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 cryptocurrency, is known for its decentralized nature and high volatility, making price prediction both challenging and essential. Since market sentiment plays a crucial role in influencing Bitcoin prices, integrating sentiment analysis into predictive models offers valuable insights.

Machine learning, particularly Long Short-Term Memory (LSTM) networks, excels in time-series forecasting by capturing sequential patterns. However, standalone LSTM models often struggle with the complex, non-linear dynamics of cryptocurrency markets. To address this, hybrid models, such as combining LSTM with Random Forest (RF), have been explored to improve accuracy and robustness.

In this project, FinBERT, a pre-trained financial sentiment model, was used to extract sentiment scores from Bitcoin-related tweets. These scores, along with historical price data and engineered features like the Relative Strength Index (RSI) and Simple Moving Average (SMA), were used to train both an LSTM model and a hybrid LSTM-RF model. The objective is to evaluate their performance in predicting Bitcoin prices and highlight the benefits of hybrid modeling for financial forecasting.

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

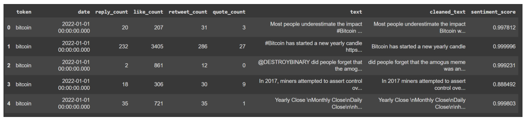
This study introduces a hybrid approach combining Long Short-Term Memory (LSTM) networks with Random Forest (RF) for Bitcoin price prediction based on the sentiment. The contribution lies in evaluating and comparing the performance of standalone LSTM models against the hybrid LSTM-RF model, aiming to 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 of this project lies in integrating deep learning with ensemble learning and leveraging financial sentiment analysis. LSTM models are excellent for capturing sequential data patterns but often struggle with noise and non-linearities in complex financial datasets. By combining LSTM with Random Forest, a robust ensemble learning method, predictions are refined, addressing the limitations of standalone models. Additionally, the use of FinBERT, a pre-trained financial sentiment analysis model, to extract sentiment scores from Bitcoin-related tweets further enriches the input features, capturing market sentiment more effec-

tively. This hybrid approach enhances the accuracy and reliability of Bitcoin price forecasting.

2. **Main Figure:**



	token	date	reply_count	like_count	retweet_count	quote_count	text	cleaned_text	sentiment_score
0	bitcoin	2022-01-01 00:00:00.000	20	207	31	3	Most people underestimate the impact #Bitcoin ...	Most people underestimate the impact Bitcoin w...	0.997812
1	bitcoin	2022-01-01 00:00:00.000	232	3405	286	27	#Bitcoin has started a new yearly candle https...	Bitcoin has started a new yearly candle	0.999996
2	bitcoin	2022-01-01 00:00:00.000	2	861	12	0	@DESTROYBINARY did people forget that the amog...	did people forget that the amogus meme was an...	0.999231
3	bitcoin	2022-01-01 00:00:00.000	18	306	30	9	In 2017, miners attempted to assert control ov...	In 2017 miners attempted to assert control ove...	0.888492
4	bitcoin	2022-01-01 00:00:00.000	35	721	35	1	Yearly Close \nMonthly Close\nDaily Close\n\nh...	Yearly Close \nMonthly Close\nDaily Close\n\nh...	0.999803

Figure 1. Sentiment Score predicted by FinBert model.

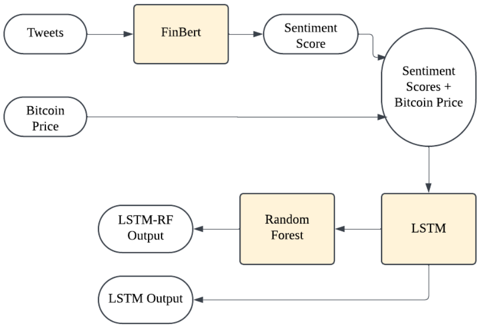


Figure 2. FinBert-LSTM-RF Model.

