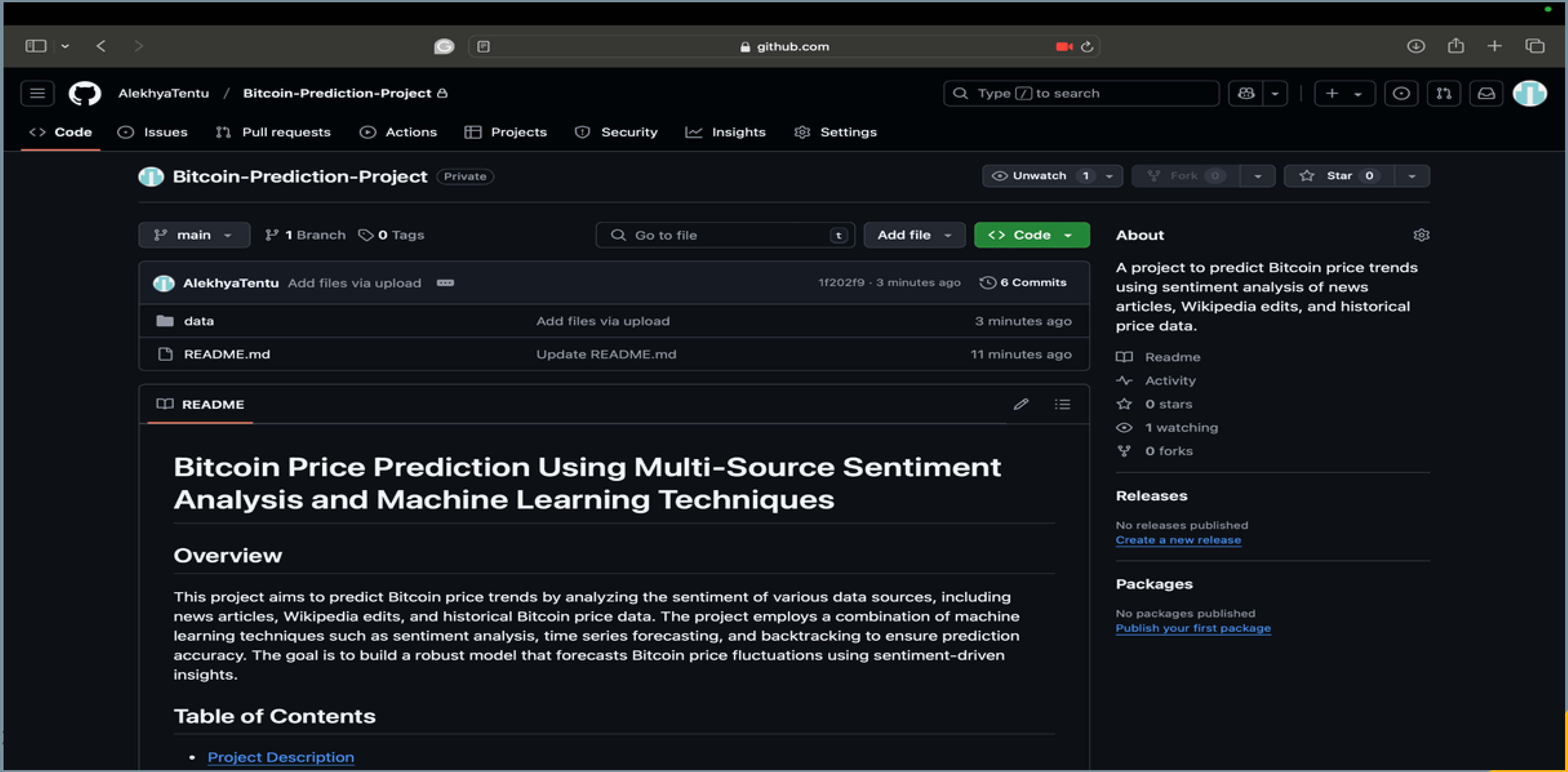
UMBC

Forecasting Bitcoin Price Trends Using Multi-Source Sentiment Analysis and Machine Learning Techniques

Team C
Members:
Alekhya Tentu (XX77459)
Vaishnavi Ratheesh Nair (MQ50131)
Adarsh Rao Akula (Jq00578)

GitHub Link: https://github.com/AlekhyaTentu/Bitcoin-Prediction-Project


github.com
AlekhyaTentu / Bitcoin-Prediction-Project
Type / to search
Code
Issues
Pull requests
Actions
Projects
Security
Insights
Settings
Bitcoin-Prediction-Project
Private
Unwatch 1
Fork 0
Star 0
main
1 Branch
0 Tags
Go to file
t
Add file
Code
AlekhyaTentu Add files via upload
1f202f9 · 3 minutes ago
6 Commits
data
Add files via upload
3 minutes ago
README.md
Update README.md
11 minutes ago
README
Bitcoin Price Prediction Using Multi-Source Sentiment Analysis and Machine Learning Techniques
Overview
This project aims to predict Bitcoin price trends by analyzing the sentiment of various data sources, including news articles, Wikipedia edits, and historical Bitcoin price data. The project employs a combination of machine learning techniques such as sentiment analysis, time series forecasting, and backtracking to ensure prediction accuracy. The goal is to build a robust model that forecasts Bitcoin price fluctuations using sentiment-driven insights.
Table of Contents
Project Description
About
A project to predict Bitcoin price trends using sentiment analysis of news articles, Wikipedia edits, and historical price data.
Readme
Activity
0 stars
1 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package

Introduction

Bitcoin, the most widely traded cryptocurrency, is highly volatile and influenced by market sentiment. Traditional price prediction models rely on historical price data, but integrating sentiment analysis can provide deeper insights into market movements.

This project aims to leverage Natural Language Processing (NLP) and Machine Learning (ML) techniques to analyze social media sentiment and predict Bitcoin price fluctuations. By extracting sentiment from financial news, and Wikipedia edits, and combining it with historical price trends, we develop a predictive model to enhance Bitcoin price forecasting accuracy.

Our approach involves:

- **Sentiment Extraction:** Using VADER, BERT, and FinBERT to analyze social media sentiment.
- **Machine Learning Models:** LSTM and Ensemble Learning for price prediction.
- **Performance Evaluation:** Assessing model accuracy using MAE, RMSE, R^2, and sentiment-price correlation.

This project bridges the gap between market sentiment and financial forecasting, helping investors make data-driven decisions.

Review of Similar Approaches

One approach used Empirical Mode Decomposition (EMD) with LSTM networks, where:

- One model processed historical price data.
- Another model integrated user sentiment and market data.

Other studies explored ARIMA and Neural Network Autoregression (NNAR) models for Bitcoin price forecasting.

- A hybrid approach combined weighted sentiment analysis from social media comments and financial news headlines with a stacked LSTM model for better predictions.
- Unlike traditional approaches that rely on a single sentiment source, we employ a multi-source sentiment fusion strategy, incorporating not only news articles and social media but also Wikipedia edits. This novel fusion of various sentiment sources provides a more comprehensive understanding of public perception, which is often a driving force behind cryptocurrency market fluctuations.

