweets and more noise in the tweets such as more fake tweets/scam tweets and more irrelevant tweets. So, to tackle this problem after training with the Ethereum dataset, we used the pretrained weights from the Ethereum model and fine tuned it on the Near dataset to get almost the same results for both.

LSTM Model for Price Prediction:

We already had the prices data and required sentiment analysis data to build a training dataset for price prediction model. After doing predictions for sentiment analysis, we added the sum of sentiments per day and Volume of tweets per day in the prices dataframe as new columns.

This data is required to be scaled first so we used minimum maximum scaler and standard scaler to scale all the columns. Close Price is used as output variable and Open Price, High, Low, Volume, Change in Percentage, Volume of tweets and Sentiment sum are used as input to the model. Few unscaled data samples are shown in figure 4.

	Open	High	Low	Close	Volume	change in %	Tweets Volume	Sentiments
0	1978.677042	2014.369526	1943.938546	1944.827845	1.336455e+10	-1.710698	0	0
1	1972.390871	1991.535499	1920.688158	1978.982754	1.305711e+10	0.334208	564	-95
2	2042.344786	2080.333377	1964.386581	1972.181889	1.643453e+10	-3.435409	600	75
3	1974.670604	2047.191436	1966.038780	2043.170126	1.094112e+10	3.468909	480	87
4	1961.317996	1985.395961	1944.265107	1974.518367	8.546822e+09	0.688437	556	-20

Figure 4

For price prediction, we have also used LSTM neural networks with a slightly different approach. We have used Mean Square Error (MSE) as a loss function. Model takes 8 features as input and returns price prediction as output. Relu is used as an activation function and the model was trained for 1500 Epochs. Rest of the model's details are the same as above.

$$\text{Relu} = \max(0, x)$$

RESULTS

Sentiment Analysis Results:

Our self trained LSTM model has performed much better than both Spacy and BERT pretrained models. In figure 5, we can see the performance of our model while training. It showed 84% accuracy on the training dataset and 72% on validation dataset.

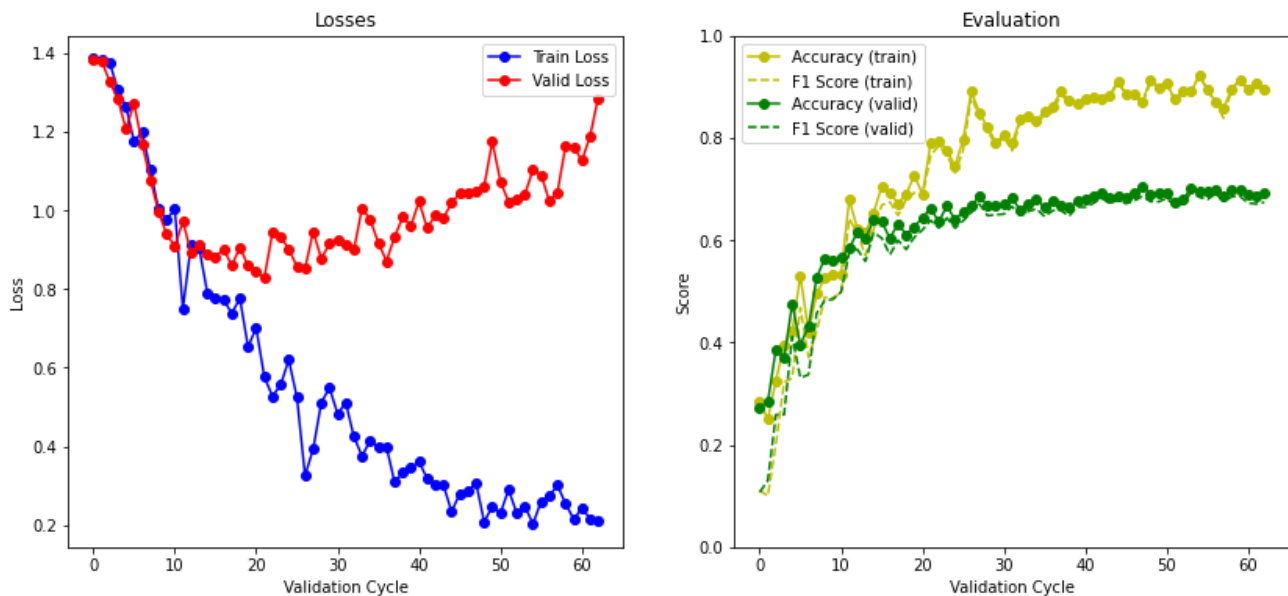


Figure 5

Figure 6 is the confusion matrix for training (Left) and testing(Right) datasets. We can see in the confusion matrix for the testing set, Neutral has least correct score and Negative has maximum. On the other hand in the dataset we have most tweets labeled as neutral and negative class has least tweets.

