IE-575 PREDICTIVE ANALYTICS

**PROGRESS REPORT**

Reddit Sentiment Analysis

for

Cryptocurrency Price Prediction

**Team 8**: NAME CENSORED, Raditya Fahritama, NAME CENSORED

horizontal line

# 

# 

# 

# 

# Project objectives

The objective of our project is to perform sentiment analysis on Reddit posts and anticipate price fluctuations in cryptocurrency (Bitcoin) over the year 2019 through time series analysis using various machine learning methods. With this project, we are trying to determine if there is a relationship between posts on Reddit and the price fluctuation in Bitcoin.

## Tasks completed

* **Study of existing solutions:** The team looked over numerous approaches for sentiment analysis and time-series analysis as part of the initial exploration for the project.
* **Survey of scientific papers and journals:** The team read a variety of scientific research publications to obtain a better understanding of the context of the tasks and approaches to be applied. It assisted us in learning about previous procedures and their outcomes, which provided us with direction in which to proceed.
* **Data acquisition and collection:** Initially, the data provided to us wasn’t in the required timeframe. Hence, we tried scraping data from Reddit however retrieving dates for posts was a tricky part. On further exploration, the dataset was available on Kaggle.
* **Data cleaning and preprocessing:** For model building, preprocessing is a critical step. After data collection, we started working on data preprocessing for the sentiment analysis which involved tasks such as removal of punctuations and stop-words, lemmatization and tokenization, bigrams and trigrams creation and TFIDF scores.

## Work in Progress

* **Data Cleaning and preprocessing of Bitcoin Price dataset:** The next stage of the project will be time-series analysis. Till the time, sentiment analysis is being done, one team member is cleaning the bitcoin price dataset for null and missing values

**Task Aligned for next week**

* **Sentiment Analysis of Reddit posts:** As we have the preprocessed data of Reddit posts for Bitcoin, we have started working on the first part of the project i.e. Sentiment Analysis. As we are using two separate methods for sentiment analysis, both the methods are divided into two team members.
* **Time Series Analysis using ARIMA, Prophet:** Once we have completed the sentiment analysis phase, we will include the sentiment scores with the price for bitcoin during the selected time period. After that, we will initiate the time-series analysis on the final merged dataset to predict the bitcoin prices and also check the impact of sentiments from Reddit posts.
* **Prediction using Random Forest:** We are using Random Forest as one of the supervised methods for price predicting of bitcoin and the impact of sentiments on bitcoin price. This method will be built in parallel to the other two time-series analysis methods to compare the performance of the models.

**Problems encountered**

* **Data collection:** Initially, there was no temporal overlap between the Reddit postings data and the bitcoin price data. As a result, we began scraping data from Reddit on our own using Python code, however extracting the data timestamp of the posts was a time-consuming operation. On further exploration, we found the data online on Kaggle [[Bitcoin Reddit posts](https://www.kaggle.com/datasets/jerryfanelli/reddit-comments-containing-bitcoin-2009-to-2019)] and used it for the initial task of Sentiment Analysis.

**Pending tasks**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Week 1 (March 21 - March 27)** | **Week 2 (March 28 - April 3)** | **Week 3 (April 4 - April 10)** | **Week 4 (April 11 - April 17)** | **Week 5 (April 18 - April 25)** |
| **Scientific Study** | **Completed** |  |  |  |  |
| **Finding relevant works** |  |  |  |  |
| **Project Proposal and Planning** |  | **Completed** |  |  |  |
| **Data Collection** |  |  |  |  |
| **Data Cleaning** |  |  | **Completed** |  |  |
| **Data Pre-processing** |  |  |  |  |
| **Sentiment Analysis - VADER** |  |  |  | **Scheduled (On Track)** |  |
| **Sentiment Analysis - TextBlob** |  |  |  |  |
| **Prediction using Facebook Prophet** |  |  |  |  |
| **Prediction using ARIMA** |  |  |  |  |
| **Prediction using RandomForest** |  |  |  |  |
| **Model Fine Tuning** |  |  |  |  | **Scheduled** |
| **Model Validation** |  |  |  |  |
| **Documentation and Presentation** |  |  |  |  |

We've set aside the final two days (April 26-27) as buffer days in case we get delayed at any point during the model building process

**Team contribution**

All the work is fairly distributed and equally contributed by all the team members. Scientific study & review and previous work done in the field were done by all the team members. We collaborated and brainstormed among ourselves to come up with a project proposal and a plan that can be executed in the next 3 weeks to present the results. While one person is working on data collection, the others worked on planning, cleaning and pre-processing steps. Data pre-processing is done by one of the team members to have standard data that can be used throughout the next steps. The work is well distributed for the coming up weeks for constructive learning and equal contribution of the team members.