Cryptocurrency Price Prediction using Forecasting and Sentiment Analysis

Shaimaa Alghamdi[1], Sara Alqethami[2], Tahani Alsubait[3], Hosam Alhakami[4]
College of Computers and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia[1,2,3,4]

***Abstract*—In recent years, many investors have used cryptocurrencies, prompting specialists to find out the factors that affect cryptocurrencies' prices. Therefore, one of the most popular methods that have been used to predict cryptocurrency prices is sentiment analysis. It is a widespread technique utilized by many researchers on social media platforms, particularly on Twitter. Thus, to determine the relationship between investors' sentiment and the volatility of cryptocurrency prices, this study forecasts the cryptocurrency prices using the Long-Term-Short-Memory (LSTM) deep learning algorithm. In addition, Twitter users' sentiments using Support Vector Machine (SVM) and Naive Bayes (NB) machine learning approaches are analyzed. As a result, in the classification of the bitcoin (BTC) and Ethereum (ETH) datasets of investors' sentiments into (Positive, Negative, and Neutral), the SVM algorithm outperformed the NB algorithm with an accuracy of 93.95% and 95.59%, respectively. Furthermore, the forecasting regression model achieves an error rate of 0.2545 for MAE, 0.2528 for MSE, and 0.5028 for RMSE.**

Keywords*—*Sentiment analysis; cryptocurrencies; forecasting; bitcoin; ethereum

I. Introduction

Recently, the prices of financial assets have been dynamically changing, which means it changes asynchronously as new information becomes available [1]. Therefore, the future prediction of finance growth in stocks, shares, and digital currency flow data is difficult for speculators and investors; these cryptocurrencies have skyrocketing and sudden fall characteristics, which means they have high price volatility over time [2].

In addition, cryptocurrencies are an alternative medium of exchange consisting of numerous decentralized crypto coin types. The essence of each crypto coin is in its cryptographic foundation. Secure peer-to-peer transactions are enabled through cryptography in this secure and decentralized exchange network based on blockchain technology [3]. Since its inception in 2009 [4], BTC has become a digital commodity of interest as some believe the crypto coins' worth is comparable to that of traditional fiat currency [5]. Unlike our usual currencies, cryptocurrencies are free from regulatory norms and do not have a central governing authority [6]. Therefore, the cryptocurrency market is investor-driven. Thus, it can be said to be affected by socially constructed opinions, and future expectations of the cryptocurrency holders and future investors [7].

Many currency users share stock price recommendations via electronic platforms. According to Abualigah et al. [8], Twitter is one of the essential sources of users' opinions on various topics. Thus, individuals can express their opinions and share alternative viewpoints on any topic. As a result, it is necessary to use modern technologies and advanced artificial intelligence methods to analyse Twitter users' opinions.

The research contributions are as follows: (i) Apply machine learning (ML) algorithms to analyse the cryptocurrency users' sentiments, (ii) Apply deep learning (DL) models to forecast the cryptocurrencies' prices, (iii) Analyse the correlation between cryptocurrency users' sentiments and price volatility in the cryptocurrency market, and (iv) Evaluate the performance of the proposed method and compare the results with those of state-of-the-art algorithms and previous studies in the same field.

This paper is structured as follows: Section II provides a literature review of the previous studies on cryptocurrency price forecasting and sentiments analysis. Section III illustrates the overall structure of the research model and the methods utilized to achieve the results. Section IV contains the collected datasets and the pre-processing techniques. Section V presents the experimental results in tables and figures and compares the results with other previous studies. In the final Section VI, we provide a conclusion, limitations, and future work.

II. Literature Review

Cryptocurrencies are a form of alternative currency comprised of various types of decentralized crypto [9]. These cryptocurrencies exhibit rapid growth and rapid downturn characteristics, implying a high degree of price volatility over time [10]. Many studies focused on predicting the "Price" of cryptocurrency by using various techniques such as the Autoregressive Integrated Moving Average (ARIMA) time-series model [11]. They aim to predict the prices using daily, weekly, and monthly time series. On the other hand, several recent studies have analyzed the sentiments of cryptocurrency users using ML [12] and DL models [5]. Additionally, These studies examined whether public sentiment measured by social media datasets are related to or predictive of cryptocurrency values. In particular, it is possible to predict the volatility of cryptocurrencies price by analyzing public sentiment on Twitter and determining the relationship between the sentiments expressed by investors on Twitter.

Accordingly, Rahman et al. [12] presented ML models based on the Twitter dataset. The researchers aim to find an association between user sentiment and BTC price. However, they use a variety of algorithms, such as Support Vector Regression, Decision Tree Regression DTR, and Linear Regression LR. As a result of the experiment, there is a discernible relationship between sentiment on Twitter and price change, based on the highest accuracy obtained from the decision tree algorithm compared with other algorithms is 75%. Similarly, Sattarov et al. [13] examined the extent to which BTC prices are influenced by investors' sentiments using a ML model. Twitter and the BTC price datasets were collected over a 60-day period, from March 12 to May 12. Additionally, the researchers used the Random Forest RF algorithm to determine a correlation between cryptocurrency users' opinions and feelings and price variation. The implemented model achieves 62.48% accuracy and a minimum error of 21.84%.

In another study conducted by Mittal et al. [14], The interrelationships between BTC price and Twitter and Google search were identified using the BTC price, tweets, and Google search patterns. According to the investigation, there is a correlation between BTC pricing, Twitter, and Google search behaviors. In addition, the authors apply Linear regression LR, polynomial regression PG, Recurrent Neural networks, and LSTM based analysis of different datasets collected from 9 April 2014 to 07 January 2019. The polynomial regression method outperforms the other techniques by achieving an accuracy of 77.01% and 66.66% of Tweet Volume and Google trends, respectively. Thus, the findings show a significant association between Google Trends and Tweet volume data and the price of BTC, but no significant correlation with tweet sentiments.

