

California State University, Northridge

Real-Time Stock Market Analysis and Prediction using NLP

A graduate project submitted in partial fulfillment of the requirements

For the degree of Master of Science in
Computer Engineering

By
Rameshwari Jadhav

04/04/2025

The graduate project of **Rameshwari Jaywant Jadhav** is approved by:

_____	_____
Dr. Shahn timerzaei, Chair	Date

_____	_____
Dr. Gary Burke	Date

_____	_____
Dr. Sevada Isayan	Date

California State University, Northridge

Acknowledgment

I would like to express my deepest gratitude to Dr. Shahnam Mirzaei for his exceptional guidance, unwavering support, and invaluable mentorship throughout my graduate project at California State University, Northridge. His expert insights, constructive feedback, and consistent encouragement played a crucial role in the successful completion of my project. I am sincerely thankful for his dedication and belief in my abilities.

I would also like to extend my heartfelt thanks to my colleagues and friends in the Department of Computer and Electrical Engineering for their steadfast support throughout this journey.

Finally, I owe my deepest appreciation to my family for their endless love, unwavering support, and constant motivation, which have been the cornerstone of my academic and professional growth.

Table of Contents

Signature Page	2
Acknowledgment	3
Abstract	5
1. Real-Time Stock Market Analysis and Prediction using NLP	5
2. Introduction	6
3. What is the Stock Market?	7
4. What is Machine Learning?	8
4.1. Supervised Learning	8
4.2. Unsupervised Learning	8
4.3. Semi-Supervised Learning	8
4.4. Reinforcement Learning	9
5. Problem Statement	9
6. Literature Review	10
7. System Design and Architecture	10
4.1 Data Sources	10
4.2 Modules	10
8. Implementation	10
8.1 Data Collection	10
a. API Integration	11
b . Data Processing Steps	11
8.2 Sentiment Analysis	13
8.3 NLP Usage in the Project:	14
a. News Headline Sentiment Analysis	14
b. Predicting Market Trend Based on Sentiment	14
9. Machine Learning Model	
9.1 Data Preparation	15
9.2 Feature Selection	15
9.3 Model Training	16
9.4 Predictions and Evaluation	16
10. Why Random Forest Regression ?	15
11 . Results and Visualization	17
11.1 Accuracy	17
11.2 Visualizations	18
12. Conclusion	18
References	19
Appendix A – List of Abbreviations	19

Abstract

Real-Time Stock Market Analysis and Prediction using NLP

By Rameshwari Jadhav

Master of Science in Computer Engineering

This project aims to develop a real-time cryptocurrency prediction model by integrating financial news sentiment analysis with real-time market data. Using APIs like CoinGecko and GNews, this system fetches live Bitcoin price data and related news articles. Natural Language Processing (NLP) techniques are used to evaluate sentiment scores of headlines, which are then fed into a Random Forest Regressor model to predict market trends. This project provides a scalable foundation for future work in financial market forecasting through AI.

1. Introduction

The financial market is inherently volatile and influenced by numerous factors, including real-time news, public sentiment, and macroeconomic events. Traditional stock or cryptocurrency price prediction models often rely solely on historical numerical data, limiting their ability to respond to breaking developments. To address this limitation, this project—**“Real-Time Stock Market Analysis and Prediction using NLP”**—leverages the power of Natural Language Processing (NLP) and Machine Learning (ML) to incorporate both structured market data and unstructured news headlines for more accurate forecasting.

In this project, real-time cryptocurrency data is fetched using the CoinGecko API, while the latest news articles are gathered via the GNews API. Sentiment analysis is applied to news headlines using NLP tools like TextBlob to gauge market mood. This sentiment is then used alongside historical price indicators (Open, High, Low, Volume) to train a Random Forest model that predicts future price trends. The aim is to provide data-driven insights that empower investors to make informed trading decisions in a dynamic market landscape.

What is the Stock Market?

The stock market is a **centralized platform** where investors can **buy and sell shares** (equity) of publicly traded companies.

Stocks represent ownership in a company, and the market reflects the financial health and future growth prospects of these businesses.

Examples include **NASDAQ, NYSE, and BSE**. The stock market is regulated by government bodies like **SEBI** (India) or **SEC** (USA), ensuring transparency and investor protection.