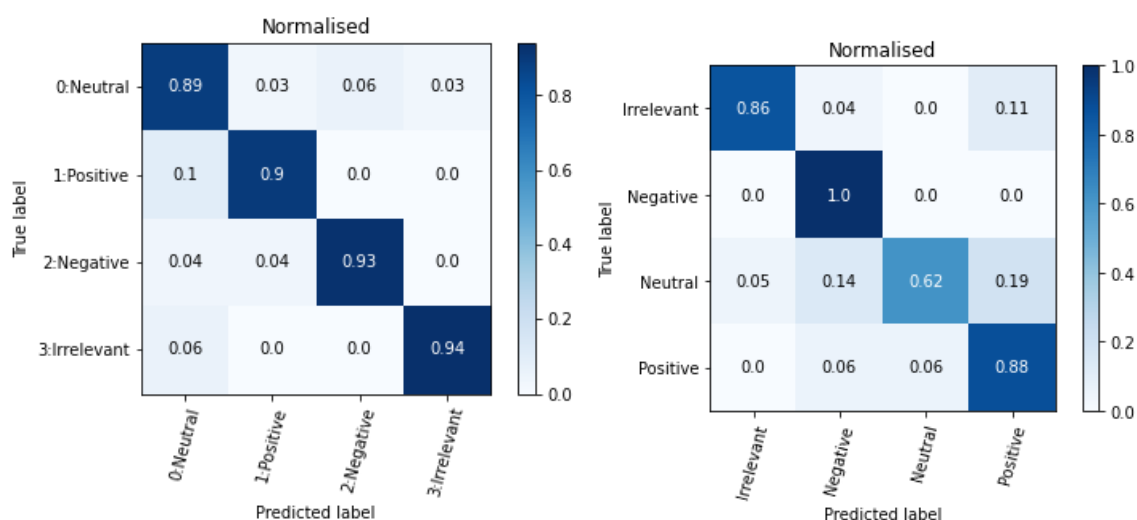
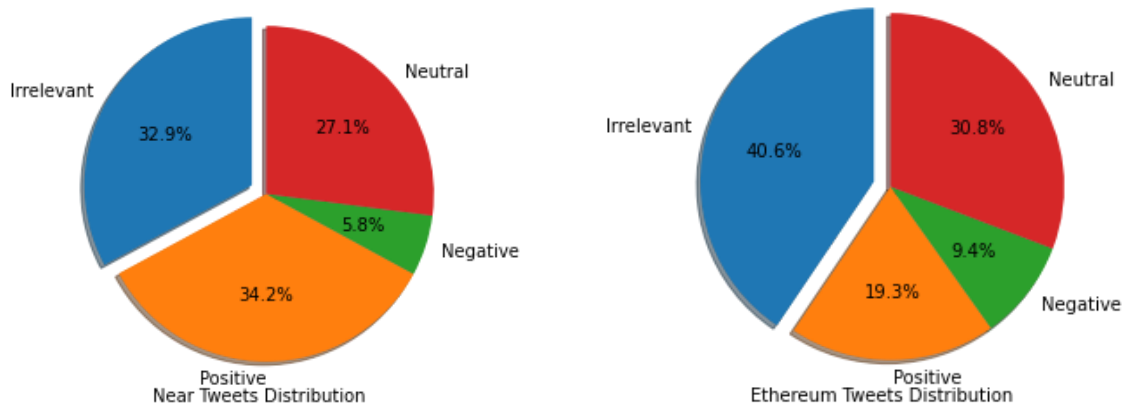


Figure 6

Figure 7 shows the distribution of prediction of our model on Ethereum and Near Dataset. Near has more positive and less negative tweets as compared to Ethereum. In general public opinion about Near is very positive. 60% of Near and 70% of Ethereum tweets were not considered as positive or negative sentiment.

**Figure 7****Accuracy:**

Our model has performed much better than baseline models (pretrained Spacy and BERT). In table 3, accuracy and F1 score for all three models are compared and LSTM has performed twice as better than other two. These LSTM results are on Ethereum tweets, Near tweets also got very similar results. But after fine tuning the Near model, results jumped by 5-6% but there are more chances of overfitting in that case.

Model	Accuracy	F1 Score
LSTM	72	0.695
BERT	0.38	0.24
SPACY	0.33	0.24

Table 3**Statistical Tests for Correlation:**

There are various features that have correlation among each other. Table 4 shows the correlation test results among Percentage in change, Sum of Sentiment Score, Volume of coin, Price of coin and Volume of tweets. We have used the Spearman Rank Correlation test and Kendall's Rank correlation test.

