IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CRYPTO CURRENCY PRICE PREDICTION USING SENTIMENT ANALYSIS

Dr. Rajeshwari.D [1], Dr. T. Ananthapadmanabha[2]

[1]Asst.Professor, [2]Director,
[1] Department of Information Science & Engineering,
[1] The National Institute of Engineering, Mysuru, India
[2]Mysore University School of Engineering, University of Mysore, Mysore, India

ABSTRACT

A developing market, crypto currencies are becoming more and more important in the financial sector. The crypto currency market is an ideal research topic because of its low entrance barrier and high data availability. By applying sentiment analysis and machine learning techniques, it is possible to gain insights into the behavior of markets for the difficult work of stock market forecasting. Although there has been some prior research, the majority have only looked at the behavior of Bitcoin.

This technique suggests using widely used machine learning tools and readilyaccessible social media data to forecast changes in the

price of the Bitcoin cryptocurrency market. The findings demonstrate that cryptocurrency markets may be predicted using machine le arning and sentiment analysis, with Twitter data being able to predict specific cryptocurrencies and outperforming other models. This system pres- ents predictive model for forecasting the Bitcoin price movements using LSTM, Bi-Directional LSTM using sentiment analysis extracted from Twitter data. A review of the relevant literature and related works which this work expands upon is presented, in addition to an analysis of relevant techniques.

Index Terms: Cryptocurrency, Bitcoin, LSTM, Bi-Directional LSTM, Twitter data

I. INTRODUCTION

Cryptocurrency is a type of digital currency that has gained popularity due to its ability to maintain value and resist inflation. It is managed by a block chain ledger system, making it difficult to counterfeit. However, the Federal Reserve's increased interest rates have had a negative impact on cryptocurrencies, with two of the most popular coins, Bitcoin and Ethereum, having declined by more than 70%. Satoshi Nakamoto published the first Bitcoin client version on January 9, 2009, creating a breakthrough electronic cash substitute.

Basic Functionality

A bitcoin can be divided into smaller units known as Satoshis and used to make purchases from businesses that accept Bitcoin. However, it differs from conventional currencies in a number of crucial ways. Wallets, the peer-to-peer (P2P) network, and most fundamentally the blockchain are the three key elements of Bitcoin. The term "Bitcoin" is capitalized when referring to the cryptocurrency's idea or protocol; otherwise, it is used to refer to a single bitcoin, as in "I acquired two bitcoins today." Either BTC or XBT is the currency's official acronym on exchange platforms.

A user must build a wallet to store bitcoins, which stores the associated private key in encrypted form, which must be kept confidential. Cryptography is used to maintain the block chain's consistency and chronological order. A transaction must be submitted to the Bitcoin network to spend a bitcoin and signed using the private key, which cryptographically validates the wallet owner's approval and prohibits tampering. The transaction is published to all other nodes and becomes part of the block chain if the validation is successful.

Sentiment Analysis

Sentiment analysis is a branch of research that evaluates people's ideas, sentiments, assessments, attitudes, and emotions in written language. And is essential for companies to collect feedback on the internet to improve their products.

Twitter

The total number of tweets per day exceeds 500 million. Users can post their tweets, retweet other tweets, comment on tweets, and like tweets. Every registered user owns their timeline where all their tweets and retweets are listed. Based on the hashtag, Twitter can index tweets in their database to prepare them for the search or trend functionality.

Scope of proposed work

The proposed system's goal is to investigate Bitcoin price series in US dollars based on the daily close price. The Daily sentiment in Bitcoin tweets, the correlation, and the value of an accurate forecast are the main topics of discussion.

Importance and relation to previous work and the present developments

With 89% of the coins in circulation as of February 2021, the market value of Bitcoin had already topped US$1 trillion. As of June 2022, CoinMarketCap lists 20,028 crypto coins, largely due to the BTC open source initiative, which enables the ongoing development of new cryptocurrencies. Through the use of text and sentiment analysis techniques, this work seeks to delve into the uncharted waters of crypto market volatility in order to forecast BTC movement on a certain day.

Motivation behind the proposed work

With increasing amount of investors barging in to cryptocurrencies, we need a strategy to get away from the herd mentality the majority of the investors follow. That's where this system comes in. It helps investors in getting to know the mindset of the general public at a given daybefore diving in.

II. Literature survey

The study stated by Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan [1] used VADER" (Valence Aware Dictionary and Sentiment Reasoner). The outcome of their work was Both Google Trends and tweet volume were highly correlated with price. But, Sentiment of tweets was determined to not be a reliable indicator when cryptocurrency prices were falling. Valencia, Franco, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. (2019) [2] using RNN LSTM data is trained on multiple decision trees and the final output is the average from these multiple trees. It has been observed that out of 4 features the historic data has the highest weight followed by the volume of Bitcoin traded. The news and Twitter sentiment scores have a minimum weight. But, result showed that the news and the tweets didn't have that much of an impact on the price of BTC, which contradicts all the other papers in here. Wołk, K. in his work [3] Advanced social media sentiment analysis for short-term cryptocurrency price prediction used VADER, SVR. GBM, MLPNN techniques. The work used two predictive models that are the least square linear regression (LSLR) and Bayesian ridge regression models. They used mean to reduce error. Usage of too many algorithms study might be time consuming. Pano, Toni, and Rasha Kashef. (2020) [4] used VADER dictionary. And they developed 13 different preprocessing strategies for BTC tweets. They use the Pearson correlation between the VADER scores of each of the 13 intermediate preprocessing steps over time with BTC's closing prices per minute. This paper didn't predict prices, but just found the correlation between the prices and the tweets. Xin Huang, Wenbin Zhang, Xuejiao Tang , Mingli Zhang, Jayachander Surbiryala , Vasileios Iosifidis , Zhen Liu and Ji Zhang [5] used LSTM- Long Short-Term Memory (LSTM) combined with a Twitter sentiment analysis outperforms other machine learning models such as Support Vector Machine in predicting. It was developed to capture investor sentiments from Chinese markets. But, Chinese investors exchange crypto information via news articles and social media platforms.

III. SYSTEM ANALYSIS

Problem Definition / Framework

The volatility is what makes long-term cryptocurrency predictions more difficult. In this work, focused on cryptocurrency price predictions using linear regression models. Here, looking at predictions over a number of time intervals by considering various model features, like current price and previous price.