Dataset Exploration

Bitcoin Price Data (5 Years)
Link: https://drive.google.com/file/d/1-1GriZRFxvyLVjKk-X0rlFz1jV3JkLkl/view?usp=drive_link

- Source: Yahoo Finance
- Shape: 2,238 rows × 6 columns

Columns:

- Adj Close: Adjusted closing price of Bitcoin.
- Close: Closing price.
- High: Highest price within the period.
- Low: Lowest price within the period.
- Open: Opening price.
- Volume: Trading volume.

```
Shape of the DataFrame: (2238, 6)
Size of the DataFrame: 13428

Data types of each column:
Adj Close    float64
Close        float64
High         float64
Low          float64
Open         float64
Volume         int64
dtype: object

Index of the DataFrame:
DatetimeIndex(['2019-01-02', '2019-01-03', '2019-01-04', '2019-01-05',
               '2019-01-06', '2019-01-07', '2019-01-08', '2019-01-09',
               '2019-01-10', '2019-01-11',
               ...
               '2025-02-07', '2025-02-08', '2025-02-09', '2025-02-10',
               '2025-02-11', '2025-02-12', '2025-02-13', '2025-02-14',
               '2025-02-15', '2025-02-16'],
              dtype='datetime64[ns]', name='Date', length=2238, freq=None)
```

```
Cleaned Bitcoin Data:
              Adj Close        Close         High          Low         Open  \
Date
2019-01-02  3943.409424  3943.409424  3947.981201  3817.409424  3849.216309
2019-01-03  3836.741211  3836.741211  3935.685059  3826.222900  3931.048584
2019-01-04  3857.717529  3857.717529  3865.934570  3783.853760  3832.040039
2019-01-05  3845.194580  3845.194580  3904.903076  3836.900146  3851.973877
2019-01-06  4076.632568  4076.632568  4093.297363  3826.513184  3836.519043

                Volume
Date
2019-01-02  5244856836
2019-01-03  4530215219
2019-01-04  4847965467
2019-01-05  5137609824
2019-01-06  5597027440
```

Wikipedia Edit Data

Link: https://drive.google.com/file/d/1sbacJS3elda251tpyiE0Hw3uSoZJVc_K/view?usp=drive_link

- Source: Wikipedia edit logs related to Bitcoin.
- API Used: Wikipedia API (MediaWiki API) to extract edits made to Bitcoin-related Wikipedia pages.
- Shape: 1,227 rows × 6 columns

Columns:

- revid: Revision ID of the Wikipedia edit.
- parentid: Parent revision ID.
- user: Username of the editor.
- timestamp: Time of the edit.
- comment: Edit summary provided by the user.
- comment hidden: Whether the comment is hidden.

1	revid	parentid	user	timestamp	comment	commenthidden
2	1.28E+09	1272296704	PiggyGull	2025-02-15 19:16:57+00:00		
3	1.27E+09	1272293121	JivanP	2025-01-28 00:11:09+00:00	/* Mining */ Edit some language for clarity	
4	1.27E+09	1272283874	JivanP	2025-01-27 23:53:29+00:00	Lead: Try to make introductory explanation of bitcoin more accessible	
5	1.27E+09	1272282659	JivanP	2025-01-27 23:05:56+00:00	Lead: It is unknown if Satoshi is a single person.	
6	1.27E+09	1272012724	JivanP	2025-01-27 22:59:14+00:00	/* Addresses and transactions */ Addresses can be linked to things ot	
7	1.27E+09	1271980138	A455bcd9	2025-01-26 19:40:08+00:00	/* Regulatory responses and environmental concerns */ Super short s	
8	1.27E+09	1271897623	Person by t	2025-01-26 16:25:08+00:00	"Ideology" was a section with only one paragraph. As it relates to Austr	
9	1.27E+09	1271891625	Adolphus7	2025-01-26 07:22:09+00:00	Dating maintenance tags: {{Obsolete source}} {{As of}}	
10	1.27E+09	1270477900	LinusShapi	2025-01-26 06:32:18+00:00	/* Use for investment and status as an economic bubble */	
11	1.27E+09	1270477437	Gjb0zWxOl	2025-01-19 18:50:50+00:00	/* 2013â€“2014: First regulatory actions */ improved phrasing	
12	1.27E+09	1270473143	Gjb0zWxOl	2025-01-19 18:48:05+00:00	/* Scalability and decentralization challenges */ 2022 protest use	
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

News Articles Data (From 2019)

Link:https://drive.google.com/file/d/1--FJI3F3keB8f0UvmB-YsGeFcLfH6ZeA/view?usp=drive_link

•**News Source & Query:** Used The Guardian API with the query **“Bitcoin OR BTC”** to fetch relevant articles.

Data Collection: Retrieved **up to 250 articles** (50 per page, max 5 pages).

Extracted Fields: Title, snippet (trailText), body, author (byline), word count, and section.

Storage & Processing: Data saved in **CSV (guardian_bitcoin_simple.csv)** and converted into a DataFrame.

Analytics: Computed article count, daily article frequency, and missing days.

Columns:

•title: Headline of the article.

•url: Link to the full article.

•content: A short excerpt or summary of the article.

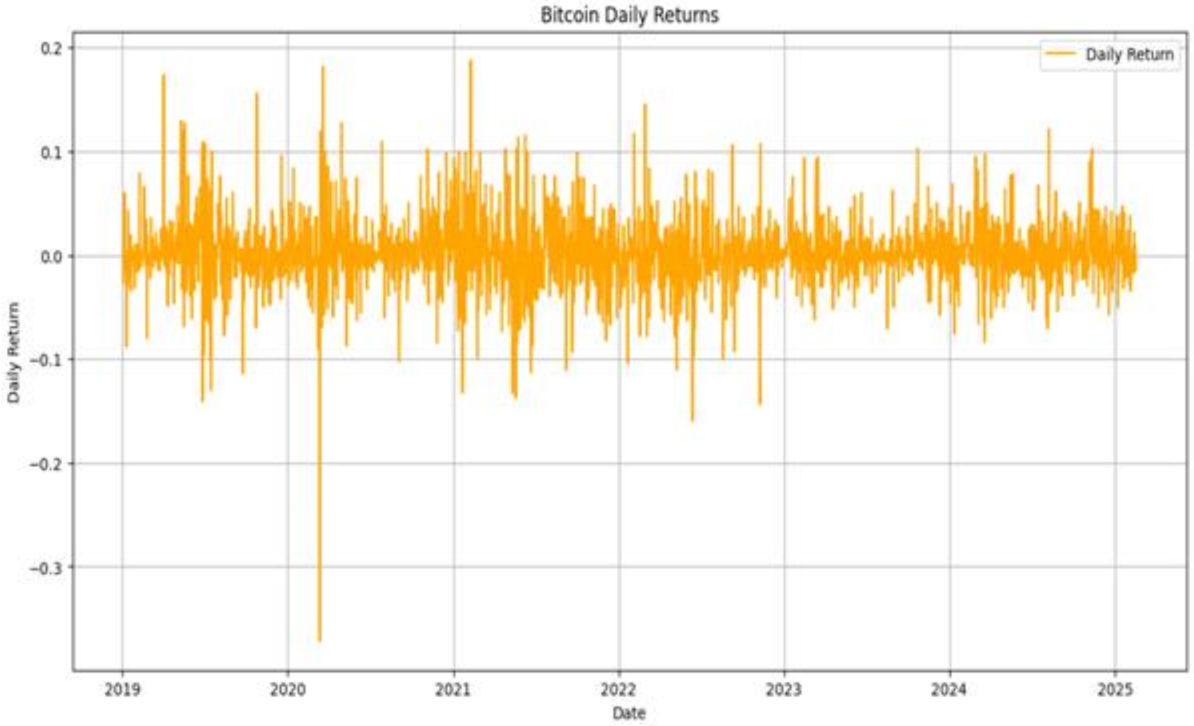
•published date: Publication date of the article.

