Social Media Impact on Crypto-Currency

Vishwanath Kulkarni
Computer Science, 5502-001B SID 109566304
University of Colorado, Boulder Colorado, USA
viku3942@colorado.edu

Venkat Arigela
Computer Science, 5502-001 SID 109568386
University of Colorado, Boulder Colorado, USA
vear9887@colorado.edu

Mahesh Parab
Computer Science, 5502-001B SID 109253392
University of Colorado, Boulder Colorado, USA
mapa4070@colorado.edu

Priyanka Umesh Pandit Tailapur
Computer Science, 5502-001B SID 109259683
University of Colorado, Boulder Colorado, USA
prpa4917@colorado.edu

ABSTRACT

Today we have massive social media Platform which changes our lifestyle and impact in ways which we cannot even imagine. Our goal is to study one such impact on the cryptocurrency today. There has been a trend that was going on social media for a longer period, which we plan to encapsulate in form of a modal and try to come up with a solution that will not only apply to one cryptocurrency but can be generalized. We plan on analyzing one social media impact - Twitter on the value of bitcoin. This paper will briefly discuss which key factors and key personnel have more effect on the value of bitcoin. Will rank the person according to the impact index and will consider the effect in value change they bring. Will then generalize the events which affect the value of bitcoin.

We will remove the outliers – non-significant events and then use the Twitter Data to determine the currency price. It is important to connect the Tweets to the bitcoin currency rate. So, we are trying to come up with a model that will predict the future prices based on the current tweets using a model developed with previous tweets and previous bitcoin currency price. We all know that how tweets by some famous people influence people and in turn influence the bitcoin currency rates. Sometimes it could increase the price and sometimes decrease it.

It is also important to consider the context of the tweet to determine its influence on bitcoin currency value than just getting the influence of sentimental analysis. So, we are doing both contextual analysis and sentimental analysis to make our model more accurate in determining the value of bitcoin currency.

KEYWORDS

Text Analysis, Index Graphs, Impact Evaluation Plot, Information Retrieval, Data Preprocessing, Sentimental Analysis, Deep Learning, Neural Networks, Logistic Regression, Support Vector Model, Page Rank Algorithm, Long Short-Term Memory, Random Forest, Linear Regression, Recurrent Neural Network

ACM Reference Format

Vishwanath Kulkarni, Venkat Arigela, Mahesh Parab and Priyanka Umesh Pandit Tailapur. 2019. Social Media Impact on Crypto-Currency, viku3942@colorado.edu, vear9887@colorado.edu, mapa4070@colorado.edu, prpa4917@colorado.edu

1 Introduction

In this era where human lives are entangles by the various Social impacts and its existence. It is hard to not have any effect of social media on the face value of any stock. We usually see a tread of company's stock price varying by key factors like, CEO's exit, new product announcement or companies court cases. Tough there is a trend in the world, humans tend to ignore and procced with a blind eye.

Specifically, in this project we give light to the impacts of social media on the crypto-currency value. We are considering the key factors and the personals involved in shaping the face value of bitcoin. Project will analyse and come up with a chart indexing the volume of tweets and its impact to the value of bitcoin. The input of our system will be bitcoin value and twitters tweet. For initial analysis of the trend we have taken past data from Kaggle and will come up with a model, which show us some co-relation with respect to tweets and value. Our input dataset has the various noise factor which we need to pre-process and then allow model to find co-relation.

Ultimate goal of the project stands with the accurate prediction of the price of bitcoin by analysing the twitter trend and monitoring key person and activities. Thus, allowing us to get a summarized picture of how does the money flow in and flow out of the system. This will not only be used for bitcoin but can be used for other crypto-currency.

We are trying to train a model which will predict the bitcoin currency prices based on the present tweets using a model developed with past tweets and past bitcoin currency prices. We all know that how tweets by some famous people influence people and in turn influence the bitcoin currency rates. Sometimes people end up buying bitcoin currencies just because some famous people tweeted about it and sometimes sell it because some famous people tweeted something against it.

We also referred to many previous related papers based on social media prediction for bitcoin currency values, prediction of bitcoin currency based on social media data using machine learning models like Support Vector Regression (SVR), Neural Networks

(NN), Long Short-Term Memory (LSTM) to evaluate each of the techniques and determine a better way to predict the prices based on social media data.

The values are also impacted by some huge news. It could be some higher management people quitting the job, new CEO joining the company, new product launch, defect in some new products etc. This news could take the values both high and low. So social media is highly influential in deciding the currency values and it would be interesting to predict the values based on tweets. But while training the model it is important to remove outliers so that model is trained well, and error rate is decreased increasing the accuracy of the model. Twitter data does come with noise. So, we have to remove the unnecessary data.

The two datasets that we have considered are – Coin Market past data and Kaggle Data. Both has different time range. So, in preprocessing we had to combine it by mapping the time range and get the data. We also had to handle missing data by replacing it with previous values so that the accuracy of the model is not decreased.

2 Related Work

During the past two years there has been significant research in this area. Some great professors, talented students and professional employee have deep dived into this field and found various meaningful insights and tried to predict the value with great accuracy.