Proposed Methodology

The system gathers tweets on the Bitcoin cryptocurrency from a Kaggle dataset that is frequently updated in order to do pre-processing on the tweets by removing extraneous columns. Using the VADER Sentiment Analyser, compute a compound score and then using this compute sentiment score for each tweet. Afterwards, using data from Yahoo, compile historical Bitcoin prices. Further, begin the model preparation as well as the NLP modelling for the tweets. Tokenizing the text from the tweets will be used to train the TensorFlow Keras Sequential Model. In order to determine the public's overall consensus, plot a co nfusion matrix. After that, forecast the price over a period of time using historical prices. The architecture of the system is depicted as in Figure 1.

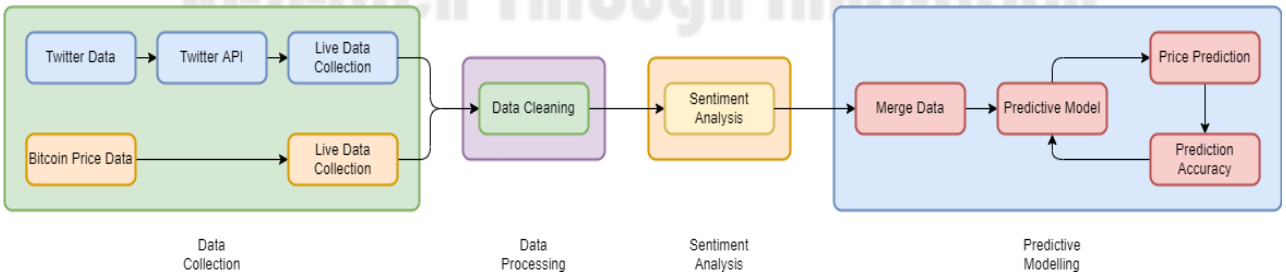


Figure 1: Block diagram of a Proposed Architecture

System Design

The six layers of Sequential Model used are Embedding Layer, Conv1D, MaxPooling Layer, Bi-directional LSTM, Dropout Layer and Dense Layer.

Embedding Layer

It performs embedding operations in input layer. It is used to convert positive into dense vectors of fixed size. Its main application is in text analysis. This is mainly used in Natural Language Processing related applications such as language modelling, but it can also be used with other tasks that involve neural networks.

Conv1D Layer

Kera's contains a lot of layers for creating Convolution based ANN, popularly called as Convolution Neural Network (CNN). All convolution layer will have certain properties, which differentiate it from other layers (say Dense layer).

Conv1D layer is used in temporal-based CNN. The input shape of the Conv1D will be in format: (batch-size, timesteps, features) where, batch size refers the size of the batch, timesteps refers the number of time steps provided in the input, features refer the number of features available in the input.

Max-Pooling Layer

It is used to perform max pooling operations on temporal data. Similarly, MaxPooling2D and MaxPooling3D are used for Max pooling operations for spatial data.

Bi-directional Layer

Bidirectional LSTMs will train two instead of one LSTMs on the input sequence. The first on the input sequence as is and the second on the reversed copy of the input sequence. They will provide context to the network and result in fast and full learning on the problem. LSTM processes information from inputs and then passed using the hidden state. The unidirectional LSTM will store information of the past because the inputs are seen from the past.

Dropout Layer

Dropout is one of the important concepts in the machine learning. It is used to fix the over-fitting issue. Input data may have some of the unwanted data, usually called as Noise. Dropout will try to remove the noise data and thus prevent the model from over-fitting.

Dense Layer

Dense layer is the regular deeply connected neural network layer. It is most common andfrequently used layer. Dense layer does the below operation on the input and return the output.

output = activation (dot (input, kernel) + bias)

where input represent the input data, kernel represent the weight data, dot represent numpy dotproduct of all input and its corresponding weights, bias represent a biased value used in machinelearning to optimize the model and activation represent the activation function. The model plot is shown Figure2.

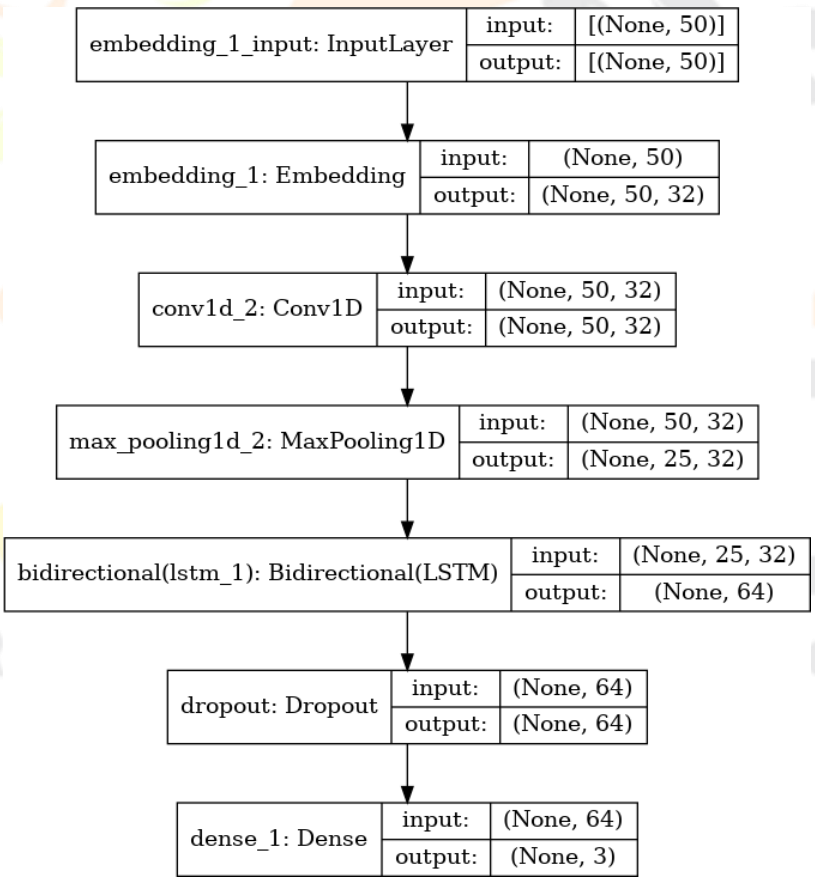


Figure 2: Dense Layer Model Plot

SYSTEM IMPLEMENTATION

The algorithm gathers tweets with the hashtag "BTC" as well as tweets about cryptocurrency markets. Use VADER to determine a compound score for it. Then, based on the amount of followers and likes a tweet has, calculate the emotion score for each tweet. Then read the Yahoo Bitcoin Price to learn how the cryptocurrency has changed over time in relation to the general opinion. Additionally, plot cross-correlation graphs for the price using Pearson, Kendall, and Spearman. Normalize the prices, and then compare them to the normalised tweet data, as well as the cryptocurrency derivative and sentiment score, before printing the resulting graphs.

