Final Report CS521: Cryptocurrency Market Analysis using Data-mining Techniques

Prasanth Reddy Guvvala, Rahul Payeli

December 2024

1 Introduction

In this project we aim to analyze the Crypto currency market using some data mining techniques. The market cap for crypto is around 1 trillion US dollars in 2023. Now (2024) the market cap is grown to 2.4 trillion US dollars. In recent years, the cyrpto currency has grown significantly, not only in size but also in the complexity of dealing with them. Investors and analysts are facing challenges in understanding and predicting the price trends, as well as in identifying patterns in different cryptocurrencies. Understanding the market helps in making smarter decisions and avoid losses. Applying advanced data mining methods to a comprehensive data of cryptocurrency from Yahoo finance. We seek to uncover some patterns and valuable insights in the crypto market, which could inform trading strategies, risk management, and market understanding. We seek to undercover how public sentiments are playing role in cryptocurrency prices, and find groups of cryptocurrencies with similar behaviors. This will encourage more people to invest in this market, and drive innovation in finance.

2 Data

2.1 How did we collect the data?

Our project focuses on the cryptocurrency market data, which is retrieved using Yahoo Finance API. We ran a python file which is using this API to retrieve data. This API widely covers financial markets, including cryptocurrencies. Although, the data provided by this api has granular range as fine as 1 minute, we have collected hourly frequency data which spans for the last 6 months due to API limitations. Our dataset includes 231 different cryptocurrency tickers, from which, some of them have 4500 records and remaining contains 8300 records on an average due to different start dates in Yahoo finance API. The total number of records sums up to 1.9 million data points.

Datetime	Open	Close	High	Low	Adj Close	Volume	Ticker
2023-09-30 19:00:00+00:00	0.268558264	0.270085573	0.270489603	0.268558264	0.270085573	0	1INCH-USD

Table 1: Sample Data Point from Yahoo crypto currrency dataset

To generate the cryptocurrency sentiments we created a data pipeline that fetches 5 articles at most related to a cryptocurency in the 6 months date range. The API used was Cryptonews-api.The API has the capability of fetching more news articles but due to rate limits we had to fetch articles that are related to top 10 cryptocurrencies which are Bitcoin, Ethereum, Binance Coin, Tether, Cardano, Ripple, USD Coin, Dogecoin, Solana, Polkadot.

ID	News URL	Title	Date	Sentiment	Score	Topics	Type	Ticker(s)	Full Text Sentiment (Pos/Neg/Neu)
1	u.today/is-cardano-ada-aiming...	Cardano Aiming at Change	30/09/2023 19:00	Positive	0.0188	ADA	Article	ADA	0.0188 / 0.5531 / 0.4281

Table 2: Sentiment Analysis and Cryptocurrency News sample Data

Field name	Description
Datetime	date and time of the data point
Open	opening price of the cryptocurrency
Close	closing price of the cryptocurrency
High	highest price of the cryptocurrency
Low	lowest price of the cryptocurrency
Adj Close	Adjusted closing price after stock splits or dividends (may be the same as Close for cryptocurrencies)
Volume	Number of units traded
Ticker	The unique identifier for the cryptocurrency

Table 3: Description of Data Fields

Field Name	Description
ID	A unique identifier for each news article entry.
News URL	The URL of the news article related to cryptocurrency.
Title	The title of the news article.
Date	The publication date and time of the news article.
Sentiment	The overall sentiment classification of the article (e.g., Positive, Negative, Neutral).
Score	The sentiment score indicating the strength of the sentiment (higher values indicate stronger sentiment).
Topics	The cryptocurrency or topic associated with the article (e.g., ADA).
Type	The type of news content (e.g., Article, Blog).
Ticker(s)	The ticker symbol(s) of the cryptocurrency related to the article.
Full Text Sentiment	Detailed sentiment analysis scores for Positive, Negative, and Neutral tones in the full text of the article.

Table 4: Field Descriptions for Sentiment Analysis and Cryptocurrency News Data

2.2 How did we Organize and Stored the data?

After getting cryptocurrency data using Yahoo finance API, we use CSV (Comma Separated Values) as our primary data storage format. Each cryptocurrency ticker has its own CSV file. A file is named based on the ticker symbol of the cryptocurrency, currency in which the crypto is dealt in dataset, and the span of time of the data in dataset (eg: ETH-USD-6m-2024-09-30.csv). Each CSV has columns like 'Datetime', 'Open' 'Close', 'High', 'Low', 'Adj Close', 'Volume', and 'Ticker'. All CSV files are stored in single directory, as shown in Fig. 1, which allows us for easy access of the data. Whenever needed, we can easily load and combine data from multiple files for cross-cryptocurrency analysis. This implementation can accommodate analysis on additional cryptocurrencies by simply adding some new CSV files in the directory.

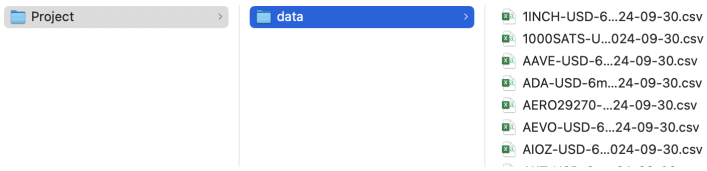


Figure 1: Data directory

From collected articles we implemented a web scraping technique using beautiful soup 4 to fully process the body of article and we used Fin Bert extract the positive , negative , neutral sentiment probabilities in the article that was published regarding a certain cryptocurrency. Then our pipeline stores the data in a Postgres database. We had to make a design choice of persisting data in the Postgres data-base because pandas and numpy are found to be extremely slow when the data size gets larger.

The pipeline we implemented also has the feature to split the enormous text into chunks of 512 tokens so that it can be further processed to extract sentiments. Finally here's the snap shot of the data that is ingested in the Postgres database.