Other studies applying sentiment analysis technology to cryptocurrency trading had similar results. For example, Pant et al. [5] developed a novel DL algorithm by combining the BTC sentiment score and historical price. Their objective is to establish a link between user emotions and BTC price volatility using the Recurrent Neural Network RNN model. The datasets chosen covered the same time period, from January 1 to December 31, 2015, in order to investigate the correlation between them. According to the experiment results, the distinction between positive and negative tweets is 81.39 percent accurate. The same proposed model achieves a 77.62 percent accuracy in predicting BTC prices. Moreover, Pathak and Kakkar [15] present DL trained model using LSTM networks. They intend to forecast the overall price trend by analyzing BTC and Twitter data spanning 450 hours. Their study established a link between Twitter users' sentiments and the BTC trading currency's pricing, as the relationship between them was 77.89% accurate.

Furthermore, Aggarwal et al. [16] utilized a symmetric-deep learning approach with value parameters to assess the impact of BTC price prediction on socioeconomic indicators. In particular, they investigated the effect of Gold price and investors' sentiments on the price of BTC by using DL models such as Convolutional Neural Network (CNN), LSTM, and Gated Recurrent Unit (GRU). To enhance the results, they used a variety of datasets from January 15, 2017, to May 12, 2017. Researchers have noticed the significant effect between Twitter users' sentiments and BTC price volatility as the posting of tweets containing positive feelings leads to an increase in the price of BTC. And vice versa. In another study conducted by [17], DL models based on the StockTwits platform and cryptocurrencies datasets are utilized. The authors applied efficient language modeling tools such as recursive neural networks (RNNs) using datasets from March 2013 to May 2018. The findings show a significant association between speculators' posts and cryptocurrency volatility prices.

Most studies used either deep neural network models such as LSTM or conventional ML models such as RF to forecast cryptocurrency price volatility. Additionally, they employ various preprocessing strategies when developing a cryptocurrency models. However, only a few studies conducted a comparative analysis of DL to identify an optimal preprocessing strategy. These studies were conducted over various timeframes, making them relatively old in this rapidly evolving cryptocurrency market. Thus, additional research is required on forecasting cryptocurrency prices and analyzing investor sentiments using DL approaches, as the literature has not been extensively exploited. Therefore, new studies must be conducted to ensure that these results remain valid in 2022 and to discover new patterns.

III. Methodology

This section explains the proposed cryptocurrency forecasting model using DL and ML. The sentiments analysis model pre-processes the tweets, calculates the sentiment score using ML algorithms, and classifies them into (Positive, Negative, and Neutral). In contrast, the forecasting model uses the historical price dataset and works on predicting the next three months using the DL algorithm. The methodology of the classification and regression models used to forecast the prices and find the correlation between investors' sentiments and cryptocurrency historical price is presented in Fig. 1.

A. Sentiment Analysis Model

This step aims to apply the primary goal of the research, which is sentimental characteristics tweets measurements such as polarity and subjectivity. The TextBlob3 Python library was utilized to process the Twitter dataset by providing the natural-language processing NLP features [18]. This strategy classifies the polarity of the tweet into positive, neutral, and negative groups as '1', '0', and '-1'. Table I presents the sentiments divided process into three groups (positive, negative, and neutral) depending on the tweets' polarity.

TABLE I. Polarity Classification

Value of Polarity	Sentiment
>0	Positive
0	Neutral
<0	Negative

The sentiments analysis model of cryptocurrency users was utilized using supervised ML approaches, including SVM, and NB chose these approaches for modeling because it is faster and more lightweight in the classification processes.

1) Support Vector Machine (SVM): is a supervised ML algorithm utilized for classification and regression purposes. Both linear and nonlinear classification is executed by SVM. It is created the line between two classes. That means all points in the same part has the same category. Moreover, It can be more than two lines to separate the categories. The vectors near the hyperplane are the support vectors. In addition, to solve classification problems, there are four kernel functions (linear, polynomial, radial-based, and sigmoid) [19]. The advantages of the SVM algorithm are the better classification accuracy and the best analysis performance if the input data is correctly labeled before the process [20].

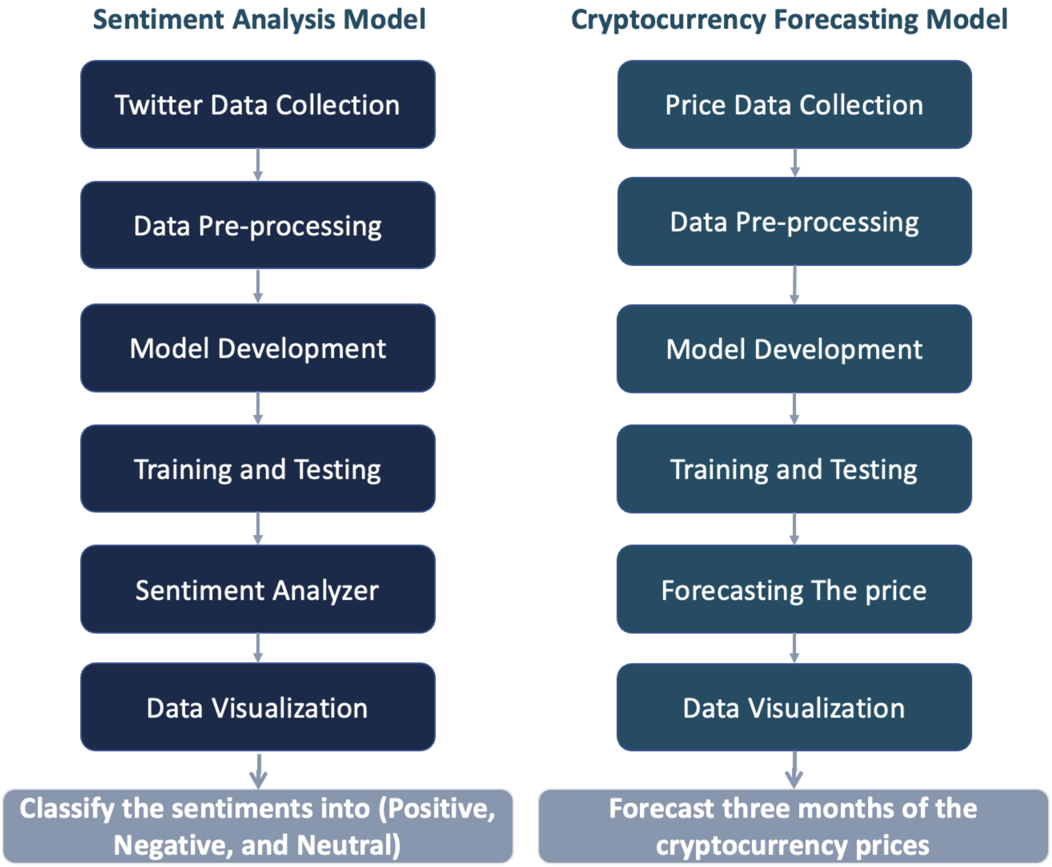

Fig. 1. Forecasting Model Structure.