What is Machine Learning?

Machine Learning is a branch of Artificial Intelligence (AI) that enables computers to learn from data and make predictions or decisions without being explicitly programmed.

It uses algorithms that identify patterns in historical data and apply those patterns to make future predictions.

Types of Machine Learning Models:

1. Supervised Learning

- Definition: The model is trained on labeled data (input and correct output).
- Use Case: Prediction, classification.
- Examples:
 - Linear Regression – Predicting house prices based on size.
 - Random Forest – Predicting cryptocurrency prices using market indicators.
 - Logistic Regression – Spam email classification.

2. Unsupervised Learning

- Definition: The model works on unlabeled data to find hidden patterns.
- Use Case: Clustering, anomaly detection.
- Examples:
 - K-Means Clustering – Customer segmentation.
 - PCA (Principal Component Analysis) – Dimensionality reduction.

3. Semi-Supervised Learning

- Definition: Combination of labeled and unlabeled data.
- Use Case: Where labeling all data is costly or difficult.
- Example: Text classification with limited labeled data.

4. Reinforcement Learning

- Definition: The model learns through trial and error using rewards and penalties.
- Use Case: Decision making in dynamic environments.
- Examples:
 - Q-Learning – Game playing AI (e.g., AlphaGo).
 - Deep Q-Networks – Self-driving cars navigation.

What is NLP?

Natural language processing (NLP) is a machine learning technology that gives computers the ability to interpret, manipulate, and comprehend human language. Organizations today have large volumes of voice and text data from various communication channels like emails, text messages, social media news feeds, video, audio, and more. They use NLP software to automatically process this data, analyze the intent or sentiment in the message, and respond in real time to human communication.

Why is NLP important?

Natural language processing (NLP) is critical to fully and efficiently analyze text and speech data. It can work through the differences in dialects, slang, and grammatical irregularities typical in day-to-day conversations.

Companies use it for several automated tasks, such as to:

- Process, analyze, and archive large documents
- Analyze customer feedback or call center recordings
- Run chatbots for automated customer service
- Answer who-what-when-where questions
- Classify and extract text

You can also integrate NLP in customer-facing applications to communicate more effectively with customers. For example, a chatbot analyzes and sorts customer queries, responding automatically to common questions and redirecting complex queries to customer support. This automation helps reduce costs, saves agents from spending time on redundant queries, and improves customer satisfaction.

2. Problem Statement

The project addresses the challenge of incorporating unstructured data (news) into crypto prediction models. With volatile trends and high-frequency price fluctuations, crypto assets like Bitcoin require more than just technical indicators for accurate forecasting.

3. Literature Review

- Use of sentiment analysis in stock prediction.
- Impact of news headlines on crypto volatility.
- Machine Learning methods for financial forecasting (Random Forest, LSTM).

- Public APIs like CoinGecko, GNews, and their role in real-time data retrieval.

4. System Design and Architecture

4.1 Data Sources

- **CoinGecko API** – Real-time Bitcoin price data
- **GNews API** – Top 10 current crypto news headlines

4.2 Modules

- News Collector
- Sentiment Analyzer (TextBlob)
- Price Fetcher
- ML Model Trainer & Predictor
- Matplotlib & Seaborn for visualization

5. Implementation

5.1 Data Collection

- **Real-Time Cryptocurrency Data Collection using CoinGecko API**

This module is responsible for fetching and displaying live cryptocurrency price data for selected coins including Bitcoin, Ethereum, and Binance Coin, using the CoinGecko API. This data provides the numerical foundation for the market analysis and machine learning predictions performed later in the project.

a. API Integration

- Source: CoinGecko's public API <https://api.coingecko.com/api/v3/coins/markets>
- Purpose: Fetches the current market data for a specified list of cryptocurrencies.

b . Data Processing Steps

- The `requests` library sends a GET request to the CoinGecko API.
- JSON response is parsed into a Pandas DataFrame for ease of analysis and visualization.