First paper mentioned mainly discuss about price fluctuations of crypto currencies based on news and social media data rather. While the second and third paper discuss about the models used for training the model and price prediction of the model.[1]

“Deep Neural Networks for Cryptocurrencies Price Prediction” by Bruno Spilak has some great prediction system using neural networks. This has the accuracy score of 54 % with a variance of 0.06. In this paper various MLP architecture have been compared like hyper-parameter tuning. His work in this field is tremendous and has given us many ways into look into the problem and get meaningful insight.[2]

Second paper that we read was - “Bitcoin Volatility Forecasting with a Glimpse into Buy and Sell Orders” [3] by Tian Guo, Albert Bifet, and Nino Antulov-Fantulin. This paper talked about the volatile property of bitcoin and the various factors that should be considered for the changes in the priceIn this paper, they used the Order book to get the history of buying and selling order which is determined to predict the future trends.

Third paper we referred was - “Bitcoin Price Prediction with Neural Networks” [6] by Kejsi Struga and Olti Qiric in which they talk about LSTMS recurrent neural networks to train the model and predict the prices and also consider the influential factors for the change in bitcoin currency prices.

Fourth paper that we referred was - “Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning” [7] by Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre which talks about the usage of various Machine Learning Techniques namely – Random Forest (RF), Support Vector Model (SVM), and Neural Networks (NN) and compares all models to determine which gives better accuracy in prediction of bitcoin currency prices based on Social Media Data

“Bitcoin Price Prediction Using Machine Learning” [8] by Neha Mangla, Akshay Bhat, Ganesh Avabratha, Narayana Bhat also talks about Machine Learning in bitcoin currency prediction. But it also talks about various criteria that come into picture determining the price fluctuations that needs to be considered.
“Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis” [10] by Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra talks about the Tweet Volumes correlation with Sentimental Analysis and its influence in determining the bitcoin currency prices [12].

“Predicting the Price of Bitcoin Using Machine Learning” [11] by Sean McNally, Jason Roche, Simon Caton talks about using Long Short-Term Memory model, Support Vector Machine (SVM) and how Long Short-Term Memory model performs with better accuracy when compared to SVM

There were many similar papers that used keyword sentiments and predicted the bitcoin currency values based on Social Media Data. Very few papers spoke about the rank the importance of the Tweets and that is what we will be considering in our project and then determine the values of bitcoin currencies.

Our approach will focus not only on keywords sentiments, but we will rank the importance of Tweets. We aim to filter out Tweets from less reputed users, based on previous posts, re-tweets, etc. Rather than focusing on deep machine learning techniques (Long Short-Term Model, Recurrent Neural Network, Support Vector Models etc), we will aim to apply simple learning algorithms, and focus on setting up a system to provide concise and relevant information about the datasets. We will focus on the context along with the sentiments. This is to make sure the model is not just based on sentimental analysis

3 Motivation

Bitcoins are becoming very popular these days and it will be interesting to know the effect of social media on bitcoins. Many papers have come up with different ideas on determining the

trend of increase or decrease in the value of bitcoin-based on social media influence. It all drills down to the sentimental analysis of social media posts and its effect on fluctuations of bitcoin values.

Social Media Platform versus the Value of Cryptocurrency - Today most of us get the world news from Social Media. There may be cases where a person gets to know of a news from Twitter or Facebook before watching news forums or channels. That is how famous and influential social media is. Public sentiment on social media website can influence the price of cryptocurrency. In this project, we want to explore the influence of Social Media on the value of bitcoin and see if we can predict future value using the same.

We will be determining the influence of Social Media on cryptocurrency value and try to use it predict future values. Based on news in Social Media on say, exit of CEO, new product launch can have a huge influence on cryptocurrency prices, and this is what we would want to explore and extrapolate it to determine the future trends

Famous personalities versus the Value of Cryptocurrency - There may be cases where based on tweets of great personalities, the value of bitcoins has fluctuated. Celebrities have an influence on millions of people all around the globe. Since they have a huge number of followers, their tweets about bitcoins influence people and in turn influence bitcoin value fluctuations. Celebrities who are active on social media have huge number of followers and any tweet or post or status they post will be viewed/read by millions of people.

Sometimes a fan of celebrity may invest into bitcoin only because celebrity has tweeted about it and may not really know about bitcoins. This is how famous personalities influence value of Cryptocurrency and we are trying to use it to determine future trends. We have tried to determine the ranking algorithm based on their retweets, followers, likes and use it to predict the prices based on their tweets.

Events influence of Cryptocurrency price – As we saw in the above how famous personalities can influence the price of cryptocurrency, there can be other event like government ban, major company's investment into cryptocurrency startups, etc. So, doing event analysis we get to know what all event influenced the bitcoin and what was their extent of effect on the price of bitcoin. This helps us in the future to anticipate how much such events can affect the bitcoin price and enables investors to take decisions promptly.

Money Factor - There is a huge amount of money that is involved in the entire bitcoin transactions. It would be interesting to predict the future value of bitcoin and decide on whether to invest or not.

