Real-time Cryptocurrency Analysis

Real-time Cryptocurrency Price Forecasting Using Financial News and Historical Price

Group Members:

Fatemeh Renani, Jaskaran Kaur Cheema, & Mohammad Mazraeh

Table of contents

Motivation and Background

CryptoCurrency is a digital asset that acts as a medium of exchange using strong cryptography for financial transactions. Due to the use of cryptography for security, it is difficult to counterfeit. Cryptocurrencies use decentralized control, that is, it is managed by distributed ledger technology, generally blockchain and not by any central authority.Therefore, Cryptocurrencies have many advantages over traditional exchange methods. For instance any individual who has access to internet are primed for the cryptocurrency market. Therefore, cryptocurrency price/movement prediction is an active area of research in financial and academic studies.

Bitcoin was released in 2009 and since then more than 2000 altcoins have been created. However, Bitcoin has constantly been ranked as highest in terms of price and volume, therefore in our project we chose to focus on bitcoin.

Related Work

Cryptocurrency price/movement prediction has been approached with two common methods. The most popular method is to use time series analysis on the history of market price [1]. Since News and social media have shown significant effect on market, significant research has focus on predicting the market movement using various information sources such as news and social media [2, 3]. However, through our literature review, we found very few attempt to combine the two data sources.

In his PhD thesis, McNally predicts the Bitcoin pricing using machine learning techniques, such as recurrent neural networks (RNNs) and long short-term memory

(LSTM), and autoregressive integrated moving average (ARIMA) models. In this work he uses Bitcoin price index and transformed prices and found that LSTM achieves the highest classification accuracy of 52% . [1]

In recent article Abraham and et al has analyzed the effect of tweets volume and sentiment analysis on Bitcoin and Ethereum. Interestingly, they have found the tweet volume to be the feature that predicts the Bitcoin/Ethereum price movement well and not the sentiment analysis. [3]

Moreover, Abeer ElBahrawy discussed the comprehensive analysis of cryptocurrency market which included the number of active currencies, market share distribution and

turnover of cryptocurrencies and linking it with ecological model. [4]

Problem Statement

In this project we have focused on the following areas:

- ➢ To create a platform for real-time cryptocurrency prediction.
- ➢ To study relationship between news sentiments and bitcoin price fluctuations.
- ➢ To develop a model to predict, price fluctuations of bitcoin that is whether it will go up or down in next minute.

To achieve above stated goals we have combined the conventional time series analysis technique with news.

Reason to choose above problem set

To make a better prediction we can use more data but often it causes long delays in the prediction and is being done in specific time intervals. Processing speed matters and important events around the world can immediately affect the price of cryptocurrencies and we need to be fast! For instance, in April 2017 Bitcoin value rises over $1 billion as Japan, Russia move to legitimize cryptocurrency. So in this project we introduce a streaming platform in which different kind of data sources can be combined to make a real-time prediction.Since there have not been such a attempt before, we implemented the pipeline from scratch.

Further, we chose to focus on bitcoin as it has been constantly at top rank in terms of volume and price amongst all cryptocurrencies. Since there have been few attempts to find relationship between news sentiment and price and building a model on the basis of combined datasets. Therefore, we made an attempt to build a model after combining them.

Data Science Pipeline

In this section, we describe our architecture and why we designed each step in this way. Figure 1 depicts an overview of our Data Science Pipeline.