news_url text	image_url text	title text	text text	source_name text	date timestamp without time	sentiment text	sentiment_score numeric	topics text[]	type text	tickers text[]	full_text text	full_text_sentiment_pos numeric	full_text_sentiment_nega numeric	full_text_sentiment_neutral numeric
https://u.to...	https://cry...	Is Cardan...	Cardano (...	UToday	2023-09-30 19:00:00	Positive	[null]	{}	Article	{ADA}	Disclaime...	0.01875678045325912	0.5531275570392609	0.428115660091862
https://thec...	https://cry...	Cardano's...	In recent ...	The Currency Analytics	2023-09-30 08:16:26	Positive	[null]	{}	Article	{ADA}	The Curre...	0.01911689846843955	0.27339364049782944	0.707489457621704
https://daily...	https://cry...	6 Cardan...	Which Ca...	Dailycoin	2023-09-30 07:00:00	Neutral	[null]	{}	Article	{ADA}	Error fetc...	0.001525893923826551	0.9984108209609985	0.000063345301896333
https://www...	https://cry...	The Most ...	Behind th...	NewsBTC	2023-09-30 06:00:43	Positive	[null]	{}	Article	{ADA}	Behind th...	0.000013102801252291	0.8539412021636963	0.146045680566203
https://thec...	https://cry...	Cardano's...	In the eve...	The Currency Analytics	2023-09-30 05:51:03	Positive	[null]	{}	Article	{ADA}	The Curre...	0.00000044759482875	0.000020903178779008	0.999978661537170
https://thec...	https://cry...	The Ultim...	In the eve...	The Currency Analytics	2023-10-01 08:45:57	Positive	[null]	{}	Article	{ADA}	The Curre...	0.00002678397201331	0.6892825166384379	0.310690666055031
https://cryp...	https://cry...	Massive ...	Cardano-...	CryptoPotato	2023-10-01 08:28:12	Positive	[null]	{}	Article	{ADA}	Cardano-b...	0.00100869929883629	0.1438482701778412	0.855143070220947
https://coin...	https://cry...	Price anal...	Bitcoin an...	Cointelegraph	2023-10-02 12:05:00	Neutral	[null]	{tanalysis}	Article	{ADA,BNB...	Error fetc...	0.00005463500565383	0.9999358654022217	0.000009544998647470
https://u.to...	https://cry...	Cardano (...	Cardano (...	UToday	2023-10-02 09:43:00	Positive	[null]	{}	Article	{ADA}	Cardano (...	0.00000438187157669	0.9999880194664001	0.000007609390195284
https://zycr...	https://cry...	Cardano ...	Cardano (...	Zycrypto	2023-10-02 09:38:06	Positive	[null]	{}	Article	{ADA}	Cardano (...	0.00000153827876658	0.000123524216178338	0.999874949455261
https://amb...	https://cry...	What wen...	The 25-29...	AMBCrypto	2023-10-02 09:30:02	Positive	[null]	{}	Article	{ADA}	The 25-29...	0.00021388098321040	0.6452771574258804	0.354508986696600
https://www...	https://cry...	$ADA: Cr...	On 1 Octo...	CryptoGlobe	2023-10-02 04:55:47	Positive	[null]	{}	Article	{ADA}	On 1 Octo...	0.00000308844937535	0.9996538162231445	0.000343053718097507
https://thec...	https://cry...	Cardano ...	In a surpri...	The Currency Analytics	2023-10-03 15:50:48	Positive	[null]	{tanalysis}	Article	{ADA}	The Curre...	0.1787669883281821	0.46040316224098204	0.360829830914735
https://amb...	https://cry...	ADA bulls...	The mass...	AMBCrypto	2023-10-03 15:30:15	Positive	[null]	{tanalysis}	Article	{ADA}	The massi...	0.06764952652156353	0.8737863302230835	0.058564085047692
https://coin...	https://cry...	Cardano ...	Ardana cl...	Cointelegraph	2023-10-03 14:59:00	Negative	[null]	{stablecoi...	Article	{ADA}	Error fetc...	0.00137631292454898	0.9985535740852356	0.000070147747464943
https://u.to...	https://cry...	Cardano (...	Bears are ...	UToday	2023-10-03 12:00:00	Negative	[null]	{tanalysis}	Article	{ADA}	Disclaime...	0.00009798346218303	0.9997815787792206	0.000120445392894907
https://www...	https://cry...	Cardano ...	On-chain ...	NewsBTC	2023-10-03 11:00:39	Negative	[null]	{}	Article	{ADA}	On-chain ...	0.02527035905279717	0.9669191042582194	0.007810527684711
https://thec...	https://cry...	Cardano (...	In the eve...	The Currency Analytics	2023-10-04 16:48:06	Positive	[null]	{}	Article	{ADA}	The Curre...	0.00175200036028400	0.7834719307720661	0.214776075561530
https://daily...	https://cry...	Cardano ...	Despite ki...	Dailycoin	2023-10-04 14:26:31	Negative	[null]	{}	Article	{ADA}	Error fetc...	0.12277854979038239	0.8762537240982056	0.000967746018432086
https://coin...	https://cry...	Price anal...	Bitcoin is ...	Cointelegraph	2023-10-04 11:43:30	Positive	[null]	{tanalysis}	Article	{ADA,BNB...	Error fetc...	0.00005501011037267	0.9999393224716187	0.000005765898962007
https://u.to...	https://cry...	Cardano's...	Midnight, ...	UToday	2023-10-04 08:37:00	Positive	[null]	{}	Article	{ADA}	Midnight, ...	0.00008140932209244	0.9999016523361206	0.000016919389508984
https://www...	https://cry...	Cardano ...	Cardano (...	NewsBTC	2023-10-04 07:00:41	Positive	[null]	{}	Article	{ADA}	Cardano (...	0.00033476168118795	0.8219616131488389	0.177703628665919
https://amb...	https://cry...	Cardano: ...	Cardano's...	AMBCrypto	2023-10-05 15:30:42	Neutral	[null]	{tanalysis}	Article	{ADA}	Cardano's...	0.01249887701123952	0.9870128035545349	0.000488296354888
https://u.to...	https://cry...	Cardano's...	Input Out...	UToday	2023-10-05 11:48:01	Positive	[null]	{}	Article	{ADA}	Input Out...	0.00075695456234825	0.9947234988212585	0.004519518060988
https://finb...	https://cry...	DeFi prot...	The total ...	Finbold	2023-10-05 11:27:58	Positive	[null]	{}	Article	{ADA}	By submit...	0.00020249633515353	0.7823965052763621	0.217400990145956

Figure 2: Database table articles latest in schema Crypto-news

2.3 Data Pre-processing

We begin our pre-processing by checking if there are any missing values in our cryptocurrency data we retrieved and found out there are **no missing values**. But there are **null values** for sentiment proablilities whenever there was an error scraping the information of the full text.

2.3.1 Date-Time Conversion

Converted the given date information into a **standardized date-time format** to ensure consistency across all the data samples. This makes the time based analysis easier and helps in sorting the data easily.

2.3.2 Normalization

We applied normalization by selecting certain features like closing price, based on the tasks. By doing this, all cryptocurrencies are brought on to similar scale, and preventing the large numbers dominating the analysis. This will help us in maintaining fair comparison between different cryptocurrencies. For instance, if we look at Figure 3, the closing price distribution of BTC-USD and POL28321-USD are seen to be way far in scale. From figure 4a, the same currencies can be seen that large difference, BTC-USD is near to 8000 but POL28321-USD is almost near to 0.1. After normalization (fig.4b), both are spread from 0 as a percentage change. This is very important to get a fair compaision between them.

2.3.3 Seasonal Decomposition and Anomaly Detectection

We have generated Seasonal Decomposition graphs, as shown in Figure 5a, for some cryptocurrencies with different features. This gave visual representation of trends and patterns in the data. With the help of these graphs, we identified the seasonality, cyclical patterns, or anomalies that cannot be observed directly from the raw data. This also helped us in selecting features for a given data-mining task. For instance, in the figure 5a, if we look at the seasonality it shows that the BTC-USD has no seasonality which is an indicator that this currency is more volatile in nature and risky. This graph also helped us in visually conforming if any given two cryptocurrencies are similarly behaving.

In figure 5b, this anomaly detection using 3-standard deviation method, helped us identify and potentially remove or flag extreme outliers that could skew our analysis. This highlights interesting market events or potential data errors. It also helped us in cross-referencing with external data to understand their causes.

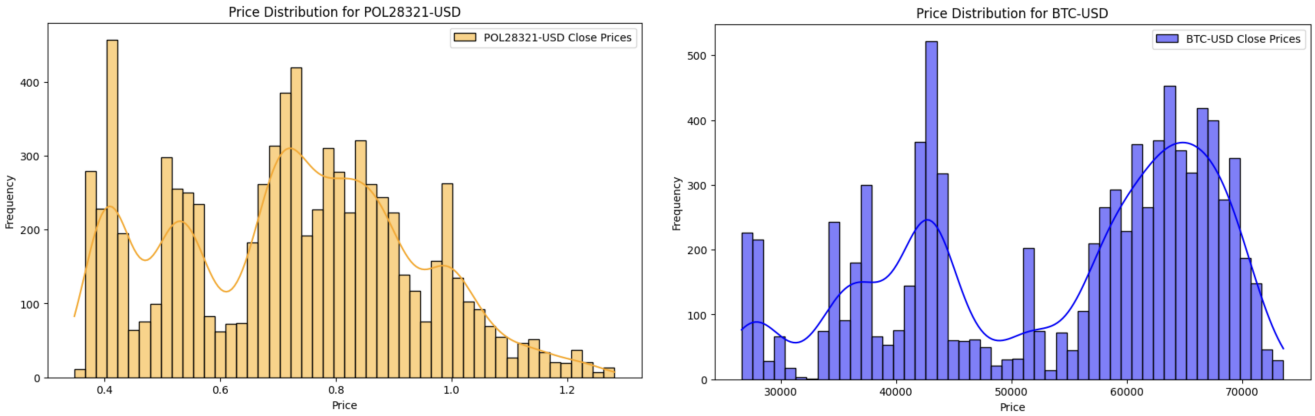
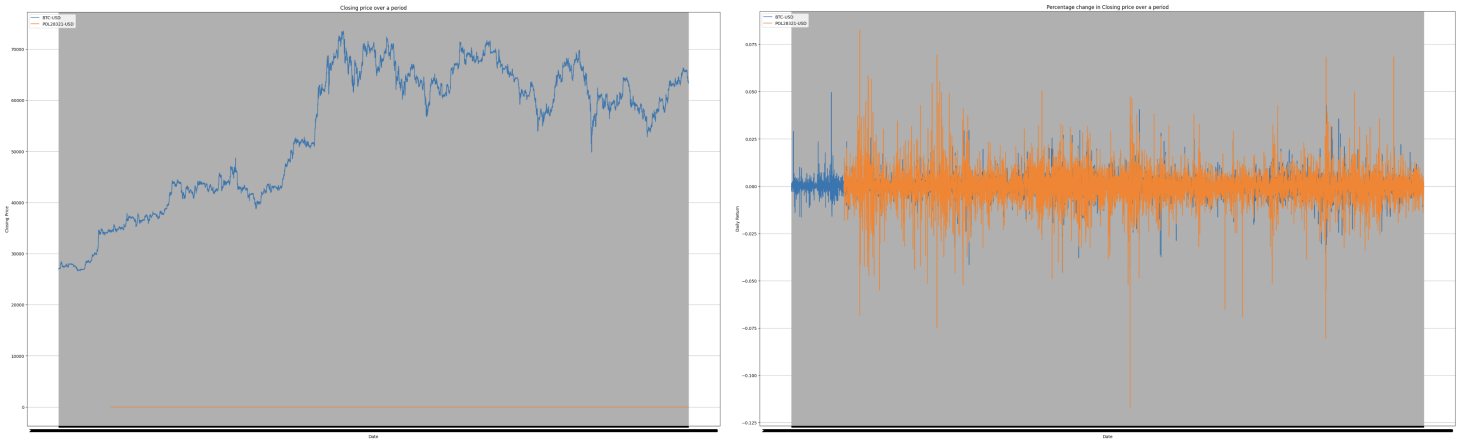


