COSE474-2024F: Final Project Proposal “Bitcoin Price Prediction Using Sentiment Analysis with FinBERT: A Comparison of LSTM and LSTM-RF Models”

Muhammad Nazim Fu

1. Introduction

Bitcoin, the world’s first and most prominent cryptocurrency, is known for its decentralized nature and high volatility. Predicting its price is challenging yet crucial for investors and researchers. Since market sentiment significantly impacts Bitcoin prices, incorporating sentiment analysis into predictive models has proven valuable.

Machine learning has become a powerful tool for financial forecasting, with Long Short-Term Memory (LSTM) networks excelling in time-series prediction by capturing temporal dependencies and sequential patterns. However, standalone LSTM models can struggle with the complex, non-linear dynamics inherent in cryptocurrency markets. To address this, hybrid models such as the integration of LSTM with Random Forest (RF) have been proposed to improve predictive accuracy and robustness.

In this project, FinBERT, a pre-trained financial sentiment analysis model, was used to extract sentiment scores from Bitcoin-related tweets. These sentiment scores were combined with historical Bitcoin price data and engineered features such as the Relative Strength Index (RSI) and Simple Moving Average (SMA). This enriched feature set was then used to train both the standalone LSTM model and the hybrid LSTM-RF model. The objective is to evaluate their performance in predicting Bitcoin prices, identify their strengths and weaknesses, and contribute insights into hybrid modeling for cryptocurrency forecasting.

The findings presented in this report demonstrate the value of incorporating sentiment analysis and hybrid machine learning approaches to enhance the accuracy and reliability of financial time-series predictions.

2. Problem definition & challenges

Bitcoin’s price is highly volatile, influenced by market sentiment, economic events, and historical trends. Accurately predicting its price requires models that can handle complex, non-linear relationships. While LSTM networks excel at sequential data modeling, they often struggle with noise and overfitting. This study explores the effectiveness of combining LSTM with Random Forest to improve prediction accuracy and robustness, comparing the standalone LSTM model with the hybrid approach.

Challenges:

- High Volatility: Sudden price changes make predictions difficult.
- Overfitting in LSTM Models: LSTM models, while powerful, are prone to overfitting, especially when the dataset is small or lacks diversity.
- Hybrid Model Complexity: Combining LSTM with Random Forest introduces additional computational and implementation challenges. Aligning the outputs of LSTM with the inputs required by Random Forest requires careful feature engineering and data transformation.

This study focuses on building and comparing models to see if a combined approach can better predict Bitcoin prices and provide more reliable results for users.

3. Concise Description of Contribution

This study introduces a hybrid approach combining Long Short-Term Memory (LSTM) networks with Random Forest (RF) for Bitcoin price prediction based on the sentiment. The contribution lies in evaluating and comparing the performance of standalone LSTM models against the hybrid LSTM-RF model, aiming to improve prediction accuracy and robustness in the volatile cryptocurrency market. By integrating sequential data processing (LSTM) with ensemble learning (RF), the research offers insights into enhancing predictive models for financial time-series forecasting, with potential applications in cryptocurrency market analysis.

4. Methods

1. Significance/Novelty:
The novelty lies in combining deep learning with ensemble learning. LSTM models are adept at modeling sequential

data but may struggle with noise and non-linearities. Random Forest complements LSTM by refining predictions through its robustness to noise and ability to model complex relationships.

This hybrid approach addresses the limitations of standalone models, offering a potentially more accurate and reliable forecasting system.

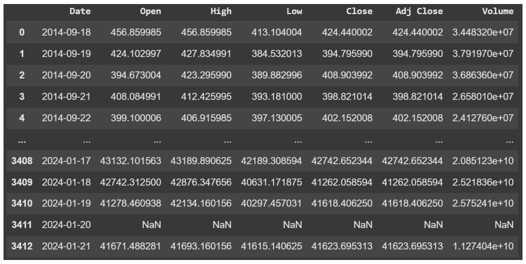
3. Reproducibility:
The methodology for this project involves several key steps to ensure reproducibility and effectiveness. Historical Bitcoin prices (open, high, low, close) were collected, along with sentiment scores derived using FinBERT from Bitcoin-related tweets. The price data was normalized using MinMaxScaler, sequences were generated for time-series modeling, and the dataset was split into training and testing sets. To enhance predictive accuracy, additional features such as the Relative Strength Index (RSI) and Simple Moving Average (SMA) were calculated and included in the input data.

The LSTM model was trained with 50 units, a batch size of 32, and 50 epochs to predict Bitcoin prices. Following this, an LSTM-Random Forest (LSTM-RF) hybrid model was developed, where LSTM predictions were used as input features for the Random Forest model to refine predictions by learning residual errors. Model performance was evaluated and compared using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), demonstrating the robustness of the proposed approach.

