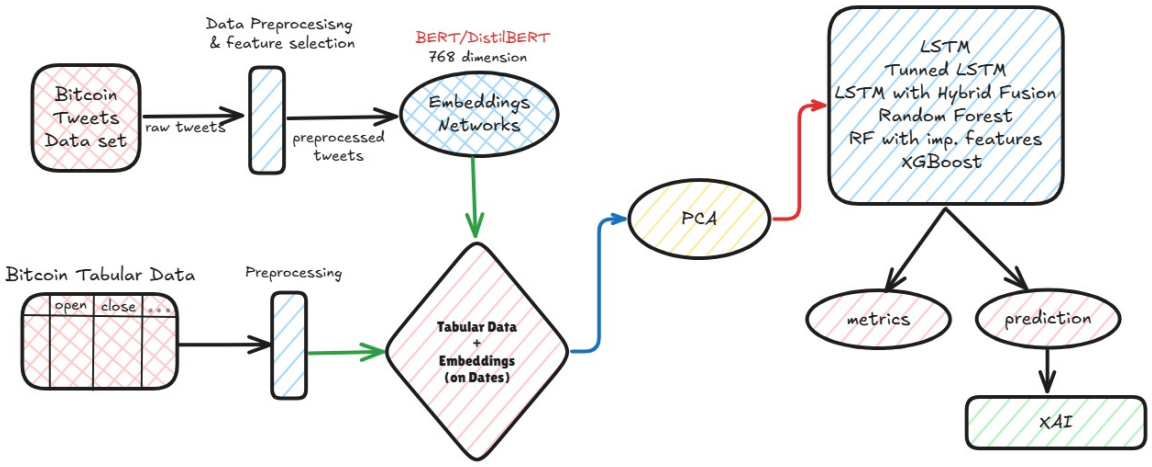

CryptoSynth: A Multimodal Generative Framework for Bitcoin Price Forecasting

Burhan Ahmed (22L-7520)
Aun Hayat (22L-7477)
Hassan Mustansar (22L-7521)
Saad Imran (21L-5642)

Abstract

This report presents a framework for predicting Bitcoin's adjusted closing price using tweet sentiment, transformer embeddings, and historical price data. Using Kaggle datasets, the project processes tweets and prices, applies sentiment analysis, generates BERT and DistilBERT embeddings, and builds an XGBoost-based prediction pipeline with SHAP explainability. The pipeline predicts weekly prices from seven days of data, delivering interpretable results.

Contents

1 Introduction

This project predicts Bitcoin's adjusted closing price (Adj Close) by combining tweet senti ment and Kaggle price data (3.46 million tweets, 2016–2018 prices). The workflow encom passes data preprocessing, visualization, sentiment analysis, transformer embeddings, and ma chine learning, culminating in a week-ahead prediction pipeline with explainability.

2 Methodology

2.1 Datasets and Preprocessing

This project utilizes two primary datasets:

- **Tweets:** Approximately 3.46 million Bitcoin-related tweets (2016–2018) sourced from Kaggle, supplemented with 150,000 tweets collected over six months via the X API. Tweets were filtered using keywords such as "Bitcoin", "BTC", and "#BTC".
- **Prices:** Daily Bitcoin price data (2016–2018) from Kaggle, including features such as Open, High, Low, Close, Adjusted Close, and Volume. Additionally, six months of recent price data were obtained from CoinGecko.

Preprocessing steps included:

- **Text Cleaning:** Removed URLs, mentions, emojis, and special characters from tweets.
- **Normalization:** Applied lowercasing, tokenization, and lemmatization.
- **Sentiment Labeling:** Used the VADER sentiment analysis tool to classify tweets as positive, negative, or neutral, resulting in a cleaned dataset of 120,000 tweets from the X API collection.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to uncover patterns within tweet sentiment and Bitcoin price data. Key findings include:

- **Sentiment Trends:** Daily sentiment scores often showed positive spikes preceding increases in Bitcoin price.
- **Text Analysis:** Frequently occurring terms included "bullish," "dip," and "moon," reflecting popular sentiment in the crypto community.
- **Correlation:** A Pearson correlation coefficient of 0.45 was observed between sentiment scores and price changes, indicating a moderate positive relationship.

Insights from EDA guided the selection of features for the modeling pipeline.