The bitcoin currencies are also impacted by some big news. It could be someone at a high position quitting the job, Change of CEO to the company. This news could take the values high.

Social media is highly influential in deciding the currency values and it would be interesting to predict the values based on tweets. Sometimes there could be defect in the product that the company launched which could lead to a lot of negative tweets making the bitcoin currency values all less.

Social Media and human actions are tightly coupled. So, it is not surprising that the stock prices are affected by Social Media. We are trying to explore this trend and predict the currency prices. There are still so many theories on factors of Social Media influencing bitcoin prices and it would be interesting to explore all such factors and use it to determine future trends

4 Problem Statement

The use case would be analyzing the correlation between the bitcoin value fluctuations and tweets thereby and predicting the value of bitcoin, based on the tweets. Here are some interesting facts that need to be considered before starting on the data mining.

The difference in price/value formats in various trading platforms - The units/value formats of bitcoins will be different in the different trading platforms. A classic example would be Coinbase and Binance. So, the difference in formats must be taken into account

Data Preprocessing - The two datasets are in different formats: one of them is a web page that needs to be scrapped. The datasets will be cleaned and stored as CSVs. Next, the timelines of the datasets need to be correctly matched since their time frames and granularities are different. The Tweets dataset has to be hashed/grouped by the username to identify popular users. The datasets have different time ranges (2013-2019 and 2016-2019) and time period. This needs to be considered for the creation of a single dataset. We also need to clean the data by removing unnecessary items not needed for analysis. We also have to remove noise data and also handle missing data. We handled missing data by taking into consideration the immediate previous and immediate next values to make our model more accurate

Chronological analysis and prediction - We know the value of Bitcoin changes in a short amount of time. We will consider a time window-based approach to analysis groups of tweets and their

impact on the price trend in the future. Considering a time frame (s, t), and the price trends in a similar time frame, we will aim to predict the price trend up to k time steps into the future to construct a correlation.

Synchronizing geographically distributed tweets - Tweets are geographically distributed and we have to synchronize it to get a correlation between bitcoin value and geographical locations. There are various bitcoins values in various countries and tweets may also be geographically distributed and we have to consider it and use it appropriately when determining the future trends of bitcoin currencies

5 Popularity Prediction

Bitcoin volatility is dependent on 3 factors, one among them is top twitter users in the field of cryptocurrency. We have analyzed that bitcoins price varices with the intervention from popular personalities or certain events or political influential ban or new crypto currency. In this paper we focus on popular personalities influence as that holds more weightage in the variation of bitcoin price.

Due to influence of twitter on crypto currencies there has been a negative usage of the platform, which led to currency manipulation and impact on the crypto market. There were popular persons whose one tweet about bitcoin would affect both positively and negatively on the price of bitcoin. In this paper, we have provided a brief analysis of identifying key persons who responsible for price manipulation. We have limited our finding to only one crypto currency due to the limitation of the dataset. This model can be applied to the different crypto currencies with appropriate data related to the crypto currency. This will not only predict the top influential figures but also predict top users who were followed, liked and replied.

5.1 Top Users

Our model predicts a list of users who are most influential by calculating popularity index on each tweets and time frame. A Popularity index is assigned to each user based on many factors.

1. Retweets
2. Likes
3. Replies
4. Timeframe
5. Impact index

Above mentioned factors have different weightage towards contributing for the popularity index of the user. Retweets, Likes and replies have the same weightage, but timeframe and Impact index varies.

$$popularity_i = \sum(P(retweets) + P(likes) + P(replies)) + \int_{start}^{end} impact$$

$$P(x)) = \sum(x_i)$$

Fig 5.1 popularity prediction

Popularity index has few attributes which are calculated accordingly. Based on the formula above we have calculated each user rank with giving equal priority to retweets, likes and replies and giving boosted weights to impact index w.r.t positive or negative impact.

Top Users	Rank
22loops	1
APompliano	2
6BillionPeople	8

Fig 5.2 Temporary Top User for time ranging from 2013-2016

5.2 Influential Users

Another set of users who were most influential in Bitcoin price manipulation were present with the greatest number of mentions to them. These users were most influential to general public who were trading and following them. These influential persons have impacted the readers and followers in a positive manner.

We created a directed graph from each user who has mentioned other user's handle and obtained a complex directed graph among all tweeter users who tweeted related to bitcoin. Later we ran **PageRank algorithm** [14] on the graph and assigning weightage to each receiving edge node. Once the weight of each node is calculated, we remove nodes with less than 10 mentions or less than 10 weightages. This allows us to reduce the graph significantly and handle it in a highly configured VM.

$$PR(u) = \sum v \epsilon B_u \frac{RP(v)}{L(v)}$$

Fig 5.3 PageRank algorithm

the PageRank value for a node(user) u is dependent on the PageRank values for each node(user) v contained in the set Bu (the set containing all mentions linking to node(user) u), divided by the

number L(v) of edges from node(user) v. The algorithm involves a damping factor for the calculation of the PageRank [15]

