Cryptocurrencies Prices Prediction Using Sentiment Analysis

Abdallah Ragab

araga093@uottawa.ca

Marwa Hamdi Boraie

mmahm073@uOttawa.ca

Hosam Mahmoud

hmahm074@uottawa.ca

Yomna Mohammed

yahme022@uOttawa.ca

Khaled Mohamed Mohamed

kelsa047@uottawa.ca

***Abstract*— Bitcoin is a Cryptocurrency that became more famous in the last few years. Like other cryptocurrencies, it is not controlled by the central bank or government. This made bitcoin prices be fluctuating, and hard to predict the price in the feature. This paper aims to enhance the prediction of Bitcoin prices by applying a sentiment analysis on Bitcoin-related tweets and using bitcoin financial history. VADER and pre-trained Roberta have been used as sentiment analyzers and several regression models.**

Keywords— Bitcoins prices, Sentiment analysis, Cryptocurrency, Regression

I. Introduction

This is the era of online banking and cryptocurrency soon, many countries will accredit cryptocurrencies as their official way of trading. More and more investors are investing in cryptocurrencies but many of them neither know how to analyze them nor make the right decisions so they follow many professionals on social media or read the news and follow their instructions.

This project aims to help investors to predict the cryptocurrency price by making good use of Sentiment Analysis over social media (Twitter) then this information is fed to the Machine Learning models to predict the next day's prices.

Bitcoin [1] is the chosen cryptocurrency for the experiment as it is the most famous, dominant, and highly affects the other cryptocurrencies when it rises and lowers. "The main difference between Bitcoin and traditional currencies lies in the fact that no one controls Bitcoin as it is decentralized. It allows Bitcoin to be an independent peer-to-peer money system that can function regardless of anyone's wishes" [2]. The same methodologies and techniques could be applied to any other cryptocurrency.

Sentiment Analysis [3] is a natural language processing (NLP) technique used to determine whether the text is positive, negative, or neutral, but how Sentiment Analysis could help or affect the prediction of a cryptocurrency's price? Remember the famous Elon Musk (The CEO of Tesla and SpaceX Company) when he tweeted about how he will still continue to hold his Bitcoin (BTC), Ethereum (ETH), and Dogecoin (DOGE). That tweet affected the prices of these cryptocurrencies and their prices increased by 3-4 percent at the time of the tweet post [4]. People are highly affected by others' thoughts and opinions. Moreover, it affects the decisions regarding their investments. To extract the Sentiment Analysis values we applied two techniques:

- Vader [5]: Lexicon and rule-based Sentiment analysis.
- Twitter-roBERTa-base [6]: Retrained Machine Learning model for Sentiment analysis

Combining the Bitcoin price with Sentiment Analysis information as a part of our feature engineering processes, these features are then fed to a couple of Machine Learning models to select the champion model that best predicts the next day's price. These models are

- Random Forest Regressor [7].
- Linear Regression [8].
- LSTM (Long Short Term Memory) [9] [10].
- XGBoost [11].

II. RELATED WORK

There have been previous attempts to utilize sentiment from twitter to predict fluctuations in the price of bitcoin, Authors In [12] have worked on 92550 tweets from 12th of March 2018 to 12th of May, they have used twitter API and web scraping to collect these tweets, they have chosen this interval because they have noticed that the bitcoin prices in this interval was skyrocketing and falling down and that would help them to evaluate they model effectively, they have cleaned their data and lost more than 10000 rows after the cleaning process. after that they have used VADER (Valence Aware Dictionary and Sentiment Reasoner) to compute the sentiment scores of the tweets to use these scores to help them in the price prediction process, their final data was the combination of sentiment analyzing score and history prices of Bitcoin, after that they have built Random Forest model to predict the bitcoin prices and they found that the Maximum value of error is equal to 43.83% and the Minimum value of error is equal to 21.81%.

