Multimodal Quantitative Finance Hackathon

By
Manish Nayak - CE22B069
Om Shende - CE22B083
Sarvesh - CE22B103

Text Data Analysis

Text Feature Selection

Initially, we analyzed various columns from Twitter data, but many had very low counts or minimal contribution to the predictive power. To streamline the analysis, we removed these low-value columns and retained 4 that were most relevant to understanding the market sentiment.

Feature Concatenation:

Instead of treating each feature separately, we combined key elements —such as the tweet content, reply count, likes, and retweets—into a single aggregated feature. This allowed us to process the text data in a more unified manner, capturing a broader sense of engagement and sentiment.

Sentiment Analysis with RoBERTa

We implemented the **RoBERTa** model, a powerful NLP tool, to assess the sentiment of the aggregated Twitter data. The model calculated the probabilities that each input was Bullish (positive sentiment), Bearish (negative sentiment), or Neutral. These probabilities helped us understand how Twitter sentiment might correlate with price movements and were also used as additional features.

Visual Data Analysis

Planned Approach with Vision Transformers:

Our initial plan involved using **Vision Transformers** to extract insights from visual data, such as images and video thumbnails related to cryptocurrency trends on social media. The goal was to create **embeddings** (vector representations) of these visuals, which could serve as additional features in our prediction models.

Challenges with Visual Data

Despite our efforts, the visual data we had access to was quite random and lacked consistency. This randomness meant that it was difficult to derive meaningful patterns or trends from the images, making the data less useful for prediction. As a result, we decided to drop this aspect of the analysis to focus on more promising approaches.

Price Data Analysis:

ARIMA Model to handle discontinuities

A major challenge we faced was that the training data had discontinuities, meaning that certain time periods were missing in the training set but appeared in the test set. This inconsistency made it difficult to model trends directly, as the lack of continuous data could lead to inaccuracies in predicting future prices. To address this, we implemented a time-series forecasting method called **ARIMA (AutoRegressive Integrated Moving Average)**. ARIMA helped us predict the closing prices for the missing time periods, effectively filling in the gaps. By doing so, we were able to create a continuous dataset that captured the overall price trends more accurately. This continuous data was crucial for building a robust model that could better understand and anticipate future price movements.

Final Features

1. **Tweet Engagement**
2. **Sentiment Score**
3. **Predicted Close Prices**
4. **Financial Indicators**

Calculation of Financial Indicators:

With a more complete price dataset, we calculated several key financial indicators such as **Simple Moving Average (SMA)**, **Exponential Moving Average (EMA)**, **Relative Strength Index (RSI)**, **Average True Range (ATR)**, and **Moving Average Convergence Divergence (MACD)**. These indicators helped to capture market momentum, volatility, and trend strength, providing valuable inputs for our final model.

Prediction using Shallow Learning Models

After generating the financial indicators, we used **shallow learning models** like **regression** and **decision trees** to predict the closing prices of cryptocurrencies. The sentiment features from Twitter and the financial indicators from price data together provided a robust set of inputs, making our predictions more accurate even when dealing with incomplete datasets.