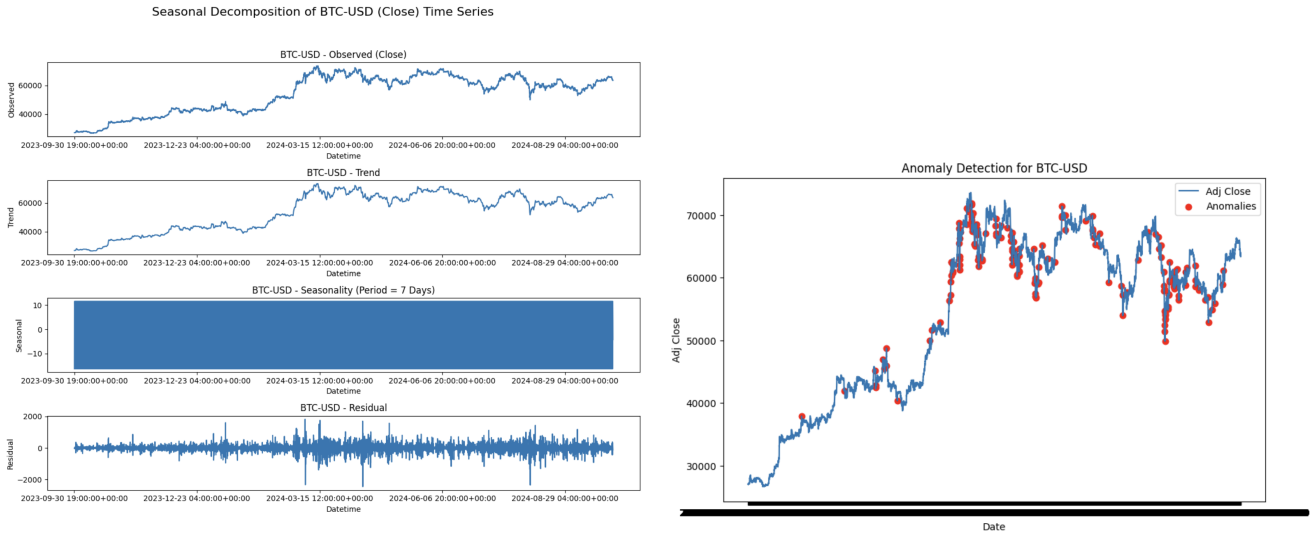
Figure 3: Price Distribution of two different cryptocurrencies



(a) Comparison of closing prices between cryptocurrencies.

(b) Comparison of normalized closing prices between cryptocurrencies

Figure 4: Describing before and after normalization



(a) Seasonal Decomposition

(b) Anomalies in BTC-USD

Figure 5: Seasonality and Trend

2.3.4 Data preprocessing for LSTM

The steps performed to preprocess the cryptocurrency and sentiment data inorder to train the LSTM model are listed as follows:

Preprocessing features

- **Datetime**
  - *hour*: Extracted the hour values from the `Datetime` feature.
  - *day_of_week*: Extracted the day's number in the week (0 = Monday, 6 = Sunday).
  - *is_weekend*: Extract a binary feature to represent whether the current date is a weekend or not.
- **Volume**
  - *Volume_log*: In-order to normalize skewed information we used logarithmic transformation to transform the Volume column.
- **Computing 24 hours rolling values for sentiment based features**
  - For the following features we calculate the rolling values on a 24 hours window:
    - *full_text_sentiment_positive*
    - *full_text_sentiment_negative*
    - *full_text_sentiment_neutral*
  - the extracted features are stored as *full_text_sentiment_*_rolling*.
- **Lag calculation for Close price**
  - Extracted lagged values of the *close* price for i=1, 2, 3 time steps:
    - *Close_lag1*, *Close_lag2*, *Close_lag3*.

Target variable

- fixed the target variable as the *Close* price.

Normalization

- predictors and the target were scaled to the range $[0, 1]$ using *MinMaxScaler*.
- predictors included in normalization:
  - *Open*, *High*, *Low*, *Volume_log*.
  - *full_text_sentiment_positive_rolling*.
  - *full_text_sentiment_negative_rolling*.
  - *full_text_sentiment_neutral_rolling*.
- Target feature: *Close*.

Preparing sequential data for LSTM

- We transformed the dataset into sequences for LSTM input:
  - Each sequence contains *sequence_length* (default = 24) time steps.
  - The target is fixed as the *Close* price of crypto-currency by the end of sequence of 24 hours.
- Resulting shapes:
  - **X**: (*number of samples*, *sequence_length*, *number of features*).
  - **y**: (*number of samples*,).

Data Cleaning

- Rows containing *NaN* values are removed .

Train-Test Split

- We have split our dataset into training(80%) and testing (20%).

3 Methodology

3.1 Task 1: LSTM for Crpyto-currency price prediction

Algorithm: We used a Long Short-Term Memory (LSTM) neural network to predict cryptocurrency prices based on historical price movements and sentiment data. The LSTM model that we implemented is expected to capture temporal dependencies in sequential information therefore expecting it to work on time-series prediction tasks. By depending on feature engineering and extracting features such as rolling averages for sentiments, lagged prices, and log-transformed trading volumes, the model learns from both price trends and market sentiment to perform better for prediction accuracy based tasks. This methodology therefore makes it possible for a deeper understanding of the relationships between market changes and sentiment, therefore helping in prediction and decision-making.

Algorithm 1 LSTM-Based Cryptocurrency Price Prediction

Require: Dataset $\mathcal{D}$ with features $\{x_1, x_2, \ldots, x_m\}$ and target y.
Ensure: Predicted cryptocurrency prices $\hat{y}$.
Feature extraction:
Extract temporal features (e.g., *hour*, *day_of_week*, *is_weekend*).
Compute rolling sentiment averages and lagged price features.
Normalization:
Scale features and target to the range $[0, 1]$ using *MinMaxScaler*.
Sequential data Preparation:
Divide $\mathcal{D}$ into sequences of length T and corresponding targets.
Model Declaration:
Define an LSTM model with parameters: input size, hidden size, number of layers, and output size.
Training:
for each epoch in {1, 2, ..., N} **do**
  for each batch (X, y) in training data **do**
    Forward pass: Predict $\hat{y}$ using the LSTM model.
    Compute loss using Mean Squared Error (MSE).
    Backpropagate the error and update model parameters.
  end for
end for
Evaluation:
Use the trained model to predict prices $\hat{y}$ on the test set.
Denormalization and Visualization:
Inverse transform predictions and actual values.
Plot true vs. predicted prices for evaluation.

Steps:

1. **Data preparation and Feature extraction:**
   - Extracting temporal features like *hour*, *day_of_week*, and a binary feature *is_weekend*.
   - Extracting rolling averages for sentiment features: *full_text_sentiment_positive*, *full_text_sentiment_negative*, and *full_text_sentiment_neutral*.
   - Extracting lag features for the *Close* price with lags of 1, 2, and 3 time steps.

- Transformed *Volume* using lograthimic transformation inorder to normalize skewed data.

2. **Normalizing the Features**
   - Scaled all features and the target (*Close*) to the range $[0, 1]$ using *MinMaxScaler* using Sklearn library.
3. **Sequential data Preparation:**
   - generated sequences of length T (e.g., 24 time steps) from the normalized features.
   - every sequence has the feature matrix for T time steps as input and the corresponding *Close* price as the target.
4. **Splitting dataset into train-test**
   - seperated the sequences into training (80%) and testing (20%) sets.
5. **Model**
   - Define an LSTM model with specific parameters:
     - Input size: Number of features 7 in our case.
     - Hidden size: Number of LSTM units (e.g., 50).
     - Number of layers: Depth of the LSTM model (e.g., 2).
     - Output size: 1 which is basically the prediction for the *Close* price.
6. **Training the model**
   - train the model using data that was made into batches.
   - compute the loss using the Mean Squared Error (MSE) for every epoch.
   - store the loss value inorder to use in the future.
   - back-propagating the error and updating the model parameters using the Adam optimizer.
   - Repeat for a specified number of epochs (e.g., 50).
7. **Evaluation**
   - Use the trained model to predict prices on the test set.
   - Compare the predicted prices with the actual prices.
8. **re transformation and Visualization**
   - By inverse transforming the scaled predictions and actual values to their original scale.
   - We plot the true and predicted *Close* prices to evaluate the model's performance.