Authors In [13] have collected their data from Bitfinex Exchange website from Apr 28, 2013, to Feb 28, 2018, they have focuses on the closing price of the bitcoin to develop their predictive model, they applied different data cleaning techniques like dealing with missing values and normalizing their data, after that they have trained their LSTM model and ARIMA model, and they have compared between these two models and they found that compilation time of LSTM is 61 milliseconds and for ARIMA model it is 4 milliseconds and for RMSE (Root Mean Square Error) they found that the LSTM Model is 456.78 where the ARIMA is 700.69, they did not use the sentiment analysis technique in their research, they just rely on bitcoin data like (open, close, high, low, close) to predict the bitcoin prices.

III. System Architecture

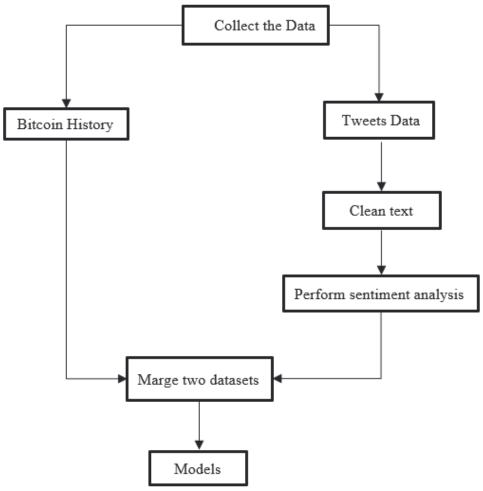


Figure 1 System Architecture

IV. METHODOLOGY

A. Collect The Data

- *Bitcoin History*, Data downloaded from "finance. Yahoo"[13] between 2016 to 2019
  Data contain Date, Open price, High, Low, Close, Adj Close and Volume features.
- *Tweets Dataset*, Data from Kaggle[14] that contain 16 million tweets from 2014 to 2019 about bitcoin has been used

B. Clean text Data

To get accurate results preprocessing methods have been done as the following:

- Select tweets from 2016 to September 2019
- Remove hashtags from the text
- Remove links
- Remove empty lines
- Remove Non-English tweets
- Sample choice 200 tweets if the number of tweets is greater than 200 every day to reduce the time of training models.

The total number of tweets after this step is 265,186 tweets.

C. Perform sentiment analysis on tweets

After cleaning the text 2 sentiment analysis have been applied as follows:

- *Vader* which is a lexicon analyzer used in sentiment analysis in social media and returns 1 for the positive tweet, 0 for natural, and -1 for negative.
- *Twitter-roBERTa-base* which is a pretrained model based on transformers.

We apply these methods on the tweets and combine the results for reducing the number of natural classes we select the positive or negative class if one of them classifies it as positive or negative.

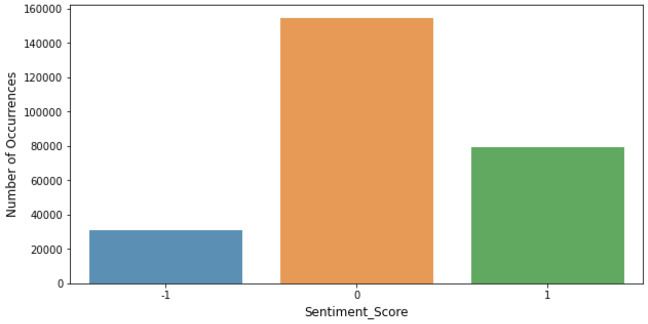
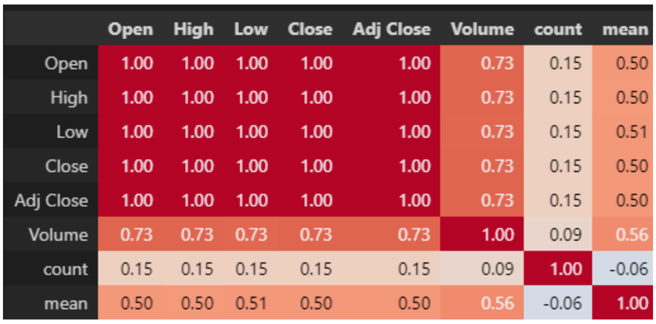


