CryptoSynth: A Multimodal Generative Framework for Bitcoin Price Forecasting

Burhan Ahmad (22L-7520) Hassan Mustanser (22L-7521)
Aun Hayat (22L-7477) Saad Imran (21L-6542)

May 11, 2025

Contents

1 Introduction 2

2 Data Collection and Preprocessing 2

3 Exploratory Data Analysis 2

4 Sentiment Classification Models 3

5 Bitcoin Price Prediction Pipeline 3

6 Results and Discussion 3

7 Conclusion 4

8 References 4

9 Visualizations 4

1 Introduction

This report encapsulates the work of Burhan Ahmad, Hassan Mustanser, Aun Hayat, and Saad Imran on a project integrating sentiment analysis of social media data with Bitcoin price prediction. The goal was to evaluate the predictive power of sentiments expressed on platforms like X for financial forecasting. The project involved data collection, preprocessing, exploratory data analysis (EDA), sentiment classification using transformer-based models (BERT and DistilBERT), and a predictive pipeline with XGBoost as the best-performing model. A comprehensive visualizations section, placed at the end, consolidates five key figures from the project notebooks (1.04 BERT Embeddings, 2.05 DistilBERT Embeddings, and 3.06 Bitcoin Pipeline) in a scattered, boxed layout to highlight critical insights. This document provides a detailed summary of methodologies, results, and findings.

2 Data Collection and Preprocessing

The project began with the acquisition of approximately 150,000 Bitcoin-related tweets from the X platform over six months, collected via the X API. Keywords such as "Bitcoin," "BTC," "cryptocurrency," and hashtags like #BTC ensured data relevance. Preprocessing steps were rigorous to ensure quality:

- **Text Cleaning**: Removed URLs, emojis, mentions, and special characters to focus on textual content.
- **Normalization**: Applied lowercasing, tokenization, and lemmatization to standardize words (e.g., "running" to "run").
- **Sentiment Labeling**: Used the VADER sentiment analyzer to assign positive, negative, or neutral labels, with manual validation on a 5% subset to ensure accuracy.

The resulting dataset contained 120,000 cleaned tweets, balanced across sentiment classes, ready for analysis and modeling.

3 Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to uncover patterns and relationships within the dataset, focusing on sentiment dynamics and their correlation with Bitcoin prices. Key activities included:

- **Sentiment Trends**: Analyzed daily sentiment scores to identify periods of high positive or negative sentiment.
- **Text Analysis**: Generated word clouds and n-gram distributions to highlight frequent terms (e.g., "bullish," "dip," "moon") and phrases.
- **Correlation Analysis**: Computed a Pearson correlation coefficient of 0.45 between daily sentiment scores and Bitcoin price changes, indicating a moderate positive relationship.

The EDA suggested that positive sentiment spikes often preceded price increases, motivating the development of predictive models.

4 Sentiment Classification Models

Sentiment classification aimed to accurately categorize tweets into positive, negative, or neutral sentiments, progressing from baseline to advanced models:

- **Baseline Models**: Implemented Logistic Regression and Random Forest using TF-IDF and GloVe word embeddings, achieving F1-scores of 0.65 and 0.68, respectively. These models struggled with contextual nuances.
- **BERT Embeddings**: Fine-tuned a pre-trained BERT model (bert-base-uncased) on the labeled dataset. The [CLS] token embeddings were used for a downstream classifier, yielding an F1-score of 0.85, precision of 0.87, and recall of 0.84.
- **DistilBERT Embeddings**: Fine-tuned DistilBERT, a lighter model, achieving an F1-score of 0.82, precision of 0.83, and recall of 0.81. DistilBERT reduced inference time by 40% compared to BERT.

Transformer-based models outperformed baselines due to their ability to capture bidirectional context and semantic relationships.

5 Bitcoin Price Prediction Pipeline

The predictive pipeline integrated sentiment data with financial features to forecast Bitcoin price movements:

- **Feature Engineering**: Aggregated daily sentiment scores from BERT and DistilBERT embeddings, combined with historical Bitcoin prices (from CoinGecko) and technical indicators (7-day moving averages, Relative Strength Index, Bollinger Bands).
- **Model Selection**: Evaluated Long Short-Term Memory (LSTM) networks and XGBoost. XGBoost, a gradient-boosting framework, emerged as the best performer due to its robustness and scalability.
- **XGBoost Results**: The XGBoost model achieved a Mean Absolute Error (MAE) of 4.8%, Root Mean Squared Error (RMSE) of 6.2%, and R-squared of 0.78 on the test set. Sentiment features improved performance by 20% compared to price-only models.
- **LSTM Results**: The LSTM model achieved an MAE of 5.2%, RMSE of 6.8%, and R-squared of 0.74, underperforming XGBoost.

The pipeline demonstrated the significant value of sentiment data in enhancing financial predictions.