5. Datasets

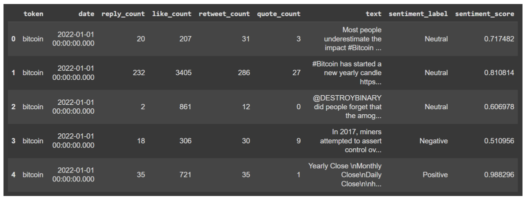
This project uses two datasets for Bitcoin price prediction:

1. Tweets Dataset (`tweets.csv`) This dataset contains tweets related to Bitcoin and their sentiment scores, sourced from social media platforms. It includes fields such as:
   - **Tweet ID**: Unique identifier for each tweet.
   - **Timestamp**: Date and time of the tweet.
   - **Tweet Text**: The content of the tweet.
   - **Sentiment Score**: A numerical value representing the tweet's sentiment (positive, negative, or neutral).
2. Bitcoin Price History Dataset (`BTC-USD.csv`) This dataset provides historical Bitcoin price data from 2014 to 2024. It includes daily details such as:
   - **Date**: Date of the record.
   - **Open, High, Low, Close**: Bitcoin's daily price metrics.
   - **Volume**: Total Bitcoin traded on that day.
   - **Adj Close**: Adjusted closing price, accounting for factors like stock splits.



	Date	Open	High	Low	Close	Adj Close	Volume
0	2014-09-18	456.859985	456.859985	413.104004	424.440002	424.440002	3.448320e+07
1	2014-09-19	424.102997	427.834991	384.532013	394.795990	394.795990	3.791970e+07
2	2014-09-20	394.673004	423.295990	389.882996	408.903992	408.903992	3.686360e+07
3	2014-09-21	408.084991	412.425995	393.181000	398.821014	398.821014	2.658010e+07
4	2014-09-22	399.100006	406.915985	397.130005	402.152008	402.152008	2.412760e+07
...	...	...	...	...	...	...	...
3408	2024-01-17	43132.101563	43189.890625	42189.308594	42742.652344	42742.652344	2.085123e+10
3409	2024-01-18	42742.312500	42876.347656	40631.171875	41262.058594	41262.058594	2.521836e+10
3410	2024-01-19	41278.460938	42134.160156	40297.457031	41618.406250	41618.406250	2.575241e+10
3411	2024-01-20	NaN	NaN	NaN	NaN	NaN	NaN
3412	2024-01-21	41671.488281	41693.160156	41615.140625	41623.695313	41623.695313	1.127404e+10

Figure 1. Tweets Dataset.



	token	date	reply_count	like_count	retweet_count	quote_count	text	sentiment_label	sentiment_score
0	bitcoin	2022-01-01 00:00:00.000	20	207	31	3	Most people underestimate the impact #Bitcoin ...	Neutral	0.717482
1	bitcoin	2022-01-01 00:00:00.000	232	3405	286	27	#Bitcoin has started a new yearly candle https...	Neutral	0.810814
2	bitcoin	2022-01-01 00:00:00.000	2	861	12	0	@DESTROYBINARY did people forget that the amog...	Neutral	0.606978
3	bitcoin	2022-01-01 00:00:00.000	18	306	30	9	In 2017, miners attempted to assert control ov...	Negative	0.510956
4	bitcoin	2022-01-01 00:00:00.000	35	721	35	1	Yearly Close \nMonthly Close\nDaily Close\n\nh...	Positive	0.988296

Figure 2. Bitcoin History Price Dataset.

Both datasets are synchronized by timestamp, allowing sentiment data and price information to be used together for training the predictive models.

6. Computer Resources

Platform: Google Colab Pro **Operating System**: Managed by Google Colab Pro Programming Language: Python 3.7 **Libraries and Frameworks**: Pandas, NumPy, scikit-learn, TensorFlow **Machine Learning Framework**: scikit-learn (for Random Forest) **Deep Learning Framework**: TensorFlow (for LSTM)

7. Experimental Design and Setup

7.1. Data Collection and Preprocessing

7.1.1. Data Sources

The datasets used for this project are sourced from Kaggle:

- **Tweets Dataset (`tweets.csv`)**: Contains Bitcoin-related tweets with sentiment scores.
- **Bitcoin Price History Dataset (`BTC-USD.csv`)**: Provides historical Bitcoin prices from 2014 to 2024.

7.1.2. Data Alignment

The datasets are aligned based on timestamps to ensure that sentiment scores and Bitcoin price data are correctly synchronized.