Influential Users	Rank
Youtube	1
yabtcl	2
Bitcoin	4

Fig 5.4 Influential users list predicted by our system

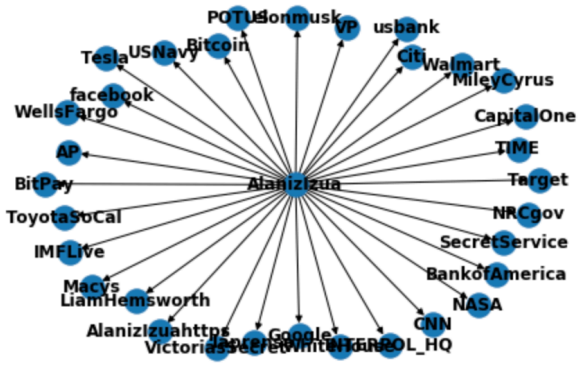

Fig 5.5 Directed graph of users with more than 20 edges

5.3 Tweet Sentiment Analysis

Sentiment analysis is the process of guessing the attitude or emotion of a piece of text – positive, negative or neutral. The dataset that we obtained had a variety of Tweets about the Bitcoin cryptocurrency – some Tweets were positively aligned towards the use of Bitcoin and cryptocurrency in general, others were opposed to it due its highly volatile nature. Several Tweets were also advertisements that sometimes had no sentiment towards the product. Thus, we also took into account the sentiment polarity of each Tweet in the range [-1, 1] where 1 is highly positive and –1 is highly negative, and the subjectivity of the classification in the range [0, 1] where 0 is highly objective and 1 is highly subjective. These values were appended to the dataset to serve as another feature for training our model.

We used NLTK's TextBlob API to perform sentiment analysis. This API internally performs text pre-processing tasks like tokenization, lemmatization and parts-of-speech tagging. It even performs language translation which was helpful to classify Tweets or parts of Tweets that were not in English. It then uses a Naïve Bayes classifier to predict the classification of the sentiment of the text. We decided to go ahead with this simplistic classification technique since it is usually effective on small pieces of text, like Tweets.

6 System Architecture

Our system is completed based on microservice architecture [16] which allows us to develop the system in different module and have API signature scheduled to interact with each other. We have multiple modules set up in GCP which allows us predict Realtime and display om dashboard.

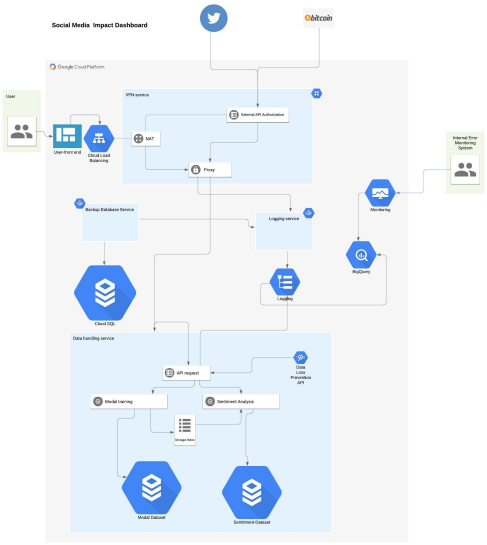

Fig 6.1 System Architecture of the Bitcoin price prediction hosted on GCP

We created a VM instance on Google Cloud Platform and hosted the Flask Application. Flask is a Python Web Framework that allows easy visualization of data. We have used APIs exposed by Twitter and Coin Market Database to get the recent Tweets and recent Bitcoin values. We have also exposed our own APIs from the prediction tool and made use of it while generating the model graph on the dashboard.

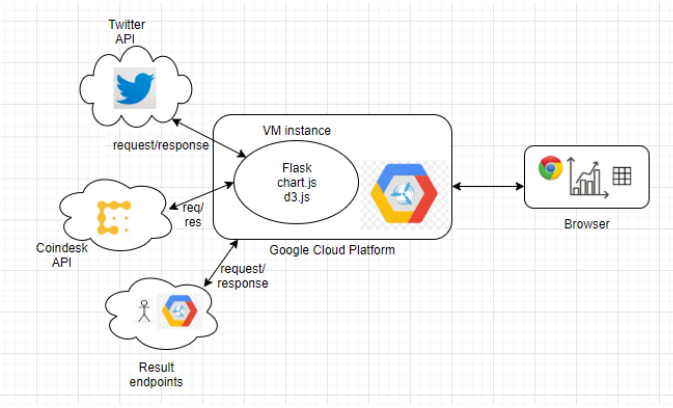

Fig 6.2 Visualization Dashboard Architecture

7 Proposed Work Dataset

We did research on finding an appropriate dataset for training our model. We found two such datasets that we need to merge to use as a single dataset for training the model. We have considered following two data sets. One is from Kaggle and one is from Bitcoin Historical Data. Below are the brief details of the datasets and its fields.

Kaggle

Data Set from Kaggle was from 2016 to 2019.

The Data Fields included Username (Unique ID of the person who Tweeted), Tweet-ID (Unique ID for the Tweet), Timestamp (Time of the tweet posting), URL (Link associated with the Tweet), Likes (Likes associated with the Tweet), Replies (Replies associated with the Tweet), Retweets (Retweets associated with the Tweet), Text (Tweet text).