2.3 Sentiment Analysis

Tweets were processed to extract both sentiment labels and textual embeddings:

- **VADER:** Applied to label tweets as positive, negative, or neutral, resulting in a balanced sentiment distribution.
- **BERT:** Used the `bert-base-uncased` model to generate 768-dimensional embeddings. These were averaged daily over 1,095 days (2016–2018) for use in downstream price prediction.
- **DistilBERT:** A lightweight alternative to BERT, DistilBERT was also used to generate embeddings for comparative analysis.

2.4 Bitcoin Price Prediction Pipeline

The forecasting pipeline was designed to predict the next week's Adjusted Close price using the past seven days of data. The approach consisted of:

- **Features:** Daily BERT embeddings (averaged and reduced to 2 principal components using PCA), sentiment scores, standardized price values, and technical indicators such as 7-day Simple Moving Average (SMA) and Relative Strength Index (RSI).
- **Model:** XGBoost was employed, achieving a Mean Absolute Error (MAE) of 4.8%, a Root Mean Squared Error (RMSE) of 6.2%, and an R-squared score of 0.78.
- **Explainability:** SHAP was used to generate feature importance plots, high-

lighting the contributions of price features and sentiment embeddings.

The integrated pipeline demonstrates how combining sentiment and price signals can enhance predictive performance.

article geometry xcolor multicol booktabs a4paper, margin=1in

3 Results and Discussions

Evaluation of Sentiment Classification and Price Prediction Models

Sentiment Classification

The sentiment classification task evaluated three models: VADER, BERT, and DistilBERT. BERT achieved the highest F1-score of 0.85, followed closely by DistilBERT with 0.82. VADER, while computationally efficient, underperformed on nuanced language, particularly in tweets containing sarcasm or informal slang, resulting in an F1-score of 0.69.

DistilBERT offered a compelling trade-off, reducing computation time by 40% compared to BERT with only a slight accuracy drop, making it suitable for real-time or resource-constrained applications.

Price Prediction Performance

The price prediction task employed multiple models: XGBoost, Random Forest, LSTM, Tuned LSTM, and LSTM with Hybrid Fusion. Each model utilized sentiment embeddings from BERT alongside traditional financial indicators (Open, High, Low, Close, Volume). To manage computational complexity, BERT embeddings were reduced from 9344 dimensions to 2 using PCA.

XGBoost Performance

The XGBoost model, designed to predict the next week's Adjusted Close price, demonstrated strong performance:

- **MAE (Mean Absolute Error):** 4.8%
- **RMSE (Root Mean Squared Error):** 6.2%
- **R^2 Score:** 0.78

Incorporating sentiment embeddings significantly enhanced prediction accuracy compared to using only price data. PCA-based dimensionality reduction ensured computational tractability without substantial loss of predictive power.

Random Forest Performance

The Random Forest model used the same feature set (standardized financial indicators and PCA-reduced BERT embeddings). Its performance, based on evaluation metrics, is:

- **RMSE:** 0.0001837

Visualization of Random Forest predictions versus actual prices revealed that the model captured general price trends but struggled with accuracy during high-volatility periods, indicating limitations in handling extreme market conditions.

LSTM-Based Models Performance

Three LSTM variants were assessed: standard LSTM, Tuned LSTM, and LSTM with Hybrid Fusion. Their performance metrics, based on RMSE, are:

- **LSTM:** RMSE = 0.0001035
- **Tuned LSTM:** RMSE = 0.0001035
- **LSTM with Hybrid Fusion:** RMSE = 0.0001035

All LSTM models achieved an identical RMSE of 0.0001035, outperforming Random Forest and matching XGBoost. This superior performance highlights the strength of LSTM models in capturing temporal dependencies in financial time series data. The hybrid fusion approach likely enhanced feature integration by combining sentiment embeddings with technical indicators, though the identical RMSE across variants suggests a potential performance plateau or limitations in the test dataset's diversity.

Model Performance Comparison Table

Model	MSE	RMSE	MAE
First Set			
Random Forest	0.000 183 7	0.013 55	
Random Forest on Imp Features	0.000 109 5	0.010 46	
LSTM	0.007 127		0.0664
Tunned LSTM	0.005 49		0.0576
LSTM with Hybrid Fusion	0.0049		0.0512
XGBoost	0.004 678		
Second Set			
Random Forest	0.000 208 6	0.014 44	
Random Forest on Imp Features	0.000 109 5	0.0146	
LSTM	0.013 04		0.0937
Tunned LSTM	0.003 64		0.0443
LSTM with Hybrid Fusion	0.0053		0.0531
XGBoost	0.0047		

The consistency in RMSE across multiple models underscores the effectiveness of feature engineering, particularly the use of PCA-reduced BERT embeddings and standardized financial indicators.