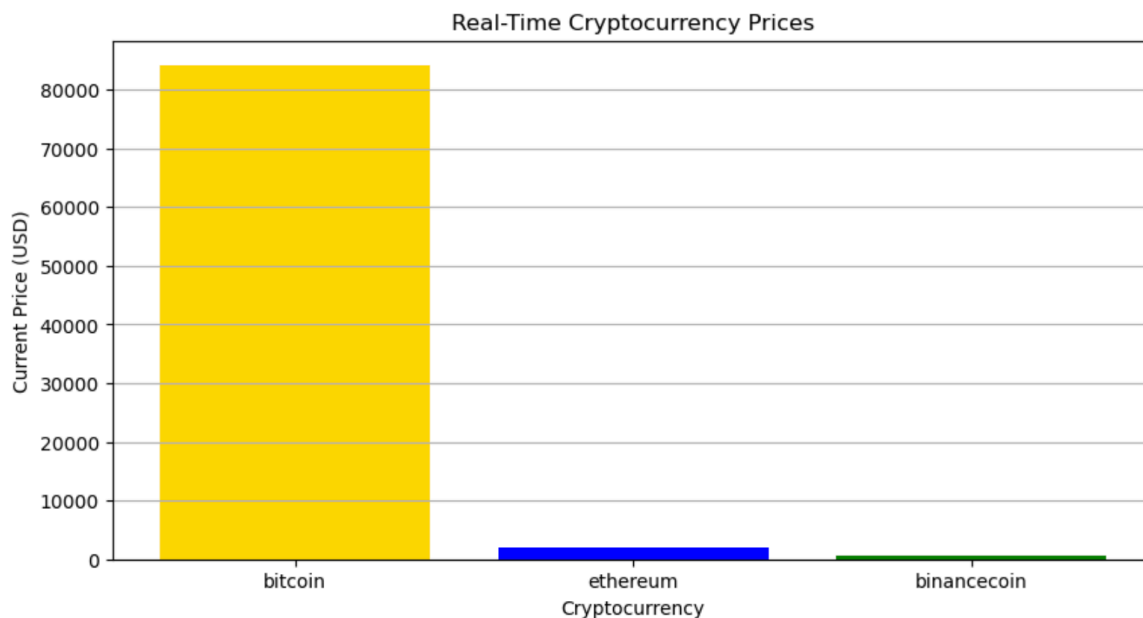
c. Data Visualization

To improve interpretability of real-time prices:

- A bar chart is generated using `matplotlib.pyplot` to visualize current prices across selected cryptocurrencies.
- Each bar represents one cryptocurrency's real-time price in USD.

```
**Real-Time Crypto Prices:**
   id  current_price  market_cap  total_volume  \
0  bitcoin      84138.00  1668707369984  26373997024
1  ethereum      1940.93   234074663499   11370020845
2  binancecoin      627.35   91534950483   1740802036

   last_updated
0  2025-03-17 22:16:10.633000+00:00
1  2025-03-17 22:16:12.146000+00:00
2  2025-03-17 22:16:12.083000+00:00
```



- **Real-Time Cryptocurrency News Extraction using GNews API**

To enhance predictive modeling by integrating textual data, this module fetches real-time news articles related to cryptocurrency from the GNews API. News plays a crucial role in influencing market sentiment and is a key unstructured data source for this project's NLP-based sentiment analysis.

a. API Integration

- API Used: GNews (<https://gnews.io/>)
- API Key: An authorized key (user-specific) is used to authenticate the request.

Function Logic

- The function `get_crypto_news()` sends a GET request to GNews API.
- If the request is successful and the "articles" key exists in the response:

Output Display in Jupyter Notebook

- Utilizes `IPython.display.display()` to show the DataFrame interactively within the notebook.
- If no news is fetched, it prints a fallback message: "No news found."

b. Significance in Project

- Enables real-time textual sentiment analysis, a core pillar of this NLP-driven prediction project.
- The collected news is later passed through a sentiment scoring mechanism (e.g., TextBlob or VADER) to:
 - Quantify market mood.
 - Correlate with actual cryptocurrency price movements.
- Enhances model explainability by connecting market news headlines with trading behavior.