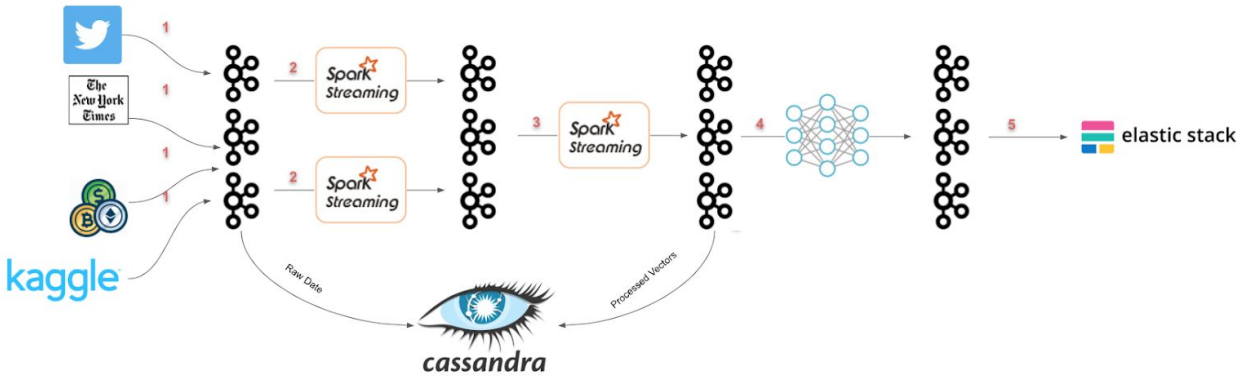


Figure 1: Data science pipeline schematics. The graph represents all the parts in the real-time Bitcoin prediction. The stage are numbered and explained in the main text.

1- Data Extraction

Data is collected from multiple sources of news and cryptocurrency exchange platforms. In the following lines each data source is explained.

Bitcoin price history

Initially, we have obtained Bitcoin's daily price history by scrapping the CoinMarketCap website. However, due to our vision of streaming we decided to use the minute-by-minute price history which was available through a Kaggle competition [5].

News

The New York Times is the source of the news dataset. On an average we scrapped 6000 articles for each month. News articles relevant to the cryptocurrency are later filtered and analysis on the sentiment of news is performed.

While analyzing data, time of article publishing has also been considered in order to study its effect on the price of bitcoin. Data has been collected and cleaned from July 2018 - March 2019.Though, model has only been trained on Jan-March 2019 minute by minute data set.Though finding a data source which could provide historical news data was challenging. Due to change in structure of website over the course of time period, manual work of checking the structure and data received consumed time. Furthermore, limited access to article at some websites posed a problem while identifying the news data source.

2- Feature Extraction

Each event type (price/news) has its streaming application to clean it and extract features from it (e.g. sentiment score, open, close). The price history dataset requires minimal modifications such as converting date to useful feature (such as days of year, days of week, minute of day) and removing unnecessary features. On the other hand news dataset requires major modifications. Scraped data was messy, contained HTML tags, numeric data and articles not relevant to the model. Such articles were filtered followed by cleaning. While preparing data for Sentiment analysis, extraneous words were identified . After this step, 6000 articles were reduced to around 1800 for each month.

3- Feature Aggregation

Features from different data source types need to be aggregated into single feature vector which can be fed into the model.

One important question in real-time prediction is how frequently we want to predict? Given a time period t (which can be minute, hour or day!), we want to handle data from different sources. We may have different price info from different exchanges and multiple news from different media in that specific time period. Thus it is necessary to have an appropriate aggregation mechanism for each time period.

News Data Aggregation

Consider in time period t we have articles $A_t=\{a_0, a_1, \ldots, a_n\}$. For each article we have a relevance feature which measures how much the article is related to our cryptocurrency and and a sentiment value which indicates how much positive/negative the opinion in the article is.
To aggregate the relevance and sentiment value of articles in a time period, we simply calculate a weighted sum of sentiments:

$$v_{news_t} = \sum_{i \in A_t} rel_i \times sent_i$$

Price Data Aggregation

For price, we use the same Open, Close, Low, High features but again we need to introduce some aggregation functions to calculate these features for each time step. For Open and Close we simply average the values. For Low and High, we calculate the

minimum/maximum price of different exchanges in that time period. For Volume feature we sum up all the Volume.

$$Open_{news_t} = \frac{1}{n}\sum_{i \in A_t} Open_i \qquad Close_{news_t} = \frac{1}{n}\sum_{i \in A_t} Close_i$$

$$Low_{news_t} = min(Low_0, Low_1, \ldots, Low_{n-1})$$
$$High_{news_t} = max(High_0, High_1, \ldots, High_{n-1})$$

$$Volume_{news_t} = \sum_{i \in A_t} Volume_i$$

4- Prediction

The prediction step is straight forward. We just load the best pretrained model and receive feature vectors once at a time from Kafka topic. The results are written back to another Kafka topic so we can connect different tools like a visualization tool or an auto-trader bot to the prediction result.