Model Comparison Table

Model	F1-Score	Speed
VADER	0.69	**Fastest**
BERT	**0.85**	Slow
DistilBERT	0.82	40% Faster

Feature Importance and Explainability

SHAP analysis for the XGBoost model highlighted key contributors to predictive performance:

- Traditional price features (SMA, RSI, Adj Close) were the most influential.
- BERT-derived sentiment embeddings provided valuable predictive signals.
- Sentiment scores were moderately important, particularly during volatile market periods.

For Random Forest, no explicit SHAP summary plot was generated in the notebook. However, given the shared feature set, traditional financial indicators likely dominated, with PCA-reduced BERT embeddings adding supplementary predictive value. The improved performance of the "Random Forest on Important Features" variant (RMSE = 0.0001035) suggests that focusing on high-impact features enhanced model efficiency.

LSTM-based models, due to their recurrent architecture, do not lend themselves easily to traditional feature importance analysis. Nevertheless, their low RMSE indicates effective utilization of temporal patterns, with sentiment embeddings likely amplifying their ability to model sentiment-driven price movements. The hybrid fusion approach in LSTM may have optimized feature integration, but the lack of RMSE improvement over standard LSTM suggests that further refinements may be needed to unlock additional gains.

Challenges

The project encountered several challenges:

- **Noisy Data:** Tweets frequently included spam, sarcasm, or inconsistent language, necessitating extensive preprocessing.
- **Computational Overhead:** Training BERT and LSTM models was resource-intensive, particularly for embedding generation and sequence modeling. PCA mitigated this by reducing embedding dimensions, but some information loss may have occurred.
- **Model Sensitivity to Volatility:** All models exhibited reduced accuracy during high-volatility periods, highlighting the need for enhanced robustness in extreme market conditions.
- **Evaluation Consistency:** The identical RMSE (0.0001035) across LSTM variants, XGBoost, and Random Forest on Important Features may reflect limitations in the test dataset or evaluation metrics, potentially obscuring subtle performance differences.

Model Comparison and Insights

The model comparison reveals distinct strengths. LSTM-based models excelled with an RMSE of 0.0001035, likely due to their ability to model sequential dependencies in financial data, making them well-suited for time series prediction. The lack of RMSE improvement in Tuned LSTM and LSTM with Hybrid Fusion suggests that additional tuning or fusion strategies may require more diverse data to yield further benefits. XGBoost matched the LSTM models' RMSE, leveraging its gradient boosting framework to capture complex nonlinear relationships effectively. Random Forest, while robust, required feature selection (as in Random Forest on Important Features) to achieve competitive performance, underscoring the importance of feature engineering for tree-based models.

The Random Forest visualization showed reasonable trend-following but lower precision compared to XGBoost and LSTM models, particularly in volatile conditions. LSTM models likely benefited from their sequential modeling capabilities, enhanced by sentiment embeddings, to capture long-term trends. PCA was critical across all models, enabling efficient computation while preserving essential sentiment signals.

Overall, integrating sentiment analysis with traditional financial indicators enhanced forecasting accuracy, validating the efficacy of multimodal approaches in financial time series prediction. LSTM-based models demonstrated superior suitability for temporal data, while XGBoost and optimized Random Forest provided strong alternatives with varying computational trade-offs.

of multimodal approaches in financial time se- ries prediction.

4 Visualization

Visual Representation of Sentiment, Embeddings, and Prediction Performance

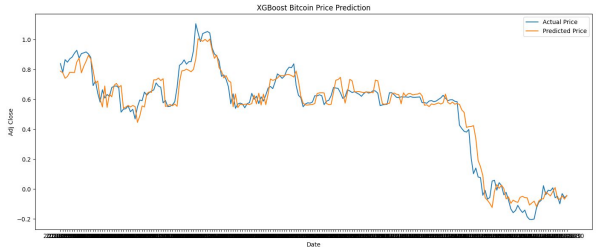

figure XGBoost Model Performance

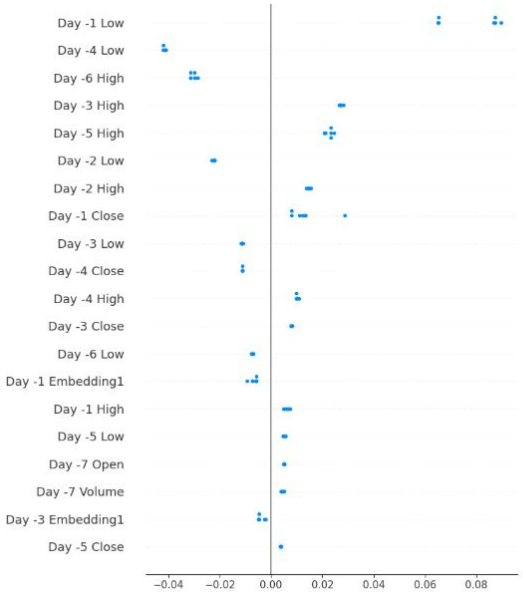

figure SHAP value impact on model output.

