# Analyzing the Impact of Tweets on Cryptocurrency Prices

**Parth Bhatt[1], Poojan Thakkar[2]**

Project Report for Course DS5220[1,2], Khoury College of Computer and Information Sciences, Northeastern University
bhatt.pa@northeastern.edu[1], thakkar.po@northeastern.edu[2]

**Abstract**

Cryptocurrency is a highly volatile market, a small change of even 2% in the prices is in the volume of a few thousand dollars. A lot of people who hold a high influence on the prices are avidly active on Twitter, a social media platform. Through this project, we aim to analyze the correlation between the tweeting pattern of the users and changes in cryptocurrency prices using Sentiment Analysis and Long Short-Term Model (LSTM). Our model forecasts the prices considering the recent tweets with an accuracy of 75.31% and RMSE of 867.23 (~2% of the recent Bitcoin Price)

## Introduction

The rise of Cryptocurrency is a largely discussed and polarizing topic. This tremendously volatile digital asset is designed to work as a medium of exchange, often touted as the currency of the future. At its core, it utilizes technologically superior and secure blockchain technology or public ledger to secure currency transactions, control the creation of additional units, and verify their transfer. This volatility over a period of months, and in some cases even days, causes an upheaval in the stock market since conventional stock market wisdom may not necessarily translate into wise investment decisions for cryptocurrency. This coupled with the fact there was an increase in stock trading activity fueled by the COVID-19 pandemic, and mobile trading applications like Robinhood, cryptocurrencies have become a major talking point in our culture.

The growth of cryptocurrency was also fueled by social media. Sites like Twitter, Reddit, Facebook were heavily used as discussion forums for cryptocurrency trade, especially during the COVID-19 pandemic. Twitter is a social networking platform, where communication takes place through short messages and posts, commonly referred to as "Tweets". It is one of the major primary channels of communication for cryptocurrency, with any price changes, forecasts, being reported on Twitter. The influence of these tweets cannot be overlooked since it has been well documented at times, how a thread of tweets alone was sufficient to bring major changes in the market.

Through this project, we have tried to assess the influence of Twitter on the prices of cryptocurrencies like Bitcoin. For this, we have used a dual approach of Sentiment Analysis, in addition to training a Long Short-Term Model (LSTM) on our data.

## Background

Sentiment Analysis is at the intersection of various fields of study such as Natural Language Processing (NLP), Biometrics, Computational Linguistics, and Text Analysis. Its aim is to effectively study, analyze, and quantify the meaning of subjective information, which may usually be interpreted in different ways based on context.

We used an approach of combining three different types of Sentiment analysis, i.e., TextBlob Sentiment Analysis, AFINN sentiment lexicon analysis, and VADER analysis.

Text Blob Sentiment Analysis: This approach requires categorized words to be predefined. This set of words can be uploaded from existing libraries, for example, the NLTK database. This analysis model outputs the polarity and subjectivity. In terms of polarity, it categorizes tweets as positive, neutral, or negative with a score range of -1 to 1, with -1 being the most negative. In terms of subjectivity, it identifies very objective sentences as 0, and very subjective sentences as 1. One drawback of this model is it considers only the literal sense of the words and not the context.

AFINN sentiment analysis: It is a wordlist-based approach for sentiment analysis. The AFINN dictionary is derived from a list of 3300+ words rated for valence with a score between -5 to 5, constructed using a combination of crowdsourcing and author's works, and were validated through reviews, crowdsourcing and Twitter data.

VADER sentiment analysis: Valence Aware Dictionary for Emotion and Reasoning (VADER), uses lexicon to provide weightage to sentiment and categorize words as positive, negative, neutral, and compound. The compound score is the normalized sum of the other three scores, with scores close to 1 being positive, while scores close to -1 being negative. Considers the context as well as the meaning. VADER usually provides decent results when used for social media data. This is because it does not require any data to be trained on. Since social media has high usage of emoticons, slang, and other non-conventional sources of communication, it is important for a model to weigh these in for their results. VADER can do that.

