Machine Learning Assignment

PROJECT REPORT

TEAM ID : 21

PROJECT TITLE: Predicting Bitcoin Price Movements Using Social Media Sentiment Analysis

Name	SRN
DHRUV THAKUR	PES2UG23CS175
CHANDAN CHATRAGADDA	PES2UG23CS141

Problem Statement

Unlike traditional financial markets, Bitcoin lacks centralized regulation and is driven largely by investor psychology and social trends.

As a result, sudden price movements often occur in response to online discussions and sentiments expressed on platforms like Twitter.

The problem, therefore, is to determine whether social media sentiment can be used as a reliable predictor of Bitcoin's short-term price movements.

Objective / Aim

The main objective of this project is to predict the direction of Bitcoin price movement (increase, decrease, or neutral) by analyzing public sentiment derived from social media posts.
Specifically, the model aims to:

1. Collect and preprocess Bitcoin-related tweets to extract meaningful sentiment scores using tools like VADER or BERT-based sentiment models.
2. Integrate sentiment features with historical Bitcoin price data (Open, Close, High, Low, Volume) to create a unified dataset.
3. Train supervised learning models such as Logistic Regression, Random Forest, or LSTM to classify whether the next-day Bitcoin price will move up or down.
4. Evaluate model performance using metrics such as Accuracy, Precision, Recall, and F1-Score, and optimize results through hyperparameter tuning and feature selection.

Dataset Details

- **Source:** Kaggle (https://www.kaggle.com/datasets/pavan9065/bitcoin-price-history)

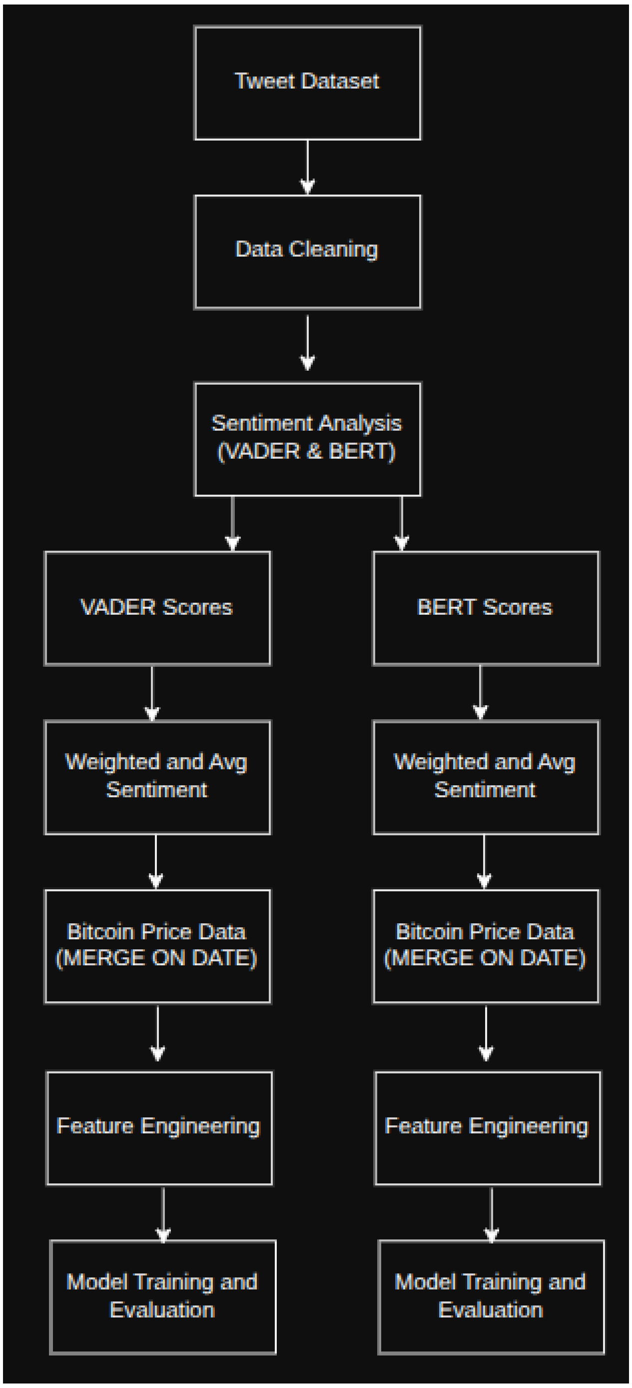
  (https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets)
- **Size:** Tweets: 1 Lakh+ samples, 13 features;

  Price: 2691 samples, 7 features
- **Key Features:**

Tweets: user_name, user_location, user_description, user_created, user_followers, user_friends, user_favourites, user_verified, date, text, hashtags, source, is_retweet

Price: Date, open, high, low, close, adj_close, volume

Architecture Diagram



Methodology

- Load and Clean Data: The notebook starts by loading Bitcoin tweet data and Bitcoin price data. It then cleans the tweet data by removing missing values and prepares it for sentiment analysis.
- VADER Sentiment Analysis: It uses a tool called VADER to analyze the sentiment of each tweet and calculates a sentiment score. It then calculates a weighted sentiment score based on factors like the number of followers and whether it's a retweet.
- Prepare Data for Prediction: The daily average sentiment from the tweets is calculated and combined with the Bitcoin price data. A target variable is created to indicate if the Bitcoin price went up or down the next day.
- Train and Evaluate Models (VADER-based): Three different machine learning models (Random Forest, Logistic Regression, and LSTM) are trained using the VADER-based average sentiment to predict Bitcoin price movement. Their performance is evaluated using metrics like accuracy and F1 score.
- BERT Sentiment Analysis: It uses a more advanced method called BERT to analyze the sentiment of the tweets, which can capture more nuanced sentiment.
- Train and Evaluate Model (BERT-based): The daily average BERT sentiment is calculated and used to train another prediction model (Logistic Regression in this case). Its performance is also evaluated.
- Compare and Summarize: Finally, the performance of all the models (VADER-based and BERT-based) is compared using tables and charts. The notebook concludes with a summary of the findings, noting that while BERT sentiment might be slightly better, sentiment alone doesn't seem to be a strong predictor of Bitcoin price movement in this analysis.