Correlation Tests Results for Ethereum					
Test	Data 1	Data 2	Stat	P	Result
Spearman R	Sum of sentiment score per day	change in price	0.26	0.0	Dependent
Kendall's R	Sum of sentiment score per day	change in price	0.18	0.0	Dependent
Spearman R	Sum of sentiment score per day	Daily price	-0.064	0.147	Independent
Kendall's R	Sum of sentiment score per day	Daily price	-0.049	0.099	Independent
Spearman R	Sum of sentiment score per day	Daily Volume	0.435	0.0	Dependent
Kendall's R	Sum of sentiment score per day	Daily Volume	0.31	0.0	Dependent
Spearman R	volume of tweets	Daily price	0.089	0.044	Dependent
Kendall's R	volume of tweets	Daily price	0.064	0.032	Dependent
Correlation Tests Results for Near					
Spearman R	Sum of sentiment score per day	Daily Volume	0.605	0.0	Dependent
Kendall's R	Sum of sentiment score per day	Daily Volume	0.420	0.0	Dependent
Spearman R	Sum of sentiment score per day	change in price	0.107	0.015	Dependent

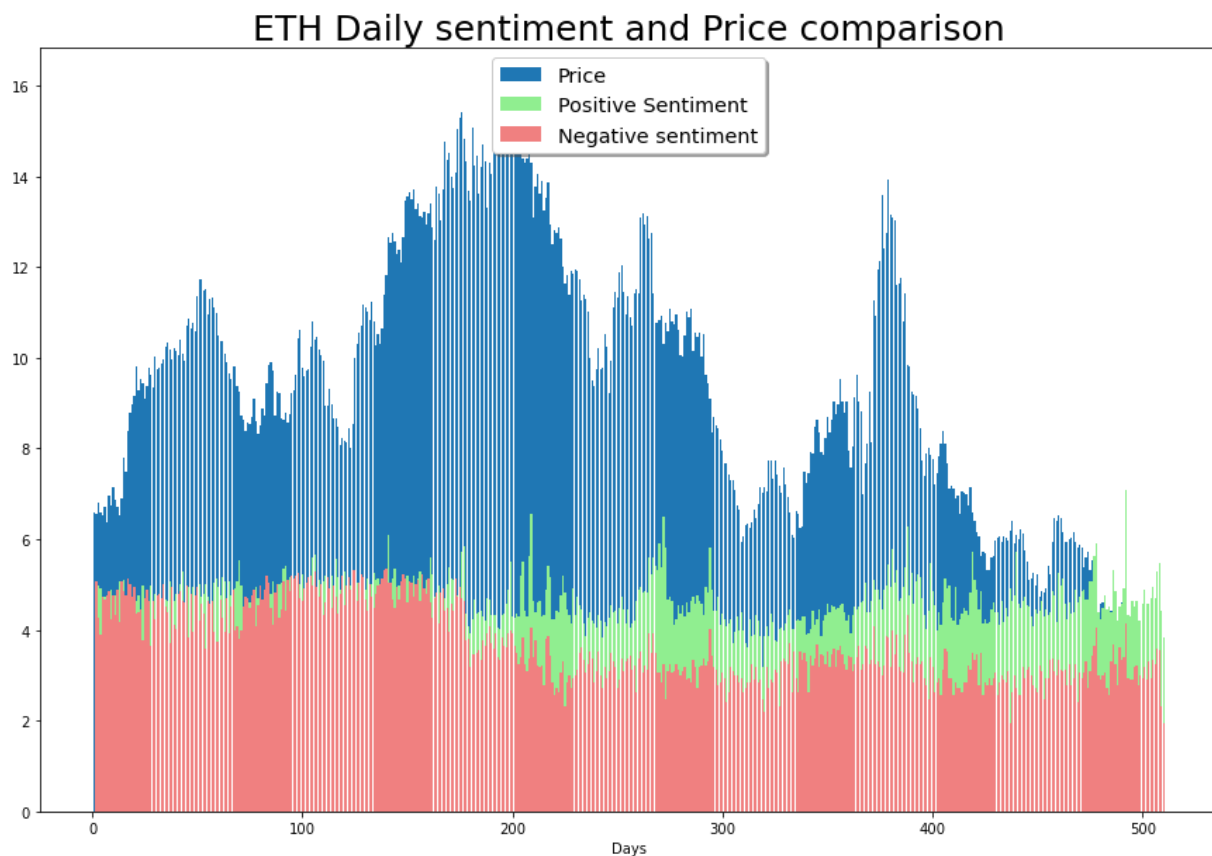
Kendall's R	Sum of sentiment score per day	change in price	0.074	0.014	Dependent
Spearman R	Sum of sentiment score per day	Daily price	0.58	0	Dependent
Kendall's R	Sum of sentiment score per day	Daily price	0.4	0	Dependent
Spearman R	volume of tweets	Daily price	0.089	0.044	Dependent
Kendall's R	volume of tweets	Daily price	0.064	0.032	Dependent

Table 4

All of the features compared for correlation tests were dependent, only Ethereum's 'Daily Price' and 'Sum of sentiment score per day' are independent.