Figure 2 number of classes in the dataset

D. Marge the dataset and prepare the final training set

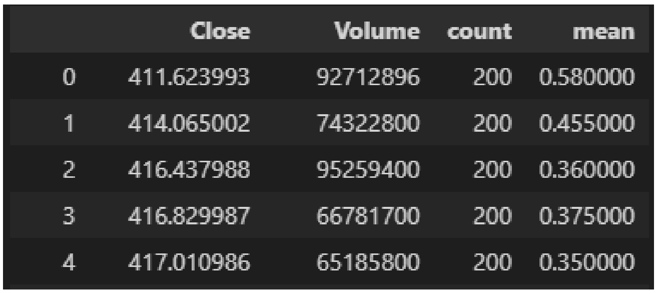
We aggregate tweets by day and calculate the count of the tweets and sentiment score mean for every day. The total number of days for training and testing is 1348 days. A correlation plot between the features has been plotted and shows that there is a strong correlation between Open price, High, Low, Close, Adj Close, and Volume.



	Open	High	Low	Close	Adj Close	Volume	count	mean
Open	1.00	1.00	1.00	1.00	1.00	0.73	0.15	0.50
High	1.00	1.00	1.00	1.00	1.00	0.73	0.15	0.50
Low	1.00	1.00	1.00	1.00	1.00	0.73	0.15	0.51
Close	1.00	1.00	1.00	1.00	1.00	0.73	0.15	0.50
Adj Close	1.00	1.00	1.00	1.00	1.00	0.73	0.15	0.50
Volume	0.73	0.73	0.73	0.73	0.73	1.00	0.09	0.56
count	0.15	0.15	0.15	0.15	0.15	0.09	1.00	-0.06
mean	0.50	0.50	0.51	0.50	0.50	0.56	-0.06	1.00

Figure 3 correlation plot between the features

Close, Volume, mean and count have been chosen as input and the target is the close price for the next day.



	Close	Volume	count	mean
0	411.623993	92712896	200	0.580000
1	414.065002	74322800	200	0.455000
2	416.437988	95259400	200	0.360000
3	416.829987	66781700	200	0.375000
4	417.010986	65185800	200	0.350000

Figure 4 Dataset example

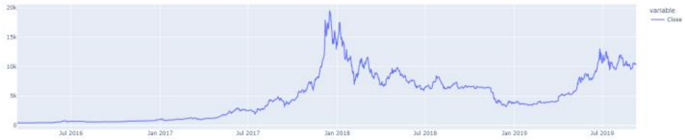


Figure 5 close Price with days

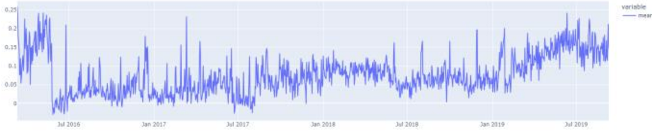


Figure 6 sentiment score mean with days

Dataset has been splitted 80% for train and 20 % for text.

E. Model

For every used model we have trained and tested model on bitcoin features only (close, volume) and sentiment features (mean, count) with the bitcoin features to determine if sentiment analysis will reduce the error or not. We will discuss in the next section.

- *Linear Reagression*

  We used linear regression model from Sklearn library we plotted train and test vs actual prices.

- *Random Forest Regressor*

  We used Random Forest Regressor model from Sklearn library. We applied hyperparameters tuning on Random Forest to choose the best parameters we tuned number of estimators, max_depth, min_samples_split, min_samples_leaf, Bootstrap.

- *GradientBoostingRegressor*

  We used GradientBoostingRegressor model from Sklearn library. We applied hyperparameters tuning on GradientBoostingRegressor to choose the best parameters we tuned number of estimators, max_depth, learning_rate, subsample, max_features.

- *BI-LSTM*

  Long Short-Term Memory networks is kind of RNN model it is used predict output based on the input and previous output. We used Bidirectional LSTM to learn from input and the reverse of this input.

V. PERFORMANCE EVALUATION

The main two metrics for evaluating our Regression models are Mean Square Error MSE [15] and the Root Mean Square Error (RMSE) [16].