•section: category

```
Article Distribution Analysis:
Total Articles: 2263
Total Days: 2250
Days With Articles: 1183
Days With No Articles: 1067
Max Articles Day: 11
Min Articles Day: 1
Avg Articles Day: 1.9129332206255283
Median Articles Day: 1.0

Days with most articles:
published_date
2024-11-11    11
2021-05-19    10
2022-12-13    10
2024-11-12     9
2022-11-15     8
Name: count, dtype: int64
```

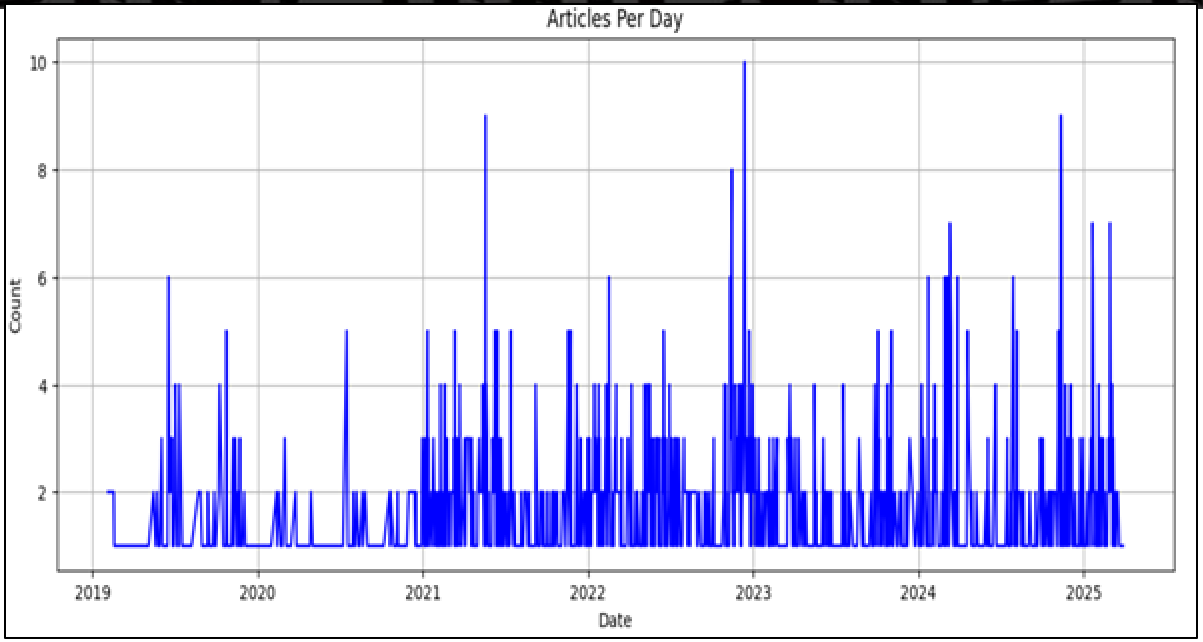
Exploratory Data Analysis

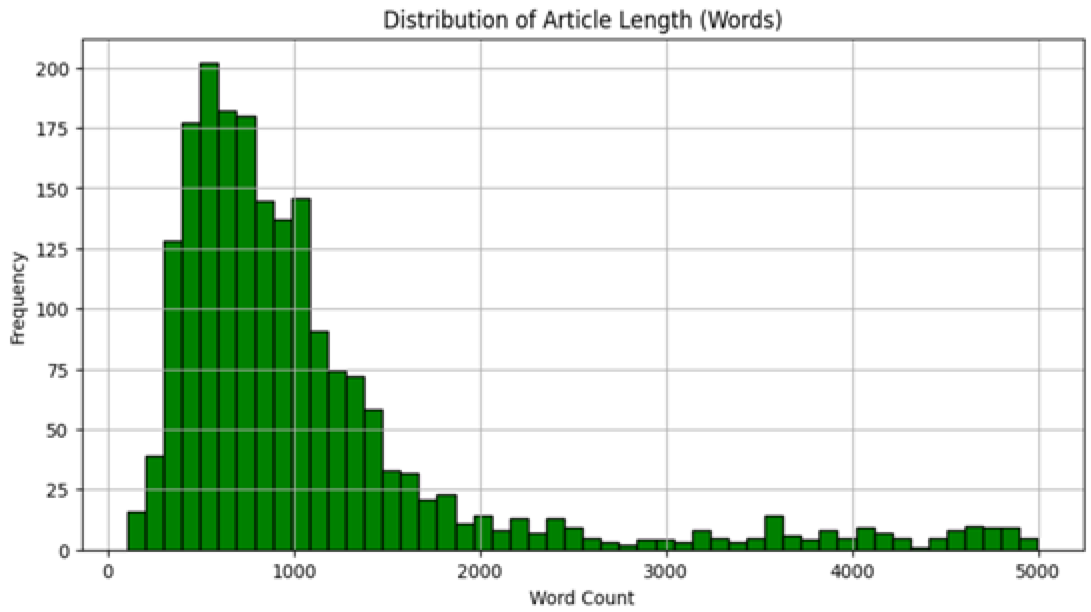


The **"Bitcoin Closing Price and 30-Day Moving Average"** graph reveals long-term trends and key price cycles between 2019 and 2025, helping smooth short-term volatility.

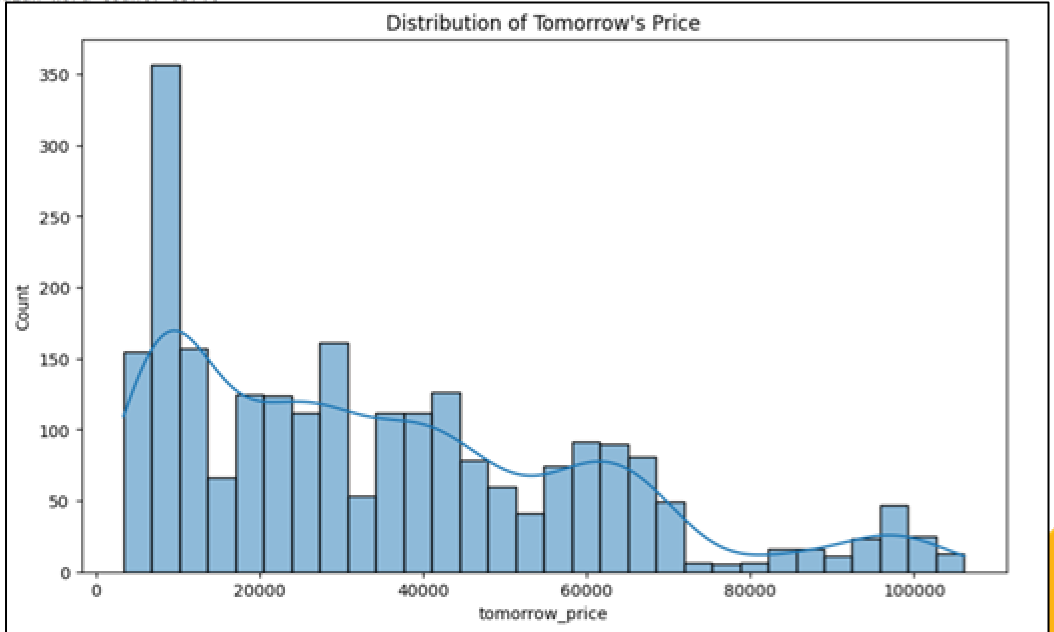
The **"Bitcoin Daily Returns"** graph highlights extreme fluctuations in return values, reflecting the high volatility characteristic of the cryptocurrency market.