We have also performed correlation analysis with our results. Correlation analysis is usually done using 4 types of correlation, namely, Pearson, Spearman, Kendall, and Point Biserial correlation. We will look at Pearson, Spearman, and Kendall.

1. Pearson correlation: It is one of the most used correlation metrics. For two given linearly related variables in a normal distribution, and given that the property of homoscedasticity is satisfied, Pearson correlation is used to find the degree of relationship between variables.

2. Spearman correlation: Given data that is ordinal with unknown distribution, and a monotonic relation between the two variables, Spearman uses a non-parametric test to quantify the level of the existing relationship between the two variables.

3. Kendall correlation: Kendall rank correlation is a non-parametric test that is used to measure the degree of dependency between two variables.

The second aspect of our project is utilizing LSTM. While RNNs do a great job of using previous information to correctly give out predictions with an increase in long-term dependency of data, the accuracy of RNNs tends to decrease. Since the data in this project does have long-term dependencies and implications, it was decided that using the LSTM model would be the right fit for this dataset. LSTM is a special kind of RNN model, which is designed to remember and handle long-term data. It does this by using a combination of multiple gates and mathematical states in its recurrent nodes, allowing information to flow unchanged.

Many optimizers can be used in an LSTM model, but we decided to implement the Adam optimizer. Adam combines the benefits of AdaGrad and RMSProp and works by calculating the squared gradient and calculating the exponential moving average of the gradient. It then uses two parameters to control the decay rate of these moving averages. Due to its enhanced results and ease of usage, Adam is the widely used optimizer for LSTMs.

The tanh function is used as the gating function in our LSTM model. It is used as its second derivative can be sustained for the long term, hence helping the flow of information to be sustained and moderated, rather than being 0 or completely flowing through.

Our proposed LSTM model is trained over 3000 epochs. Epoch simply suggests the number of times the learning algorithm goes over the training dataset. This takes place by passing the entire dataset forward and backward through the neural network.

## Project Description

### Data Collection

The data required for or analysis had two aspects. The twitter data was gathered by scraping individual user posts, respective post statistics viz. comments count, retweets count, and likes, users' statistics viz. users' followers count, and account creation date, and the date of post in form of relational data. To collect this data, the cryptocurrency name as well as the short handle for the same were used as search terms. The data collection was enabled through Twitter Developer Account, and TweePy Python Library. While the developer account gives access to an individual to scrape the tweets, there is a limit to scrape up to of 500,000 tweets per month. The timeframe of the above data was from 5$^{th}$ Feb'21 to 8$^{th}$ Aug'21. The final data-frame collated after this exercise contains the following fields:

- user_name: a unique twitter handle for the user
- user_created: time when the user created their account on Twitter
- user_followers: number of followers of the user at the time of twitter post
- date: timestamp at which the post was created
- text: tweet posted
- hashtags: hashtags included in the tweets
- is_retweet: indicator of whether the data is pertaining to a retweet or original post

The bitcoin prices data was collected through the Bitfinex API. This API enabled gathering of the data at a resolution of 60 seconds interval. The data gathered in form of table

contained following details for the respective cryptocurrency:

- time: unix timestamp
- open: the price of cryptocurrency at the start of the minute
- close: the price of cryptocurrency at the end of the minute
- high: the highest value peaked in the minute
- lowest: the highest value in the minute
- volume: total trading happened in the timeframe

The pricing data was gathered for following cryptocurrencies, starting from their initiation, until 8th Aug'21: Bitcoin, Ethereum, Ripple, Dogecoin, Chainlink, Cardano, and Uniswap.