- MSE measures how close the data points are from the predicted line by averaging the summation of the square of the distance between the data points and the predicted points.

MSE formula = (1/n) * Σ (actual – forecast)^2
Where:

- n = number of items.
- Σ = summation notation.[15]
- Actual = original or observed y-value.
- Forecast = y-value from regression.

RMSE measures the standard deviation of the prediction errors Indicating how spread the prediction errors from the predicted line.

- RMSE measures the standard deviation of the prediction errors Indicating how spread the prediction errors from the predicted line.

$$\mathrm{RMSE}_{fo} = \Big[\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N\Big]^{1/2}$$

Figure 7 RMSE Formula. [16]

Where:

- Σ = summation (“add up”).[16]
- (zfi – Zoi)2 = differences, squared.
- N = sample size.[16]

The comparison between the models is applied using the MSE and RMSE metrics. Also comparing the models trained using the Bitcoin history data only and using both Bitcoin and Sentiment analysis data as explained in table 1.

Models Comparison using data containing Bitcoin data only, and both Sentiment analysis and Bitcoin

Models	***Bitcoin only***		***BTC & Sentiment***	
	MSE	RMSE	MSE	RMSE
LR	134353.62	366.54	135241.35	367.75
RF	237984.07	487.83	192979.29	439.29
XGBoost	474237.91	688.65	351862.82	593.18
LSTM	115244.37	339.48	130535.40	358.32

Table 1 Models Comparison.

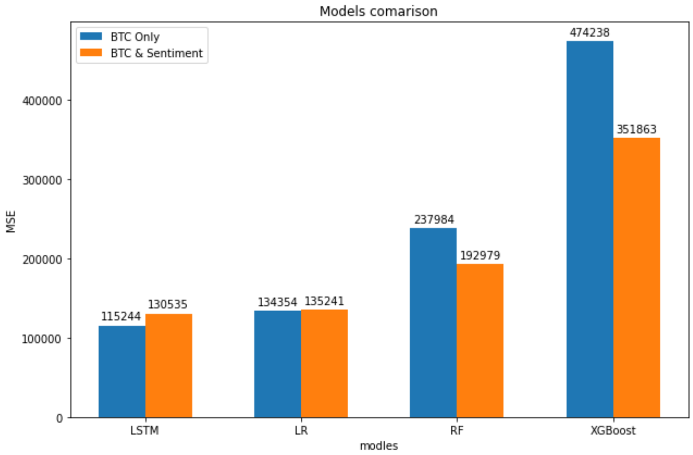


Figure 8 MSE comparison.

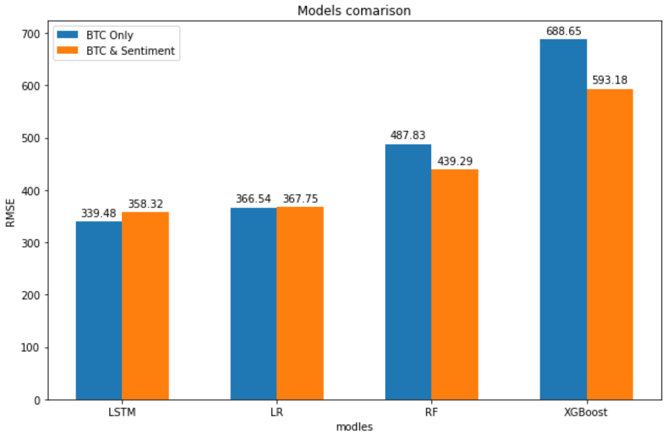


Figure 9 RMSE comparison.

From Table 1, figure 8, and figure 9, we can conclude that LSTM is the champion model with the lowest MSE and RMSE. It's also clear that it performs better with fewer features (BTC only) than the other traditional ML models like Random forest and XGBoost.

The MSE for Linear Regression with BTC only and BTC with Sentiment Analysis is almost the same.

LSTM and Linear Regression are so close to each other, that if we applied error analysis for both of them and pay more attention to outliers, Linear Regression may outperform LSTM and we could obtain better results.