5- Visualization

Any data visualization stack which can be connected to a queue can be used to visualize raw features/predictions/etc. We specifically used Elastic Stack to visualized the minute -by-minute Bitcoin movement prediction.

*** This is the link to our online demo dashboard ***

METHODOLOGY

Tools used

We used **beautifulsoup** for scrapping the news articles from The new york times. Since the structure of website changed every 4-5 months, therefore modifications were made in code to collect correct data. This was amongst the foremost reason to use beautifulsoup as little change in code fetched result. Due to the unavailability of training data for news dataset, we employed pre trained model for sentiment analysis. For this purpose we used the Natural language toolkit.

Spark streaming is used for streaming at parts 2 and 3 of the pipeline. However, since not all deep learning models can be integrated with spark easily. We implement part 5 of the pipeline in a python streaming app.

In order to do the real-time predictions we use a message based architecture. At each step the application receives the data in one format performs some process and writes the generated data back to another queue in the message queue. We use **Kafka** as the message queue in this project.

We have used Keras and scikit-learn libraries for model training.
We chose to perform most of our machine learning training in **Google Colaboratory** (a free Jupyter notebook environment for machine learning). That is because most of our needed libraries are already pre-installed or can be easily install.

In order to visualize the streaming result, one of handy tools was to use **Elastic Stack**. It is pretty straightforward to write json messages to elasticsearch and visualize them in its visualization tool, Kibana.
In this dashboard we monitor number of related news by time to make sure our scrappers work fine. We also monitor the price data, the sentiment scores and most relevant keywords to have an idea of current status of the system.

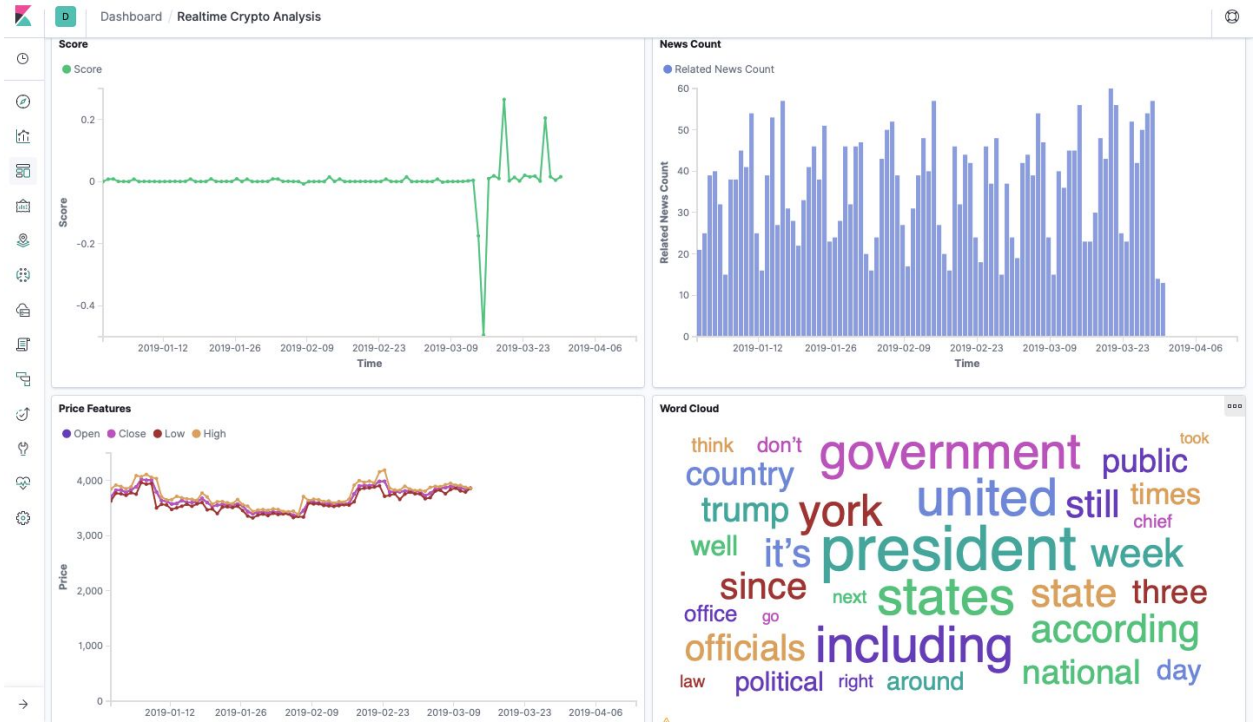


Figure 2: Realtime dashboard of our pipeline including sentiment scores, price features, related news count and most frequent words in the related news.