2) Naive Bayes (NB): is a generative learning algorithm that solves text classification and sentiment analysis ML models based on Bayes' theorem. All features assumed as independent thought give the class value [21]. The term "Naive" directs to data points that are unrelated to one another. The advantages of the NB algorithm, multinomial classifier NB is commonly utilized in text categorization cases, and capable to build, use, train, and ignoring useless variables [22]. The (1) presents in the mathematical equation of NB.

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)} \tag{1}$$

Where:

- P(c | x) is the probability posterior of the given class value.
- P(c) is a prior probability of class.
- P(x) is a prior probability of value.
- P(x | c) is the probability of value given class.

B. Cryptocurrency Forecasting Model

LSTMs were chosen for modeling the cryptocurrency forecasting model since it is an ideal algorithm for time series forecasting and works well with historical data by storing memory. Furthermore, it solves complex problems that earlier recurrent network algorithms have never been able to solve. Therefore, Table II presents the hyperparameter, the values that control the learning process for the LSTM model.

TABLE II. HYPERPARAMETER VALUES

Activation Function	Sigmoid
Epochs	150
Hidden layers	One hidden layer
Batch size	256
Optimizer	Adam
Learning rate	0.00050

1) Long-Term-Short-Memory (LSTM): is a form of RNN with extra elements for memorizing sequential input. The cell state, which transmits information across the sequence chain, is a crucial component of LSTM. It serves as the network's memory. Because information can be withdrawn or added via gates, the cell state can truly hold only the necessary information in the sequence. During training, the gates learn what information is important to keep or forget. As a result, information from previous stages now influences later stages in the sequence [23].

C. Evaluation Metrics

There is a need to evaluate the performance of the models involved.

- **Regression**
  In the regression model, the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model. Scale-dependent errors are commonly used, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [24] as follows:
  Mean Absolute Error (MAE): calculated the average of the original and forecasted values [24]. This is expressed in mathematical terms as (2):

$$MAE = \frac{1}{n}\sum_{n=1}^{t=1} |e_t| \tag{2}$$

  Mean Squared Error (MSE): calculated the square average of the difference between original and forecasted values [24]. This is expressed in mathematical terms as (3):

$$MSE = \frac{1}{n}\sum_{n=1}^{t=1} e_t^2 \tag{3}$$

  Root Mean Squared Error (RMSE): The forecast errors standard deviation. That means the residuals measurement of far points from the regression line data [24]. This is expressed in mathematical terms as (4):

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{t=1} e_t^2} \tag{4}$$

  Where:
  - n is the sample's forecast total number.
  - e is the real value of the sample.
  - t is the forecasting value of the sample.

- **Classification**
  On the other hand, the evaluation metrics used for classification models the performance of the algorithms are Accuracy, Precision, Recall, and F1-Score [25] as follows.

  Precision The positive quantification, which are True Positives (TP) and False Positives (FP) number of predictions [25]. This is expressed in mathematical terms as (5):

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

  Recall The number of positive class predictions made from all positive cases in the dataset that can be counted. And when there is a large cost related to False Negatives (FN) [25]. This is expressed in mathematical terms as (6):

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

  F1-Score Precision and Recall have a harmonic mean [26]. This is expressed in mathematical terms as (7):

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

  Accuracy The intuitive performance measure, which is the observed the correctly forecasted ratio to the total observations [26]. This is expressed in mathematical terms as (8):

$$Accuracy = \frac{(TP + TN}{TP + TN + FP + FN} \tag{8}$$

IV. Experiments

This section illustrates the selected dataset in this research and the pre-processing steps.

A. Dataset

According to Coinmarketcup [27], there are 10111 various types of cryptocurrencies in circulation, such as BTC, ETH, Binance (BNB), and Cardano (ADA). Furthermore, BTC is still the prevalent cryptocurrency that has been widely used since its creation. Then ETH followed it as the top cryptocurrency compared to other currencies. Thus, this study focuses on the two most-traded currencies: BTC and ETH.

- **Twitter Dataset**
  An open-source Python library known as Tweepy [28] is used to access the Twitter API and extract the Twitter dataset. Therefore, tweets were collected using three hashtags and three keywords for two cryptocurrency types: The BTC cryptocurrency collection process used keywords "Bitcoin", "BTC", and "BTCUSD" along with their respective hashtags ["#Bitcoin", "#BTC", "#BTCUSD"]. And, the ETH cryptocurrency collection process used keywords "Ethereum", "ETH", and "Etherum" along with their respective hashtags ["#Ethereum", "#ETH", "#Etherum"]. Furthermore, the period we focus on is five months, and assigned to extract (2000 tweets) every seven days for 150 days, between January 1, 2022, and May 9, 2022. The final dataset was around 37,998 for BTC cryptocurrency tweets and around 37,997 for ETH cryptocurrency and separately saved in a CSV file. Hence, the dataset for all tweets with different cryptocurrencies is around 75,995 posted in the English language.

- **Cryptocurrency Dataset**
  The dataset of the two currencies was collected from CoinMarketCap [27], the world's most-referenced price-tracking website for crypto assets in the rapidly increasing cryptocurrency market [10]. The BTC dataset starts from September 17, 2014, to March 31, 2022. On the other hand, the ETH dataset starts from November 9, 2017, to March 31, 2022. However, in this research, the 1-minute interval trading exchange data rate in USD is focused on. In addition, the two datasets consist of the Date, Opening, Closing, Low, Adj Close, and Volume of transactions which increase over time. Additionally, the BTC and ETH datasets are saved in CSV files separately.