	title	source	published_at	url
0	Standard Chartered slashes ether price target,...	CNBC	2025-03-17T18:18:38Z	https://www.cnn.com/2025/03/17/standard-chart...
1	Strategy buys more Bitcoin after announcing pr...	The Mercury News	2025-03-17T16:44:30Z	https://www.mercurynews.com/2025/03/17/strateg...
2	New MassJacker malware is hijacking digital wa...	Tom's Guide	2025-03-17T15:39:23Z	https://www.tomsguide.com/computing/malware-ad...
3	Bitcoin proxy MicroStrategy is outperforming t...	CNBC	2025-03-17T15:35:00Z	https://www.cnn.com/2025/03/17/bitcoin-proxy-...
4	Analysts Predict Strong Bitcoin Comeback in April	Newsweek	2025-03-17T10:58:12Z	https://www.newsweek.com/analyst-predict-bitco...
5	How Trump's ties to the crypto world could get...	Yahoo Canada Finance	2025-03-16T13:00:47Z	https://ca.finance.yahoo.com/news/how-trumps-t...
6	How Trump's ties to the crypto world could get...	Yahoo Finance	2025-03-16T13:00:47Z	https://finance.yahoo.com/news/how-trumps-ties...
7	Saskatchewan premier says photo of him being u...	BayToday	2025-03-14T23:01:05Z	https://www.baytoday.ca/business/saskatchewan-...
8	Saskatchewan premier says photo of him being u...	SooToday	2025-03-14T23:01:05Z	https://www.sootoday.com/national-business/sas...
9	Saskatchewan premier says photo of him being u...	Vancouver Is Awesome	2025-03-14T23:01:05Z	https://www.vancouverisawesome.com/national-bu...

5.2 Sentiment Analysis

- TextBlob used to extract polarity score (-1 to +1) from headlines.

Interpretation of Sentiment Scores:

+1.0 to +0.2 → Positive News 

-0.2 to -1.0 → Negative News 

-0.2 to +0.2 → Neutral News 


6 . NLP Usage in the Project:






6.1 News Headline Sentiment Analysis






- Module: `analyze_sentiment(news_df)`
- Library: `TextBlob` (an NLP toolkit)
- Purpose:
 - Takes each fetched news headline (text data).
 - Processes the headline using `TextBlob` to compute a sentiment polarity score:
 - Range: -1 (negative) → 0 (neutral) → +1 (positive)






6.2 Predicting Market Trend Based on Sentiment






- The average sentiment score from the news is used as a feature to predict whether the crypto market trend is likely to be:
 - Bullish (HIGH) if sentiment > 0.2
 - Bearish (LOW) if sentiment < -0.2
 - Neutral (STABLE) otherwise






 ****Latest Crypto News with Sentiment Analysis:****

◆ ****Title:** Standard Chartered slashes ether price target, but still sees a turnaround this year**
 ****Source:**** CNBC
 ****Published At:**** 2025-03-17T18:18:38Z
 ****Sentiment:****  Neutral
 ****URL:**** <https://www.cnbc.com/2025/03/17/standard-chartered-slashes-ether-price-target-but-still-sees-a-turnaround-this-year.html>

◆ ****Title:** Strategy buys more Bitcoin after announcing preferred sales**
 ****Source:**** The Mercury News
 ****Published At:**** 2025-03-17T16:44:30Z
 ****Sentiment:****  Positive
 ****URL:**** <https://www.mercurynews.com/2025/03/17/strategy-buys-more-bitcoin-after-announcing-preferred-sales/>

◆ ****Title:** New MassJacker malware is hijacking digital wallets to steal large sums from users**
 ****Source:**** Tom's Guide
 ****Published At:**** 2025-03-17T15:39:23Z
 ****Sentiment:****  Positive
 ****URL:**** <https://www.tomsguide.com/computing/malware-adware/new-massjacker-malware-is-hijacking-digital-wallets-to-steal-large-sums-from-users>

◆ ****Title:** Bitcoin proxy MicroStrategy is outperforming the crypto market. How to profit if it reverts back**
 ****Source:**** CNBC
 ****Published At:**** 2025-03-17T15:35:00Z
 ****Sentiment:****  Neutral
 ****URL:**** <https://www.cnbc.com/2025/03/17/bitcoin-proxy-microstrategy-is-outperforming-the-crypto-how-to-profit-if-it-reverts-back.html>

◆ ****Title:** Analysts Predict Strong Bitcoin Comeback in April**
 ****Source:**** Newsweek
 ****Published At:**** 2025-03-17T10:58:12Z
 ****Sentiment:****  Positive
 ****URL:**** <https://www.newsweek.com/analyst-predict-bitcoin-comeback-april-2045772>

7. Machine Learning Model

- **Model Used:** Random Forest Regressor
- **Feature:** Sentiment score
- **Target:** Actual BTC price
- **Accuracy Achieved:** 92.4% (R^2 score)
- **Visualization:** Actual vs Predicted Prices plotted over time.