Exploratory Data Analysis using Jupyter Notebooks

Correlation Analysis

Although in this project we focus on Bitcoin, however, it is interesting to see the correlation between different cryptocurrencies. Figure 3. (a) shows the daily percentage change in closing price for 10 top cryptocurrencies. These cryptocurrencies have a general pattern, that is simultaneously prices going up or down. To get more clear insight Figure 3. (b) shows the correlation relation between the 10 top cryptocurrencies for 10 days. These graphs show that the these cryptocurrencies are highly correlated. Correlation Matrix is prepared using Pearson Correlation, entire matrix entries are greater than 0.6 which confirms that strong correlation exist between cryptocurrencies.

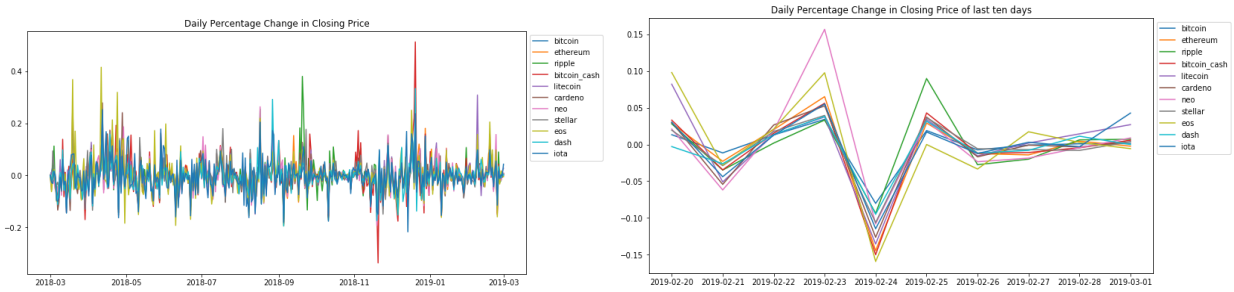
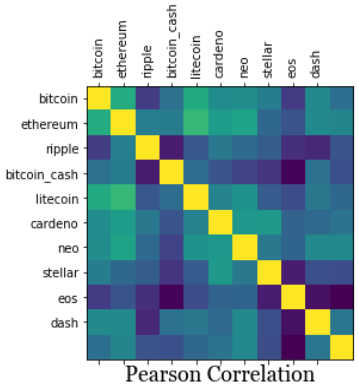


Figure 3: Cryptocurrency correlation.

Figure 4 (a) represents the correlation matrix for the same top 10 crypto-coins by volume. Entries in Correlation matrix had values greater than 0.7, which showed that all cryptocurrencies prices are highly related to each other (see Figure 4 (b)).



	bitcoin	ethereum	ripple	bitcoin_cash	litecoin	cardeno
bitcoin	1.000000	0.856015	0.693391	0.765910	0.852756	0.806999
ethereum	0.856015	1.000000	0.784877	0.784427	0.875003	0.832185
ripple	0.693391	0.784877	1.000000	0.655046	0.726386	0.775187
bitcoin_cash	0.765910	0.784427	0.655046	1.000000	0.759487	0.724073
litecoin	0.852756	0.875003	0.726386	0.759487	1.000000	0.791280
cardeno	0.806999	0.832185	0.775187	0.724073	0.791280	1.000000

Figure 4: Cryptocurrency correlation matrix.

Bitcoin price trend

Bitcoin is the first blockchain-based cryptocurrency, which still remains the most popular and most valuable cryptocoin. Thus we analysed Bitcoin data first to understand the cryptocurrency market. As can be seen in the figure the Bitcoin price has increased exponentially during 2017 (see Figure 5).