Beginning the NLP modelling. Using WordNetLemmatizer, start eliminating the stop words from the tweets. After the tweets have been cleaned, create a polarity and subjectivity score using the Text Blob module and add it to the dataset. Furthermore, print the sentiment score-derived bar graph of the overall bitcoin market sentiment. Get the Sequential Model ready for training. For this phase, utilize a typical LSTM with an epoch size of 10. After training, print the classification report with the accuracy, precision, and F1 score as well as the model accuracy and model loss graphs. Print a confusion matrix to make it simpler to comprehend. Tokenize the tweets to create an integer sequence for the main model, and then pad each one to the same length. However, for this model, utilise Bi-Directional LSTM with an epoch size of 50. In addition to printing the confusion matrix, the classification report with the accuracy, precision, and F1 score is also printed here when all the epochs have been completed. After that, you can print the price graph that was predicted using a second Bidirectional LSTM model that was trained using the price history.

Modules/algorithms Developed

Sequential Model and Bi-directional LSTM Code Snippet

```
model= Sequential()
model.add(Embedding(vocab_size, embedding_size, input_length=max_len))
model.add(Conv1D(filters=32, kernel_size=1, padding='same', activation='relu'))
model.add(MaxPooling1D(pool_size=2))
model.add(Bidirectional(LSTM(32))) model.add(Dropout(0.4))
model.add(Dense(3,activation='softmax'))
model.compile(loss='categorical_crossentropy',optimizer=sgd,metrics=['accuracy',Precision(),
Recall()])
history = model.fit(X_train, y_train, validation_data=(X_val, y_val), batch_size=batch_size,
epochs = epochs, verbose = 1)
```

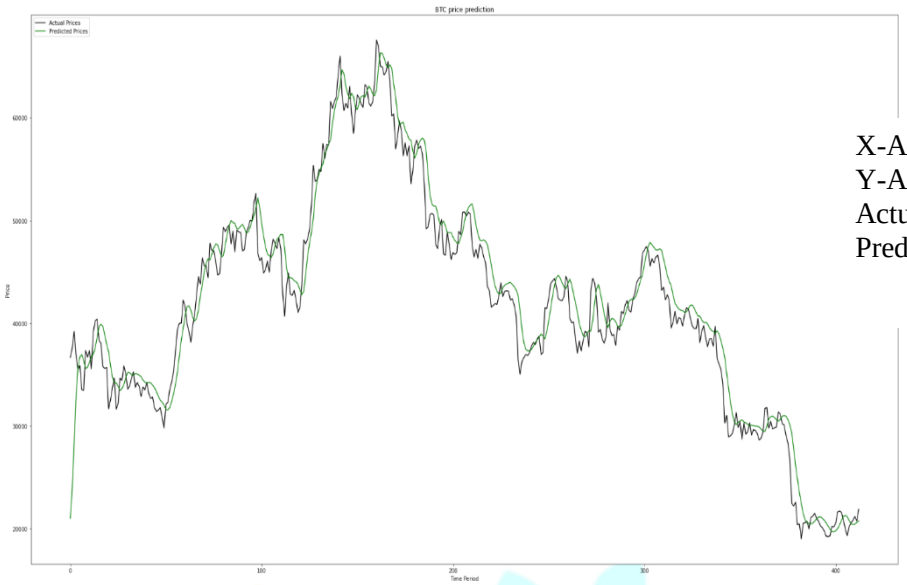
Sentiment Score Formula

```
(s["compound"] * ((int(s["user_followers"]))) * ((int(s["user_favourites"])+1)  /
                int(s['user_followers']+1)) * ((int(s["is_retweet"])+1)))
```

IV Results

Test Cases

Tested our price prediction model. Gathered the actual Bitcoin Price History from February 6th, 2021 till date and then used that



Data to train our model and test it for a period from October 1st, 2021 till now. The output was very satisfactory and has been as plotted in figure 3.

Tweet Dataset

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags
0	DeSota Wilson	Atlanta, GA	Biz Consultant, real estate, fintech, startups...	2009-04-26 20:05:09	8534.0	7605	4838	False	2021-02-10 23:59:04	Blue Ridge Bank shares halted by NYSE after #b...	['bitcoin']
1	CryptoND	NaN	🤓 BITCOINLIVE is a Dutch platform aimed at inf...	2019-10-17 20:12:10	6769.0	1532	25483	False	2021-02-10 23:58:48	🤓 Today, that's this #Thursday, we will do a "...	['Thursday' 'Btc', 'wallet', 'security']
2	Tdlmatias	London, England	IM Academy : The best #forex, #SelfEducation, ...	2014-11-10 10:50:37	128.0	332	924	False	2021-02-10 23:54:48	Guys evening, I have read this article about B...	NaN
3	Crypto is the future	NaN	I will post a lot of buying signals for BTC tr...	2019-09-28 16:48:12	625.0	129	14	False	2021-02-10 23:54:33	$BTC A big chance in a billion! Price: \487264...	['Bitcoin', 'FX', 'BTC', 'crypto']
4	Alex Kirchmaier ATSE #FactsSuperspreader	Europa	Co-founder @RENJERJerky | Forbes 30Under30 | I...	2016-02-03 13:15:55	1249.0	1472	10482	False	2021-02-10 23:54:06	This network is secured by 9 508 nodes as of t...	['BTC']

Figure 4: Sample Tweet Dataset

Modified tweet dataset by adding Compound Score and Sentiment Score is shown in Figure 5.

user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet	compound	score
2021-06-14 16:19:03	580.0	143.0	575.0	False	2021-11-12 11:19:49	Bitcoin Price (USD): 63998.36 \nEthereum Price...	['Bitcoin', 'Dogecoin', 'Ethereum', 'Safemoon'...	TwitterBot-6/15/21	False	0.0000	0.000000
2020-05-12 14:29:18	732.0	1.0	3.0	False	2021-08-21 03:49:47	$EVLO /T189cHCL3G Evelo Biosciences Inc Our an...	['investing', 'bitcoin', 'smallbusiness']	intratio	False	-0.5719	-2.284479

Figure 5: Compound and Sentiment Score

Crypto currency evolution compared to twitter sentiment is plotted and shown as in Figure6.

