# Data Preparation and Analysis (CSP 571)

1. **Project Group:**


| Name | Hawk-id | Email-id |
|---|---|---|
| **Sai Yeshwanth (lead)** | A20514173 | Sbolikonda@hawk.iit.edu |
| Ashlesh Khajbage | A20517913 | akhajbage@hawk.iit.edu |
| Chandana poloju | A20527866 | cpoloju@hawk.iit.edu |
| Swastika Pandit | A20503284 | spandit3@hawk.iit.edu |
| Deepthi pinninti | A20513400 | dpinninti@hawk.iit.edu |


**Project Topic:**

## Facilitation of Cryptocurrency price prediction by Sentiment analysis.

Application subject area – Cryptocurrency

# Project Proposal

### Project Description

For all investors, how to approach 4,000%, 88,000% or even 15,390,000,000% R.O.I. Does that sound like a return on investment? These figures are YTD returns for Litecoin (LTC), Ethereum (ETH) and Bitcoin (BTC). Since the creation of Bitcoin, the first decentralized cryptocurrency, in 2009, more than 1,500 new coins have been launched. It has built an empire with a market capitalization of over $1 trillion. Millionaires are made in years, months, even days. But with great victories come great losses. With over 1,500 coins to invest in, there are definitely huge winners and losers to choose from in the crypto world. Each coin is usually associated with a project to bring value. Other coins are scams plain and simple. Even if a coin is legitimate, there is no guarantee that the project backing it will succeed, nor is there any guarantee that the project will attract market interest. Adding to the ambiguity of a crypto investment strategy is the decentralized nature of the space. Since there is no governing body to control the value or flow of these currencies, price fluctuations are purely market driven by free market forces. Simply put, the market is driven by supply and demand. An algorithm that scores opinions on daily news articles about money trading and/or related projects would help investor R.O.I. On the other hand, it can also minimize the loss of a bad investment.

**What we seek to address**

- Does historical data of BTC, LTC, status or ratings influence this result?
- How keywords in daily news articles impact coin's trading performance.
- Is the Global sentiment library suitable for crypto sentiment analysis? If not, do we need to generate our own library ?
- How does the Density Distribution of Deviation of High Price differ from Open Price?
- Can we discover if there is a correlation among investors and R.O.I.
- Do total traded Quantity and Close Price of coin hold any sort of relation between them ?
- Which analysis model will be best for analysis of dataset ?
- Can we analyze price movements and trading volumes and give next day predictions?

**Proposed Methodology/Approach**

Our methodology consists of the following parts:

1. **Data Acquisition**

There are two main categories of data used for this project. The first type is Investing.com's historical price data. No API is required to extract this data, as the site provides an easy-to-use function to download the desired data in CSV format. The CSV data columns are: price, open, high, low, volume and percentage of daily change.

The ultimate goal would be to develop an API that would learn an algorithm for any coin of interest. However, Ripple (XRP) was used as the coin of interest for this project as a basis and proof of concept. Most exchanges do not allow cryptocurrencies to be traded directly with fiat1. Therefore, the most common currency to buy cryptocurrencies is another cryptocurrency. Regarding the use of XRP historical data, historical data was also taken from Bitcoin (BTC) and Ethereum (ETH). These coins were chosen because they are the most used currency to buy all other cryptos. Also, the market generally follows the trend of Bitcoin.

The second type of data was 360 different news articles from five different crypto news sites. Between May 16, 2017 and June, 2018, these articles were scraped by Bitcoin, CNBC, Coindesk, Cointelegraph, and Forbes.

2. **Data Cleaning and Preprocessing**
    - **Historical Price Data Preprocessing**
      The price data for all 3 coins were represented by a multi-index data frame. Although the bulk of the data was numerical, there were a lot of non-numerical values that needed to be converted, namely “Change %” and “Vol.”.

- **Text Preprocessing**
  When dealing with scraped text, the first issue to combat was encoding. As the text was parsed and written to CSV files, the encoding was set as “utf-8”2. However when reading back from CSV, with the encoding set as “utf-8”, python still manages to misrepresent some characters. Thankfully all characters encoded incorrectly were the same and contained the following string: \xao. This string is non-breaking space in Latin1 (ISO 8859-1). A simple replace function fixed all these incorrect strings.

3. **Sentiment Analysis**
   Here we will be extracting overall sentiment from each article. That would be fed into multiple machine learning algorithms, both supervised and unsupervised.

4. **Preparation for Exploratory data Analysis EDA**
   The above sentiment score in conjunction with historical prices of the coin of interest will be paired with historical prices of Bitcoin and Ethereum. We plan to perform EDA on - XRP Correlation Analysis, Time Series Analysis, XRP Response to Sentiment, price range and exact price analysis  etc.

5. **Predictive Models with Deep Neural Networks**
   Due to the extremely random nature of the crypto world, we are planning to implement deep learning models that aim to accurately predict tomorrow’s price.

**Matrices to Analysis Results**

As we are trying to develop a predictive classifier model, we can use following metrics to validate the results

- Correlation
- Confusion matrix
- F1 score
- Test set accuracy and loss

**Project Outline**

1. **Literature Survey**

- **https://github.com/Mooseburger1/Springboard-Data-Science-Immersive/tree/master/Capstone%201%20Project**

- **Jeffrey A. Ryan and Joshua M. Ulrich (2020). quantmod: Quantitative Financial Modelling Framework. - http://www.quantmod.com**
- **https://portfolios.cs.earlham.edu/wp-content/uploads/2021/09/ACM_Conference_Proceedings_Primary_Article_Template_1_-3.pdf**
- **https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9904351**
- **https://onlinelibrary.wiley.com/doi/full/10.1111/exsy.12493**

**Dataset Description**

**Data Source**: CNBC, FORBES, BITCOIN, COINDESK, COINTELEGRAPH
**Dataset Type**: CSV files

## Data Sources with description

| File name | Description | No. of Attributes | No. of rows |
|---|---|---|---|
| BTC Historical Data.csv | Historical data price of Bitcoin from investing.com in CSV format | 7 | 254 |
| ETH Historical Data.csv | Historical data price of Ethereum from investing.com in CSV format | 7 | 254 |
| Ripple_CNBC.csv | Articles from CNBC written to CSV | 6 | 35 |
| Ripple_Forbes.csv | Articles from Forbes written to CSV | 6 | 139 |
| Ripple_bitcoin.csv | Articles from Bitcoin written to CSV | 6 | 330 |
| Ripple_coindesk.csv | Articles from Coindesk written to CSV | 6 | 792 |
| Ripple_cointelegraph.csv | Articles from Cointelegraph written to CSV | 6 | 122 |

## Analysis Model

- **Correlation Analysis**
- **Time Series Analysis**
- **Sentiment Analysis**
- **Deep Learning**

## Tools & Software

1. **Software:**
    - RStudio
2. **R Libraries:**
    - Ggplot2
    - Rvest
    - Plotly
    - Sentimentr
    - Dplyr
    - Tidyverse
    - gridextra
    - corrplot
3. **Project Management and Source Control:**
    - GitHub
    - Slack
4. **Data Visualization tool**
    - Tableau
    - Power BI