## Data Cleaning and Preprocessing

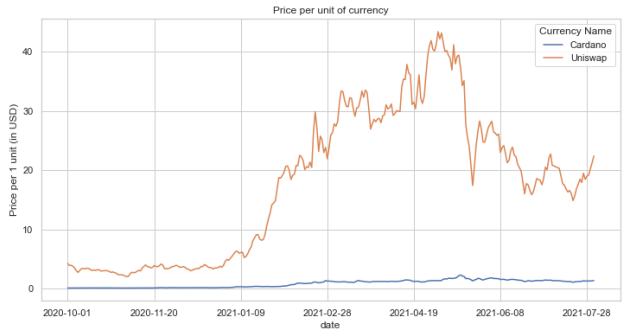
The text field from raw twitter info collated contained emojis in form of Non-Unicode Sequences, hashtags, username mentions, URLs, and multiline characters. To prepare our data for our model input, the text needed to be in form of relevant text tokens that could give us relevant Sentiment Information. The unwanted character sequences and non-alphabetic characters were removed from the text through a thorough regex pattern match and replace process. Post cleaning the text, the text was tokenized through TF-IDF tokenization and sentiment scores for the texts were calculated. Assessing the efficacy of the calculated sentiment scores, it was observed that VADER sentiment compound score is the best amongst the metrics used to project the sentiment of our text. This compound score was then multiplied with the user follower counts, as well as tweet's likes and comments to get a new score, sentiment compound score. This score was then averaged over every hour.

After converting the UNIX timestamp to python datetime format, the Cryptocurrency pricing data was also rolled up at hourly resolution. The rolled-up data projected the maximum of "high" and minimum of "low" over the period, the earliest "open" and the latest "close", and the sum of the "volume". The pricing data as well as the tweets data was then aggregated into a single table. The aggregated data was then deprived of duplicates as well as no-impact instances.

## Exploratory Data Analysis

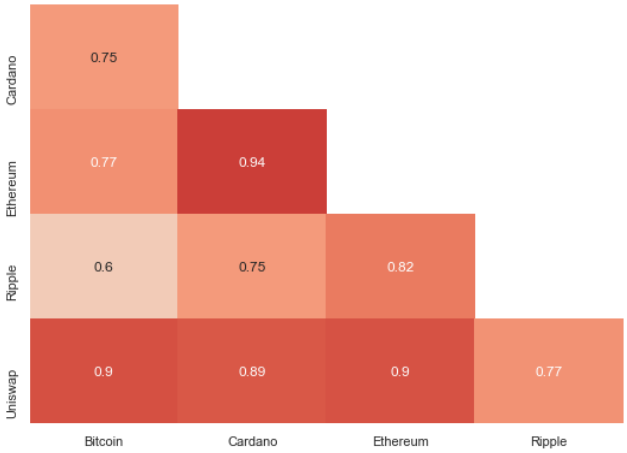
To figure out the which of the cryptocurrencies had the highest market impact, we plotted various graphs for the comparison. On analyzing the pricing trend over the matter of recent months as well as complete timeframe (from cryptocurrency initiation) we observed that "Bitcoin" had the highest volatility as well as volume. Few of the major reasons for the same was the age of the currency, its popularity and ease of Trade.


Trend comparison between Bitcoin, Ethereum, and Ripple



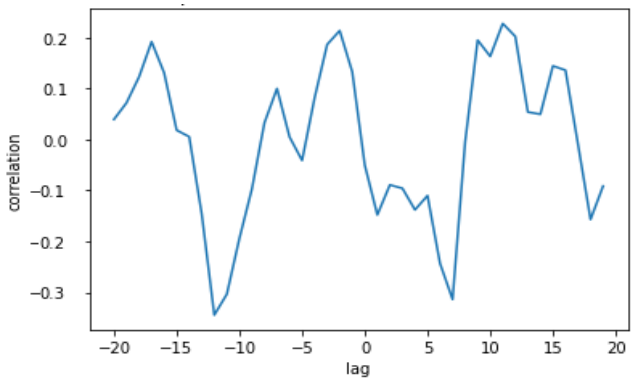

Trend comparison between Cardano and Uniswap

While Bitcoin had the highest impact, we observed that there was a significant correlation between the prices of various currencies. Based on this observation we realized, that observing impact of twitter on any of the cryptocurrency would be sufficient to postulate our initial hypothesis.