These patterns support the integration of **technical indicators and sentiment scores** to enhance predictive performance by capturing both trend momentum and public sentiment shifts.





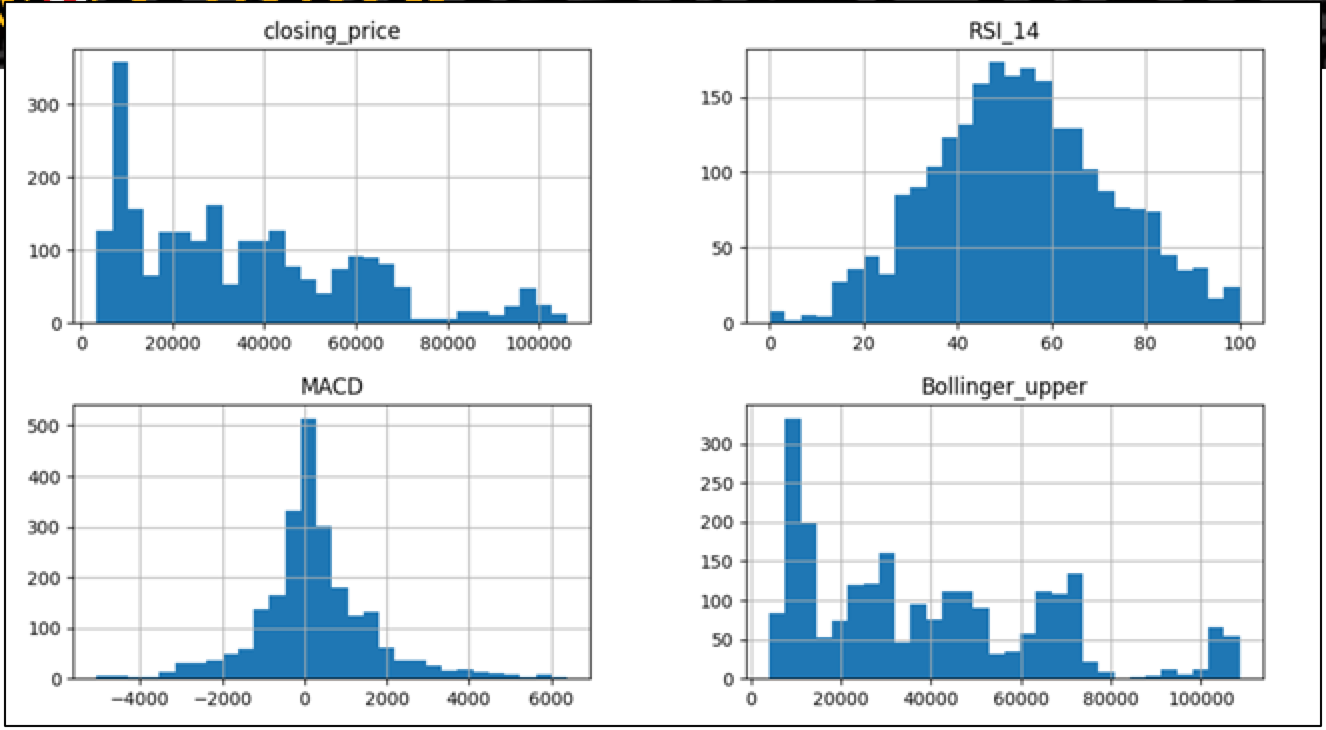
Median word count: 837.0

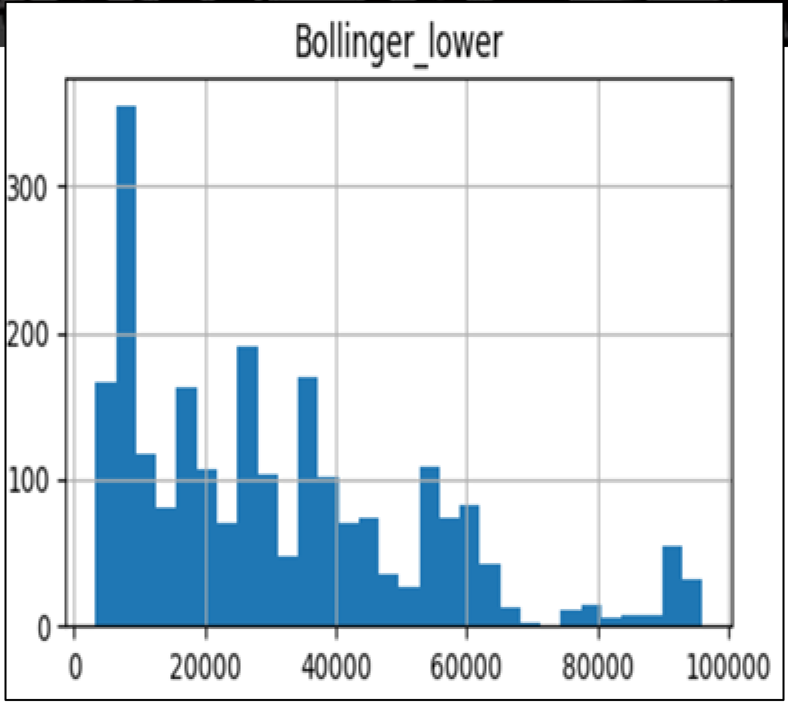


- The **"Articles Per Day"** graph shows the frequency of Bitcoin-related articles from 2019 to 2025, revealing an increase in media coverage during key price surge periods.

- The **"Distribution of Article Length (Words)"** plot highlights that most articles are between **500 to 1000 words**, with a median word count of **837**, suggesting consistent content depth for sentiment extraction.

- The **"Distribution of Tomorrow's Price"** graph is right-skewed, indicating Bitcoin's closing prices on the following day often fall within the **$0–$20,000** range, helping define the target variable's behavior for modeling.





Closing Price:-Shows the daily Bitcoin closing price. Most values fall between **$0 and $20,000**, indicating Bitcoin traded most frequently within this range.

RSI (14-day):-The **Relative Strength Index** measures momentum. The bell-shaped curve suggests most values lie between **40 and 60**, indicating a generally neutral market (neither overbought or oversold).

MACD:-The **Moving Average Convergence Divergence** captures trend strength. The centered peak near **0** implies a balanced mix of bullish and bearish momentum periods.

Bollinger Upper & Lower Bands:-These define the volatility range of Bitcoin prices. The distribution skew shows that prices were often closer to the lower band, indicating market caution or downward volatility.

Sentiment Analysis

In financial forecasting, especially for assets as volatile as Bitcoin, market sentiment plays a critical role. Our objective in this analysis was to quantify and integrate public sentiment into our Bitcoin price prediction model.