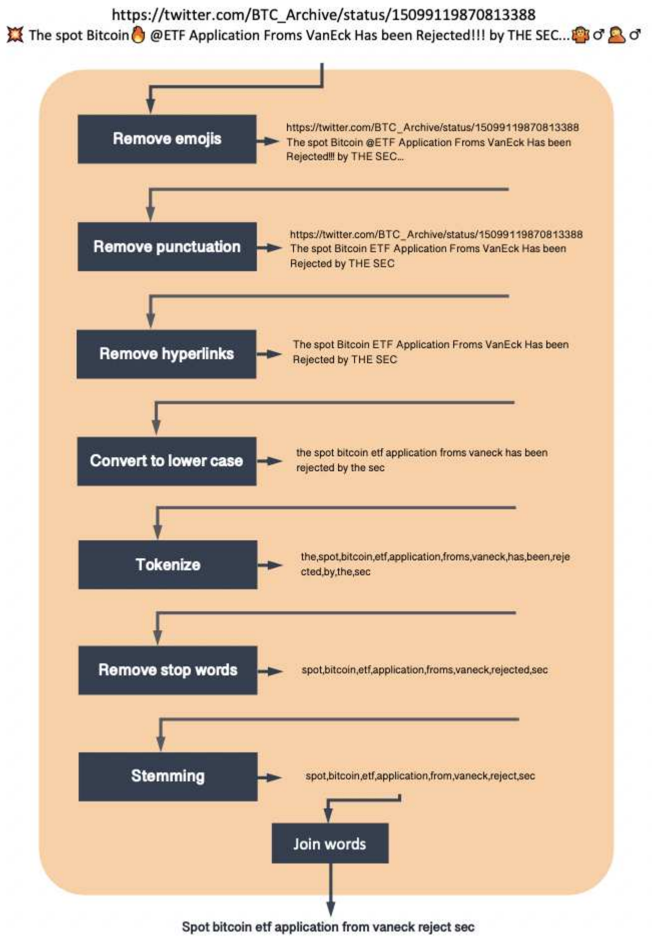
https://twitter.com/BTC_Archive/status/15099119870813388
💥 The spot Bitcoin🔥 @ETF Application Froms VanEck Has been Rejected!!! by THE SEC...🙈♂ 🤦♂
Remove emojis
https://twitter.com/BTC_Archive/status/15099119870813388 The spot Bitcoin @ETF Application Froms VanEck Has been Rejected!!! by THE SEC...
Remove punctuation
https://twitter.com/BTC_Archive/status/15099119870813388 The spot Bitcoin ETF Application Froms VanEck Has been Rejected by THE SEC
Remove hyperlinks
The spot Bitcoin ETF Application Froms VanEck Has been Rejected by THE SEC
Convert to lower case
the spot bitcoin etf application froms vaneck has been rejected by the sec
Tokenize
the,spot,bitcoin,etf,application,froms,vaneck,has,been,rejected,by,the,sec
Remove stop words
spot,bitcoin,etf,application,froms,vaneck,rejected,sec
Stemming
spot,bitcoin,etf,application,from,vaneck,reject,sec
Join words
Spot bitcoin etf application from vaneck reject sec

Fig. 2. The Pre-Processing of the “Cleaned” Function over BTC Sample Tweet Text.

B. Data Pre-Processing

Pre-processing the data is an important step in enhancing the model’s predictive accuracy and getting better results. The pre-processing techniques were applied in the Twitter and cryptocurrency datasets.

- **Twitter Dataset Pre-Processing**
  The pre-processing data phase contains multiple procedures that remove parts of tweets that may excessively or unnecessarily impact the sentiment score. To achieve the goal, Python’s String methods and the library Natural Language Toolkit (NLTK) was utilized [29]. NLTK library allowed the text to process for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Fig. 2 presented the pre-processing steps applied in the BTC dataset with a tweet example. In addition, ETH follows the same pre-processing steps. As a result, the total number of tweets decreases from 37,998 to 17,608 on the BTC dataset and 37,997 to 17,373 on the ETH dataset.

- **Cryptocurrency Dataset Pre-Processing**
  In the stage of pre-processing, as given in Table III, seven features were selected out of eight to forecast the cryptocurrencies’ prices. The cryptocurrency dataset is updated daily. Therefore, trades where the date value is from September 17, 2014, to March 31, 2022, for BTC and November 9, 2017, to March 31, 2022, for ETH are considered. In addition, according to the collected dataset period, the Close column is modified in the Adj Close column. Thus, the Close column ware removed and depending on the Adj Close column as the final result of the latest trade recorded day.

TABLE III. FEATURES SELECTION

Features	Definition
Date	recorded time of the price
Open	Opening trade (Open price on recorded day)
High	Opening trade (Open price on recorded day)
Low	Lowest trade (Least price on recorded day)
Adj Close	Adj Close price on recorded day
Volume	Volume of transactions

C. Experimental Environment

The experimental environment is Google Colaboratory [30], known as “Colab”, a suitable Python environment for ML and data analysis purposes. It provided free accessibility to GPU computing resources. Among the used tools and technologies in the experimental environment are statistical libraries used are NumPy [31] and Pandas [32]. Furthermore, the experiment is conducted on a MacBookPro laptop with 8 GB RAM, an M1 chip, and Macintosh 12.0.1 operating system. In addition, Open-source Neural Network libraries TensorFlow [33] and Keras [34] were implemented to cryptocurrencies’ future price prediction model.

V. RESULTS AND DISCUSSION

This section presents the results of the sentiments analysis model and the cryptocurrency forecasting model, implemented using BTC and ETH cryptocurrency datasets.

A. Result of Sentiment Analysis Model

To better understand the public opinion towards cryptocurrencies and find the relationship with the price volatility. The sentiment is classified into three groups. Fig. 3 presents the distribution of BTC and ETH users’ sentiments: (A) shows the bar graph of the BTC result. The tweets expressed an observed positive opinion that might have happened because most of the tweets contained sentences that did not express negative or neutral emotions. In addition, the rest are due to the BTC currency having a large number of transactions compared with the other cryptocurrencies. On the other hand, (B) presents the bar graph of the ETH result. The tweets expressed an observed neutral opinion towards ETH currency.

Moreover, Table IV presents the macro average of the employed SVM and NB methods to classify the BTC sentiments. While Table V presents the macro average of the employed SVM and NB methods to classify the ETH sentiments. The SVM method outperforms the NB by achieving an accuracy of 93.95% and 95.59%, respectively.