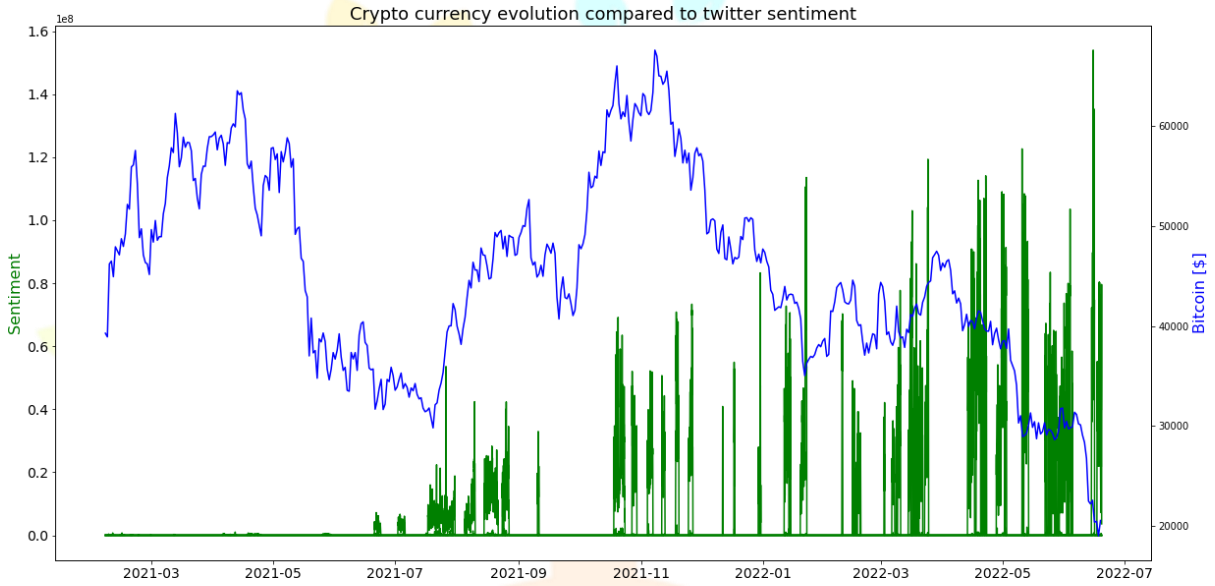


Figure 6: Crypto currency evolution compared to twitter

sentimentBitcoin Price: X-Axis: Time Period

Sentiment: Y-Axis: Sentiment

Normalized Crypto currency evolution compared to normalized twitter sentiment is shown as plotted in Figure 7.

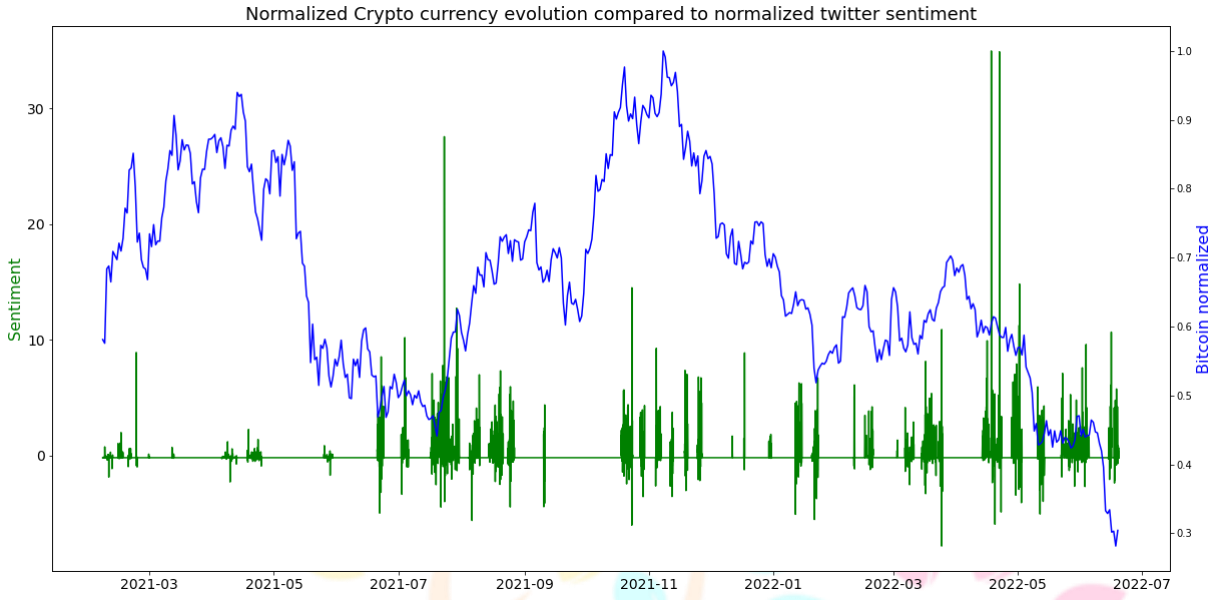


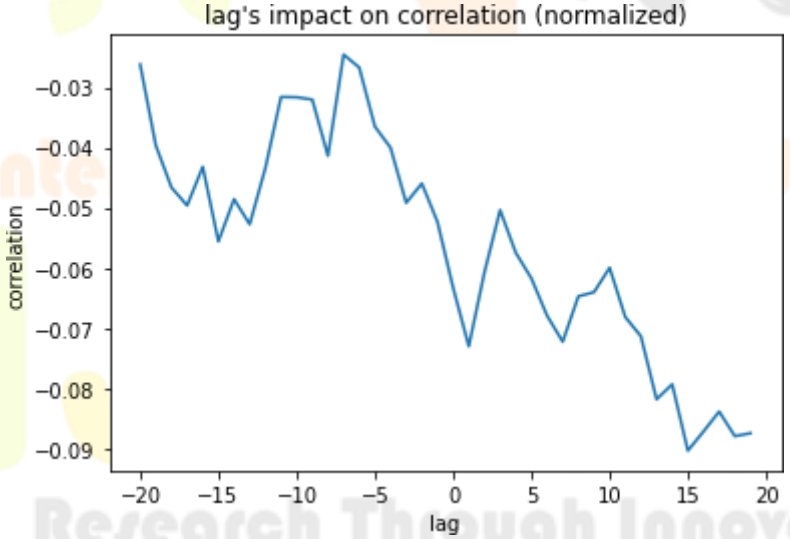
Figure 7: Normalized Crypto currency evolution compared to normalized twitter sentiment

Bitcoin Price:

Sentiment:

X-Axis: Time PeriodY-Axis: Sentiment

Lag's Impact on Correlation (normalized) is also shown in the following Figure.