Interestingly, from these statistical tests we can infer that we have more parameters to train a price predictor for Near as compared to ethereum. And this is because Near has a small market cap, twitter tweets have more influence on the price of Near. Ethereum has a very large market cap and to move even a few percent of its price, billions of Dollars of equivalent transaction are required.

We can confidently say that tweets sentiment analysis with coin price works much better for lower market cap coins as compared to high market cap coins. And less popular a coin is, a more strong impact can be made by tweets on its price. Visualization of this can be seen in Figure 8 and Figure 9.

**Figure 8**

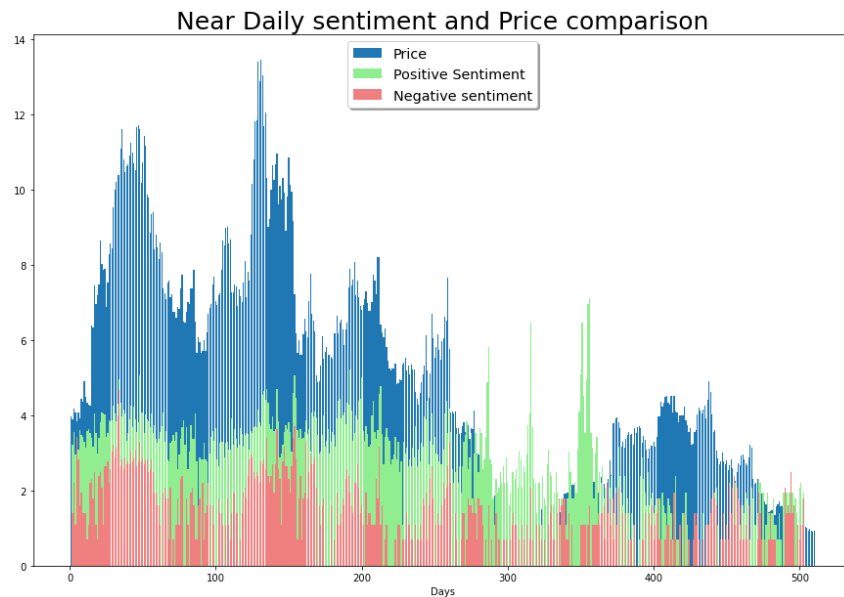


Figure 9

Volume of tweets was also found as an important factor and has a strong correlation with prices. In crypto markets, normally investors keep themselves silent when the market is going down and start tweeting a lot when the market is going up. Still there are few people who keep posting positive content in bearish markets to keep themselves and their followers motivated enough to not entirely leave the market. But overall, when the prices are going up, the volume of tweets also rises and vice versa. This is also presented in Figure 15.

Price Prediction Results:

To compare our price prediction with a baseline, we trained two models, one with sentiment analysis and prices data (Model 1) and one with only prices data (Model 2). And results are very clear that adding sentiment and volume of tweets data into the training set has increased model performance and also the model has predicted better on the testing set. Figure 10 is showing the actual data in blue lines, and predicted data in orange lines. And the red vertical line is separation between training and testing data. On the right side of the red line, the data was not exposed to the model while training. We can see that the model is predicting an almost exact pattern but with slightly different values.

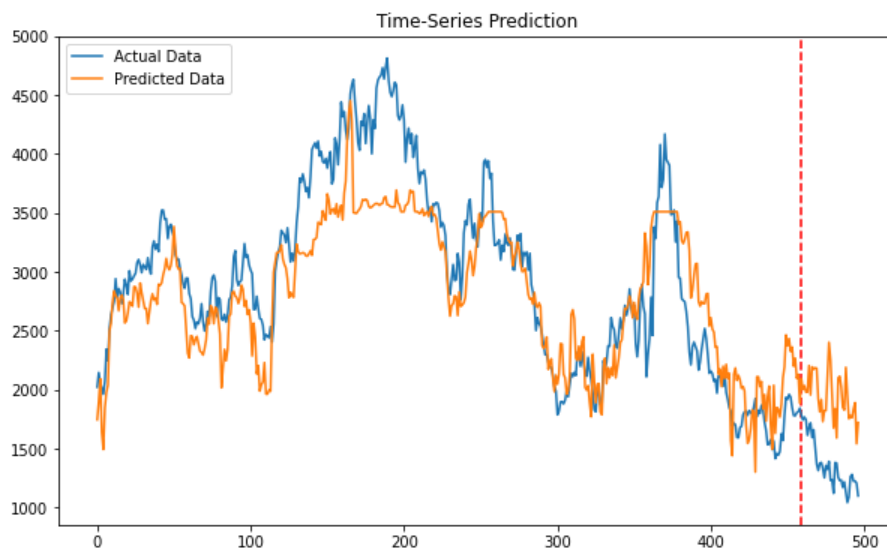


Figure 10 - Model 1

Figure 11 and 12 shows better visualization of patterns predicted by both models. Figure 11 is the predictions made by a model 1 and. Figure 11 shows prediction by the model 2. Model 1 is predicting a much better pattern than model 2.

