**IE 575: Predictive Analytics Course Project**

Reddit Sentiment Analysis for Cryptocurrency Price Prediction

**Team 8:**

* NAME CENSORED
* **Raditya Fahritama**
* NAME CENSORED

**School of Graduate Professional Studies**

Data Analytics

*Spring Semester, 2022*

# 

# 

# **Table of Contents**

[**Introduction**](#_ke2aaxwtcbnd) **3**

[**Problem Statement**](#_qf4n8o14c36a) **3**

[**Data Description**](#_o5jegsus8193) **3**

[**Methodology Proposed**](#_64iirtfygncp) **4**

[**Challenges**](#_wlt514ppzr3) **7**

**Related Works 8**

**Expected Results 8**

[**References To Data Sources**](#_bmcpbu7eeo3y) **8**

# 

# 

# 

# 

# 

# 

# 

# **Introduction**

A *cryptocurrency* is a digital asset that employs cryptography to encrypt transactions, control the generation of new cryptocurrencies, and verify the secure transfer of assets. Cryptocurrencies are digital currencies that differ from traditional currencies in that they are based on the notion of decentralized control, as opposed to traditional currencies' reliance on central banking institutions. The first cryptocurrency was created in 2008 when an unknown individual using the pseudonym Satoshi Nakamoto published a paper titled Bitcoin: A Peer-to-Peer Electronic Cash System in the public domain. In January 2009, Nakamoto released the bitcoin program as open-source code on SourceForge. The contributions of Satoshi Nakamoto sparked a wave of public interest, prompting others to build rival cryptocurrencies based on the same underlying technology but with different purposes.

There is a massive volume of unstructured data produced in the form of tweets, Reddit posts, internet articles, text messages, emails, and other formats. As a result, "natural language processing" (NLP) has emerged as a field of research or development. Web data beyond Twitter and social media has been an affluent area of research. Our project aims to integrate Reddit posts and cryptocurrency prices to understand if price fluctuations can be studied using social media behavior.

# **Problem Statement**

The objective of our project is to perform sentiment analysis on Reddit posts and anticipate price fluctuations in cryptocurrency (Bitcoin) over a period of time through time series analysis using various machine learning methods.

# **Data Description**

The Reddit CryptoCurrency data consists of the following variables:

* **title** - the title of the posts
* **score** - a score based on impact, number of comments for relevant posts
* **id** - unique id given for posts/comments
* **url** - URL of post thread
* **commns\_num** - number of comments made to this post
* **created** - creation date
* **body** - text of the post or comment
* **timestamp** - timestamp

The BitCoin price data has the following variables:

* **Date**: The date for which the price of BTC is posted
* **Open** - The opening price of BTC for the given date
* **High** - The highest price of BTC for the given date
* **Low** - The lowest price of BTC for the given date
* **Close** - The closing price of BTC for the given date
* **Volume** - The volume transacted i.e. for BTC this is in BTC amount

# **Methodology Proposed**

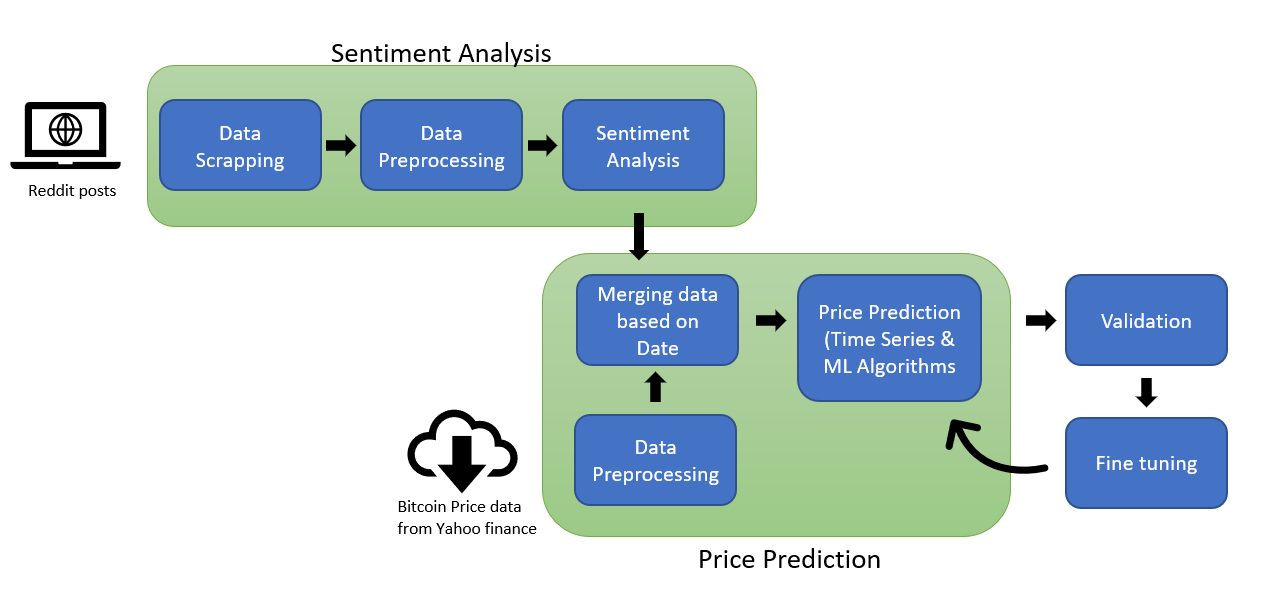


Fig: Architecture

1. **Data Preprocessing (text data):**

## **Removal of punctuations and Stopwords:** Stop words and punctuations provide minimal semantic information. Removing these words and punctuation makes it easier to isolate and find the words required to differentiate the topics.