Derivative of crypto currency and sentiment's score is plotted as shown in Figure9.

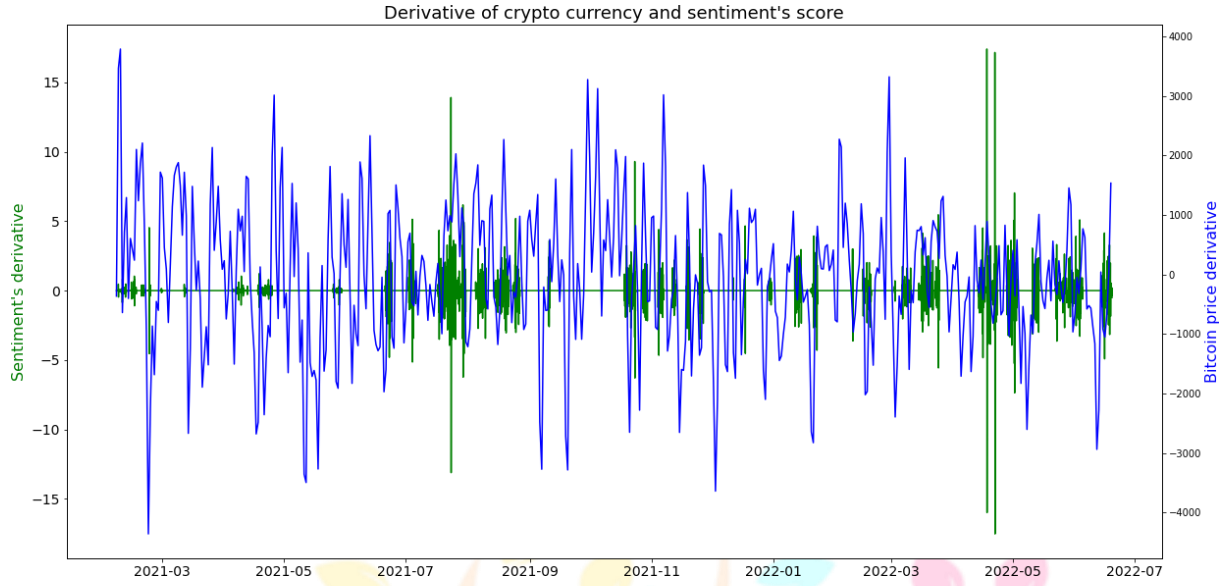


Figure 9: Derivative of crypto currency and sentiment's

scoreBitcoin Price: ——— X-Axis: Time Period

Sentiment: ——— Y-Axis: Sentiment

Cleaned Tweet Dataset and after adding subjectivity and polarity score are shown in Figure10 & Figure11 Respectively.

	tweets	cleaned_tweets	date_clean
0	Bitcoin Price (USD): 63998.36 \nEthereum Price...	Bitcoin Price USD Ethereum Price USD Dogecoin ...	2021-11-12
1	$EVLO /T189cHCL3G Evelo Biosciences Inc Our an...	EVLO T cHCL G Evelo Biosciences Inc Our analys...	2021-08-21
2	There we will be no more BTC selling I assure ...	There BTC selling I assure The way	2021-12-30
3	digitalmarketing tumblr twitter facebook insta...	digitalmarketing tumblr twitter facebook insta...	2021-11-06
4	Laikacoin is Back communitycoin charitycoin t...	Laikacoin Back communitycoin charitycoin chang...	2021-08-06

Figure 10: Cleaned Tweet Dataset

	tweets	cleaned_tweets	date_clean	crypto_sentiment	subjectivity	polarity
0	Bitcoin Price (USD): 63998.36 \nEthereum Price...	Bitcoin Price USD Ethereum Price USD Dogecoin ...	2021-11-12	negative	0.0	0.000
1	$EVLO /T189cHCL3G Evelo Biosciences Inc Our an...	EVLO T cHCL G Evelo Biosciences Inc Our analys...	2021-08-21	negative	0.4	-0.175
2	There we will be no more BTC selling I assure ...	There BTC selling I assure The way	2021-12-30	positive	0.0	0.000
3	digitalmarketing tumblr twitter facebook insta...	digitalmarketing tumblr twitter facebook insta...	2021-11-06	positive	0.0	0.000
4	🌈Laikacoin is Back communitycoin charitycoin t...	Laikacoin Back communitycoin charitycoin chang...	2021-08-06	negative	0.0	0.000

Figure 11: Dataset with Subjectivity and Polarity Score

The total Positive and Negative Sentiment counts is plotted in Figure12.