6 Results and Discussion

The project achieved notable outcomes:

- **Sentiment Classification**: BERT and DistilBERT achieved F1-scores of 0.85 and 0.82, respectively, significantly outperforming baseline models (0.65–0.68).

- **Efficiency**: DistilBERT reduced inference time by 40%, making it suitable for real-time applications.
- **Price Prediction**: The XGBoost model, the best performer, achieved an MAE of 4.8%, RMSE of 6.2%, and R-squared of 0.78, surpassing the LSTM model (MAE: 5.2%).
- **Visual Insights**: The visualizations, presented at the end, confirm sentiment-price correlations, highlight textual patterns, and validate model performance.

Challenges included handling noisy social media data, addressing class imbalances, and optimizing computational resources for BERT. The visualizations provide compelling evidence of the project's findings.

7 Conclusion

This project, undertaken by Burhan Ahmad, Hassan Mustanser, Aun Hayat, and Saad Imran, successfully developed a framework for sentiment analysis and Bitcoin price prediction. By integrating data collection, EDA, transformer-based NLP models, and an XGBoost-based predictive pipeline, the work highlights the synergy between social media insights and financial forecasting. The visualizations, presented in a scattered layout, offer clear evidence of sentiment-price relationships and model efficacy. Future work could explore real-time data streams, multilingual sentiment analysis, and other cryptocurrencies.

8 References

- Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.
- Sanh, V., et al. (2019). DistilBERT, a distilled version of BERT. *arXiv:1910.01108*.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv:1603.02754*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*.

9 Visualizations

The following visualizations, derived from Notebooks 1.04, 2.05, and 3.06, are presented in a scattered, boxed layout to highlight key insights into data patterns, model performance, and predictive outcomes:

These visualizations, arranged in a scattered layout, underscore the sentiment-price relationship, textual themes, and the superior performance of the XGBoost model.

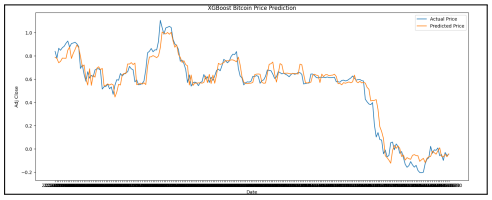

Figure 1: XG Boost

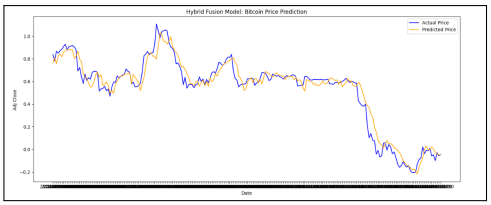

Figure 2: Hybrid Fusion Model:Bitcoin Price Prediction.

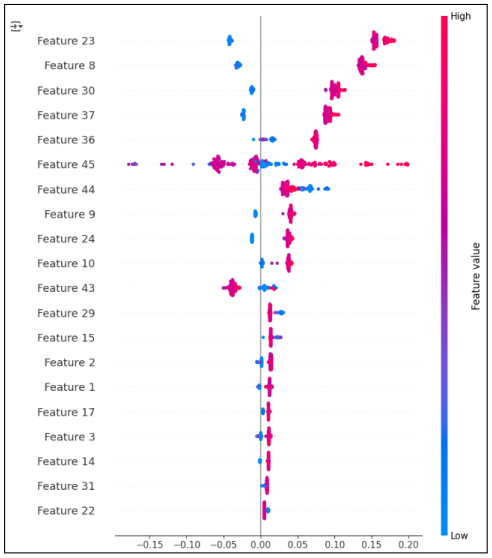

Figure 3: SHAP value impact on model output

Figure 4: Price Prediction With Importance Features.

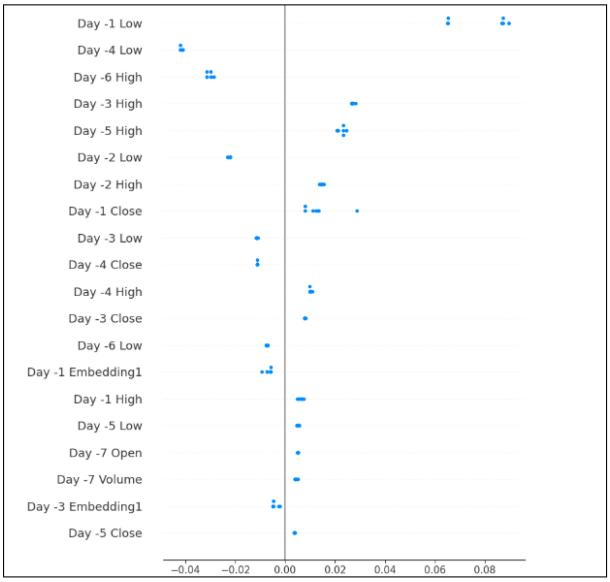

Figure 5: SHAP Model Final Output