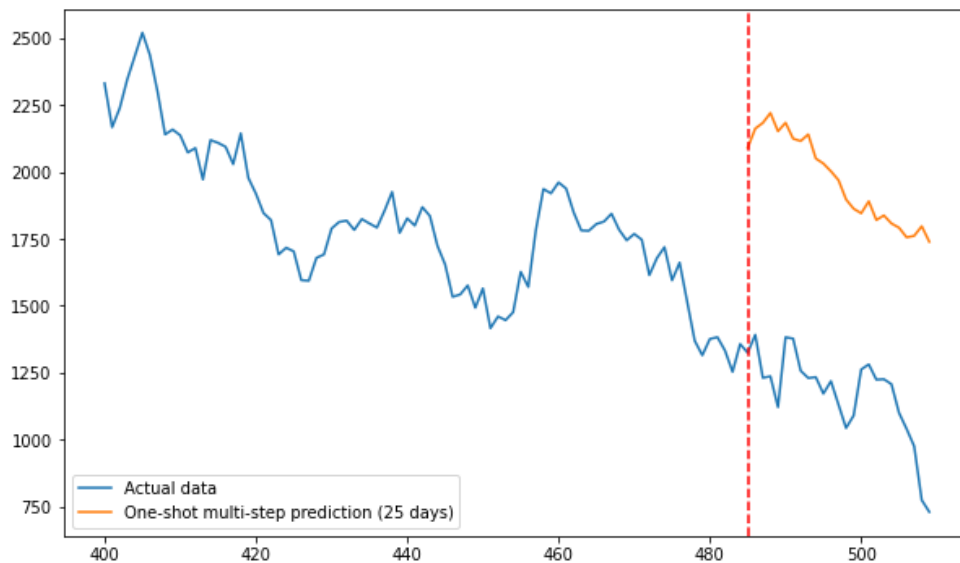


Figure 11 - Model 1

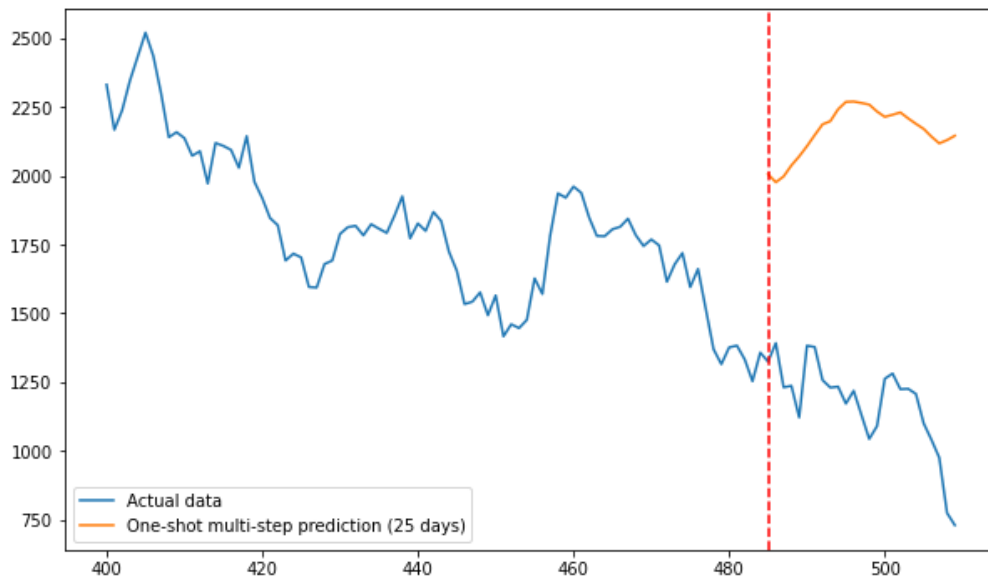


Figure 12 - Model 2

If we compare RMSE loss values in testing, model 1 has less loss with exact value of 0.135 and model 2 has 0.202, which indicates the performance of both models. Model predictions for Near are shown in figure 13.

DISCUSSION

Sentiment analysis is a difficult task because of noise, unfiltered and uncleaned data. The approaches we have used to preprocess data, labeling of data (carefully by more than one human) and training our own LSTM based model has produced good results. But still, there is room for further improvements. We labeled roughly 3000 tweets but to train a very good classifier it will need a handsome amount of labeled tweets.