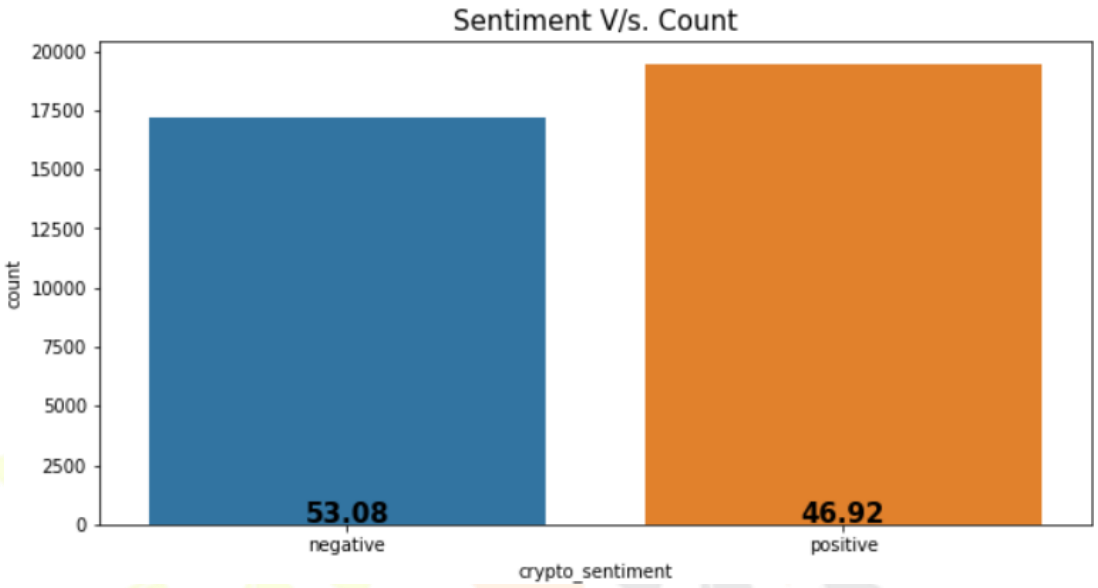


Figure 12: Positive v/s Negative Sentiment

The type of the Training Model used is sequential and its details are shown in Figure13.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 30, 100)           2000000
_________________________________________________________________
conv1d (Conv1D)              (None, 30, 32)            9632
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 15, 32)            0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 15, 32)            3104
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 7, 32)             0
_________________________________________________________________
lstm (LSTM)                  (None, 100)               53200
_________________________________________________________________
dense (Dense)                (None, 3)                 303
=================================================================
Total params: 2,066,239
Trainable params: 2,066,239
Non-trainable params: 0
_________________________________________________________________
```

Figure 13: Training Model

Validation of the Training Model is done with respect to accuracy and Loss and is as show in figure 14.