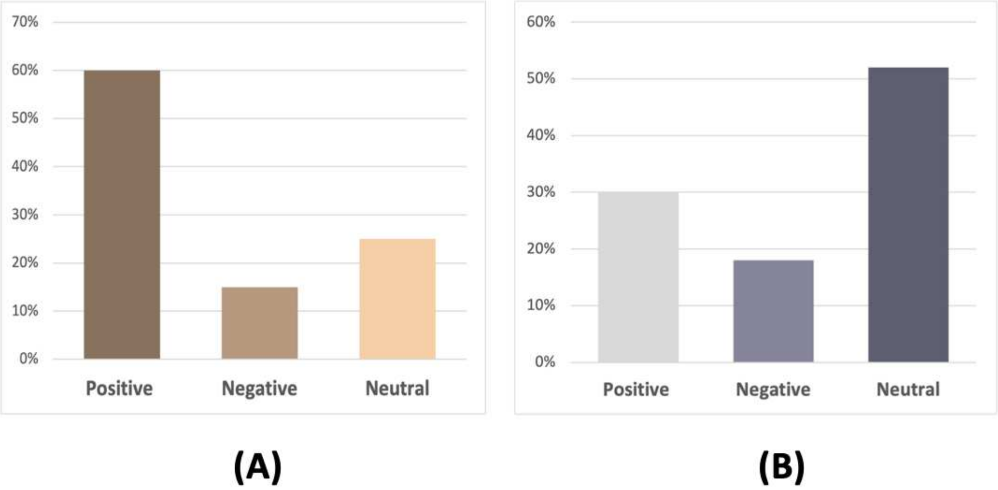

Fig. 3. Distribution of BTC and ETH Sentiments.

TABLE IV. Comparison between SVM and NB Methods on BTC Dataset

Models	Accuracy	Precision	Recall	F1-score
SVM	**93.95%**	**0.90**	**0.94**	**0.91**
NB	80.56%	0.67	0.84	0.69

TABLE V. Comparison between SVM and NB Methods on ETH Dataset

Models	Accuracy	Precision	Recall	F1-score
SVM	**95.59%**	**0.91**	**0.95**	**0.93**
NB	83.74%	0.69	0.88	0.72

B. Result of Cryptocurrency Forecasting Model

The results of the LSTM regression model for the BTC and ETH cryptocurrencies are presented in Fig. 4 and 6. Fig. 4 presented the BTC model, trained and tested from September 17, 2014, to March 31, 2022, to forecast the three following months (April, May, and June). As a result, BTC prices fall in the predicted months, as shown in Fig. 5. On the other hand, Fig. 6 presented the ETH model, trained and tested from November 9, 2017, to March 31, 2022, to forecast the three following months (April, May, and June). As a result, ETH prices fall in the predicted months, as shown in Fig. 7.

The MAE, RMSE, and MAPE are used to evaluate the performance of regression models. Table VI summarizes the results of training and testing errors. We observed that the BTC forecasting model outperformed the ETH model in terms of MAE, MSE, and RMSE, whereas the LSTM model's error in ETH is worse due to the smaller dataset size. It is noted that the small resample of time series data may get the worst result on MAE, MSE, and RMSE tests.

Fig. 8 and 9 introduced the study of the correlation between cryptocurrency users' opinions and price volatility in the cryptocurrency market from January 1, 2022, and May 9, 2022. Fig. 8 illustrates the relationship between BTC volatility prices and BTC users' sentiments. As well, Fig. 9 shows the relationship between ETH volatility prices and ETH users' opinions. We observe that the correlation in the BTC is close compared with ETH, which means when the BTC investors have a positive sentiment, the price of BTC cryptocurrency will be increased. Vice versa, when they have a negative emotion toward the BTC cryptocurrency, the price will be decreased. On the other hand, the ETH prices and users' sentiments do not have a relationship. Therefore, we may conclude that emotion is not always related to investor satisfaction.

TABLE VI. Performance of LSTM Regression Models with Different Cryptocurrencies for Predicting the Daily Closing Price

Cryptocurrency type	MAE	MSE	RMSE
Bitcoin (BTC)	**0.2545**	**0.2528**	**0.5028**
Ethereum (ETH)	0.3838	0.4677	0.6839

C. Comparing with Related Works

Several researchers used various ML [12], [13] and DL [15], [5] techniques to analyze the cryptocurrency users' sentiments. As a result, Table VII compared the result of the current study with the previous studies which used the same dataset in different periods and techniques. In addition, They aim to analyze the sentiments of cryptocurrency users. The SVM appears to be a more appropriate method of classifying the sentiments depending on the accuracy result. To the researchers' knowledge, the reason for the superiority of this study's results is the followed pre-processing techniques that were used to clean the Twitter dataset [13], [15]. In addition, the classification of cryptocurrency sentiments is divided into negative and positive, and neutral, unlike the classification process used in other studies [5], [12], they excluded the presence of normal feelings for cryptocurrency users.

On the other hand, ML and DL techniques were applied to forecast cryptocurrency prices. Table VIII presents the error rate result of two cryptocurrencies compared to Alahmari et al. [11]. The current research outperforms another study and achieves good results for BTC and ETH cryptocurrencies. To