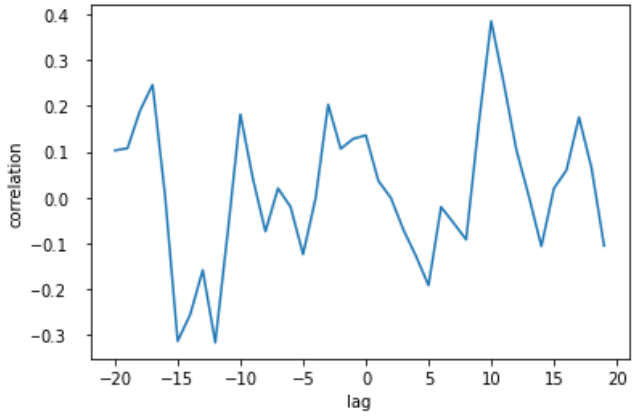



Correlation between various cryptocurrency prices

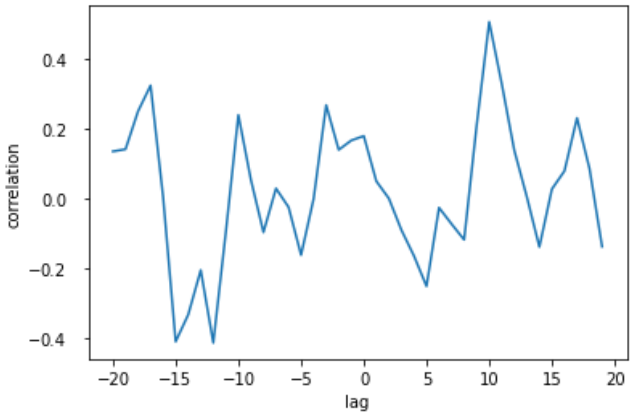
To further bolster this statement, we conducted cross-correlation analysis between the cryptocurrency price rate change and the sentiment score in the hour. We calculated Pearson correlation coefficient, Kendall correlation coefficient, as well as Spearman correlation coefficients between the cryptocurrency prices and the sentiment scores across various time lag between price and tweet, ranging from -20 hours to 20 hours. We found that there was significant correlation between price change and the sentiment score within a lag of 10 hours. This ensured that the tweets impact the cryptocurrency prices.




Pearson cross-correlation between cryptocurrency prices and compound sentiment score




Kendall cross-correlation between cryptocurrency prices and compound sentiment score




Spearman cross-correlation between cryptocurrency prices and compound sentiment score

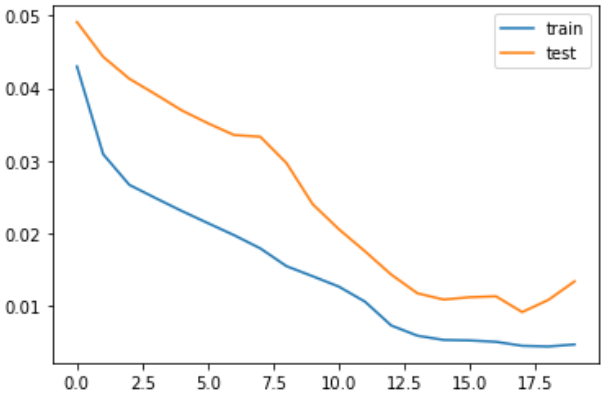
### Applying LSTM

To gauge how much would the twitter posts be impacting the cryptocurrency prices; we ran two Stacked LSTM models and compared the outputs. While both the models gave us favorable results, the results from multivariate LSTM were better.