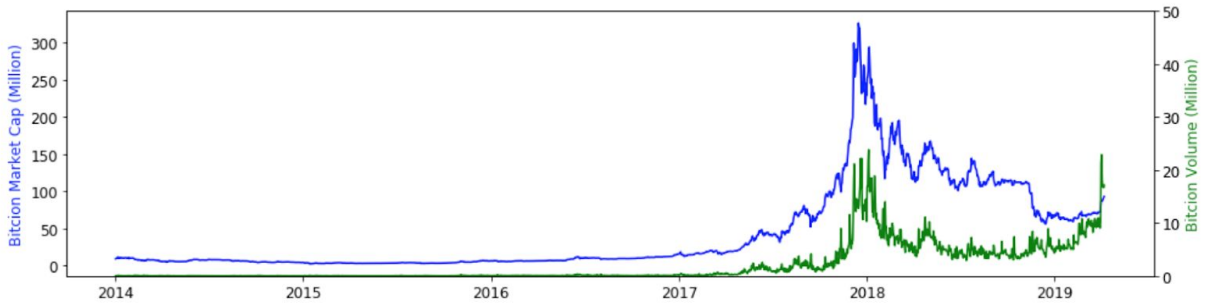


Figure 5: Bitcoin price and Volume is blue and green respectively.

There are many cryptocurrencies with various functions or specifications. Some of these are clones of Bitcoin. For example, Bitcoin was split to two coins on August 01, 2017. Figure 6 shows the history price of the Bitcoin Cash since its birthday at August 2017. It seems that the introduction of the Bitcoin Crash into the market had negative effect on Bitcoin (Market cap. Decreased at August 1, 2017).

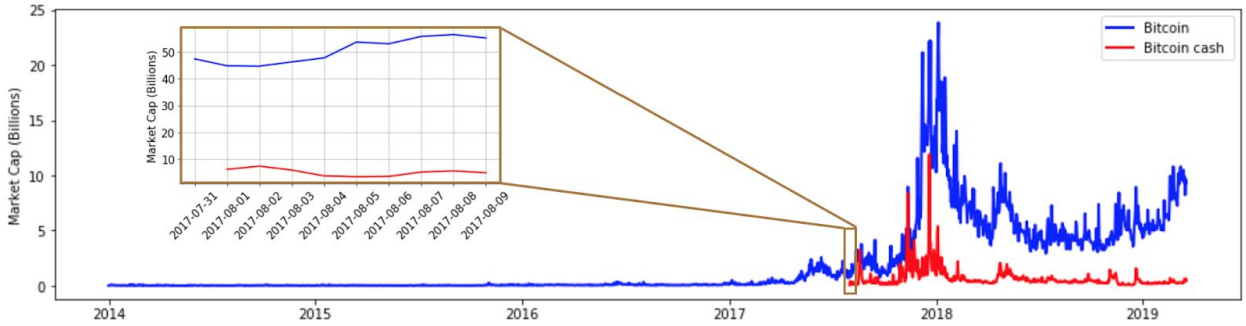


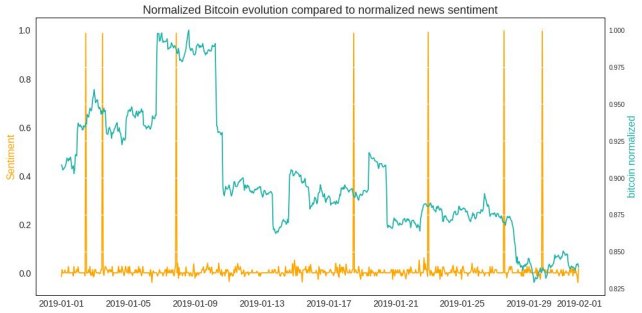
Figure 6: Bitcoin and Bitcoin cash prices represented by blue and red respectively. The top-left insert shows the close look at the price changes at August 2017.