As we can see it in few BERT models millions of labeled data has been used to produce very good results. Hand labeling of that much data is not possible but if an organization wants to build a crypto price predictor they can invest into sentiment analysis by manually labeling around 50k tweets, that will produce very good results.

Our approach with a fourth label 'Irrelevant' for tweets also showed some positive results. As, otherwise the model would have predicted those irrelevant tweets as positive or negative or neutral. Which can make overall predictions very wrong.

The statistical tests and price prediction model results both indicate that there is a strong correlation between tweets sentiment and price of cryptocurrency. And that is also the case in other recent studies done on the similar topic. We have used LSTM for both sentiment analysis and price prediction models, although LSTM is one of the state of the art deep learning architecture for time series data but still there is a room for improvement by trying other strong models such as transformers.

As in our study we observed that low market cap coins show more correlation between sentiment analysis and price prediction, which opens the further research direction to look deeply into other small market cap coins. As there are 20k total coins and by market cap rank Near is still in the top 25. So, there are plenty more coins which can be studied.

CONCLUSION

We collected tweets dataset for two of well known cryptocurrencies, Ethereum and Near Protocol. Preprocessed tweets by using a well structured pipeline to get clean and filtered data that we required for our models. We hand labeled 3000 tweets and categorized them into 'Neutral', 'Positive', 'Negative' and 'Irrelevant' by observing the writer's opinion or context of the tweet.

We trained an LSTM model for sentiment analysis and fed it with our cleaned hand labeled dataset. We got 72% testing accuracy and 84% accuracy in training and used that model to predict all other more than 350k tweets. Model classified roughly 60% of Near's tweets and 70% of Ethereum's tweets into neutral or irrelevant. Remaining were classified into positive and negative.

We performed statistical tests to check correlation between prices and tweets sentiments of our cryptocurrencies. And results showed that they are highly correlated among different features such as Volume of tweets and prices, Sentiment and change in price etc.

We then trained two more LSTM models, one was fed with only prices data and the other with both prices and sentiment data. And results showed that the model with sentiment data has performed better and in visualization we have seen the predicted pattern was much more accurate.

We will conclude this study with a statement that accurate tweets sentiment analysis is a very important feature for accurately predicting crypto currency prices.

REFERENCES

1. Nakamoto, S., 2009. Bitcoin: A peer-to-peer electronic cash system
2. Casino, Thomas K. Dasaklis, Constantinos Patsakis, (2019). *Telematics and Informatics: A systematic literature review of blockchain-based applications: Current status, classification and open issues*, Volume 36, 55-81.
3. Bharati Mahadev Ramageri, Maithili, Arjunwadkar (2020). *International journal of Future Generation Communication and Networking: Application of Blockchain Technology in various sectors: A review*, 13(2):94-99

4. Binita Verma, Ramjeevan Singh Thakur (2018). *Proceedings of international conference on recent advancement on computer and communication*: Sentiment Analysis using Lexicon and Machine Learning-Based approaches: A Survey, pp.441-447
5. Dharini Ramachandran, R Parvethi (2019). *Procedia Computer Science*: Analysis of twitter specific preprocessing techniques for tweets, 165:245-251
6. Wong, Eugene Lu Xian (2021). Prediction of Bitcoin prices using Twitter Data and Natural Language Processing.
7. Dr. G. S. N. Murthy, Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, Mounika Belusonti, (2020). *International Journal of Engineering research and technology*: Text based Sentiment Analysis using LSTM, (IJERT) Volume 09.
8. P. Uhr, J. Zenkert and M. Fathi,(2014). *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*:Sentiment analysis in financial markets A framework to utilize the human ability of word association for analyzing stock market news reports, pp. 912-917
9. Adwan, O. Y., Al-Tawil, M., Huneiti, A., Shahin, R., Abu Zayed, A., & Al-Dibsi, R. (2020). *International Journal of Emerging Technologies in Learning (iJET)*: Twitter Sentiment Analysis Approaches: A Survey. (15), pp. 79–93
10. E. Şaşmaz and F. B. Tek, (2021) *International Conference on Computer Science and Engineering (UBMK)*: Tweet Sentiment Analysis for Cryptocurrencies, pp. 613-618
11. Raheman, A., Kolonin, A., Fridkins, I., Ansari, I. and Vishwas, M., (2022). Social Media Sentiment Analysis for Cryptocurrency Market Prediction.