Historical Data

Data Set from Kaggle was from 2013 to 2019.

The Data Fields included Date (Timestamp associated with the bitcoin currency price), Open Price (Opening Price for the day), Close Price (Closing Price for the day), High Price (Highest Price for the day), Low Price (Lowest Price for the day), Market Cap (Market Capital associated with the bitcoin currency for the day)

As seen above, both the datasets have different time ranges and also the time period of data is different. So, when we had to combine the data, we had to map the time ranges and get a single dataset formed. We also had to handle the missing data and noise in the dataset so that the data model accuracy is not lowered.

We considered data from 2016-2019 because the Kaggle Data didn't have it from 2013-2016. The time period of data generation was different. So, mapping had to be done between the rows of both datasets to combine them to form a single dataset.

8 Proposed Work Implementation

We have huge data set. So, it would be convenient to store the data files into the Cloud Storage Bucket of Google Cloud Platform. The code base and Storm architecture could be hosted in the EC2 instance of Google Cloud.

Firstly, we have two datasets. We would be processing the data in both datasets (Bitcoin and Kaggle Data) to match the time ranges. We also will remove unwanted data that is not relevant to the analysis. We will also be adding data where it is missing based on the previous data so that the model accuracy is not lowered. We then process the data in both datasets to match the time range. Doing so will remove unnecessary records that are irrelevant to analysis.

Secondly, to handle the dataset we will set up a Spark cluster in the EC2 instances of Google Cloud Platform to read data from the buckets, match the timelines of the two datasets and get a single dataset to train the model, and extract relevant information - ranking/importance of any Tweet based on its likes, re-tweets, replies; sentiment inferred from the Tweet, etc.

Thirdly, we are designing a machine learning model to construct a correlation between the two datasets mentioned earlier by having test and training set and try to increase the accuracy by removing outliers, noise, handle missing data and the like

If we infer a useful relation between the two datasets, we will write a simple machine learning script to extrapolate the information and predict the future prices of Bitcoin based on current trends that we are getting from the Twitter Streaming APIs. So, the current tweets we are getting using the Twitter APIs and the current prices using the Coin Market APIs and this is used as a comparing parameter to measure the accuracy of the model

Compare the predicted prices with real time historical prices and calculate error percent. We will aim to provision the data as an API using a RESTful web server, or optionally visualize it on a web interface (Chart.js, plot.ly) or may be use Tableau to have data coming from Coin Market Database, Twitter and Model. This is to have a dashboard which can be used to easily compare the present price and predicted price and compare the accuracy of the model and calculate the error percent.

Web Data Connectors are built to schedule the pulling of data from Twitter APIs, Coin Market APIs and model data so that dashboard can be created in Tableau and data can be visualized. We can use Web Data Connector to connect our data via HTTP if our data source doesn't have WDC. It is basically JS code that needs to run on VM (Local or Remote) and get the data. We can use GCP instance to host this WDC and get the data displayed

The live Twitter data is generated from the Twitter API and we have used Flask Application to visualize it. We have

Finally, we are thinking of improving the accuracy of the model by using various model generation techniques and comparing the model performances. We will also be trying to get the past Twitter Data and add it to the training dataset. Idea is more the data, more the accuracy and it will help in reducing the error rates.

9 Data Preprocessing

As discussed above we have two completely different data sets and two to mine data from the datasets we must do a lot of preprocessing and make them like compare. The techniques include – Data Reduction, Data Cleaning, Dimensionality Reduction and others.
We have written scripts to scrape, clean and pre-process our datasets. We will run the scripts and clean the data sets in hand before applying data mining algorithms.

Data Description - First Data Set was Twitter data on Bitcoins from Kaggle. We have scripts to handle the missing values, Nan values, timestamps and at the end we had data set with ten most important attributes which will help us in predicting bitcoin currency trends. The ten attributes chosen are –

- **Username**
- **Full Name**
- **Tweet-ID**
- **Timestamp**
- **URL**
- **Likes**
- **Replies**
- **Retweets**
- **Text**
- **HTML Link**

Supplementary Data Description – We referenced another dataset from Kaggle which contained the total number of relevant Tweets in an hourly time interval. It also contained the number of Tweets with a positive sentiment, those with a negative sentiment as well as the mean sentiment polarity for the positive and negative tweets in that hour. The six attributes chosen to form this dataset are -

- Timestamp
- Negative Tweets Count
- Positive Tweets Count
- Neutral Tweets Count
- Average Negative Tweet Sentiment
- Average Positive Tweet Sentiment

Data Description - Second Data Set was Historical Bitcoin Data. We have scripts to handle the missing values, NaN values, timestamps and at the end we had data set with eight most important attributes which will help us in predicting bitcoin currency trends. The eight attributes chosen are –

- **Timestamp**
- **Open**
- **Close**
- **High**
- **Low**
- **Volume (BTC)**
- **Volume (Currency)**
- **Weighted Price**

Data Cleaning - Once the important attributes were discussed and chosen, we had to write scripts to clean the data by filling out missing values with previous values, remove redundant rows, remove NAN values and so on. Data cleaning is very important before we apply Data Mining algorithms.