We extracted sentiment from news articles—specifically from *The Guardian*—using two approaches:

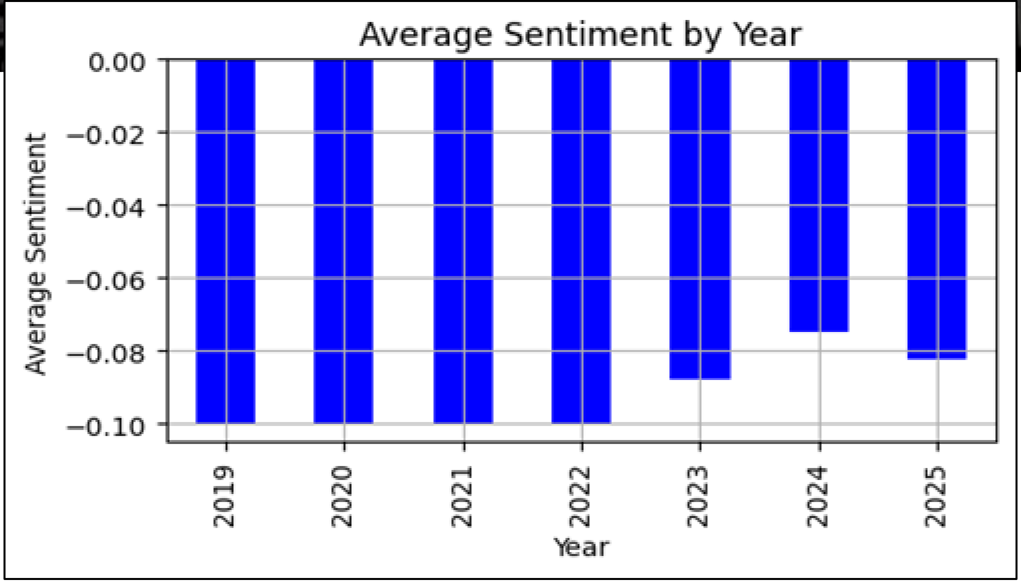
- **VADER**, a rule-based model suitable for short text sentiment.
- **BERT**, a transformer-based deep learning model capable of understanding contextual sentiment in longer narratives.

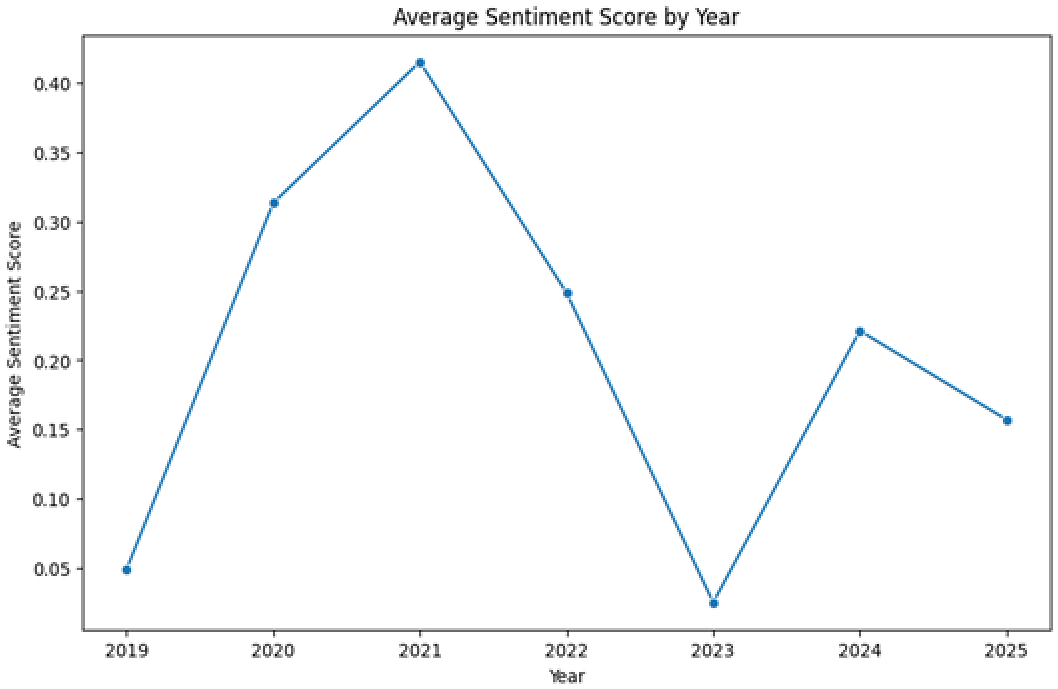
These sentiment scores were aggregated daily, allowing us to observe average sentiment **by year**, **by month**, and **by day type (weekend vs weekday)**. These visualizations helped uncover seasonal or emotional trends in public perception.

Alongside sentiment, we also analyzed several **technical indicators** such as:

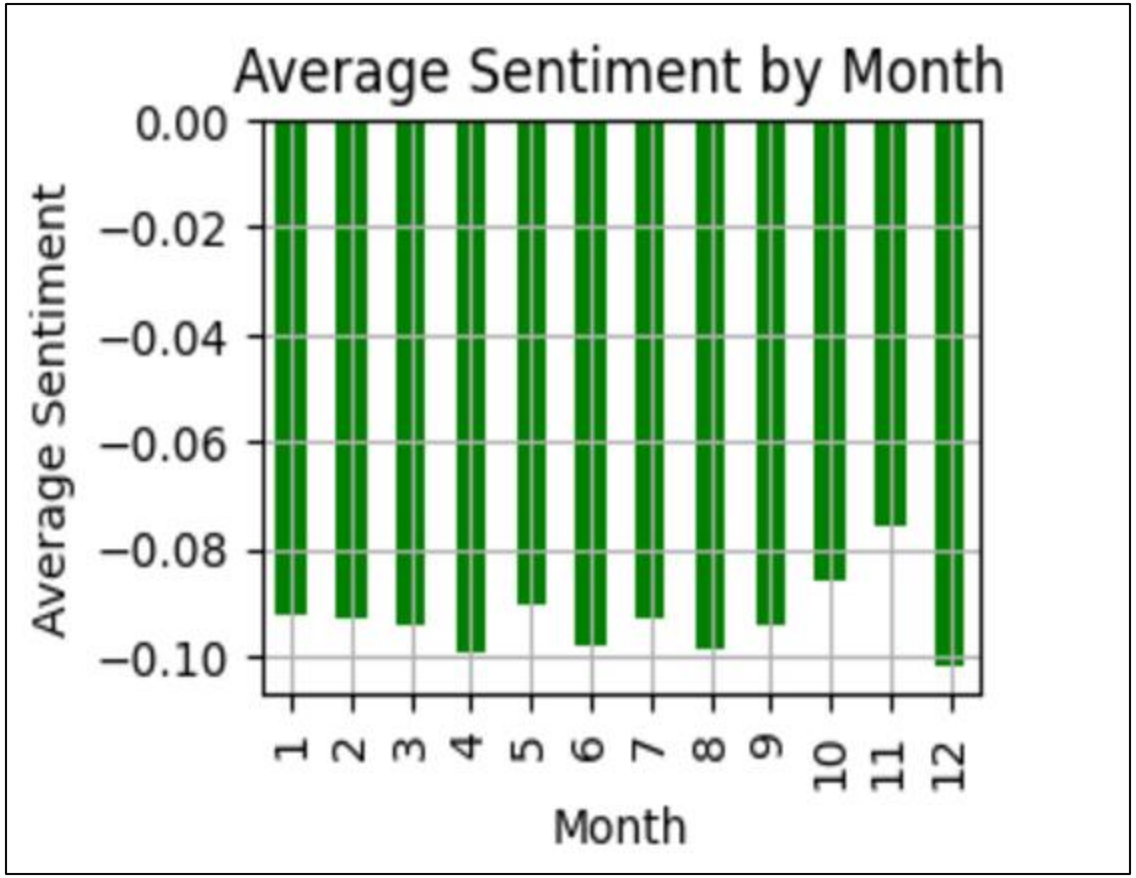
- **Closing Price** – with a concentration in the $0–$20K range,
- **RSI (14-day)** – typically showing a neutral market sentiment,
- **MACD** – indicating balance in momentum trends, and
- **Bollinger Bands** – helping measure volatility and potential reversals.