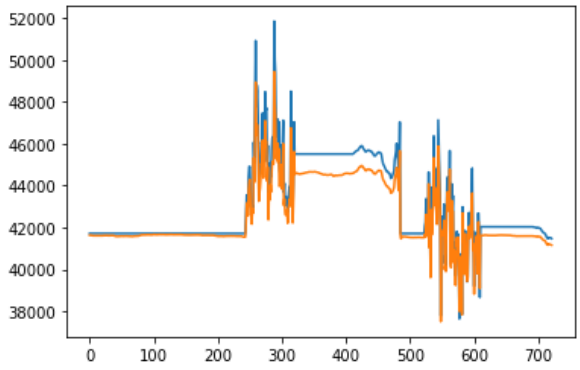
## Empirical Results

We ran two models of LSTM to forecast the prices based on the sentiment scores. Both the models were run on a test set of scale 0.3, with a standard time lag of 3 hours. The LSTM model had was stacked 5-fold with one Dense layer and "Adam" optimizer. The Evaluation metric for both the models was chosen to be Mean Absolute Error. The activation function was selected to be "tan-h". All the above hyperparameters were chosen after a trial and error over various combinations for the same. We let both the networks run for 2000 epochs with a batch size of 40 to get our final outputs.

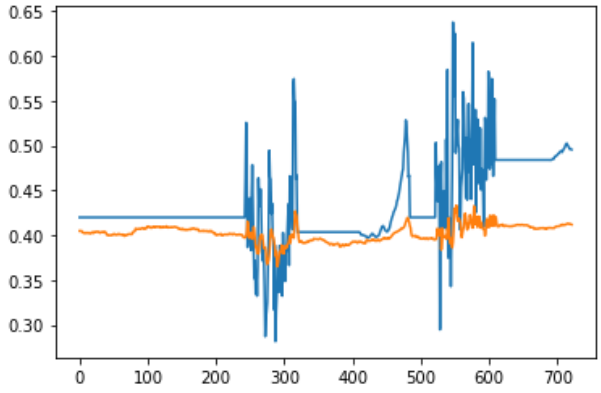
The first model we ran was using a single feature, i.e. compound sentiment score had Test RMSE of 5.71% of the current Bitcoin Price. While the second model ran over 5 features: compound sentiment score, number of tweets in the time frame, ratio of users with >1000 followers to users with <1000 followers, total number of retweets of the tweets in the time frame, and a flag for whether it contained terms related to other cryptocurrencies using the same hyperparameters. The test RMSE for our model turned out to be 2% of current Bitcoin Price. You can observe from the below graphs that, while the first model was able to correctly predict the change in the prices, the magnitude of the change wasn't perfectly predicted. A numerous time it failed to predict the magnitude. This is because of there are several other factors that influence the prices of the currency other than the social media posts. Including a few of those variables in our second model indeed improved our results.




Train vs Test Loss for Model #2




Comparison of actual vs predicted values for Model #2




Comparison of actual vs precited values for Model #1

## Conclusion

In this project, we have tried to see if any correlation exists between cryptocurrency price fluctuations, and related activity on social media, and if so, quantify it. For this, we utilized a dual approach of Sentiment Analysis and training an LSTM model. This approach gave us encouraging results, with prices being forecast based on tweets with an accuracy of close to 75.31%.

As an enhancement, we can consider analyzing more cryptocurrencies and observe the impact of tweets on the same. A caveat we faced was a limit on data scraping cap, we can get a paid subscription for the Developer API to remove the limit. Additionally, modeling techniques like Elephas can be tested to see if they give out improved results faster. This model may be useful in helping beginners in the field of cryptocurrency investment and trading gauge activity in their preferred cryptocurrency. This may help them prepare better investment strategies and contingency plans.

## References

(1997) Long Short-Term Memory. Neural Comput In: Sepp Hochreiter, Jürgen Schmidhuber.

https://doi.org/10.1162/neco.1997.9.8.1735

(2008) Pearson's Correlation Coefficient. In: Kirch W. (eds) Encyclopedia of Public Health. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-5614-7_2569

(2008) Kendall Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_211

(2008) Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_379

(2015) VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text In: Hutto, C.J. & Gilbert, Eric. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM. https://ojs.aaai.org/index.php/ICWSM/article/view/14550