Data Reduction - We had a huge amount of data with a lot of attributes some of which won't even be needed when we do data mining. So, we had to go with dimensionality reduction by removing unnecessary columns (attributes)

10 Data Analysis and Model Training

Our project is not just finding sentiments based on tweets and mapping it to bitcoin values, we will track the events, personalities and then build the model accordingly

We are proposing model to index the events and rank them based on importance. That is, we are prioritizing the events based on its importance and occurrence and then ranking it based on them and build the model.

Following steps are followed in merging both datasets. First dataset i.e., tweets.csv is loaded and removed the unnecessary attributes like 'HTML' and 'URL'. After which its timestamp is converted to UNIX time stamp. And in bitcoin price dataset, all the Nan records were discarded. Once both datasets are cleansed, they were merged based on timestamp. This dataset is used for further analysis.

Unix timestamp is way to track time as a running total of seconds. This clock started at Unix Epoch on Jan 1st, 1970 at UTC. This is the number of seconds passed from above particular date. So, this doesn't affect our analysis even if both datasets are from different geographical locations. Since one dataset has UNIX time stamp and if we convert the other dataset time also to same, we can easily avoid the time zone synchronization issue while merging.

Data Analysis - Initially, we considered building an FP-growth tree. But we rejected the idea because of the worst-case space complexity which increased with the depth of the tree. Next step was to calculate the mean among the correlated data. Each Tweet in the dataset will be assigned a weight / importance score based on the "reputation" of the user. This will be calculated based on the average popularity of their Tweets, judged by the number of likes and re-tweets they received. Additionally, if we can retrieve user profile information like their followers and following from the Twitter API, we can factor that into their reputation score. The Tweets itself were analyzed to extract the sentiment of the text. This was done using NLTK's TextBlob API, which internally uses a Naïve-Bayes classifier to determine the sentiment.

While correlating the two datasets, these two scores will be used to train an influence score. We will build a time-series model, training the model to correlate the Tweets to the Bitcoin price trends. We are still looking at various techniques to do this.

Model Training – The data obtained from the aforementioned datasets was preprocessed and converted to time series data with different time window values. We created datasets with window values 5, 10 and 20. We trained a Long short-term memory (LSTM) neural network model with this data. The advantage of this network is that it can selectively remember past datapoints and use inferences from these to predict values in the future. A common LSTM unit consists of an input gate, output gate and a forget gate. These gates are used to selectively forward parts of the input, output and memory of previous datapoints to the next cell. By selectively remembering inferences, these networks improve over vanilla Recurrent Neural Networks (RNNs) by avoiding the problem of exploding or vanishing gradients and is also less likely to overfit the data. By normalizing and standardizing the data before feeding it to the network, we were able to achieve a loss to output mean ratio as low as 5%. However, this model did not fit well to live Twitter data as the dataset wasn't expansive enough to cover all types of Tweets, and the Bitcoin closing prices in the dataset were different from the ones that the model was trained with.

Event Analysis – Through event analysis based on time series anomaly detection with series forecast [18], we are trying to find the sudden fluctuation in bitcoin price and correlate with events that caused it. For that we take the weighted price attribute and find the objects with deltas more than 200 from previous object price. This helps us to capture the approximate date on which this major fluctuation happened. After which we use external API's to find the bitcoin news event happened on those particular dates. But here we have few constraints, Google news API which allows to pull only the months for default version is not a feasible option. So, we chose NYTimes API endpoints which takes in keyword (to filter the news articles), start date and end date. So, this gives the list of new articles that are published around that timeline when bitcoin price got influenced. In our code we only fetch the URL of news article as output from the resultant json response.

After running the event analysis code on the entire processed dataset, there were 27 dates observed on which there was a price delta of 200$. When these 27 dates were used to hit the NyTimes API endpoints, we found news articles related to bitcoin on 22 dates.

We had to synchronize the time zones of tweets since sources were based out of different geographical locations and it would make no sense to merge them and analyze the trends without synchronizing the time zones.

11 Model Integration and Data Visualization

Once the model is built, we will be integrating the model prediction, real-time Twitter data and real-time data from bitcoin site and visualize the data using web frameworks and host it in GCP

System Setup – We will be leveraging Google Cloud Platform features. We will be using "Google Big Query" to run queries on two huge data sets, fetch and store. We will be using streamlining techniques on the stored data set. We are thinking of using Apache Kafka for this. Finally, we will be using EC2 instance on Google Cloud Platform to host our dashboard and free domain from github.io. We set up Flask Application in EC2 Instance of Google Cloud Platform and hosted our application dashboard.