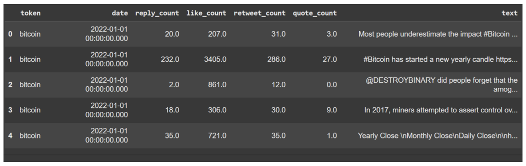
3. **Algorithm:**
Data Preprocessing: Normalize Bitcoin price data and sentiment scores using MinMaxScaler, generate time-series sequences, and split the data into training and testing sets. **Feature Engineering:** Add RSI for market momentum, SMA for smoothing trends, and integrate sentiment scores derived using FinBERT. **LSTM Model Building:** Build an LSTM model with input, LSTM, dropout, and Dense layers. Use the Adam optimizer, train with 50 units, a batch size of 32, and 50 epochs.**Hybrid LSTM-RF Model:** Use LSTM predictions as input for a Random Forest model to refine predictions by learning residual errors. **Model Evaluation:** Compare LSTM and LSTM-RF models using MAE and RMSE to assess accuracy and trend capture. **Results Analysis:** Highlight strengths, limitations, and the role of sentiment scores and features in improving accuracy. **Challenges and Considerations:**Address overfitting, noisy data, and non-linear market dynamics. Improve model generalization. **Report Findings:** Summarize LSTM vs. LSTM-RF performance, discuss implications, and suggest future improvements like advanced features or models.

5. Datasets

This project uses two datasets for Bitcoin price prediction:

The datasets used for this project are sourced from Kaggle:

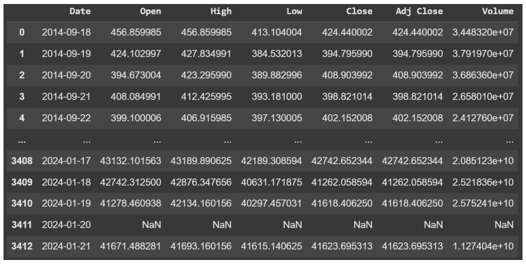
- **Tweets Dataset (`tweets.csv`)**: Contains Bitcoin-related tweets. This dataset will be used to capture market sentiment by predicting the sentiment scores from text using FinBert (pre-trained model) (Kapadnis, 2021) .



	token	date	reply_count	like_count	retweet_count	quote_count	text
0	bitcoin	2022-01-01 00:00:00.000	20.0	207.0	31.0	3.0	Most people underestimate the impact #Bitcoin ...
1	bitcoin	2022-01-01 00:00:00.000	232.0	3405.0	286.0	27.0	#Bitcoin has started a new yearly candle https...
2	bitcoin	2022-01-01 00:00:00.000	2.0	861.0	12.0	0.0	@DESTROYBINARY did people forget that the amog...
3	bitcoin	2022-01-01 00:00:00.000	18.0	306.0	30.0	9.0	In 2017, miners attempted to assert control ov...
4	bitcoin	2022-01-01 00:00:00.000	35.0	721.0	35.0	1.0	Yearly Close \nMonthly Close\nDaily Close\n\nh...

Figure 3. Tweets Dataset.

- **Bitcoin Price History Dataset (`BTC-USD.csv`)**: Provides historical Bitcoin prices from 2014 to 2024 (Kapturov, 2023).



	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 4. Bitcoin History Price Dataset.

Both datasets are synchronized by timestamp and combined, 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 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 study itlizies two main datasets: Tweets Dataset (`tweets.csv`) that contains Bitcoin-related tweets with features like the date, engagement metrics (reply, like, retweet, and quote counts), and the text content. Bitcoin Price History Dataset (`BTC-USD.csv`) : It provides the date, daily price metrics (open, high, low, close), adjusted close price, and trading volume of Bitcoin.

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

The data preprocessing involved **normalizing** Bitcoin prices and sentiment scores using `MinMaxScaler` to scale values between 0 and 1. The data was then **converted into sequences**, capturing temporal dependencies with past time steps and corresponding sentiment scores. Finally, the dataset was **split** into 80% for training and 20% for testing, preserving 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.

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

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

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

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	0.0055	0.074
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

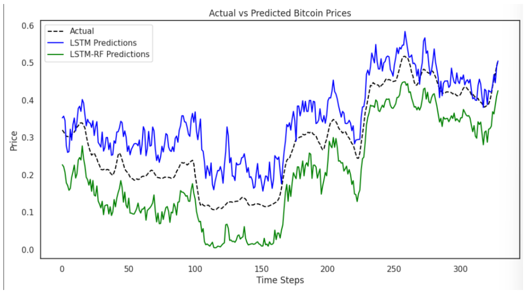


Figure 5. Actual vs Predicted BTC (LSTM LSTM-RF).

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 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

Kapadnis, S. Bitcoin tweets dataset. https://www.kaggle.com/datasets/sujaykapadnis/bitcoin-tweets, 2021. Accessed: 2024-12-10.

Kapturov, A. Bitcoin and ethereum prices from start to 2023. https://www.kaggle.com/datasets/kapturovalexander/bitcoin-and-ethereum-prices-from-start-to-2023, 2023. Accessed: 2024-12-10.