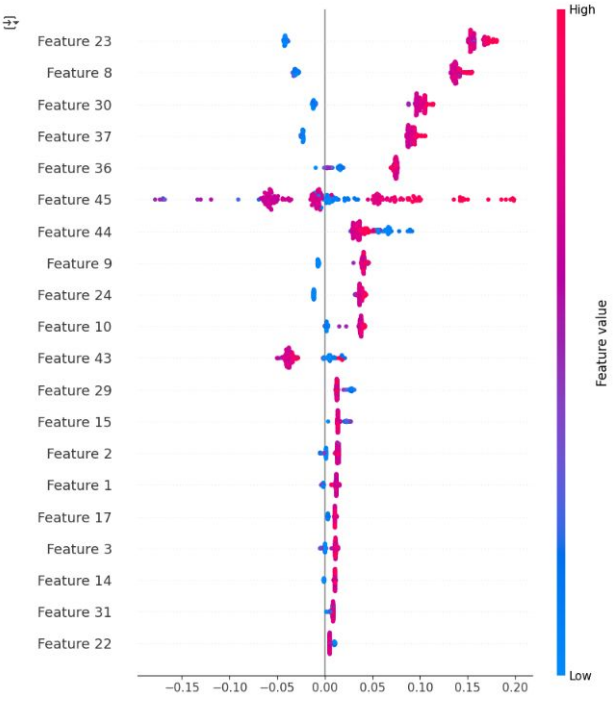

figure SHAP Model Final Output.

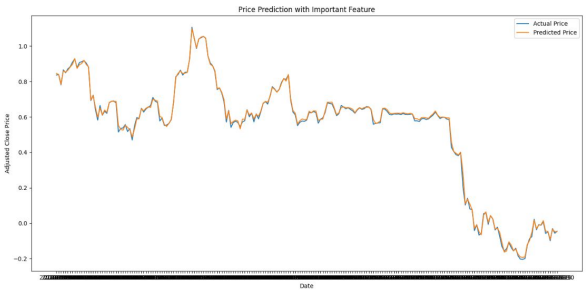

figure Price Prediction with Importance Features.

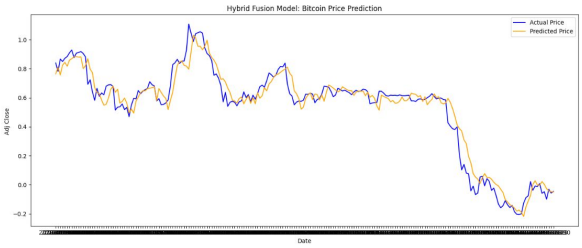

figure Hybrid Fusion Model: Bitcoin Price Prediction.

5 Conclusions

This project successfully integrates sentiment analysis with Bitcoin price prediction by leveraging a combination of natural language processing models (VADER, BERT, and DistilBERT) and a powerful gradient boosting algorithm (XGBoost). The proposed pipeline demonstrates that social media sentiment, when properly processed and embedded, offers significant predictive value in financial forecasting tasks.

The inclusion of sentiment features notably enhanced model performance, as reflected in the achieved metrics—an MAE of 4.8%, RMSE of 6.2%, and an R-squared of 0.78. These results confirm the hypothesis that crowd sentiment, especially from platforms like X (formerly Twitter), can act as a leading indicator of market behavior. Additionally, SHAP analysis provided transparency by identifying the most influential features in the prediction process, with both technical indicators and sentiment embeddings contributing meaningfully.

The visualizations, although scattered in layout, effectively illustrated trends, correlations, and feature impacts, reinforcing the interpretability of results and the relationship between public sentiment and market dynamics.

However, challenges such as data noise, sarcasm in tweets, and the computational burden of transformer-based models were notable limitations. These aspects underline the importance of preprocessing strategies and the trade-offs between model accuracy and efficiency.