## **Lemmatization and Tokenization:** Lemmatization is the task of grouping up different forms of a word so that they can be computed as a singular item.Tokenization is the process of converting a string into pieces or “tokens”. For example, converting a sentence into words or combination of words.

## **Creating Bigrams and Trigrams:** Bigrams and Trigrams plays a significant role in text preprocessing to create relevant pairs of words.

## **Using TF-IDF to remove low frequency words:** Term Frequency-Inverse Document Frequency(TF-IDF) is calculated based on two different metrics. The term frequency which is the frequency of each term in a document and inverse document frequency which is the frequency of how many documents contain that word.

1. **Data Preprocessing (bitcoin data):**

Apart from eliminating null and missing values from the dataset, we require a constant mean and standard deviation for time series analysis, which can be verified by visualizing the mean and standard deviation for a rolling window of the bitcoin dataset. We will use the Self Lag Differencing strategy to get a constant rolling mean. It is the distinction between the current series and a delayed version of it. The shift might be in the order 1,2,3,4, etc., or NULL (for items for which no lagged version exists).

Time[Price\_difference]=Time[Price]-Time[Price].shift(x)

where x=lagged version(1,2,3,4, etc., or NULL)

1. **Using Sentiment/Emotion Analysis algorithms to understand the sentiment of Reddit Posts/ Comments:**

We plan to use and compare the results of sentiment analysis algorithms VADER and TextBlob which would give polarity scores of the input corpus. VADER is a lexicon- and rule-based sentiment analysis tool that is sensitive to suppositions expressed in web-based media. Using this methodology, all the words in the vocabulary are evaluated with respect to whether it is positive or negative, and how positive or negative they are. On the other hand, TextBlob returns both the polarity and subjectivity of a sentence. Polarity lies between -1 and 1, where -1 defines negative sentiment and 1 defines a positive sentiment. Subjectivity quantifies the amount of personal opinion or the amount of facts that the text contains. Higher subjectivity depicts that text contains personal opinions rather than facts. These scores can be given as input to the further analysis/ prediction models and validate whether sentiments/ emotions play a role in predicting the prices of BitCoin.

1. **Price Prediction:**

**Auto Regressive Integrated Moving Average** (ARIMA) is a statistical analysis model that uses time series data to either better comprehend the data set or anticipate future trends. A statistical model is considered autoregressive if it predicts future values based on past values. ARIMA can be understood by outlining each of its components.

* Au[toregression (AR)](https://www.investopedia.com/terms/a/autoregressive.asp): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values
* Integrated (I): represents the differencing of raw observations to allow for the time series to become stationary
* M[oving average (MA)](https://www.investopedia.com/terms/m/movingaverage.asp): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations. This model will assist us in predicting future value based on given historical data.

**Prophet** is a time-series data forecasting process based on an additive model that fits non-linear trends with annual, weekly, and daily seasonality, as well as holiday impacts. Missing data, trend changes, and outliers are handled well using this model. A decomposable time series model with three key model components: trend, seasonality, and holidays is used in the approach. The output not only provides us with the predicted value but also the upper and lower prediction limits, which is really useful to make predictions compared to other traditional algorithms that give us one price point as the predicted value.

Since bitcoin prices data is over a period of time at day level, using the time series algorithms to predict the price of bitcoin.

**Random Forest** isa supervised learning algorithm made up of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model. Random forest is a machine learning technique that, in most cases, gives excellent results without any excessive hyperparameter adjustment. The algorithm's flexibility and convenience, which we believe will be the best fit for our model.

1. **Validation metrics of models:** We would like to perform cross-validation on the data set to ensure the model performance on unseen/ new data. We would like to compare the models (ARIMA, Facebook Prophet and RandomForests) using the values of MSE, RMSE, MAE, MAPE, R Squared and Adjusted R Squared values. Comparing these models, further model hyper-parameter tuning and feature engineering can be performed to improve the models.

# **Challenges**

1. **No validation for sentiment analysis algorithms**

There are no predefined metrics to validate the polarity scores given by the algorithms. Hence, we are using two methods for sentiment analysis in order to roughly compare the sentiment of posts predicted by each. However, due to absence of any metric, it is difficult to conclude if the predicted sentiment score is accurate.

1. **Neutral sentiment in the collected Reddit dataset**

It needs to be assessed how many Reddit posts have any sentiment at all. When the majority of the Reddit posts aren't objective, sentiment analysis of them has little value for the model. Furthermore, it must be demonstrated that there is a correlation between bitcoin Reddit posts sentiment and changes in cryptocurrency prices. The model will be filled with noise if this is not done.

**Related Works**

1. Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol. 1 : No. 3 , Article 1
2. Li TR, Chamrajnagar AS, Fong XR, Rizik NR and Fu F (2019) Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model. Front. Phys. 7:98. doi: 10.3389/fphy.2019.00098

**Expected Results**

Based on the data collected from Reddit and the bitcoin prices basic data cleaning and preprocessing would be done for further analysis. Relevant KPIs will be identified and Feature engineering will be performed to finalize the set of features that can be given as input to the time series and prediction models. Sentiment analysis will be performed on the Reddit data and the polarity for each post/ comment will be scored. The emotion score will also be further considered as input features for the prediction models. Considering all the input features, predictive models are built, validated and compared for performance. The overall goal would be to compare the price prediction using two sentiment analysis algorithms(VADER and TextBlob) and 3 predictive analysis algorithms (ARIMA, Facebook Prophet and Random Forests) and present the results.

# **References To Data Sources**

* <https://finance.yahoo.com/quote/BTC-USD/history/>
* <https://www.kaggle.com/gpreda/reddit-cryptocurrency/version/14>