APPENDIX

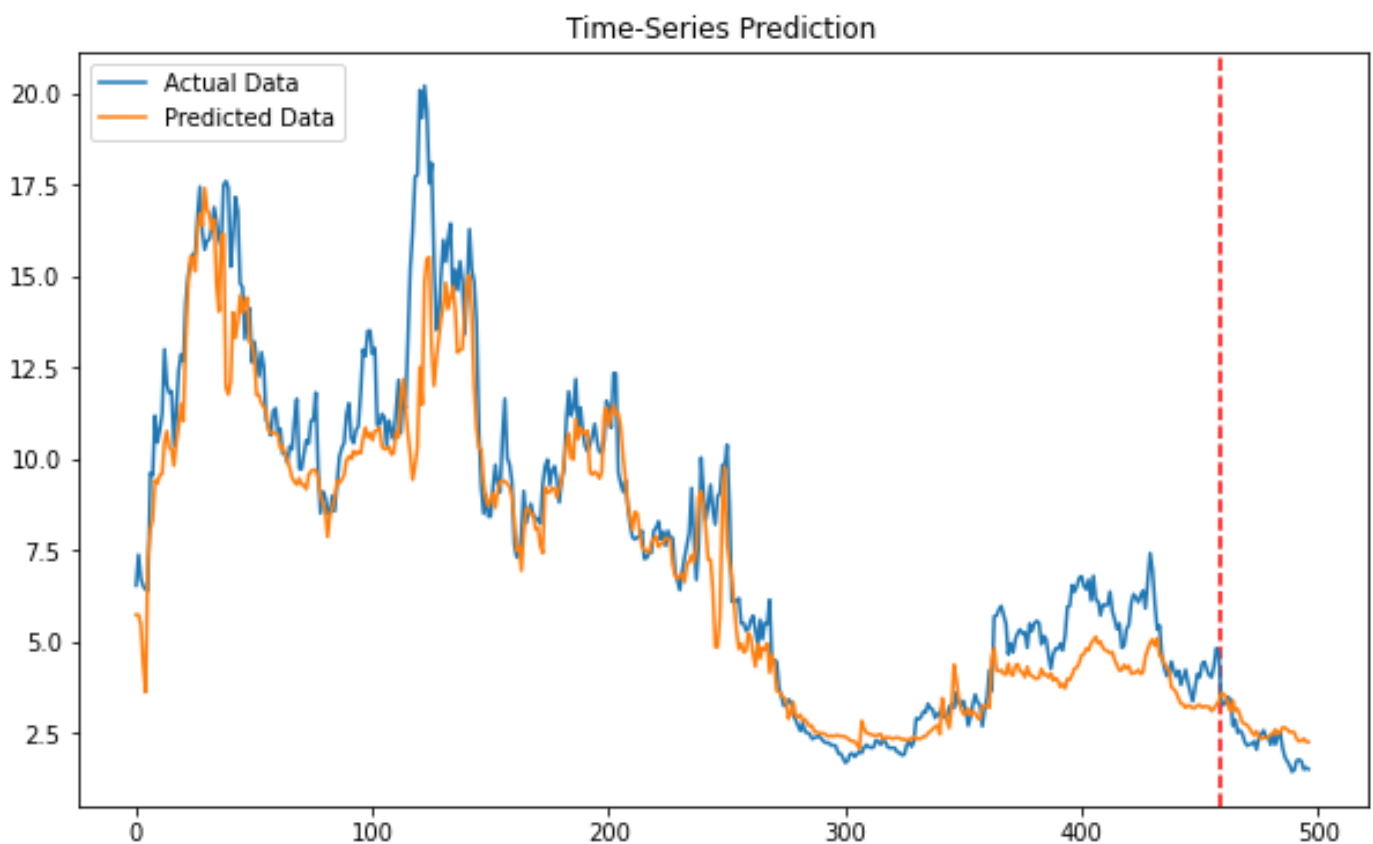


Figure 13: Price prediction for NEAR

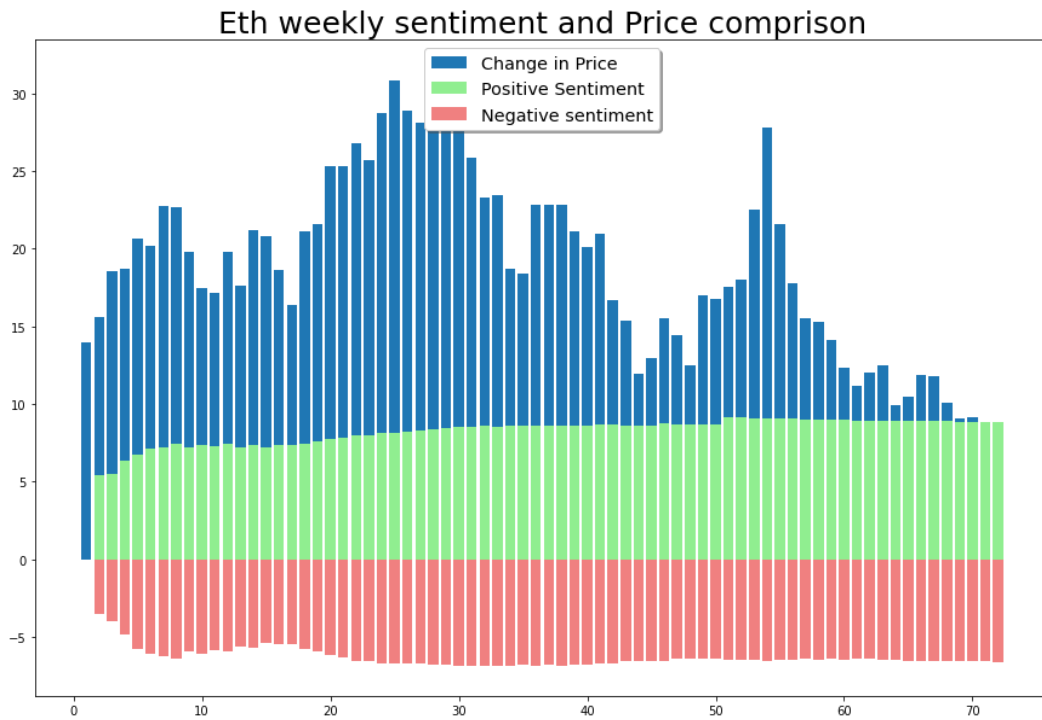