3.2 Task 2: Similarity Analysis on Cryptocurrencies

Algorithm: We used Agglomerative algorithm from Hierarchical clustering to group cryptocurrencies with similar price movements based on their pairwise correlation. This method identifies clusters of cryptocurrencies that behave similarly, helping to optimize portfolios and market analysis. The pseudo code for this task is as mentioned in Algorithm 2.
Steps:

1. **Prepare the data:**
   - Calculate the daily percentage change (daily returns) for the "Adj Close" price for each cryptocurrency and combine as a single csv file.
   - Pivot the combined CSV file to create a **time-series matrix**, where rows represent time (Datetime) and columns represent cryptocurrency tickers, with values being the daily returns.

Algorithm 2 Agglomerative Clustering Algorithm for Cryptocurrency Similarity Analysis

Require: A time-series dataset for multiple cryptocurrencies, $D = \{x_1, x_2, \ldots, x_n\}$
Ensure: Clustering similar Cryptocurrences using Agglomerative algorithm

Data Preprocessing:
Calculate the daily percentage change (daily returns) for the 'Adj Close' of each cryptocurrency. This ensures that the data we are dealing with is Normalized
Construct a pivot table and handle the missing values found, by forward-filling and backward-filling the NaN values present in the dataset to ensure consistency.
Correlation and Distance Calculation:
Compute the pairwise Pearson correlation matrix, C, for measuring the similarity between cryptocurrencies.
Transform the correlation matrix into a distance matrix, D, using the formula $D = 1 - |C|$.
Clustering using Agglomerative algorithm
Initialize each data point as its own cluster.
while more than one cluster exists **do**
    Compute the pairwise distances between all existing clusters.
    Merge the two clusters with the smallest distance.
    Update the distance matrix to reflect the newly merged cluster.
    Repeat steps 11 and 12 until there is only one cluster left
end while
Visualization:
Generate a dendrogram and identify an appropriate threshold to form the meaningful clusters.
Assign the cluster labels based on the chosen threshold.
Output the cluster labels and save them to a CSV file for further analysis.
Visualize the clusters using a heatmap of the correlation matrix and time-series plots of cryptocurrencies in key clusters.

2. **Filling Missing Values:**

    - Since some of the tickers have recent start date in the Yahoo Finance tracker website, the constructed time-series matrix can have NaN values as shown in figure.6 below. For instance, from the below plot, the MEW30126-USD topped with more NaN values because the start date is 2024-03-26, which is very recent. Similarly with other currencies.
    - In-order to deal with them, we filled the missing values using forward and backward fill, which fills in the direction of the time series and ensures consistency in clustering.

3. **Compute Correlation Matrix:**

    - Next we calculated the correlation matrix in-order to measure the similarity between the daily returns of cryptocurrencies. Also, visualize the correlation using a heatmap as shown in figure.7.

4. **Calculate Distance Matrix:**

    - Since the correlation matrix defines similarity, we transformed it to a distance(dissimilarity) matrix using $\mathbf{Distance = 1 - |Correlation|}$ formula. In this, higher values indicate greater dissimilarity.

5. **Build Linkage Matrix:**

    - We now applied the agglomerative hierarchical clustering algorithm using the above contructed distance matrix with the help of correlations matrix.
    - We used "complete linkage" as the linkage to determine clusters. We choose this because this helps in avoiding loose clusters and maintain meaningful groupings.

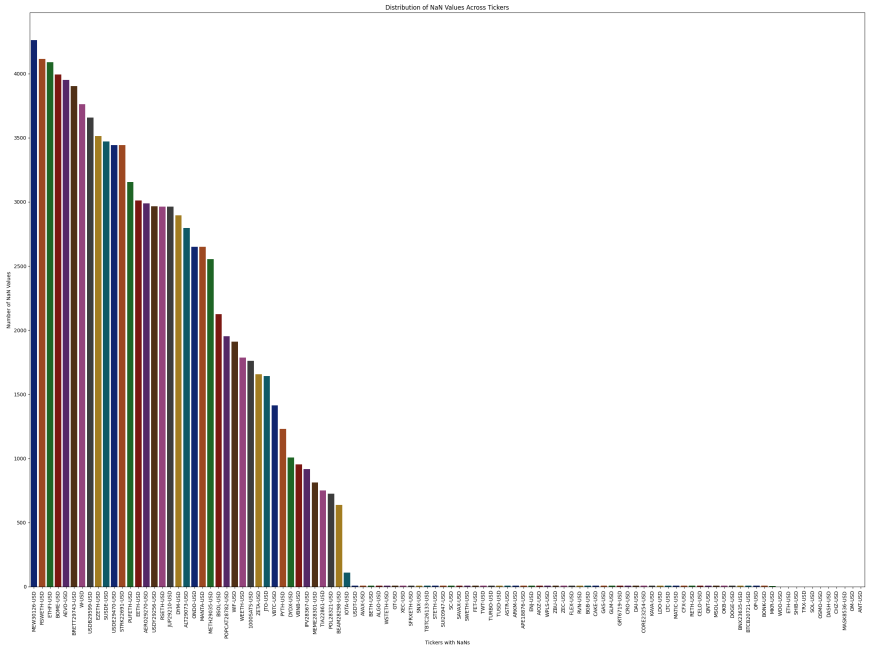


Figure 6: NaN value distribution in constructed time-series matrix

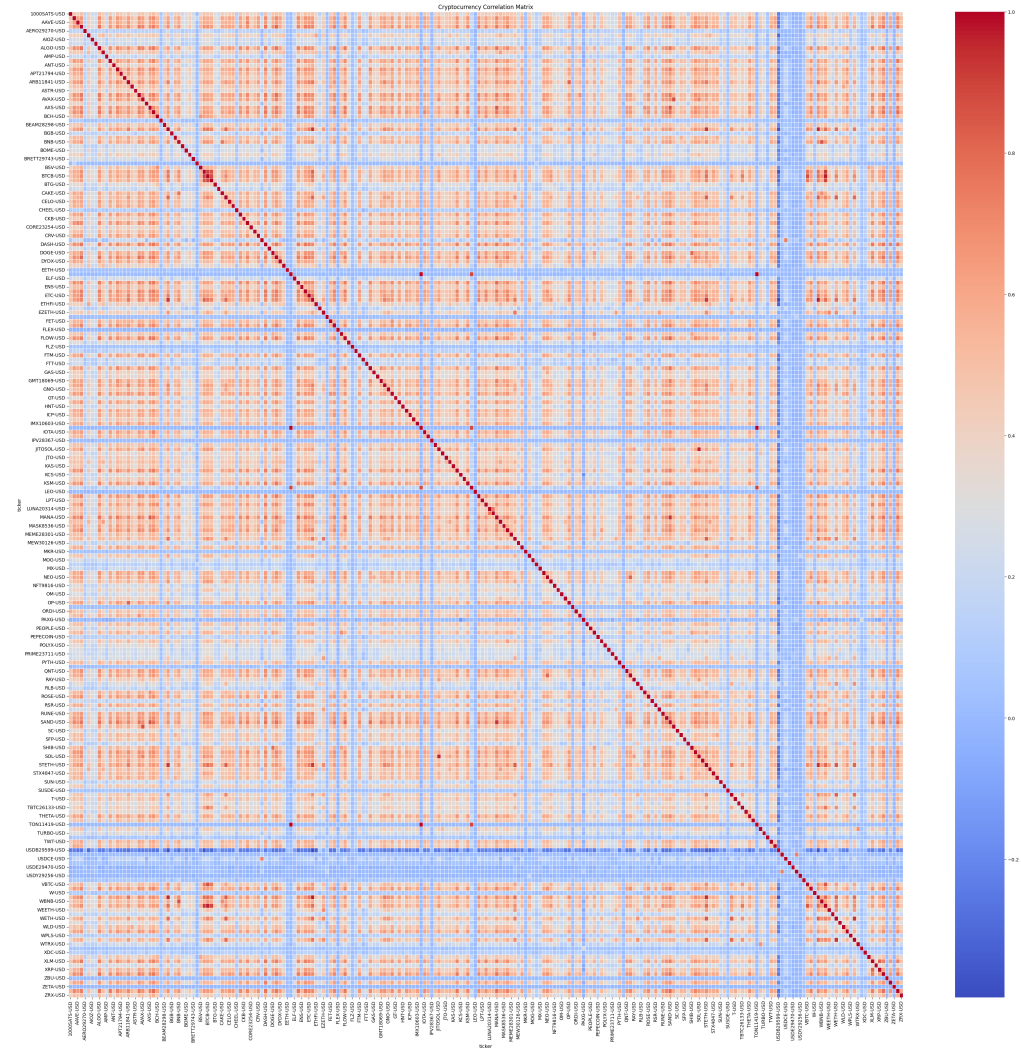


Figure 7: Correlation matrix of all 231 cryptocurrencies

6. **Visualize with Dendrogram:**

   - We plotted the hierarchical structure using a dendrogram, showing how cryptocurrencies cluster together. Then draw a horizontal threshold line on the dendrogram inorder to divide the clusters more meaningfully, as shown in figure.8.