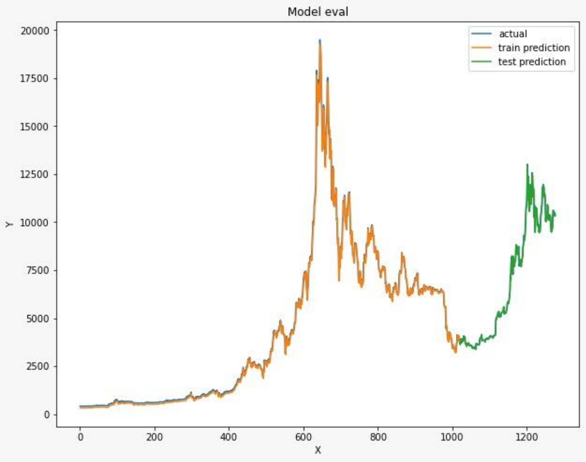


Figure 10 Train and Test prices prediction with LSTM.

The performance of the LSTM model is explained in figure 10 to show the prediction of the next day's close price with respect to the actual close price over time using the training and testing data. We can see that it wrongly predicts when a sudden change occurs in the price as it learns from the patterns of the data.

VI. SUMMARY AND CONCLUSION

The prediction of bitcoin prices became one the hottest topic these days in Artificial intelligence, because this will help the companies and the individuals to make better data-driven decision, so that's why forecasting is very trendy these days, and that's is the main purpose of this research.

And as we know the amount of data on social media is increasing every day, so we want to know the impact of social media on the bitcoin prices, and we did that by applying different technique like collecting the data from social media platform like Twitter, and clean this data to calculate the sentiment scores for the tweets, and after that combine these scores with the bitcoin data and use this features to build some machine learning models to use these models to predict the bitcoin prices.

For the future work, collecting data from different social media platforms could help to know more about the impact of the social media on the bitcoin prices, training a sentiment analysis model from scratch on labeled tweets could give better results instead of using pretrained model to get the sentiment scores like roBERTa or using VADER which is lexicon sentiment analyzer, training GRU instead of LSTM could also give better results.

VII. REFERENCES

[1] Binance https://www.binance.com/en/price/bitcoin

[2] Cryptonews https://cryptonews.com/guides/difference-of-bitcoin-from-traditional-currencies.htm#:~:text=The%20main%20difference%20of%20Bitcoin,function%20regardless%20of%20anyone's%20wishes

[3] Monkeylearn https://monkeylearn.com/sentiment-analysis/#:~:text=Sentiment%20analysis%20(or%20opinion%20mining,feedback%2C%20and%20understand%20customer%20needs

[4] Outlookindia. https://www.outlookindia.com/business/elon-musk-s-tweet-fires-up-bitcoin-dogecoin-ethereum-3-4-crypto-market-up-kleptocapture-news-186882

[5] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[6] Huggingface.co. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[7] scikit-learn.2022 https://scikitlearn.org/stable/modules/generated/ sklearn.ensemble.RandomForestRegressor.html.

[8] scikit-learn.2022 https://scikitlearn.org/stable/modules/generated/ sklearn.linear_model.LinearRegression.html.

[9] TensorFlow https://www.tensorflow.org/api_docs/python/tf/keras/ layers/LSTM

[10] machinelearningmastery https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[11] O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, "Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis," 2020 International Conference on Information Science and Communications Technologies (ICISCT). IEEE, Nov. 04, 2020. doi: 10.1109/icisct50599.2020.9351527.

[12] Anshul Saxena, "Predicting bitcoin price using lstm And Compare its predictability with arima model 1." Unpublished,2018.doi:10.13140/RG.2.2.15847.57766

[13] finance.yahoo https://finance.yahoo.com/quote/BTC-USD/history?period1=1520812800&period2=1526083200&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

[14] kaggle https://www.kaggle.com/datasets/alaix14/bitcoin-tweets-20160101-to-20190329

[15] statisticshowto https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/

[16] statisticshowto <https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error