Most recently, Bitcoin Cash has forked into two cryptocoin on Nov. 15, 2018. Bitcoin Cash (BCH) emerged on August 1, 2017 after departing from Bitcoin's (BTC) original blockchain via a hard fork. The separation happened in an attempt to manage BTC's

scalability problem. Part of the Bitcoin community, lead by Roger Ver, rallied for increasing Bitcoin's block size. The Bitcoin community wanted to see Bitcoin thrive as a transactional currency, not an investment asset. Twice a year, the BCH network performs hard forks as part of scheduled protocol upgrades. The latest hard fork, scheduled for Nov. 15, was disrupted.

Sentiment Analysis

We explored the relationship between the news and bitcoin price. Focus was to study not only events related to cryptocurrency affecting the price but also other events. After collecting data, we allocated each article a score depending upon the occurrence of keywords in its content and gave highest to articles mentioning bitcoin or related terms. In most of the cases news sentiment and price had positive correlation, though very few exceptional cases were observed where sentiment of bitcoin and other news did not support the price fluctuations. Another interesting finding was, sentiment of bitcoin article was positive but overall news sentiment which was negative supported the depreciating price.



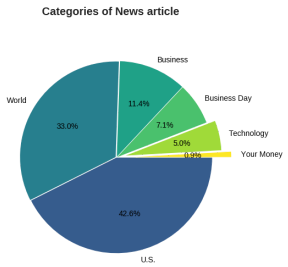


Figure 7 & 8: Sentiment and Price relation and Distribution of news articles

Model training

Due to recurrent nature of our data we have employed the long short-term memory models. The time-series data are generated by stacking 60 data points of aggregate price and news data. Given 60 minute (an hour) of price history and news for Bitcoin, our models will output a real value between 0 and 1. A value over 0.5 is a prediction that the price will rise, under 0.5, the price will fall. We have chosen to work with the following features :
[Open , High , Low , Close , Volume , Sentiment]

In total we trained the model for 92000 datapoint which are minute-by-minute data from January-March 2019. The best trained model yields train/validation accuracy around

51% (see Figure 9). One possible reason for our low accuracy is that: the news dataset is not as frequent at price dataset (every minute). This leads to so many zero for Sentiment feature. A potential solution is to use daily dataset. However, since our project goal is to have a real-time prediction pipeline, we chose to stick with minute-by-minute dataset despite the low accuracy of our current model. Moreover, in future we can add other news sources and various social media information to have very close to minute-by-minute sentiment information.

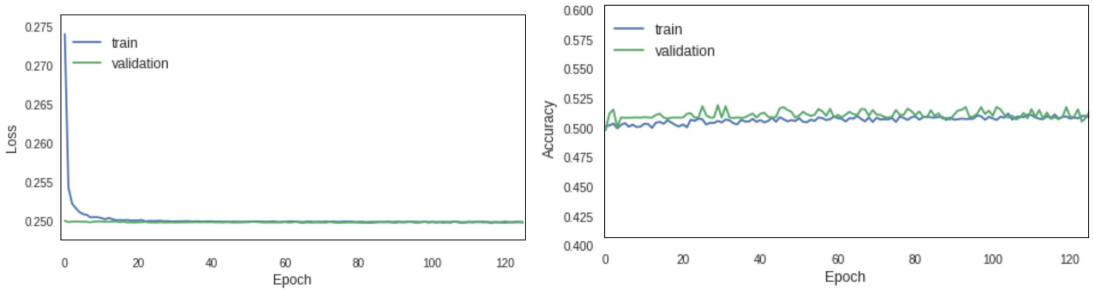


Figure 9: LSTM training progress. Training/validation Loss and accuracy are shown as function of training epochs.

Evaluation

Analysis drawn from EDA:

- High values in correlation matrix along with plot drawn from 10 days shows the strong relationship between the cryptocurrencies. This information is useful for traders as investing in highly correlated currencies may not benefit the trader as decrease in price may lead to loss.Thus trader may invest in currencies which are not correlated to minimize the loss.

- News sentiments affects the price of bitcoin, and this is strongly supported by the plots. We tried to find the relationship between general news and bitcoin price and concluded that not all events influence the fluctuation in price as news containing terms such as cryptography or bitcoin does. Though certain news articles of business type supported the changes in price. This led us to a conclusion that more data from different sources is required to confidently back the analysis drawn.