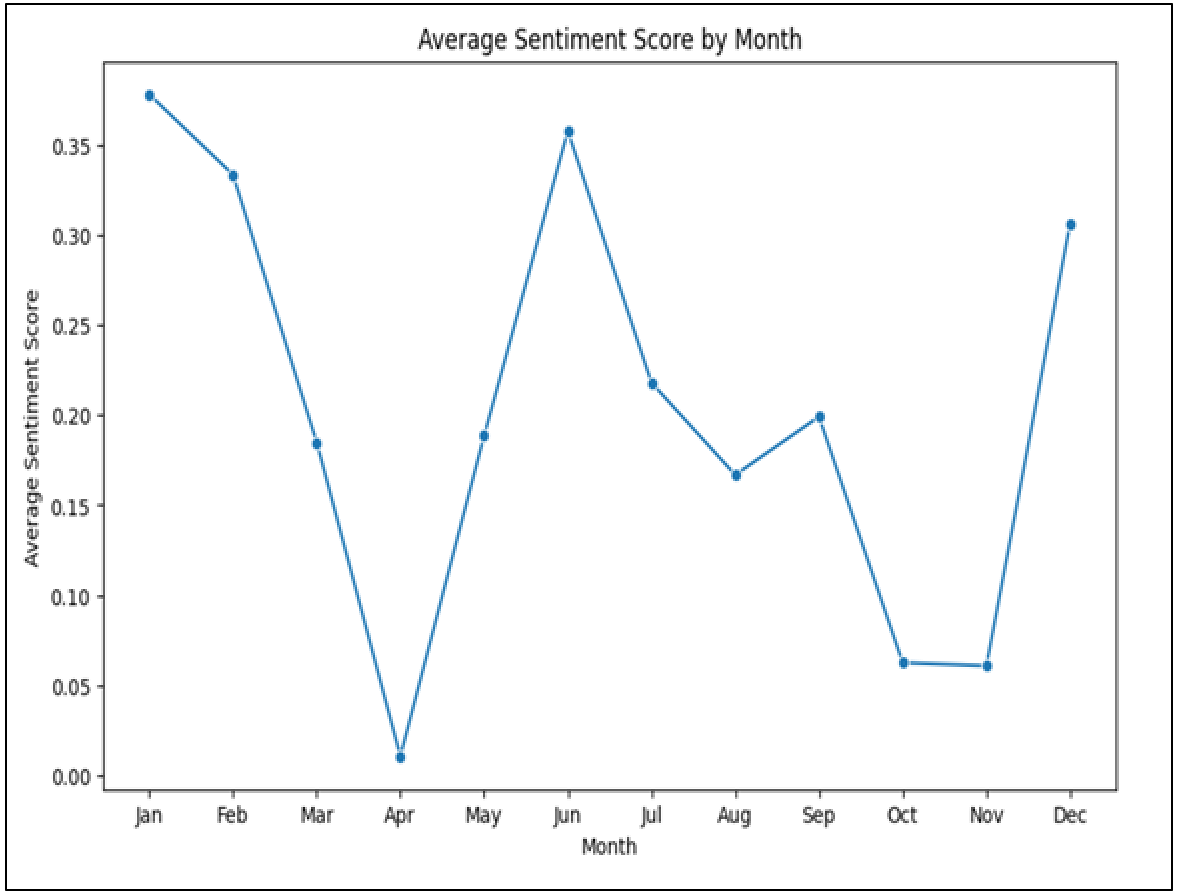
Average Sentiment by Years

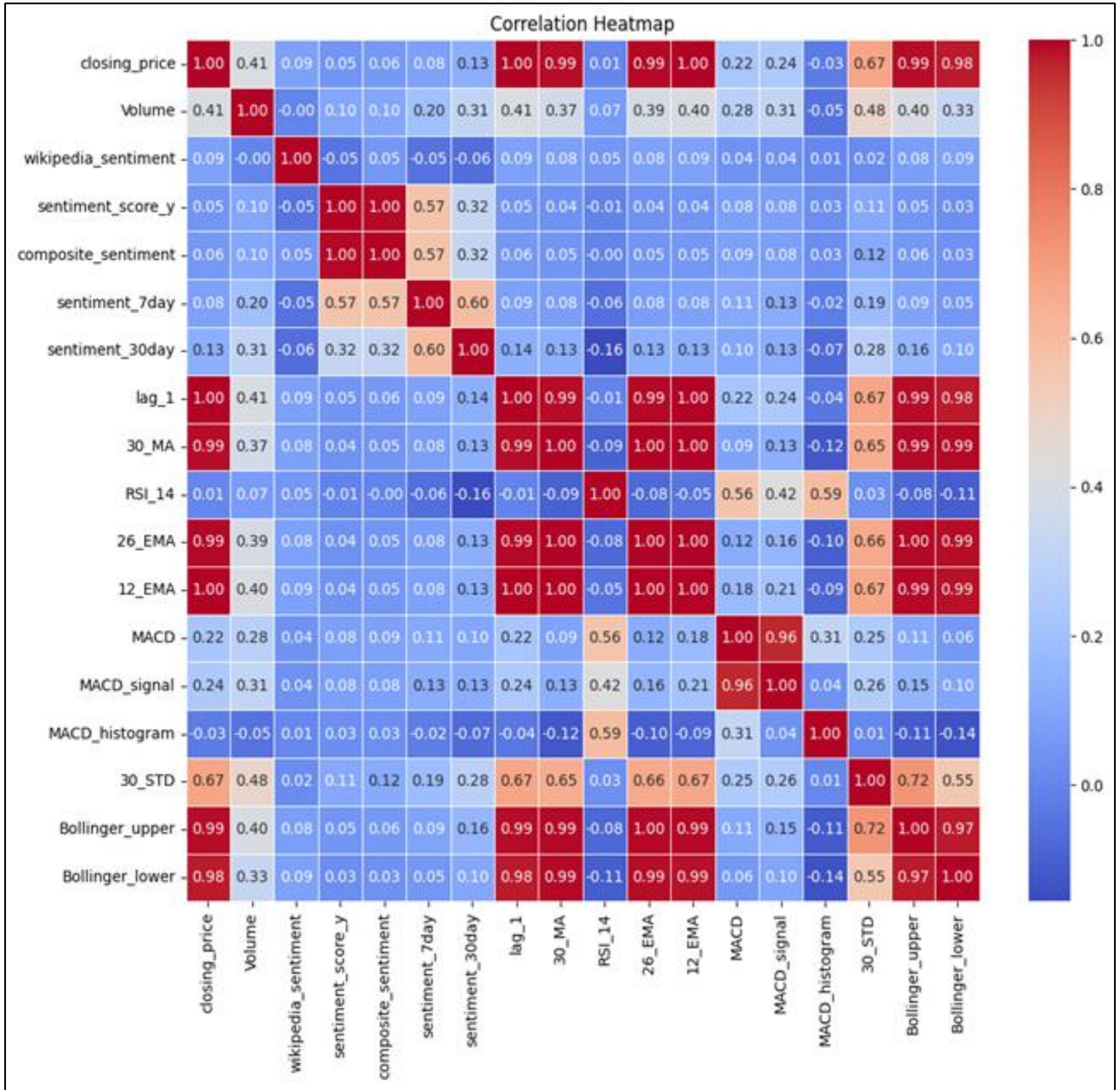




Average Sentiment by months







Sentiment & Price Correlation:

- **Composite Sentiment:** Moderately correlated (0.32) with tomorrow’s price, but weaker than technical indicators.
- **Rolling Sentiment (7-day & 30-day):** Stronger correlation (0.57 & 0.60), indicating sentiment trends impact price movements over time.