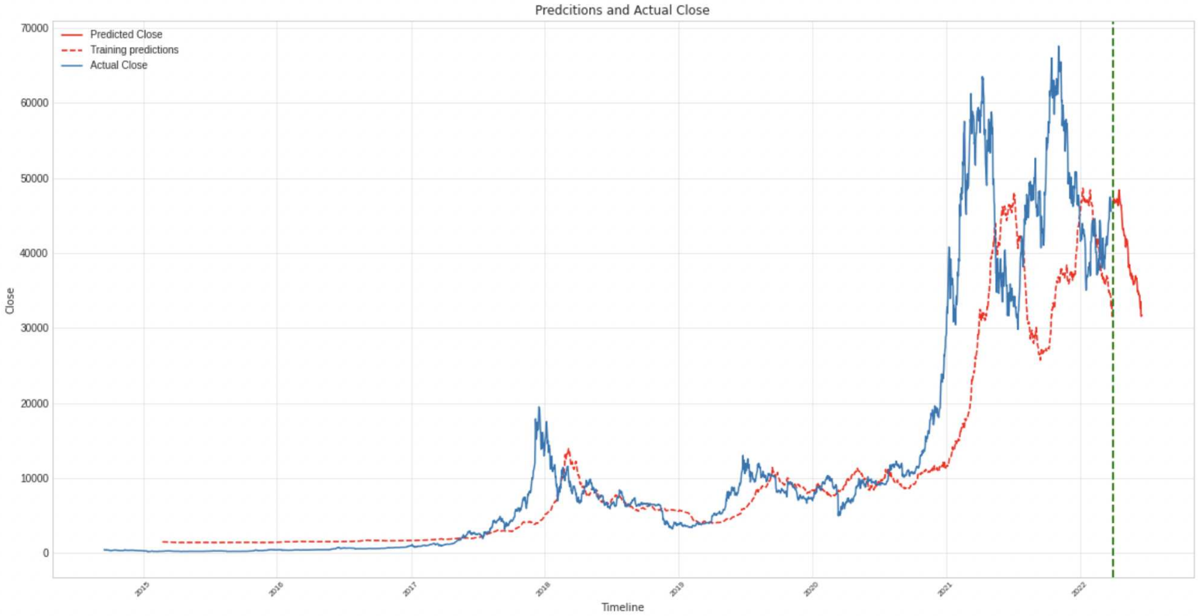

Fig. 4. LSTM Forecasting Model for Three Months for BTC Cryptocurrency Price.

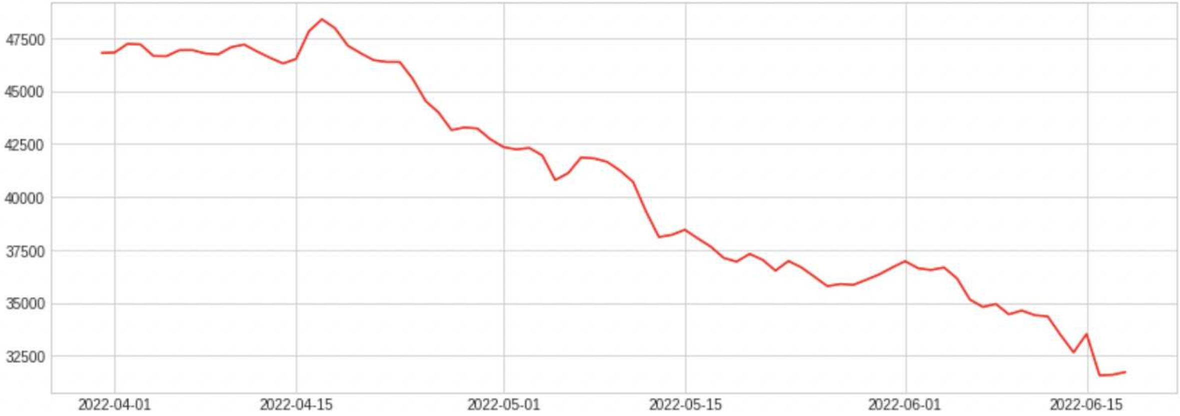

Fig. 5. Forecasting (April, May, and June) Months of BTC Cryptocurrency Price.

TABLE VII. Comparison with Related Work Studies on Sentiment Analysis

Study	Period	Methods	Accuracy
Pant et al. [5]	Jan 2015, to Dec 2015	RNN	81.39%
Rahman et al. [12]	Mar 2018, to Mar 2018	DTR	75%
Sattarov et al. [13]	Mar 2019, to May 2019	RF	62.48%
Pathak & Kakkar [15]	March 2019	LSTM	77.89%
This study	**Jan 2022, to May 2022**	**SVM**	**95%**

TABLE VIII. Comparison with Related Work Studies on Forecasting

Study	Model	Crypto Type					
		BTC			ETH		
		MAE	MSE	RMSE	MAE	MSE	RMSE
Alahmari et al. [11]	ARIMA	313.8	294.5	542.7	12.9	410.1	20.3
This Study	**LSTM**	**0.254**	**0.25**	**0.50**	**0.38**	**0.46**	**0.68**

the researchers' knowledge, the reason for the superiority of this study's results is the volume of historical data and the time-series DL model applied in the experiment. It is noted that the small resample of time series data may get the worst result on MAE, MSE, and RMSE tests.

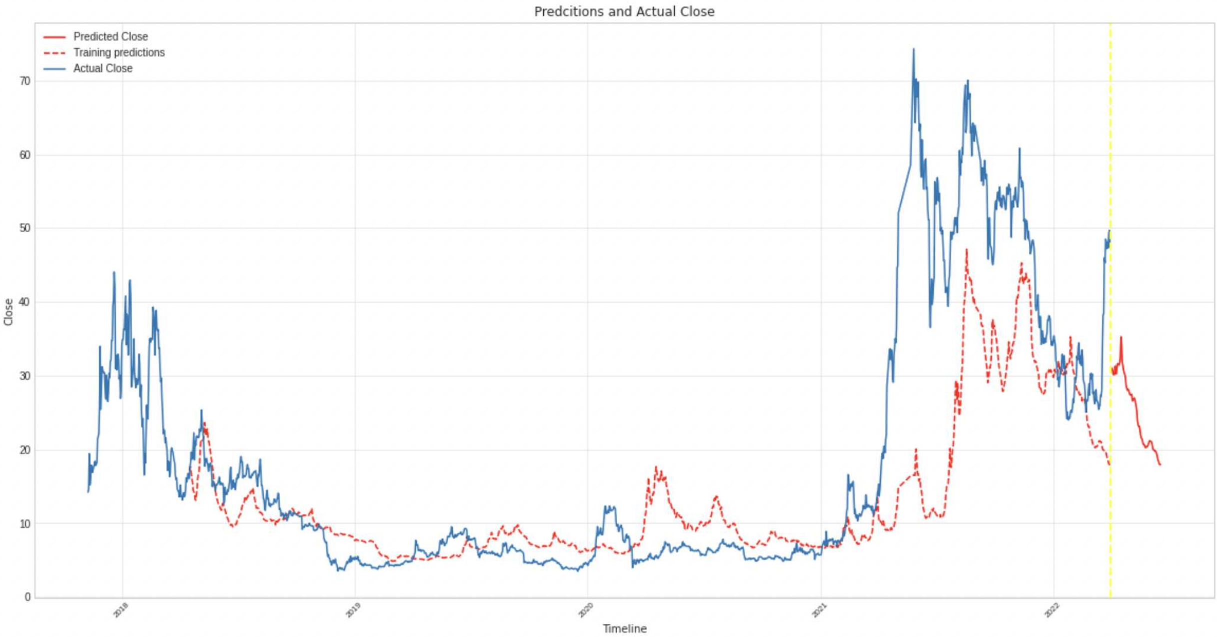

Fig. 6. LSTM Forecasting Model for Three Months for ETH Cryptocurrency Price.