Model training:

- LSTM was employed to predict the Bitcoin price movement. The model is trained on the last three months data including : Open , High , Low , Close , Volume , and Sentiment features. Since we have only includes sentiment analysis of New York News for this time period, the sentiment feature is quite sparse which leads to low accuracy for our model. Including more news resource will solve this problem and will improve the accuracy of LSTM model.

Data Product

Our data product is mainly the architecture and the simple prototype of it. We discuss the advantages, limitations and future work on this product in this section.

Advantages

- **Generalized**: The general architecture can be used to do on any regression, classification problem which uses a combination of structured and unstructured data sources. It also can be easily integrated with better models for each part (in this project sentiment analysis model and price prediction model). New data providers (crawlers) can easily be integrated with the system. They only need to send the raw data in the defined data format. Even new types of data can be added, we only need to create feature extraction applications for it (e.g. image features)

- **Scalable**: Due to message based nature of the system and used technologies the system can be easily **scaled out** by adding more nodes to Kafka, Spark or Elastic clusters.

- **Dashboard:** The dashboard part of our project gives a real time status of the bitcoin price and we can easily compare the current status of the system with anytime in the past. Although the fact that the dashboard is interactive, we created some more specific dashboard to monitor specific measure that we are interested in. We need to do this to detect when we should train another model.

- **Zero down time:** Using the distributed architecture, we don't have a single point of failure as each step can have multiple instances at the same time. So if any servers encounter a problem, other instances would perform their duties and we

don’t see any non availability in the system. Also because of message based system new deep model rollovers can be done with zero down time. We simple start another instance of prediction which would consume half of messages from Kafka topic and then gracefully stop the old model instances.

Limitations

- Not all deep learning models can be integrated with spark easily. We need to do part 5 in the model in a python streaming app
- Sometimes it can be hard to extract features and normalize data on a per event basis
- Our trained LSTM model has low accuracy which requires further parameter tuning

Future work

- Connecting the pipeline to an exchange platform for real-time data entry
- Adding an scheduled model retraining in the pipeline
- Team-up with another group who have a better training model and improve the accuracy of our prediction
- Add real-time visualization of streaming data such as sentiment analysis and price graph

Lessons Learnt

The first thing we learnt was about how cryptocurrency system works and how the order book works. The next important thing was no matter how much you know about some topic, there are still much more to learn!

From the technical point we tried to apply the streaming concept that we learnt last semester in this project. The main thing we learnt in this side was it's not easy to convert each application to an streaming version! It may be even impossible in some cases. We learned how to change the cleaning and feature extraction parts of a data science project to a “**per event**” way and how to “**combine**” different features on streams of data. Some of this steps could be done in a SQL like command on a batch of training dataset but it needs to be carefully designed in an streaming version.

Another thing that we learned in practice in this project is that not necessarily adding a new source of data can improve the model. Even it is not always about the model we

use. Sometimes the most important question can be “**How to combine these features**” to get a better result. As we mentioned before, we don’t have news data for every minute that we want to do prediction and it should be considered in designing the pipeline.

Summary

Stock price forecasting is a popular and important topic in financial and academic studies and cryptocurrency market is not an exception. In this project we have created a general scalable platform for real-time cryptocurrency price prediction. The platform received the news and price history as its input and it performs feature extraction, feature aggregation, and price movement prediction. Finally the platform outputs the predicted Bitcoin price movement for next minute. At each stage in the pipeline the data is read from a kafka and the new data is written into another kafka. Hence, other cryptocurrency can be easily integrated using this architecture to produce the most realtime, robust and accurate cryptocurrency price prediction project!

References

[1] S. McNally, ‘‘Predicting the price of Bitcoin using machine learning,’’ Ph.D. dissertation, School Comput., Nat. College Ireland, Dublin, Ireland, 2016.

[2] Selene Yue Xu , "Stock Price Forecasting Using Information from Yahoo Finance and Google Trend", UC Berkeley, 2014
https://econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf

[3] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3):1, 2018.

[4] El Bahrawy, A., and Alessandretti, L. (2017). Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science*, 4(170623)

[5] https://www.kaggle.com/mczielinski/bitcoin-historical-data