Data Visualization – The visualization data will include the below five important aspects. Firstly, Twitter Hashtag Timeline, we fetched it from List Timeline API which comes as an embedding in HTML code and it is dynamic. It fetches data dynamically from Twitter and updates it without having to refresh the page. Below is the picture depicting the same

Fig 11.1 Live Twitter Data in Dashboard

Second aspect is the population of top Twitter users who are talking about bitcoins and also the popular users among them. Below are the snapshots showing the same. The popularity helps us know the influence of them on Bitcoin prices

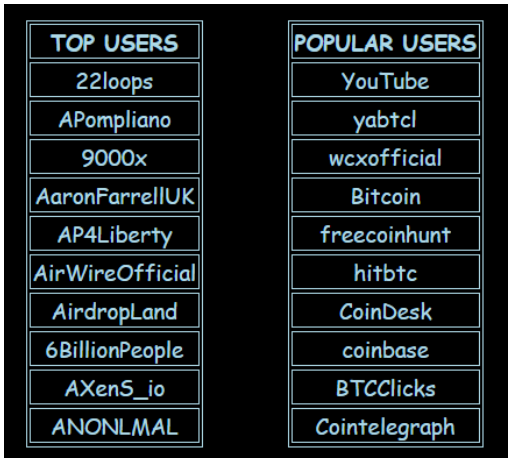

Fig 11.2 Top and Popular Users for #bitcoin

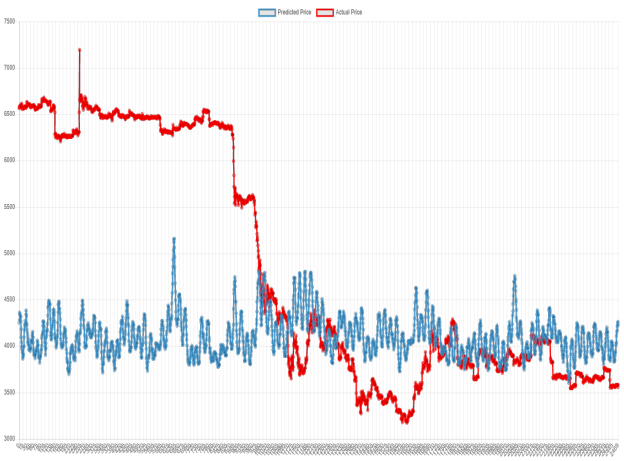

Fig 11.3 Predicted Price Vs Actual Price

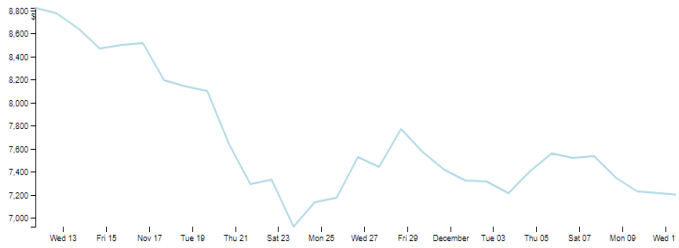

Fig 11.4 Bitcoin price for past 30 days

Model Integration – Data obtained is not just from Sentimental Analysis but also contextual Analysis (Events, Time, Personalities etc.)
The below image shows dates on which events related to bitcoin happened and their respective news articles. The image contains only few dates.

```
2017-02-16 04:27:00
2017-06-14 22:03:00
2017-08-13 04:01:00
https://www.nytimes.com/2017/08/12/technology/mikes-and-katies-week-in-tech-uber-infighting-and-the-google-memo.html
https://www.nytimes.com/2017/08/13/briefing/charlottesville-north-korea-trump.html
https://www.nytimes.com/2017/08/14/business/dealbook/bitcoin-price-virtual-currency.html
2017-10-07 20:57:00
```

Fig 11.5 Few results from the event analysis

12 Evaluation

Our project will aim to find relevant information to find the relations between the two datasets. Using a time-based approach, we will aim to predict the prices up to *k* steps into the future and provide a measure of accuracy/cost. We will then run scripts to scrape, clean and pre-process our datasets. We can apply multiple learning algorithms (Linear and Logistic Regression, Support Vector Models, Long Short-Term Memory and the like) and compare their accuracies on the datasets.

We can infer useful details from the combined dataset using Machine Leaning scripts and then extrapolate that information to predict the future bitcoin currency prices based on current Tweets from Twitter Streaming APIs. We can test our model by having a comparison between predicted prices and real time prices from Coin Market APIs and thus calculate the error rates and try to improve the model

There will be the dashboard that has bitcoin currency prices coming from the Coindesk APIs, Twitter APIs and our mode prediction prices. This is to visualize the price differences and error rates (easy visualization)

13 Conclusion

While analyzing and performing pre-processing tasks on multiple datasets, we observed that the sentiment of the aggregated Tweets was less important than the number of relevant Tweets about cryptocurrency and Bitcoin. In general, we noticed that the price of these cryptocurrencies was usually high when there was a higher number of polarizing Tweets. We also noticed during user ranking that highly polar Tweets came from a relatively small number of users who act as crypto influencers. Thus, we think that there is a lot of value in accurately ranking these users to understand their individual influence on the stock market.

Also, we looked at the correlation between the normalized Tweet sentiments and the bitcoin prices and found that the coefficient was close to 0. Aggregated data and the number of Tweets showed better correlation with the prices of cryptocurrency. Due

to the restrictions of the Twitter free API, we were not able to extract enough recent Tweets to help our model classify them better. Thus, since our model was not able to fit to current Tweets and Bitcoin data based on the training dataset, we cannot conclude with confidence that social media consistently has an impact on these cryptocurrency prices, and they probably depend on other financial and socio-economic factors. That being said, these complex financial models will benefit from taking into account the social media impact on these price trends.