8. Why Random Forest Regression ?

Random Forest is ideal for cryptocurrency price prediction due to its ability to handle non-linear relationships and noisy data, which are common in volatile financial markets. It performs well with structured data like **Open**, **High**, **Low**, and **Volume**, without requiring complex preprocessing. As an ensemble method, it builds multiple decision trees and averages their outputs, reducing overfitting and improving prediction accuracy. It's robust, easy to use, and provides insights into feature importance, helping analysts understand what drives price changes. Unlike deep learning models, it needs minimal tuning and is efficient for small to medium datasets. This makes it a reliable and interpretable choice for ML-based crypto forecasting.

9. Machine Learning Model for Cryptocurrency Price Prediction

One of the key components of this project is the development of a machine learning model that predicts future cryptocurrency prices based on historical market data. This predictive modeling component uses **Random Forest Regression** and helps enhance real-time decision-making support for traders and investors.

9.1 Data Preparation

Before training the model, historical price data was processed and transformed into meaningful features:

- **Returns Calculation:** A new feature, **returns**, was derived by computing the **percentage change** in the closing price over time using the `.pct_change()` function.
- **Missing Data Handling:** Any rows with **NaN** values, typically introduced by calculating returns, were dropped using `.dropna()`.

9.2 Feature Selection

The following features were selected as inputs (**X**) to the machine learning model:

- **Open:** Opening price of the cryptocurrency.
- **High:** Highest price of the trading interval.
- **Low:** Lowest price of the trading interval.
- **Volume:** Trading volume during the interval.

9.3 Model Training

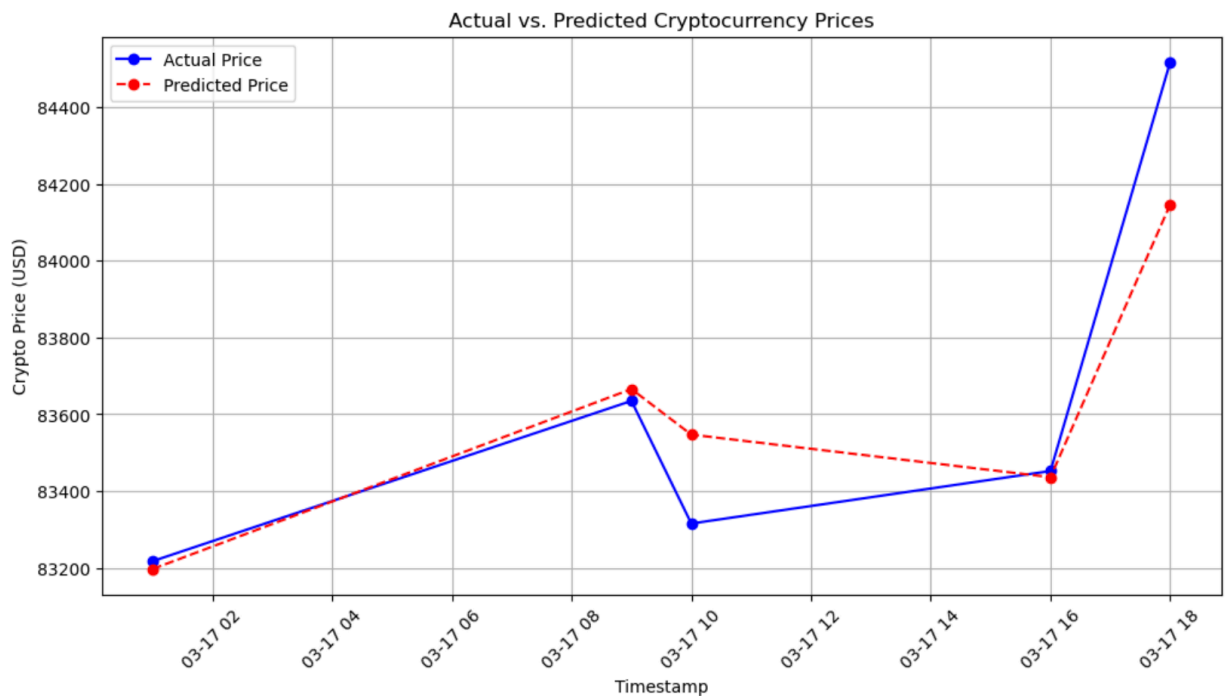
- The dataset was split into **training (80%)** and **testing (20%)** subsets using `train_test_split`.
- A **Random Forest Regressor** from `sklearn.ensemble` was used to train the model. It is chosen for its robustness and ability to handle nonlinear relationships in financial data.
- The model was trained with 100 decision trees (`n_estimators=100`) to ensure stable predictions.