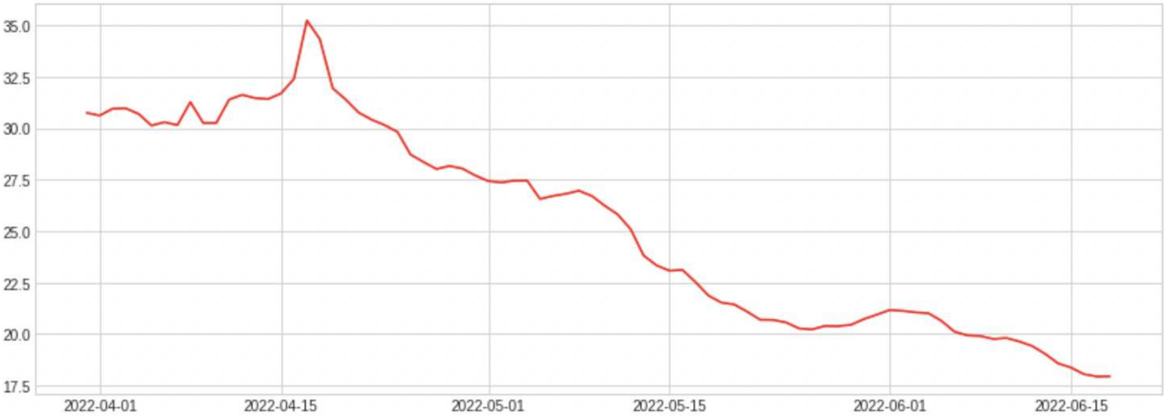

Fig. 7. Forecasting (April, May, and June) Monthes of ETH Cryptocurrency Price.

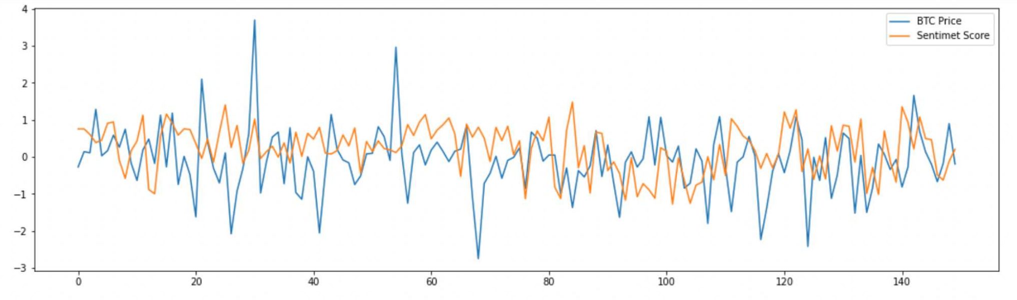

Fig. 8. Correlation between BTC Users’ Opinions and Price Volatility.

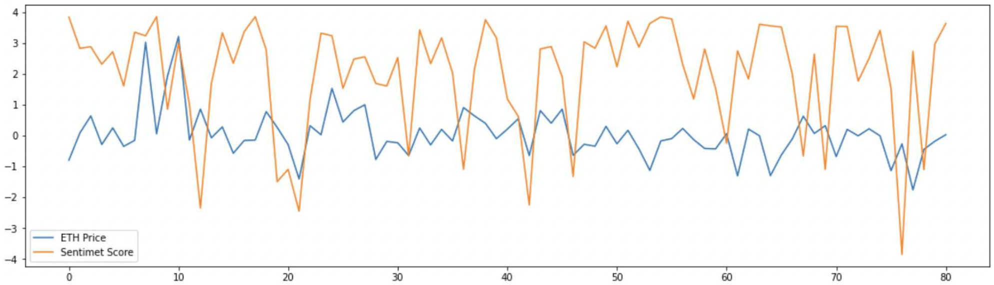

Fig. 9. Correlation between ETH Users' Opinions and Price Volatility.

VI. Conclusion and Future Work

Experiments were conducted using two datasets about two cryptocurrencies, BTC and ETH, to study the relationship between traders' sentiment and the price of cryptocurrencies. To study the impact of the proposed method and confirm its superiority, we analyzed Twitter users' sentiments about these cryptocurrencies and classified their polarity (positive, negative, and neutral) using ML classification algorithms. The SVM classification method outperformed NB with 93.95% and 95.59% accuracy, respectively. In addition, the expected price for the next three months for the two selected currencies has been forecast using the LSTM model; the BTC prediction model outperformed the ETH model with an error rate of 0.2545 for MAE, 0.2528 for MSE, and 0.5028 for RMSE, whereas the LSTM model's error in ETH is worse due to the smaller dataset size. It is noted that the small resample of time series data may get the worst result. Furthermore, the relationship between cryptocurrency volatility prices and its users' sentiments was studied to achieve the research aims. We observe that the correlation in the BTC is close compared with ETH, which means when the BTC investors have a positive sentiment, the price of BTC cryptocurrency will be increased and vice versa. On the other hand, the ETH prices and users' sentiments do not have an observed relationship. Therefore, we may conclude that cryptocurrency price volatility is not always related to investor satisfaction.