7.1.3. Data Preprocessing

- **Normalization**: Both Bitcoin prices and sentiment scores are normalized using MinMaxScaler to scale the values between 0 and 1.
- **Sequence Generation**: The data is converted into sequences to capture the temporal dependencies. Each sequence consists of a fixed number of past time steps and their corresponding sentiment scores.
- **Splitting**: The dataset was divided into 80 for training and 20 for testing, maintaining chronological order to prevent data leakage.

7.2. Feature Engineering

To improve the model's predictive power, additional features were engineered alongside raw price data. **RSI (Relative Strength Index)** was calculated to capture market momentum and identify overbought or oversold conditions, while **SMA (Simple Moving Average)** helped smooth out price fluctuations and highlight overall trends. Sentiment scores, derived using **FinBERT**, a pre-trained financial sentiment analysis model, were also integrated to capture the market's emotional outlook from Bitcoin-related tweets. Together, these features enriched the input data, enhancing the model's ability to predict future Bitcoin price trends.

7.3. Model Development

7.3.1. Standalone LSTM Model

An LSTM model is designed to capture temporal dependencies and sequential patterns in the Bitcoin price data. It includes layers such as LSTM units, dropout layers, and Dense layers. The model is trained on the preprocessed Bitcoin price and sentiment data using the training set.

7.3.2. Hybrid LSTM-RF Model

- The LSTM model is used first to predict Bitcoin prices based on historical price and sentiment data.
- The predictions from the LSTM model are then used as input features for training a Random Forest model.
- The Random Forest model refines the LSTM predictions, helping to improve prediction accuracy.

7.4. Training Process Evaluation

7.4.1. Training the LSTM Model

The LSTM model is trained using the training dataset with the Mean Absolute Error (MAE) loss function and the Adam optimizer. The model is trained for 50 epochs with a batch size of 32. The validation set is used to monitor the performance during training and prevent overfitting.

7.4.2. Training the Random Forest Model

The Random Forest model is trained using the LSTM-generated predictions and any additional features, such as the sentiment scores, from the training set. The Random Forest model uses 100 estimators (trees) and is trained to improve the overall prediction accuracy.

7.4.3. Performance Metrics

The models are evaluated using two primary metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in the predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

- **Root Mean Squared Error (RMSE)**: Provides a measure of the model's prediction error, penalizing larger errors more heavily.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

7.5. Model Comparison and Analysis

7.5.1. Quantitative Results

The performance of the standalone LSTM model is compared to the hybrid LSTM-RF model. A lower MAE and RMSE indicate better prediction accuracy and model performance.

Model Performance Comparison

Model	MAE	RMSE
LSTM	0.0044	0.066
LSTM-RF	0.0003	0.017

Table 1. Comparison of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for LSTM and LSTM-RF models.

7.5.2. Qualitative Results

Visualizations, such as line plots comparing the predicted vs. actual Bitcoin prices, are used to assess how closely each model follows the actual price trend over time.

Based on the plot above, The **LSTM predictions (blue line)** follow the general trend of the actual Bitcoin prices but show notable deviations, especially around peaks and troughs, indicating challenges in capturing complex market dynamics. In contrast, the **LSTM-RF predictions (green line)** are smoother and align more closely with the actual

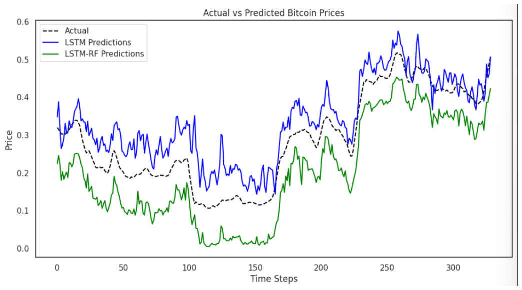


Figure 3. Actual vs Predicted BTC (LSTM LSTM-RF).

prices, demonstrating better trend-following capability and fewer large errors.

7.6. Discussion

7.6.1. Model Performance

The hybrid LSTM-RF model outperformed the standalone LSTM in capturing the non-linear dynamics of Bitcoin prices. While LSTM effectively learned temporal patterns, it struggled with market volatility. The addition of Random Forest produced smoother and more accurate predictions. However, both models faced challenges during extreme market fluctuations, highlighting the difficulty of predicting highly volatile markets.

7.6.2. Challenges and Limitations

Potential challenges, such as overfitting in the LSTM model despite regularization efforts like Dropout. The model also faced issues with noisy sentiment data, where social media sentiment doesn't always align with market trends.

7.6.3. Future Work

Based on the results, the experimental design may be expanded to incorporate additional features or other advanced models, such as Transformer networks or ensemble methods, to further improve prediction accuracy.

References