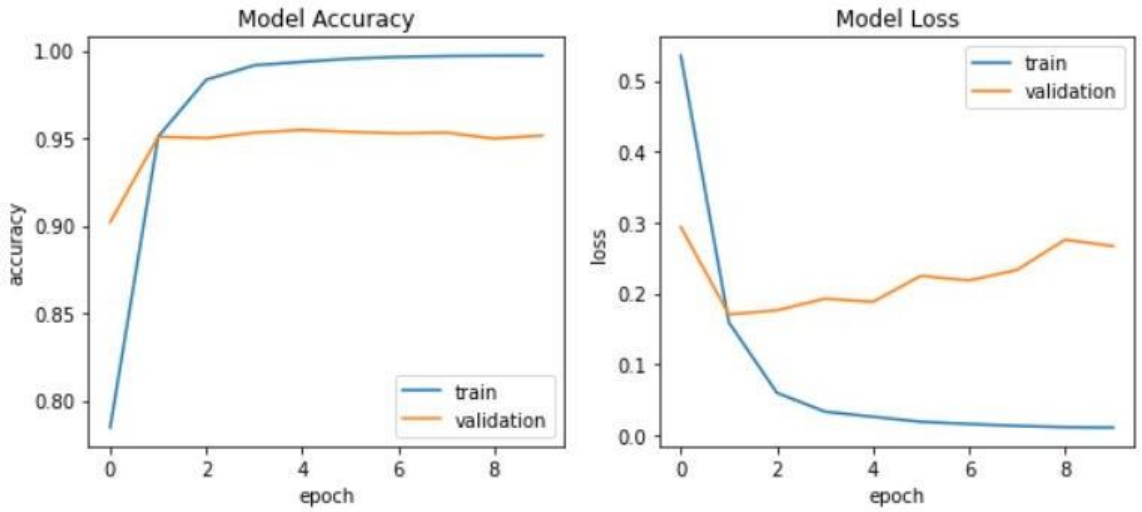


Figure 14: Training Model Accuracy and Loss

X-Axis: epoch
Y-Axis: Accuracy

X-Axis: epoch
Y-Axis: Loss

The figure 15 shows the Model Classification Report with respect to Accuracy, precision, recall f1-score and support.

Accuracy: 95.2%

	precision	recall	f1-score	support
0	0.90	0.78	0.84	808
1	0.97	0.96	0.97	2927
2	0.95	0.98	0.96	3592
accuracy			0.95	7327
macro avg	0.94	0.91	0.92	7327
weighted avg	0.95	0.95	0.95	7327

Figure 15: Classification Report

Confusion Matrix for the Training Model is as shown in figure 16.

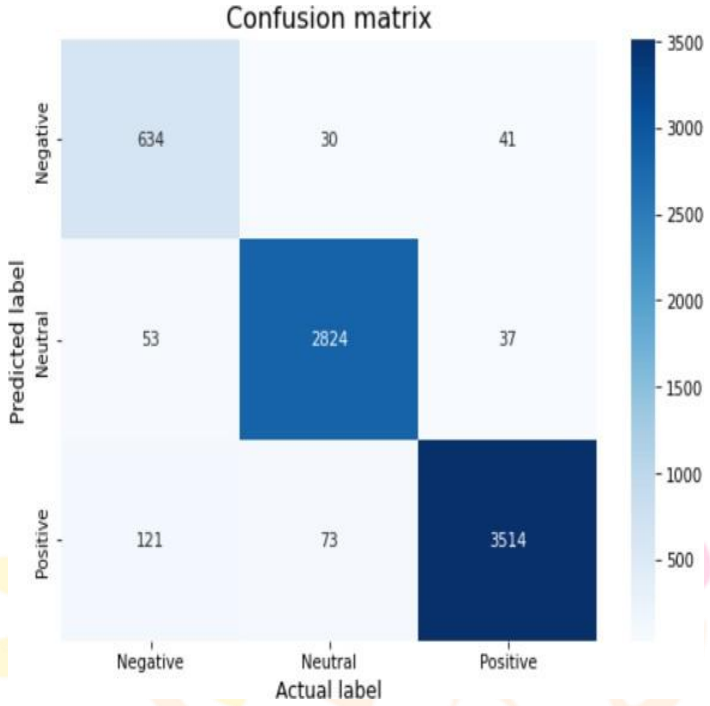


Figure 16: Training Model Confusion Matrix

The number of actual positive, neutral and negative

values are2 3785

1 3201

0 885

The number of predicted positive, neutral and negative

values are2 3891

1 3185

0 795

So, the accuracy is 94.5%, that means 7871 values are true values

	precision	recall	f1-score	support
0	0.88	0.79	0.83	885
1	0.96	0.96	0.96	3201
2	0.95	0.97	0.96	3785

Details of Testing Model Plot is shown in figure17

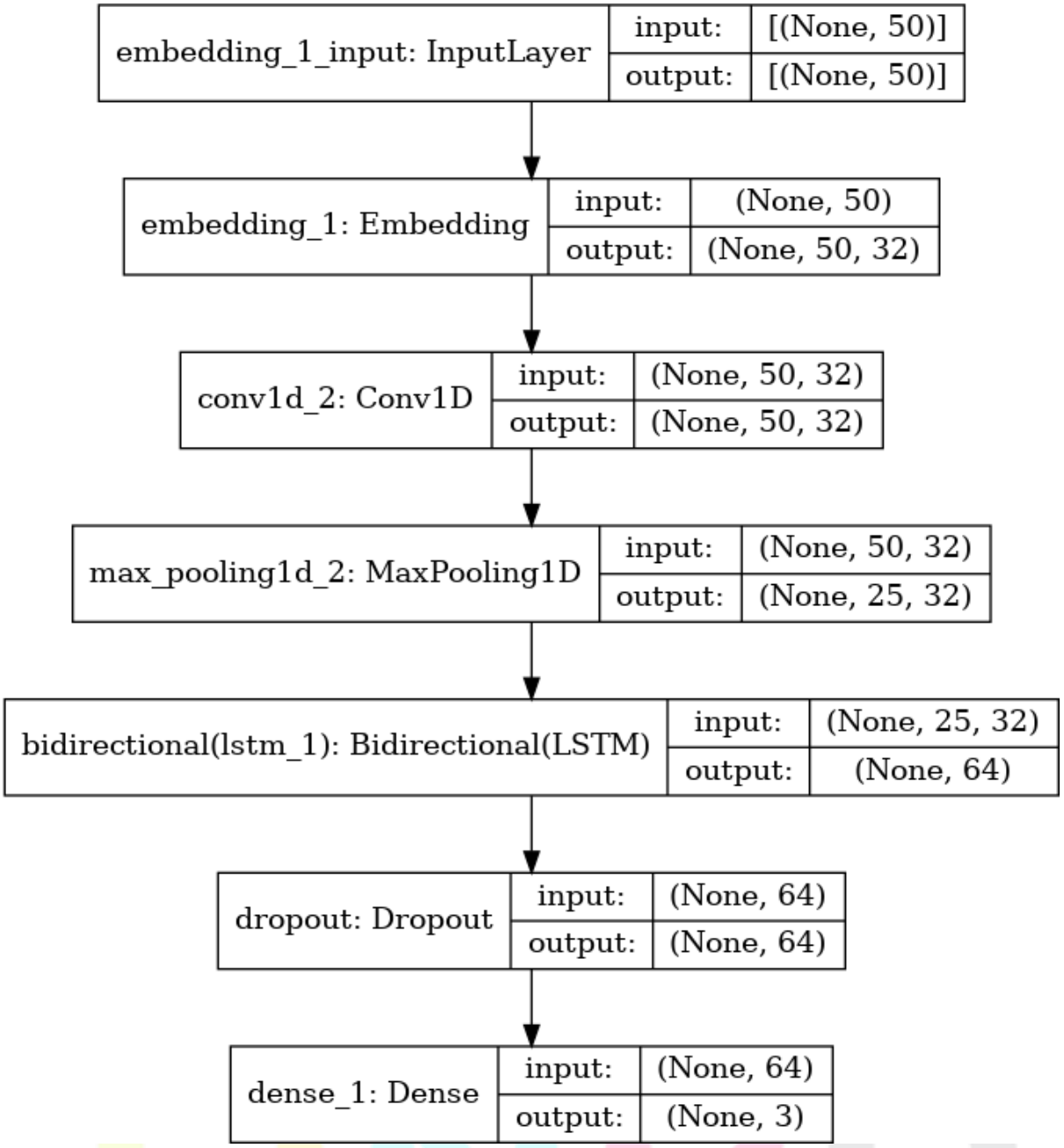


Figure 17: Testing Model Plot

Testing Model Accuracy is computed in Figure 18 & plotted in Figure 19 respectively.