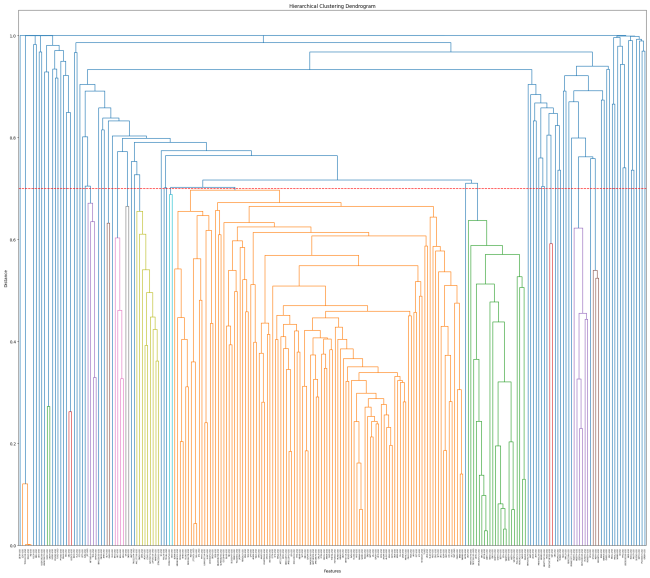


Figure 8: Dendrogram with a threshold for clustering

7. **Cluster and Visualize with Heatmap:**

   - Finally, we generated labels based on the threshold in dendrogram. Based on the labels we got, we generated a heatmap, as shown in figure.9, grouping the similarly moving cryptocurrencies. Also saved a CSV file showing the clustered currencies.

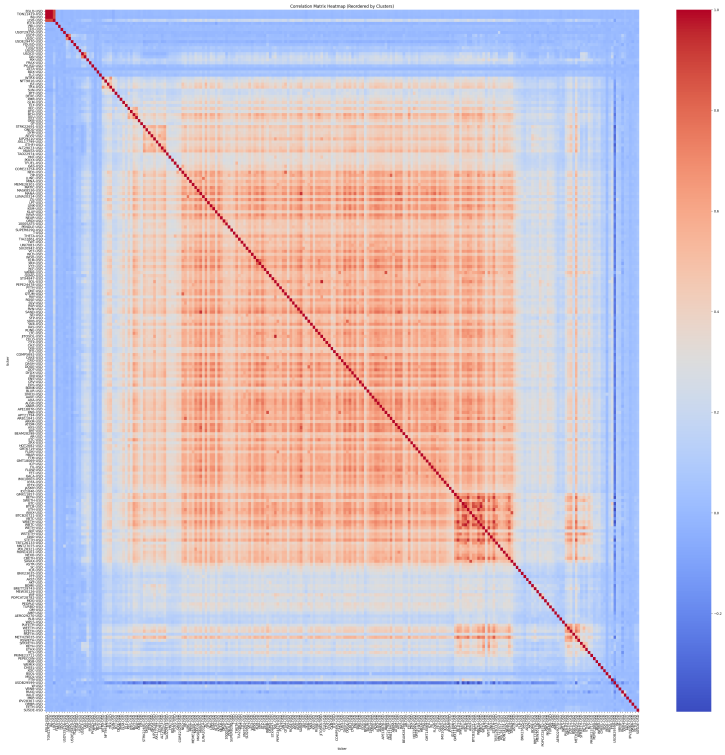


Figure 9: Heatmap showing the clustered currencies

4 Results

4.1 Task1

LSTM has been trained on sentiment and price based features for top 10 tickers.The cryptocurrencies corresponding to the tickers are Bitcoin, Ethereum, Binance Coin, Tether, Cardano, Ripple, USD Coin, Dogecoin, Solana, Polkadot.

Epoch vs Training Loss:

The training loss plots for the 10 cryptocurrencies (e.g., ADA-USD, BNB-USD, BTC-USD) show the Mean Squared Error (MSE) for all 50 training epochs. The steep decline in the loss during the starting epochs concludes that the LSTM model is rapidly learning the hidden temporal patterns that are present in cryptocurrency price and sentiment data. As the training moves forward, the loss gets stabilized near zero, reflecting convergence and inferring that the model has sufficiently captured the relationship between the input features and the target variable.

These results demonstrate the effectiveness of the LSTM architecture in modeling sequential data and its ability to adapt to different cryptocurrencies by leveraging engineered features like rolling sentiment score averages, lagged price values, and logarithmic volume transformations.

Training Loss Analysis:

Observations

The following loss plots illustrate the training dynamics for various cryptocurrency tickers.