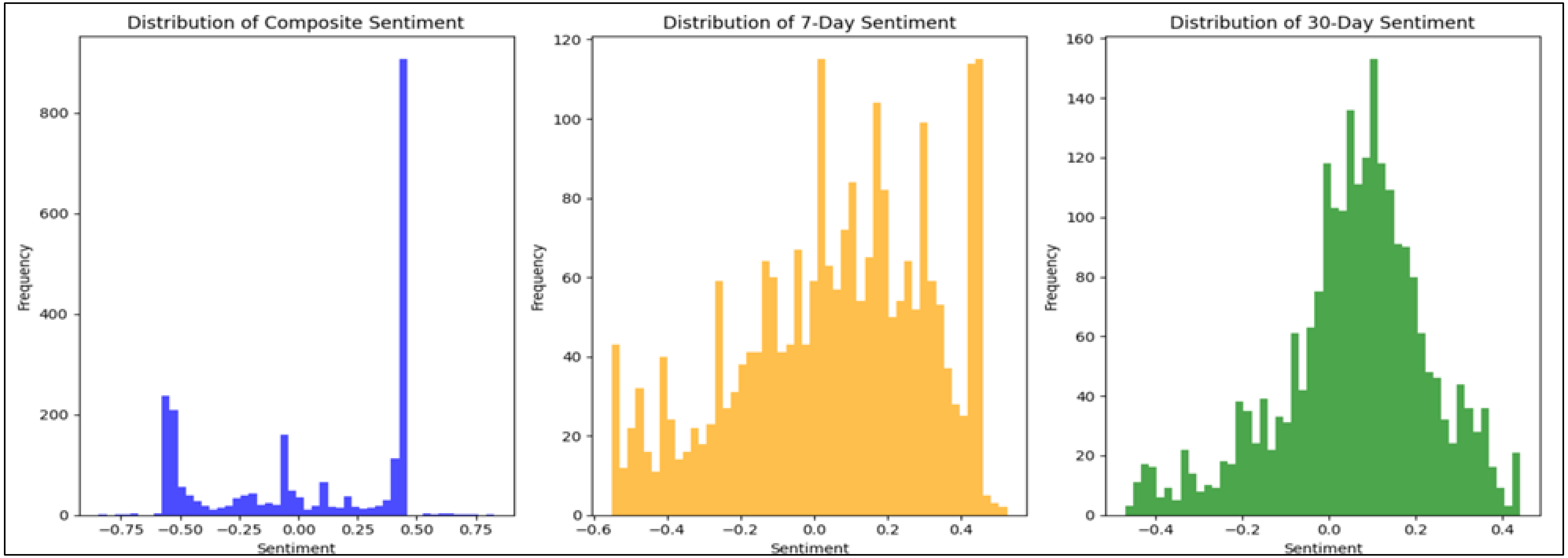
Technical Features:

- **Lag_1 (Previous Day’s Price):** Highest correlation (0.99) with tomorrow’s price, making it the strongest predictor.
- **Other Indicators (30_MA, RSI_14, etc.):** Also highly correlated, reinforcing the reliability of traditional technical analysis.

Sentiment & Technical Interaction:

- **Composite Sentiment & Technical Indicators:** Shows moderate correlation with lag_1, 30_MA, and Bollinger Upper, suggesting sentiment adds value alongside price-based features.

Distribution of Sentiment Features



- The **"Distribution of Composite Sentiment"** shows strong clustering around fixed scores, indicating consistent emotional tone across multiple sentiment sources (e.g., Guardian + Wikipedia).
- The **"Distribution of 7-Day Sentiment"** is right-skewed, capturing short-term optimism in public perception over a week's period.
- The **"Distribution of 30-Day Sentiment"** displays a bell-shaped curve centered near zero, indicating a more stable and neutral sentiment trend over longer time windows.

Methodology Overview

Goal: Predict Bitcoin price trends (increase or decrease) and forecast its future value.
Approach:

- Use Binary Classification to predict price movements (increase/decrease).
- Apply Regression to predict the actual price value.

Models Used:

- **Classification**: Logistic Regression, Random Forest, XGBoost.
- **Regression**: Linear Regression or other suitable models.

Metrics: Accuracy, Precision, Recall, F1-Score for classification; RMSE or MAE for regression.

Model Training

- Used Random Forest Classifier for classification
- Established baseline performance with default parameters

Model Improvements: Hyperparameter Tuning, SMOTE, and Feature Engineering

- **Step 1 - Hyperparameter Tuning**:
  - We performed GridSearchCV to optimize the Random Forest model by testing different combinations of hyperparameters such as n_estimators, max_depth, and others.
  - This allowed us to fine-tune the model for better performance.
- **Step 2 - SMOTE for Class Imbalance**:
  - To address the class imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic data for the minority class (price increases) to balance the dataset. This step was crucial in improving the model's ability to predict price increases.
- **Step 3 - Additional Feature Engineering**:
  - We added new features such as:
    - price_change_percentage: The percentage change in price compared to the previous day.
    - price_volatility_7day and price_volatility_30day: Rolling standard deviations over 7 and 30 days to capture volatility.
    - sentiment_momentum: The difference between 7-day and 30-day sentiment scores to track momentum.
    - Interaction Features: Multiplying sentiment features with technical indicators (e.g., sentiment_7day_x_RSI_14).

Feature Engineering Summary

- **Market-Based Features**
  • Daily OHLCV (Open, High, Low, Close, Volume)
  • Technical indicators – RSI, MACD, 30-day SMA, Bollinger Bands, lagged returns (t-1 … t-7)

- **Sentiment Features**
  • `wiki_sentiment` – VADER score of Wikipedia edit comments (daily mean)
  • `guardian_sentiment` – VADER ⇢ BERT hybrid score of news articles (daily mean)
  • `composite_sentiment` = (wiki + guardian) / 2

- **Rolling Statistics**
  • `sentiment_7day`, `sentiment_30day` – moving-average mood trends
  • `sentiment_vol_30` – 30-day standard deviation (tone volatility)

- **Target Variable**
  • `tomorrow_close` = Close price shifted −1 day (regression label)

- **Data Prep**
  • Forward-fill weekend gaps, align all series by calendar date
  • Min-Max scaling **only** for LSTM, tree models use raw values

```python
# Predict probabilities instead of classes
y_pred_prob = rf_model.predict_proba(X_test)[:, 1]

# Adjust threshold (e.g., if probability > 0.4, predict Class 1)
threshold = 0.4
y_pred_adjusted = (y_pred_prob > threshold).astype(int)

# Evaluate the adjusted predictions
from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test, y_pred_adjusted))
print(confusion_matrix(y_test, y_pred_adjusted))
```

```
              precision    recall  f1-score   support

           0       0.54      0.31      0.40       290
           1       0.44      0.67      0.53       230

    accuracy                           0.47       520
   macro avg       0.49      0.49      0.46       520
weighted avg       0.50      0.47      0.46       520

[[ 91 199]
 [ 76 154]]
```

Model Evaluation and Performance Metrics

Measured model performance using:

Accuracy:- 0.62,

- This indicates that the model correctly predicted 47% of the test cases.