Figure 14: Weekly price vs weekly sentiment

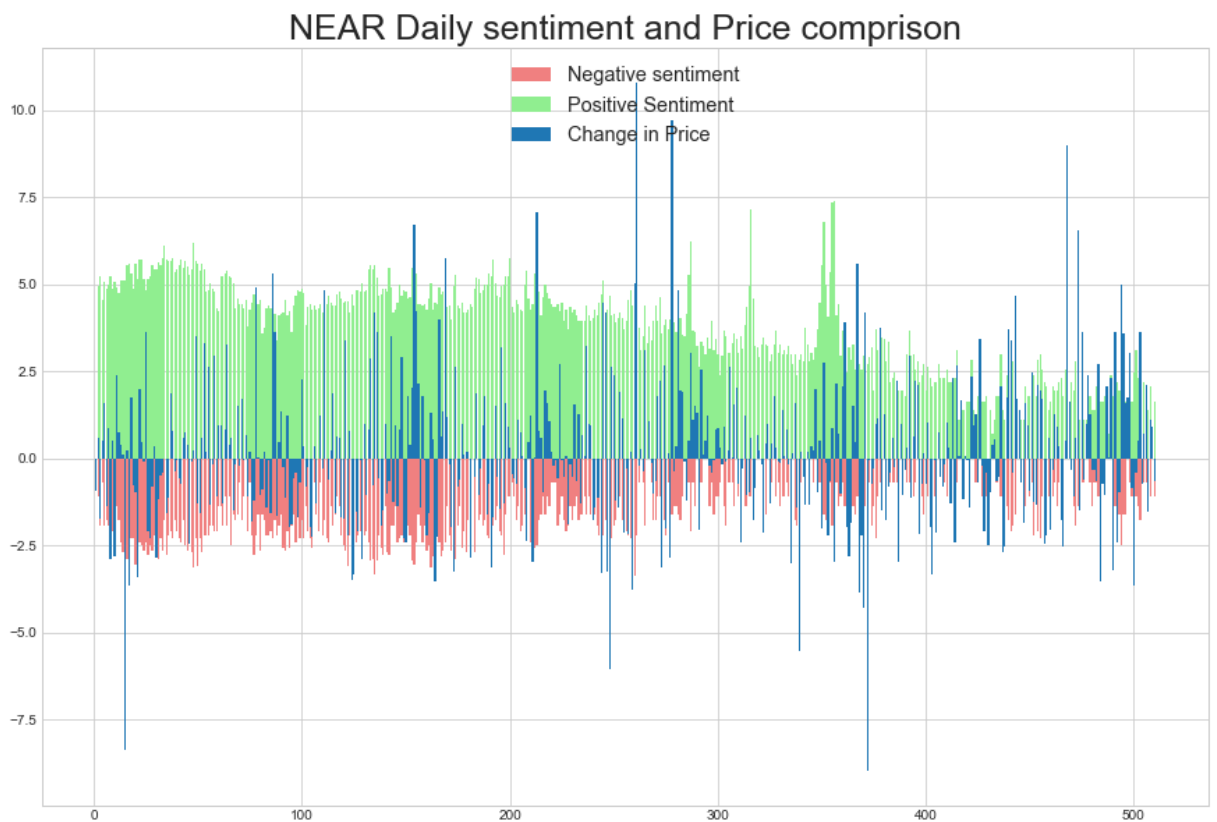


Figure 15: daily change in price vs pos and neg sentiment score