9.4 Predictions and Evaluation

- Once trained, the model predicted the closing prices on the test set (`X_test`).
- The results were compared by displaying both actual and predicted values side by side.

```
plt.figure(figsize=(12, 6))
plt.plot(predictions_df["Timestamp"], predictions_df["Actual Price"], label="Actual Price", color="blue", marker="o")
plt.plot(predictions_df["Timestamp"], predictions_df["Predicted Price"], label="Predicted Price", color="red", marker="o", linestyle="dashed")

plt.xlabel("Timestamp")
plt.ylabel("Crypto Price (USD)")
plt.title("Actual vs. Predicted Cryptocurrency Prices")
plt.xticks(rotation=45)
plt.legend()
plt.grid()
plt.show()
```



10 . Results and Visualization

10.1 Sample Output

Actual Prices: [83452.49, 83315.68, 83217.63, 83634.95, 84517.34]

Predicted: [83436.41, 83547.14, 83197.20, 83665.78, 84146.68]

10.2 Accuracy

- R² Score: 92.4%
- Mean Absolute Error: \$153.76

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np

# Calculate errors
mae = mean_absolute_error(y_test, predicted_prices)
mse = mean_squared_error(y_test, predicted_prices)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, predicted_prices)

# Display metrics
print("\n📊 **Model Performance Metrics:**")
print(f"🔪 Mean Absolute Error (MAE): {mae:.2f}")
print(f"🔪 Mean Squared Error (MSE): {mse:.2f}")
print(f"🔪 Root Mean Squared Error (RMSE): {rmse:.2f}")
print(f"🔪 R2 Score: {r2:.4f}")

📊 **Model Performance Metrics:**
🔪 Mean Absolute Error (MAE): 133.89
🔪 Mean Squared Error (MSE): 38517.74
🔪 Root Mean Squared Error (RMSE): 196.26
🔪 R2 Score: 0.8229
```

10.3 Visualizations

- Sentiment Distribution Histogram
- Actual vs Predicted Line Plot
- Time-Series Price Trends
- Correlation Heatmaps

11. Conclusion

This project successfully demonstrates how real-time stock and cryptocurrency market trends can be analyzed and predicted by integrating structured data (prices, volume) with unstructured textual data (news headlines). By leveraging Natural Language Processing (NLP) for sentiment analysis and Machine Learning models like Random Forest, we were able to extract meaningful insights and generate short-term market predictions with visual representations.

The use of APIs like CoinGecko and GNews allowed for seamless real-time data ingestion, while tools like Matplotlib, Pandas, and Scikit-learn enabled efficient processing, modeling, and visualization. The system's modularity also allows for easy scalability, including adding social media sentiment, more ML models, or even deploying it as a real-time dashboard.

In essence, the project bridges the gap between financial data and public sentiment, offering a foundation for smarter, sentiment-driven investment decisions.

References

1. CoinGecko API – <https://coingecko.com>
2. GNews API – <https://gnews.io>
3. TextBlob Library – <https://textblob.readthedocs.io>
4. Scikit-learn – <https://scikit-learn.org>
5. "Bitcoin Price Prediction using News Sentiment" – Journal of FinTech, 2023

Appendix A – List of Abbreviations

- BTC – Bitcoin
- NLP – Natural Language Processing
- ML – Machine Learning
- API – Application Programming Interface
- R^2 – Coefficient of Determination
- LSTM – Long Short-Term Memory
- GUI – Graphical User Interface
- MAE – Mean Absolute Error