Precision: -Class 0: 0.54,Class 1: 0.44

- Precision measures the percentage of correct positive predictions. A lower precision for Class 1 suggests more false positives.

Recall:- Class 0: 0.31,Class 1: 0.67

- Recall measures the percentage of actual positives correctly identified. Class 1 has higher recall, meaning fewer false negatives.

F1 Score:- Class 0: 0.40,Class 1: 0.53

- F1-score balances precision and recall. Class 1 has a slightly better overall performance.

These metrics help validate the effectiveness of using sentiment analysis from social media, Wikipedia edits, and news articles in predicting Bitcoin price movements. Although overall accuracy is moderate, the higher recall for positive predictions (Class 1) indicates that the model is more sensitive to upward price movement signals—crucial for financial forecasting and decision-making strategies in cryptocurrency investments.

UMBC

LSTM Model

Why LSTM?

- Captures sequential dependencies in price and sentiment time-series
- Able to learn long-term patterns that tree models may miss

```
Mean Absolute Error (MAE): 638.2312990191019
Root Mean Squared Error (RMSE): 1071.722256085569
R²: 0.9980048846166811
MAPE: 2.20001554678687%
```

Input Preparation

- 30-day sliding window of scaled features → one training example
- Features inside window: RSI, MACD, Bollinger bands, wiki/guardian sentiment, rolling means & volatility
- Min-Max scaling (0-1) applied to every numeric column

Strengths & Limitations

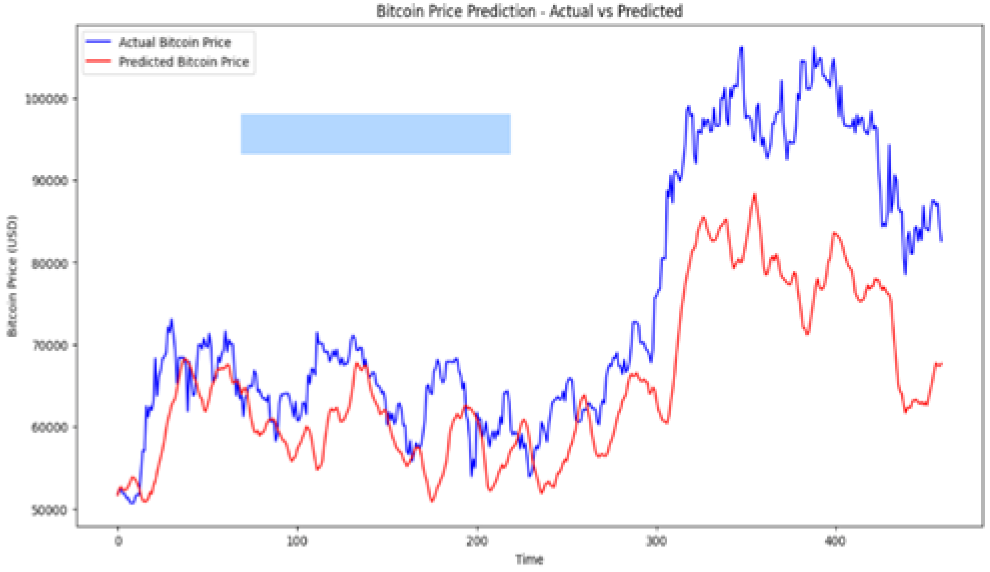
- Captures multi-day momentum & sentiment build-up

  – Needs more data to surpass tree ensemble; sensitive to extreme news spikes

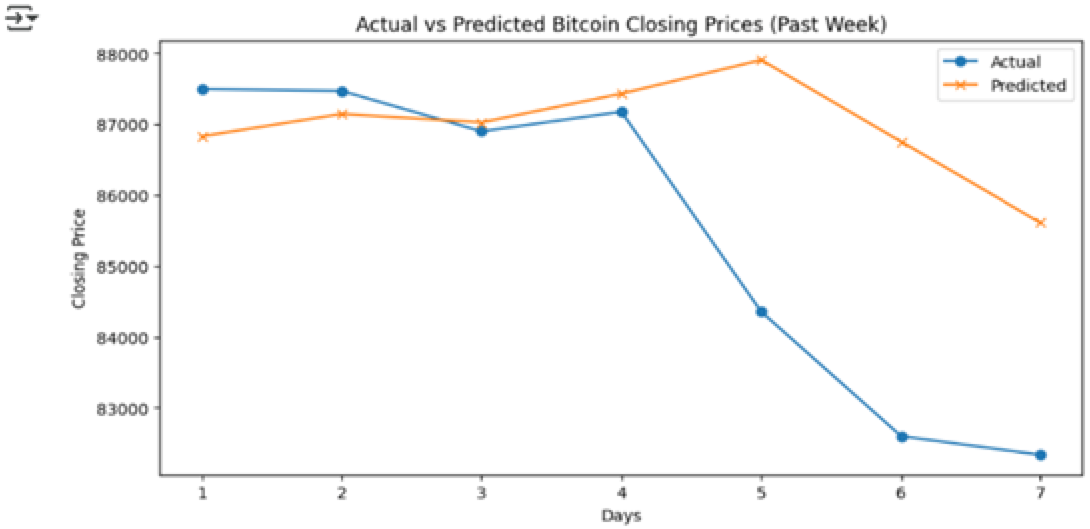
Actual v/s Predicted

Epochs

15/15 1s 27ms/step



RMSE: 12455.525300866844



Future Work

1. **Granularity Boost**
   • Incorporate intraday Guardian headlines and real-time Twitter/X sentiment

2. **Data Sources**
   • Add crypto-native news (CoinDesk, CoinTelegraph) and on-chain metrics (hash-rate, exchange flows)

3. **Model Enhancements**
   • Experiment with Gradient Boosting (LightGBM)
   • Build stacked ensemble (RF + LSTM meta-learner)

4. **Explainability**
   • Use SHAP to link specific news days to price jumps; create "news impact" dashboard

5. **Deployment Path**
   • Set up daily pipeline on AWS Lambda, push predictions + confidence bands to a Tableau or Streamlit dashboard

6. **Risk Controls**
   • Integrate Value-at-Risk (VaR) calculation to translate price forecast error into trading position limits

UMBC

References

Htay, H. S., Ghahremani, M., & Shiaeles, S. (2025). Enhancing Bitcoin Price Prediction with Deep Learning: Integrating Social Media Sentiment and Historical Data. *Applied Sciences*, *15*(3), 1554.

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018, October). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In *2018 IEEE 3rd international conference on computing, communication and security (ICCCS)* (pp. 128-132). IEEE.

Pano, T., & Kashef, R. (2020). A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. Big Data and Cognitive Computing, 4(4), 33.

Q&A session

Thank you