Figure 16: Training Loss per Epoch - SOL-USD

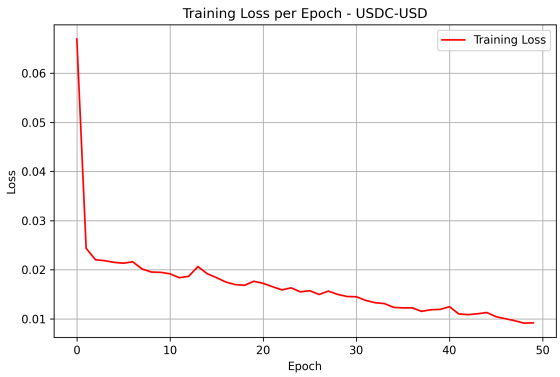


Figure 17: Training Loss per Epoch - USDC-USD

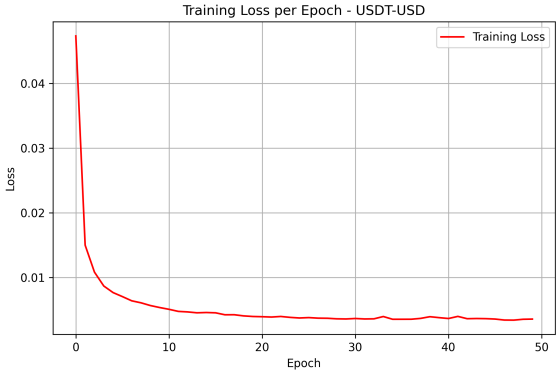


Figure 18: Training Loss per Epoch - USDT-USD

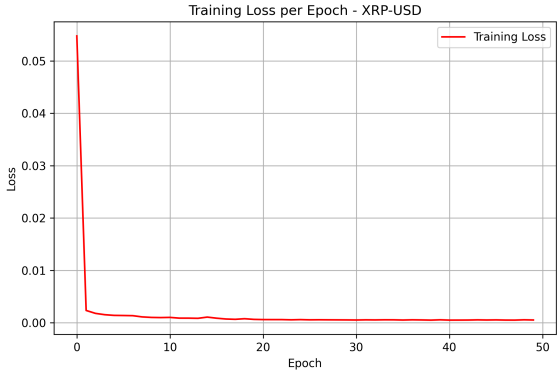


Figure 19: Training Loss per Epoch - XRP-USD

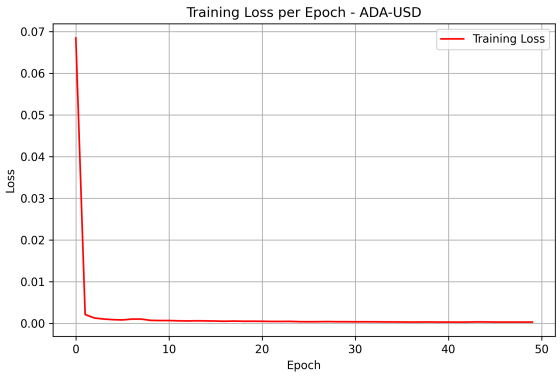


Figure 10: Training Loss per Epoch - ADA-USD

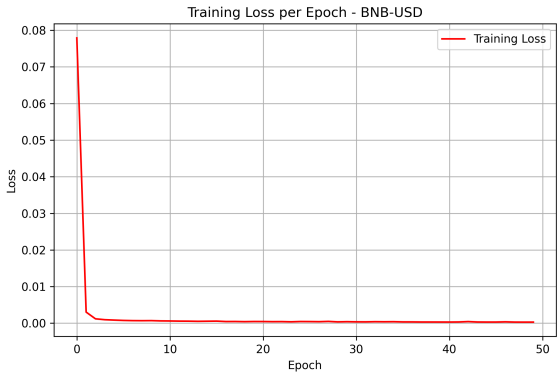


Figure 11: Training Loss per Epoch - BNB-USD

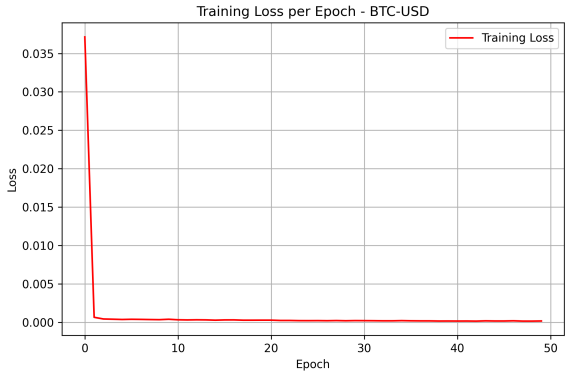


Figure 12: Training Loss per Epoch - BTC-USD

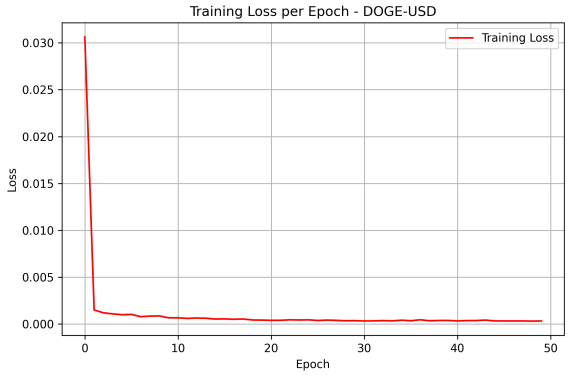


Figure 13: Training Loss per Epoch - DOGE-USD

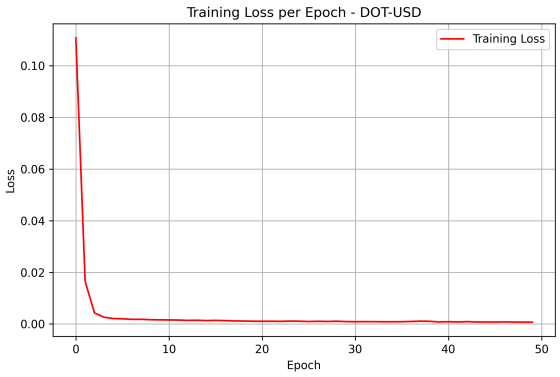


Figure 14: Training Loss per Epoch - DOT-USD

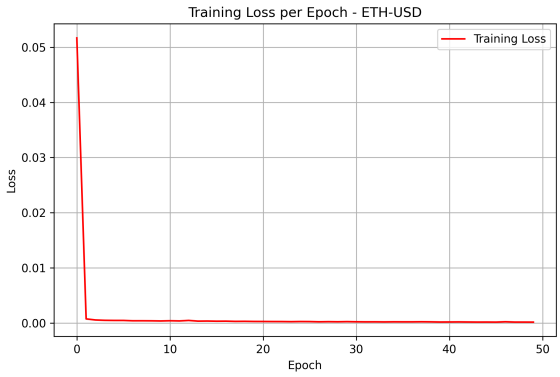


Figure 15: Training Loss per Epoch - ETH-USD

Interesting Observation: USDC-USD Training Loss:

While the training mean squared error loss for cryptocurrencies, such as BTC-USD, ADA-USD, and ETH-USD has converged rapidly within the first 15 epochs, however the plot for USDC-USD (Figure 20) follows a slightly different pattern. The loss function had lot of variance after 15th epoch and it stays the same till 50th epoch.

The explanation for this sort of behavior could be related to the nature of USDC which is a stable-coin designed to maintain a consistent value relative to the US dollar. The lack of significant sentiments and volatility in its price made it harder for the LSTM model to identify temporal patterns, leading to slower convergence and slight fluctuations during training. This also explains the fast pased and smooth convergence observed in more volatile cryptocurrencies, as shown in Figures 12 and 10.

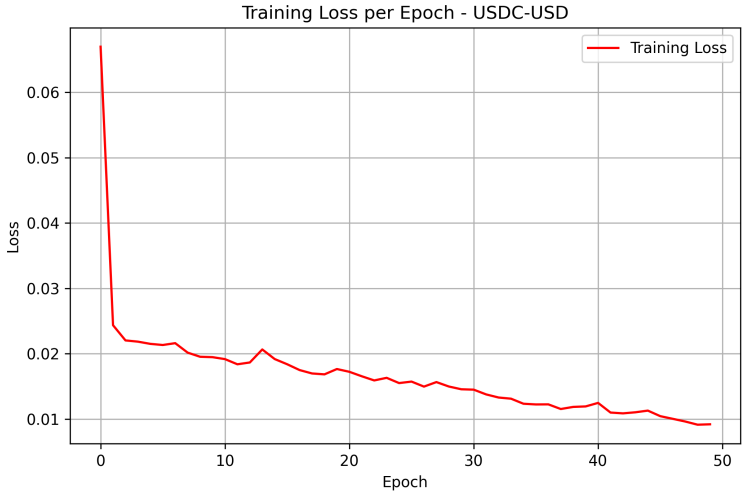


Figure 20: Training Loss per Epoch - USDC-USD

True Price , Predicted price vs Time:

The following plots illustrate the true vs. predicted prices for various cryptocurrency tickers:

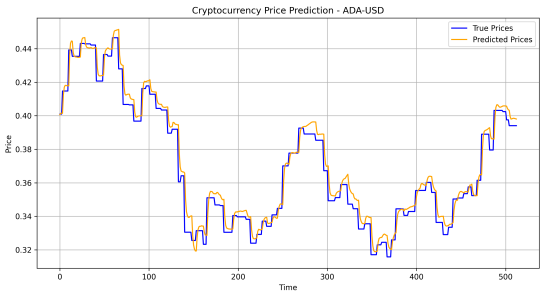


Figure 21: Cryptocurrency Price Prediction - ADA-USD

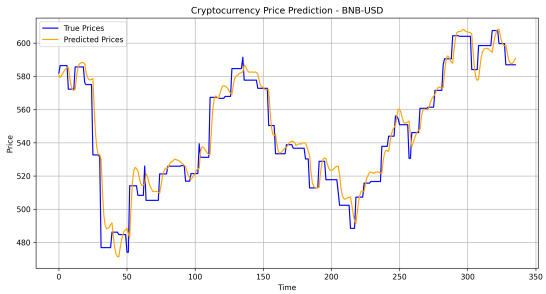


Figure 22: Cryptocurrency Price Prediction - BNB-USD

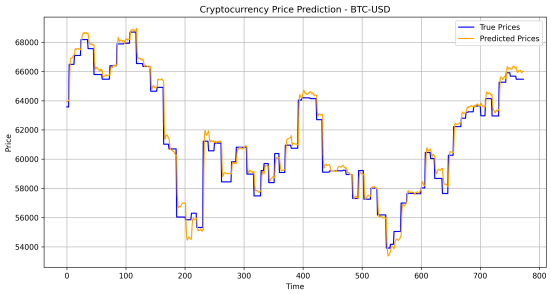


Figure 23: Cryptocurrency Price Prediction - BTC-USD

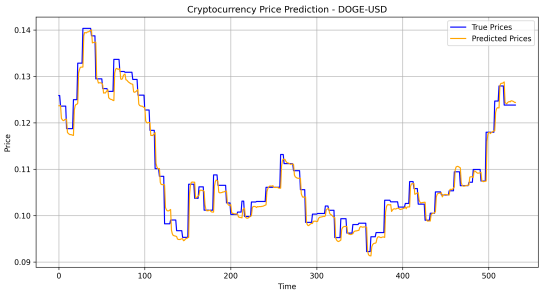


Figure 24: Cryptocurrency Price Prediction - DOGE-USD

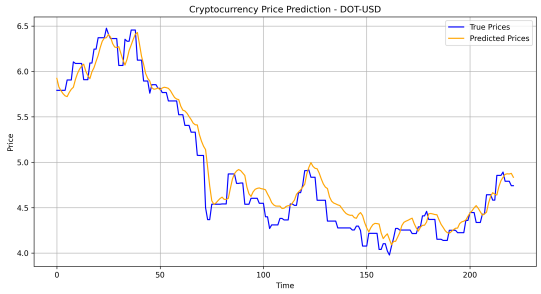


Figure 25: Cryptocurrency Price Prediction - DOT-USD

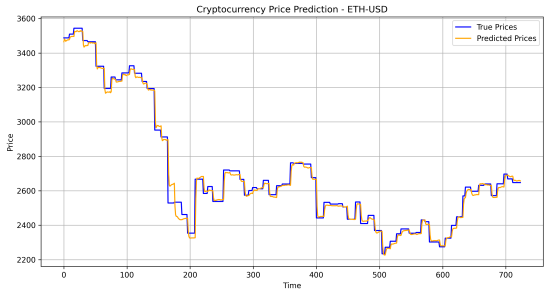


Figure 26: Cryptocurrency Price Prediction - ETH-USD

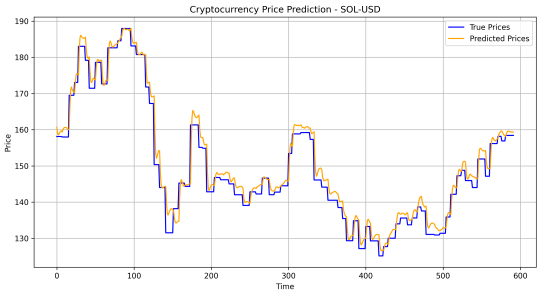


Figure 27: Cryptocurrency Price Prediction - SOL-USD

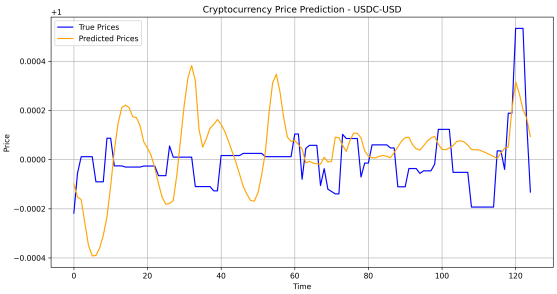


Figure 28: Cryptocurrency Price Prediction - USDC-USD

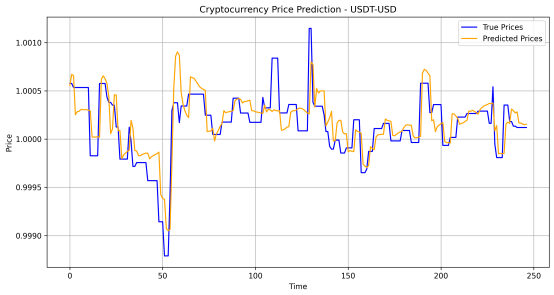


Figure 29: Cryptocurrency Price Prediction - USDT-USD

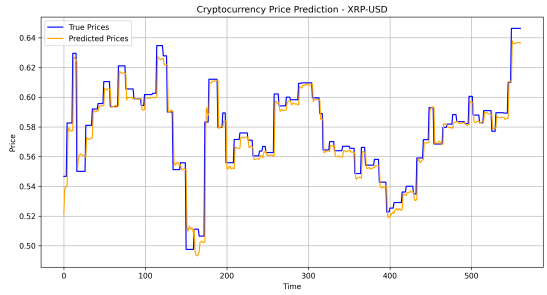


Figure 30: Cryptocurrency Price Prediction - XRP-USD

Some Observations from Cryptocurrency Price Prediction Plots:

1. **Accuracy**
   For most cryptocurrencies, the predicted prices (orange) are close to true prices (blue). This infers that the LSTM model was performing good in capturing the temporal dependencies in price movements. Especially, for **BTC-USD**, **ETH-USD**, and **ADA-USD**, the predictions are nearly not seperable from the true prices for most of the time series.

2. **Volatility vs Accuracy**
   BTC-USD and **BNB-USD** exhibit larger price variations (high volatility), show a slightly larger deviation between true and predicted prices during transitions. For example, in the **BNB-USD** plot, the predicted values are lagged slightly during hilly upward , downward trends, which explains the model not being efficient in capturing sudden market movements. Opposite to the previous scenario, stablecoins like **USDT-USD** and **USDC-USD** exhibit only slight deviations due to their estimatable price movements.

3. **Errors for sudden spikes**
   For cryptocurrencies such as **DOT-USD** and **SOL-USD**, the model predictions deviate noticeably during price spikes or drops . This behaviour be showcased to the LSTM model is being dependent on historical trends, which makes it really hard to predict surprising changes.

4. **Slight overfitting**
   In cryptocurrencies like **DOGE-USD** and **XRP-USD**, the predicted prices show minor oscillations that are not present in the true prices. This could indicate slight overfitting, where the model attempts to learn noise in the data instead of true trends.

4.2 Task2

The clustering identified 74 distinct clusters, each representing cryptocurrencies with behavioral patterns in price movements. We stored the clustering results in a CSV file, with the help of clustering labels as shown in the below table 5.

Clustered label	**Tickers**
1	['EGLD-USD', 'TON11419-USD', 'INJ-USD', 'LDO-USD']
54	['SFRXETH-USD', 'RETH-USD', 'ETHX-USD']
44	['WIF-USD', 'POPCAT28782-USD']

Table 5: some of the clustered tickers from the results CSV file

For instance, if we look at cluster 1 in the 5, the tickers ['EGLD-USD', 'TON11419-USD', 'INJ-USD', 'LDO-USD'] are clustered together, which can be also seen in the dendrogram below. The EGLD-USD is a native token of the Elrond blockchain platform, which is used to pay transaction fees on the Elrond network. TON11419-USD is also a native token of the Open Network blockchain, which is used to pay transaction fees on the TON network. INJ-USD is a utility token which powers the Injective Protocol, a decentralized exchange, which is similarly used to pay trading fees on the Injective exchange. LDO-USD is a governance token that represents ownership and voting rights in the Lido decentralized autonomous organization, and is used to vote on proposals and decisions related to the Lido protocol. But, if we look at cluster 54 in the 5 above, every ticker is a etherium based coin which makes sense on why they might be moving similarly to each other.

Although, all of these cluster-1 crypto currencies are not directly related to each other, they can be influenced by the overall market trends, investor sentiments, and regulatory developments, which is why we still see a correlation between them.

The plot, as shown in fig.31, captured the Daily return movement of the all the tickers form clutser 1 as seen in the above table.5. We can see that all of them are moving similar to each other in the plot 31.

The correlation matrix that shown in the figure.32b below reveals the correlation between those cluster 1 tickers from table 5. Since we are calculating the distance/dissimilarity with the help of this correlation matrix, it makes sense that these clusters are logically formed in the dendrogram.

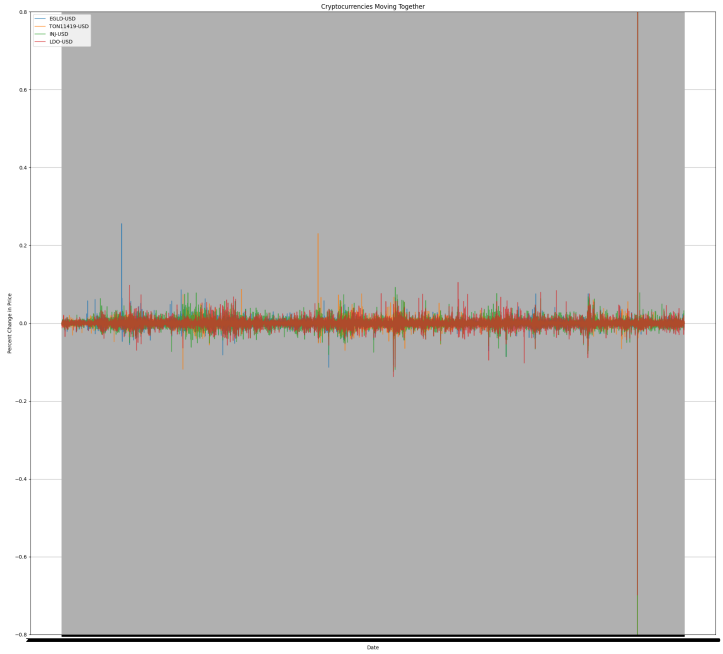
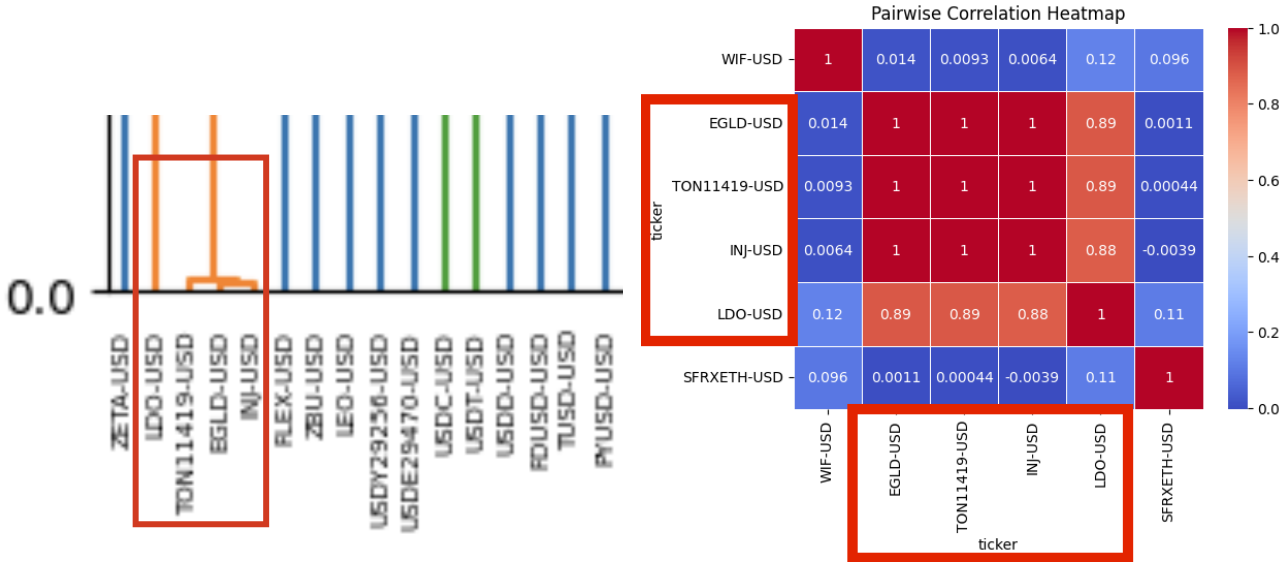