Due to a lack of resources and time, the datasets and the selected cryptocurrency types restrictions were imposed to keep the research relevant. The Twitter data set utilized to train and test the sentiment analysis model was limited to five months because of computational power limitations. Although the prices of cryptocurrency are impacted by investors' sentiments worldwide, the Twitter dataset focuses on tweets written in the English language only. On the other hand, this study focuses on forecasting the prices of the two most-traded currencies: BTC and ETH, although thousands of different cryptocurrencies are in circulation. It will be necessary to cover other factors that may impact the cryptocurrency market instead of focusing on the investors' sentiments in future works. Furthermore, conducting experiments using more than two types of the most popular cryptocurrencies and improving the pre-processing steps to study the correlation between cryptocurrency price volatility, news events, and investors' sentiments on different social media platforms that might directly relate to cryptocurrency prices. Additionally, to further study cryptocurrencies' price forecasting, we intend to examine more time-series methods; one of them is ARIMA, a statistical analysis model widely used to predict future trends through the time-series dataset.

References

[1] Y. Hua, "Bitcoin price prediction using arima and lstm," in *E3S Web of Conferences*, vol. 218. EDP Sciences, 2020, p. 01050.

[2] C.-H. Wu, C.-C. Lu, Y.-F. Ma, and R.-S. Lu, "A new forecasting framework for bitcoin price with lstm," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 168–175.

[3] F. Fang, C. Ventre, M. Basios, L. Kanthan, L. Li, D. Martinez-Regoband, and F. Wu, "Cryptocurrency trading: a comprehensive survey," *arXiv preprint arXiv:2003.11352*, 2020.

[4] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[5] D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel, and B. K. Lama, "Recurrent neural network based bitcoin price prediction by twitter sentiment analysis," in *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*. IEEE, 2018, pp. 128–132.

[6] X. F. Liu, X.-J. Jiang, S.-H. Liu, and C. K. Tse, "Knowledge discovery in cryptocurrency transactions: A survey," *IEEE Access*, vol. 9, pp. 37 229–37 254, 2021.

[7] P. Kayal and P. Rohilla, "Bitcoin in the economics and finance literature: a survey," *SN Business & Economics*, vol. 1, no. 7, pp. 1–21, 2021.

[8] L. Abualigah, N. K. Kareem, M. Omari, M. A. Elaziz, and A. H. Gandomi, "Survey on twitter sentiment analysis: Architecture, classifications, and challenges," in *Deep Learning Approaches for Spoken and Natural Language Processing*. Springer, 2021, pp. 1–18.

[9] A. Jain, S. Tripathi, H. D. Dwivedi, and P. Saxena, "Forecasting price of cryptocurrencies using tweets sentiment analysis," in *2018 eleventh international conference on contemporary computing (IC3)*. IEEE, 2018, pp. 1–7.

[10] A. M. Khedr, I. Arif, M. El-Bannany, S. M. Alhashmi, and M. Sreedharan, "Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey," *Intelligent Systems in Accounting, Finance and Management*, vol. 28, no. 1, pp. 3–34, 2021.

[11] S. A. Alahmari, "Using machine learning arima to predict the price of cryptocurrencies," *The ISC International Journal of Information Security*, vol. 11, no. 3, pp. 139–144, 2019.

[12] S. Rahman, J. N. Hemel, S. J. A. Anta, and H. Al Muhee, "Sentiment analysis using r: an approach to correlate bitcoin price fluctuations with change in user sentiments," Ph.D. dissertation, BRAC University, 2018.

[13] O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, "Forecasting bitcoin price fluctuation by twitter sentiment analysis," in *2020 International Conference on Information Science and Communications Technologies (ICISCT)*. IEEE, 2020, pp. 1–4.

[14] A. Mittal, V. Dhiman, A. Singh, and C. Prakash, "Short-term bitcoin price fluctuation prediction using social media and web search data," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–6.

[15] S. Pathak and A. Kakkar, "Cryptocurrency price prediction based on historical data and social media sentiment analysis," in *Innovations in Computer Science and Engineering*. Springer, 2020, pp. 47–55.

[16] A. Aggarwal, I. Gupta, N. Garg, and A. Goel, "Deep learning approach to determine the impact of socio economic factors on bitcoin price prediction," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–5.

[17] S. Nasekin and C. Y.-H. Chen, "Deep learning-based cryptocurrency sentiment construction," *Digital Finance*, vol. 2, no. 1, pp. 39–67, 2020.

[18] Textblob.io, "Natural Language Textblob," https://textblob.readthedocs.io/en/dev/, [Online; accessed May 2022].

[19] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science*, vol. 127, pp. 511–520, 2018.

[20] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE access*, vol. 6, pp. 33 789–33 795, 2018.

[21] K.-F. Selander, "Anomaly detection using lstm n. networks and naive bayes classifiers in multi-variate time-series data from a bolt tightening tool," 2021.

[22] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 403, 2018.

[23] H. Abrishami, C. Han, X. Zhou, M. Campbell, and R. Czosek, "Supervised ecg interval segmentation using lstm neural network," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2018, pp. 71–77.

[24] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," *IEEE Signal Processing Letters*, vol. 27, pp. 1485–1489, 2020.

[25] P. Flach and M. Kull, "Precision-recall-gain curves: Pr analysis done right," *Advances in neural information processing systems*, vol. 28, 2015.

[26] K.-F. Selander, "Anomaly detection using lstm n. networks and naive bayes classifiers in multi-variate time-series data from a bolt tightening tool," 2021.

[27] CoinMarketCap.com, "Top Cryptocurrency Spot Exchanges"," Available online: https://coinmarketcap.com/rankings/exchanges/, [Online; accessed May 2022].

[28] Tweepy.org, "An easy to use python library for accessing the twitter api," Available online: https://www.tweepy.org/, 2018, [Online; accessed March 2022].

[29] Nltk.org, "Natural Language Toolkit," Available online: https://www.nltk.org, [Online; accessed May 2022].

[30] G. Colaboratory.com, "Google colaboratory coding environment," https://colab.research.google.com/, [Online; accessed March 2022].

[31] Textblob.io, "NumPy: A guide to NumPy," https://numpy.org/, [Online; accessed March 2022].

[32] Pandas.org, "Pandas: A guide to pandas," https://pandas.pydata.org/, [Online; accessed March 2022].

[33] Tensorflow.org, "Tensorflow machine learning," https://www.tensorflow.org/, [Online; accessed March 2022].

[34] keras.io, "Keras machine learning," https://keras.io/, [Online; accessed March 2022].