```
Accuracy  : 0.8976
Precision : 0.9015
Recall    : 0.8894
F1 Score  : 0.8954
```

Fig ure18: Testing Model Accuracy

Testing Model Accuracy and Loss

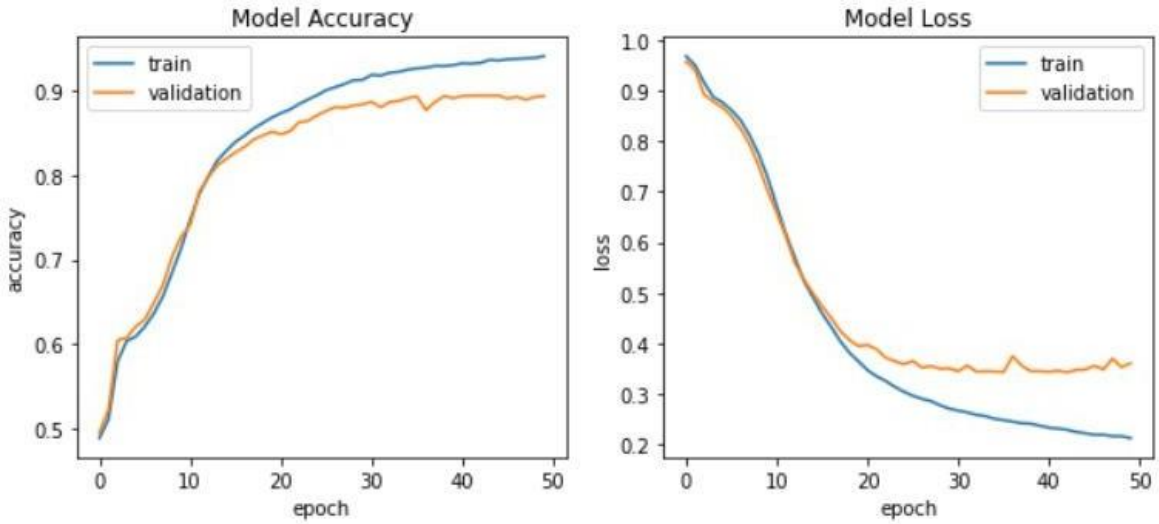


Figure 19: Testing Model Accuracy and Loss

X-Axis: epoch
Y-Axis: Accuracy

X-Axis: epoch
Y-Axis: Loss

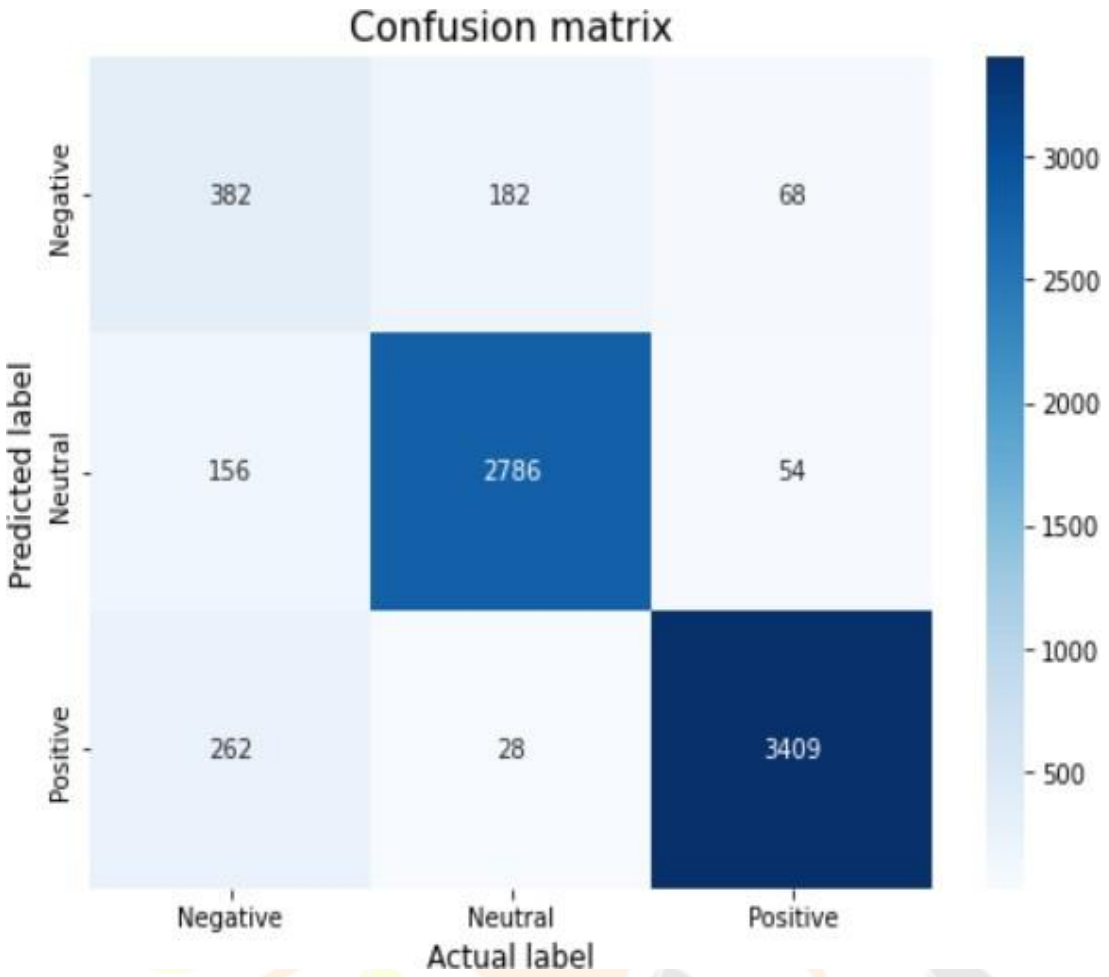


Figure 20: Testing Model Confusion Matrix

The confusion matrix of Figure 20 shows the number of actual positive, neutral and negative values are as follows

2 3777

1 3234

0 860

The number of predicted positive, neutral and negative

values are:2 3915

1 3369

0 587

And the classification

report isAccuracy:

0.9071

Precision: 0.9106

Recall: 0.9020

F1 Score: 0.9063

The following Price Dataset is considered and price prediction is done as shown in Figure22.

	High	Low	Open	Close	Volume	Adj Close
Date						
2021-02-06	40846.546875	38138.386719	38138.386719	39266.011719	71326033653	39266.011719
2021-02-07	39621.835938	37446.152344	39250.191406	38903.441406	65500641143	38903.441406
2021-02-08	46203.929688	38076.324219	38886.828125	46196.464844	101467222687	46196.464844
2021-02-09	48003.722656	45166.960938	46184.992188	46481.105469	91809846886	46481.105469
2021-02-10	47145.566406	43881.152344	46469.761719	44918.183594	87301089896	44918.183594

Figure 21: Price Dataset

Price Prediction Graph

Figure 22: Price Prediction Graph

Actual Price: _____ X Axis: Time Period

Predicted Price: _____ Y-Axis: Price

V CONCLUSION

In this study, historical price data and tweets concerning the Bitcoin crypto currency are collected, and the collected data is then cleaned to make it clearer to understand. After that, LSTM is used to train machine learning models that can forecast the general mood of the population. In order to predict price movement over the ensuing few days, the model is then trained using historical price data. This is advantageous for individuals who wish to enter the Bitcoin market but want to first measure public sentiment in order to plan their entry and minimise losses.

VI BIBLIOGRAPHY

1. Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol. 1: No. 3.
2. Valencia, Franco, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. 2019. "Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning" *Entropy* 21, no. 6: 589. https://doi.org/10.3390/e21060589.
3. Wołk, K. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Systems. 2020; 37:e12493. https://doi.org/10.1111/exsy.12493.
4. Pano, Toni, and Rasha Kashef. 2020. "A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19" *Big Data and Cognitive Computing* 4, no. 4: 33. https://doi.org/10.3390/bdcc4040033
5. Xin Huang , Wenbin Zhang , Xuejiao Tang , Mingli Zhang , Jayachander Surbiryala , Vasileios Iosifidis , Zhen Liu 5 and Ji Zhang , LSTM Based Sentiment Analysis for Cryptocurrency Prediction https://arxiv.org/pdf/2103.14804.pdf [5]
6. Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis https://cs229.stanford.edu/proj2015/029_report.pdf [6]
7. Cryptocurrency Price Prediction Using News and Social Media Sentiment https://cs229.stanford.edu/proj2017/final-reports/5237280.pdf [7]
8. Forecasting Price of Crypto-currencies using Tweets Sentiment Analysis https://ieeexplore.ieee.org/abstract/document/8530659 [8]
9. PredictingCryptocurrency Value using SentimentAnalysis https://ieeexplore.ieee.org/abstract/document/9065838 [9]
10. Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis https://ieeexplore.ieee.org/abstract/document/8586824 [10]
11. Liu, Bing (2015): Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015. [11]
12. Yi, Jeonghee; Nasukawa, Tetsuya; Bunescu, Razvan; Niblack, Wayne (2003): Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques [12]