Figure 31: time series plot for cluster 1 tickers with respect to Daily returns



(a) Part of dendrogram showing the cluster with label 1

(b) Pairwise heatmap showing cluster 1

Figure 32: cluster 1 representation in dendrogram and heatmap with respect to Daily returns

5 Insights

5.1 Related to Price prediction task

- We were only ingesting 5 articles per day because of limitation in rate limits If given a chance to redo this and we have enough funds that supports fetching all the articles regarding a crypto-currencies that would be a game changer.
- We also wanted to fetch data from Social Media platforms like Twitter , Reddit , Facebook , instagram and also use LLM search feature to identify relevant articles but LLM search is still not available thorugh API's.
- We tried Reddit but rate limits hit us hard. We even tried to use Duck-Duck-Go search features using its

news API and we even wrote code to do an incremental load but it cannot extract historical data which was a big blocker for us to move forward.

- More work can be done in the area of feature engineerig and LSTM architecture of production grade implementations we ran our workflow on T500 Nvidia Mobile GPU but working with powerful GPU's can always reduce our time in extracting sentiments over an year time period for currencies and Text mining took a lot of our time while working on this project.
- For volatile cryptocurrencies like **BNB-USD** and **BTC-USD**, incorporating additional features such as trading volume spikes or extensive market sentiment could improve the model's performance in predicting spikes or dropping price trends.
- For stablecoins like **USDC-USD** and **USDT-USD**, We thought of applying regularization techniques in order to deal with overfitting to minor fluctuations in price data.

5.2 Related to Clustering Task

We couldn't use k-means as it required number of clusters at the very beginning, it requires euclidean distance, and its performance is poor with the overlapping clusters as seen from the below figure showing PCA visualization on the data we had based on Dialy returns. Therefore, we choose Complete Hierarchical clustering, which simple and yet efficient with the overlapping clusters. Some of the clusters such as cluster 44(WIF-USD, POPCAT28782-USD) and cluster 45(MOG-USD) are far from the rest in the PCA plot. These outliers may represent unique market behaviors from the rest. The coins WIF-USD, POPCAT28782-USD, and MOG-USD are pretty new in the market and based on internet memes, which might be the factors that separated these from the rest of the coins. This overlapping PCA space, also suggests that common influences and investor sentiments are driving their movements.

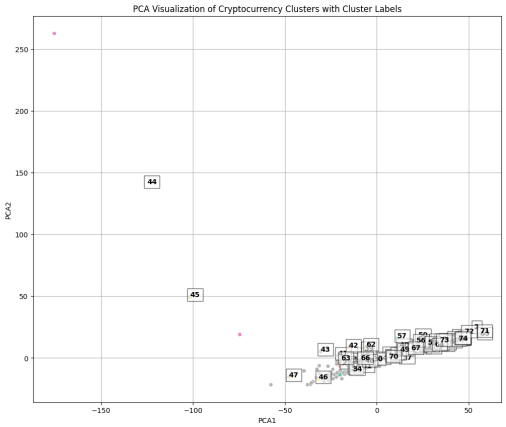


Figure 33: The PCA visualization clusters

If we have a chance of doing this all over again, we would probably include more data and also consider other features as well to see if the clustering is different everytime. We would also like to test multiple clustering methods like DBSCAN or Gaussian Mixture Models (GMM). Experiment with other similarity measures like dynamic time warping (DTW). The most important part is, we would spend more time validating data quality before clustering to minimize noise.

6 Conclusion and Contributions

Based on our analysis, we saw that the public sentiments and their investment behaviour is a mojor drive force for the cryptocurrency market. The datamining techniques allowed us to have a deeper view into the market behaviour. This will not only help the current investors but also encourage others to play with crypto market for strict educational purposes .

Prasanth Reddy Guvvala:

- Worked on Crypto price data fetching and persisting
- Worked on Task 1
- Tried different approaches like using Duck-Duck-Search API to fetch news to perform a incremental load which didn't suffice our use case
- also tried to implement a data pipe line based on Reddit
- Finally worked on building a data pipe line that fetches 5 news articles per ticker per day in a given date range
- Implemented functions and methods to scrape the body of the urls given by API
- The pipeline also included using FinBert to generate sentiment probabilities for long texts
- Finally Implemented LSTM that predicts crypto price by taking the real features and also generated/extracted features into account.
- Wrote code to generate plots for Loss vs Epoch and also Time vs Price
- Implemented methods to generate seasonality trend analysis graphs
- worked on writing report for Task 2 based content
- Participated in discussion atleast 4 hours a week regarding the project
- my total contributions total to a percentage score of 50%.

Rahul Payeli:

- Worked on data preprocessing and Normalization.
- Worked on task 2 by testing and implementing clustering tasks like k-means and Hierarchiel to pick the best performing clustering task.
- Also performed elbow method to check how it could perform on the real time data we had.
- Wrote 50% of the final report and proposal.
- Tested some scraping API tools like Twitter API and Newyork times API for collecting unstructed data to perfrom sentiment analysis.
- Tested some pre-trained bert models on our data to check which could work better for task 1.
- Implemented Agglomerative algorithm for finding similarly moving crypto currenccies.
- Wrote code to generate plots, heatmaps, and dendrogram to visually represent the data and output.
- Applied PCA to view the clustering in 2-D plots.
- Participated in the weekly meetings to discuss on the project deliverables and worked on them as per the plan.
- My contribution to the whole project is 50% of the total tasks.

References

[1] H. Dwivedi, "Cryptocurrency Sentiment Analysis using Bidirectional Transformation," in *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, 2023, pp. 140-142. doi: 10.1109/ICSMDI57622.2023.00032.

[2] S. Oikonomopoulos, K. Tzafilkou, D. Karapiperis, and V. Verykios, "Cryptocurrency Price Prediction using Social Media Sentiment Analysis," in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Corfu, Greece, 2022, pp. 1-8. doi: 10.1109/IISA56318.2022.9904351.

[3] M. H. Bin Mohd Sabri, A. Muneer, and S. M. Taib, "Cryptocurrency Price Prediction using Long Short-Term Memory and Twitter Sentiment Analysis," in *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, Pune, India, 2022, pp. 1-6. doi: 10.1109/ICCUBEA54992.2022.10011090.

[4] A. Scalia, G. Bartholdi, and L. Ferrucci, "Cryptocurrency Price Prediction and Trading Strategies Using Machine Learning and Sentiment Analysis," *arXiv preprint*, arXiv:2206.03386, 2022. Available at: https://arxiv.org/pdf/2206.03386.

[5] Yahoo Finance API, *Yahoo Finance Cryptocurrency Markets*. Available at: https://finance.yahoo.com/markets/crypto/all/. Accessed: December 9, 2024.

[6] Reddit API, *Reddit API Documentation*. Available at: https://www.reddit.com/dev/api/. Accessed: December 9, 2024.

[7] SciPy Documentation, *Pearson Correlation Coefficient*. Available at: https://docs.scipy.org/doc/scipy/reference/gener
Accessed: December 9, 2024.

[8] cryptonews-api Documentation . Available at: https://cryptonews-api.com/documentation. Accessed: December 9, 2024.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997. doi: 10.1162/neco.1997.9.8.1735.

[10] Michael Scott Brown and Micheal Pelosi, "Moving Averages Trading Method Applied to Cryptocurrencies," *ResearchGate*, 2019. Available at: https://www.researchgate.net/publication/330760754_MOVING_AVERAGES_TRADING_METHOD_APPLIED_TO_

[11] Yufeng Zhao, Haiying Che, "SkIn: Skimming-Intensive Long-Text Classification Using BERT for Medical Corpus" *arXiv preprint*, arXiv:2209.05741, 2022. Available at: https://arxiv.org/abs/2209.05741.

[12] DuckDuckGo Search Library, *Python Package Index (PyPI)*, 2024. Available at: https://pypi.org/project/duckduckgo-search/. Accessed: December 9, 2024.

[13] PyTorch Team, "PyTorch: An Open Source Machine Learning Framework," *PyTorch.org*, 2024. Available at: https://pytorch.org/. Accessed: December 9, 2024.

[14] Scikit-learn Developers, "sklearn.cluster.AgglomerativeClustering," *Scikit-learn Documentation*, 2024. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.cluster.AgglomerativeClustering.html. Accessed: December 9, 2024.