14 Appendix

14.1 Honor Code Pledge

On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.

14.2 Contributions

Vishwanath Kulkarni

- Pre-processed Tweets data (removed outliners and null records)
- Merged two raw datasets based on timestamp and pruned unrelated data
- Predicted Top User with tweets dataset and bitcoin data set
- Formulated the popularity index based on retweet popularity, reply popularity and likes popularity
- Worked on Page Ranking algorithm to predict popular users by creating Directed Graph out of the dataset and displayed the Directed Graph for visuals analysis of the event
- Created webserver for the dashboard to call and get result
- Set up the backend server VM instance on GCP
- Worked on developing architecture of the system and modularity for better communication and functioning

Venkat Arigela

- Worked on unzipping the datasets for performing event analysis.
- Worked on the segregation of dates from the dataset for prediction of events (using price delta & time series anomalies detection).
- Correlated the resultant dates to events (bitcoin events) happened on those dates.
- Worked on getting results from API endpoints using CoinDesk API and twitter API
- Worked on developing the JavaScript code for getting live and predicted bitcoin prices in the form of chart.

Mahesh Parab

- Pre-processed the merged dataset for model training
- Performed data normalization and standardization to fit the LSTM model
- Calculated the sentiment polarity and subjectivity of each tweet and augmented the dataset with these values
- Converted the dataset to time series data with different look back window lengths
- Experimented with LSTM models with different hyperparameters like activation units (ReLU, Tanh), weight update functions (SGD, ADAM) and loss functions (Mean Squared Error, Mean Absolute Error)

Priyanka Umesh Pandit Tailapur

- Worked on getting the live Tweet data using the Streaming Twitter APIs and Embedded Timeline APIs
- Worked on setting up the Google Cloud Platform Data Pipeline Architecture and VM instances
- Worked on setting up the Flask Architecture for the display of dashboard
- Worked on getting the top and popular users for #bitcoin using the APIs exposed
- Set up the dashboard using the Embedded Twitter APIs for the display of Twitter Timeline
- Set up the dashboard with bitcoin value predictions and comparing it with real values
- Set up the dashboard with top and popular users of Twitters who have contributed in #bitcoin Tweets

15 Acknowledgement

To professor Qin (Christine) Lv, for suggesting PageRank algorithm for the influential user list, explaining suggesting Sentiment analysis and using deep learning, specifically LSTM models for training.

References

[1] Cryptocurrency Price Prediction Using News and Social Media Sentiment by Connor Lamon, Eric Nielsen, Eric Redondo - http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf

[2] Deep Neural Networks for Cryptocurrencies Price Prediction by Bruno Spilak - https://d-nb.info/1185667245/34

[3] Bitcoin Volatility Forecasting with a Glimpse into Buy and Sell Orders by Tian Guo, Albert Bifet and Nino Antulov-Fantulin - https://arxiv.org/pdf/1802.04065.pdf

[4] C2P2: A Collective Cryptocurrency Up/Down Price Prediction Engine - https://arxiv.org/pdf/1906.00564.pdf

[5] Prediction of Cryptocurrency Returns using Machine Learning - https://www.researchgate.net/publication/329322600_Prediction_of_Cryptocurrency_Returns_using_Machine_Learning

[6] Bitcoin Price Prediction with Neural Networks by Kejsi Struga and Olti Qirici - http://ceur-ws.org/Vol-2280/paper-06.pdf

[7] Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning by Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre - https://www.mdpi.com/1099-4300/21/6/589

[8] Bitcoin Price Prediction Using Machine Learning by Neha Mangla, Akshay Bhat, Ganesh Avabratha, Narayana Bhat -

https://www.researchgate.net/publication/333162007_Bitcoin_Price_Prediction_Using_Machine_Learning
[9] Next-Day Bitcoin Price Forecast Ziaul Haque Munim, Mohammad Hassan Shakil, and Ilan Alon - https://ideas.repec.org/a/gam/jjrfmx/v12y2019i2p103-d241532.html
[10] Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis by Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra - https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview
[11] Predicting the Price of Bitcoin Using Machine Learning by Sean McNally, Jason Roche,Simon Caton - https://ieeexplore.ieee.org/document/8374483
[12] Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview
[13] D3 References - https://scrimba.com/g/gd3js & https://www.youtube.com/watch?v=C4t6qfHZ6Tw
[14] PageRank Algorithm https://en.wikipedia.org/wiki/PageRank
[15] PageRank Implementation https://www.geeksforgeeks.org/page-rank-algorithm-implementation/
[16] Microservice Architecture https://queue.acm.org/detail.cfm?id=1142065
[17] Chart.JS Documentation - https://tobiasahlin.com/blog/introduction-to-chartjs/
[18] https://towardsdatascience.com/anomaly-detection-with-time-series-forecasting-c34c6d04b24a