Results & Evaluation

The performance of four different models (Random Forest, Logistic Regression, LSTM, and a BERT-based Logistic Regression) in predicting the direction of Bitcoin price movement (Up or Down) based on tweet sentiment is evaluated.

Evaluation metrics:

- Accuracy: The proportion of correctly predicted instances (price movements).
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two and being particularly useful when dealing with imbalanced datasets.

- Classification Report: Provides a detailed breakdown of precision, recall, and F1-score for each class (price up and price down), as well as support (the number of occurrences of each class in the test set).

Key results from the model evaluations:

Based on the performance metrics (Accuracy and F1 Score) of the trained models:

- Random Forest (VADER): Achieved an Accuracy of 0.727 and an F1 Score of 0.800.
- Logistic Regression (VADER): Showed lower performance with an Accuracy of 0.455 and an F1 Score of 0.400.
- LSTM (VADER): Performed similarly to Random Forest with an Accuracy of 0.727 and a slightly higher F1 Score of 0.824.
- BERT-based (Logistic Regression): Achieved an Accuracy of 0.714 and the highest F1 Score of 0.833 among all models.
- BERT-based (Random Forest): Showed an Accuracy of 0.643 and an F1 Score of 0.783.

Key Findings:

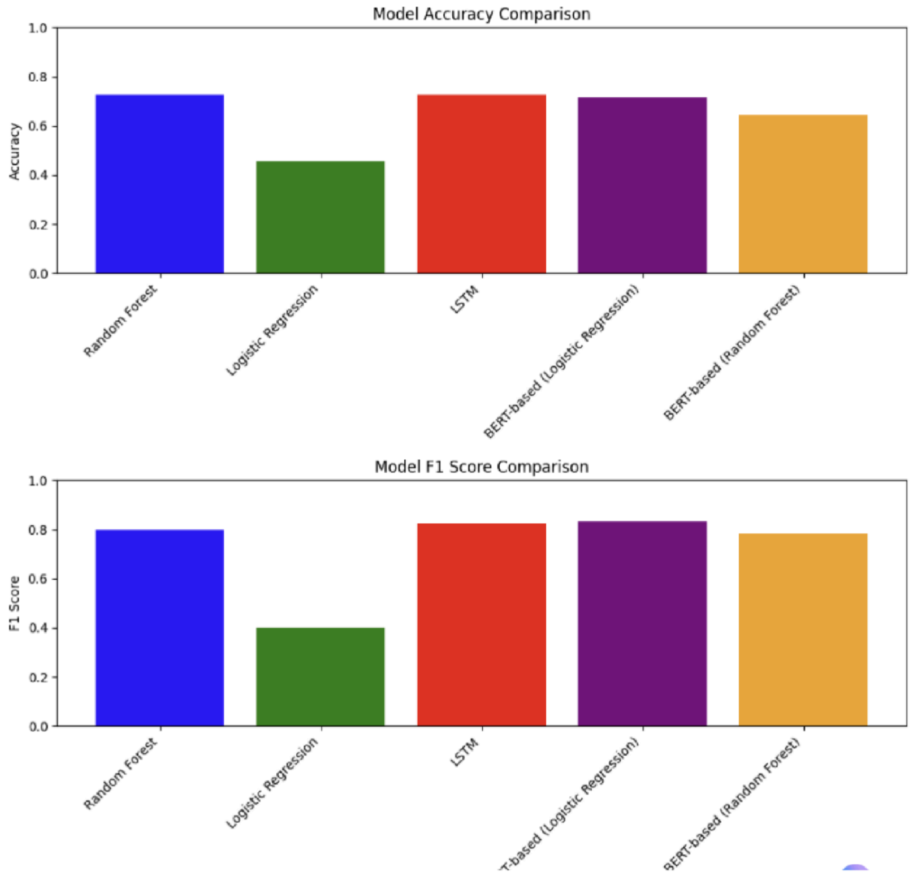
- Both the Random Forest and LSTM models using VADER sentiment performed reasonably well.
- Logistic Regression using VADER sentiment performed poorly.
- The BERT-based models, particularly the Logistic Regression classifier using BERT sentiment features, achieved competitive or slightly better F1 scores compared to the VADER-based models, suggesting that BERT's sentiment representation might be more effective for this task. However, the BERT-based models also showed a lower precision for predicting the '0' class (price decrease), indicating a higher rate of false positives for this class. This is reflected in the lower accuracy and precision for class 0 in the classification reports for the BERT-based models.
- The LSTM and the BERT-based Logistic Regression models appear to have the best overall performance based on F1 scores.

Conclusion

This study explored using social media sentiment from Bitcoin tweets to predict price direction, comparing VADER and BERT sentiment features with Random Forest, Logistic Regression, and LSTM models. Model performance varied, with LSTM and BERT-based Logistic Regression showing the best F1 scores. Both VADER and BERT sentiment showed some predictive power, with BERT-based models, particularly Logistic Regression, achieving competitive F1 scores.

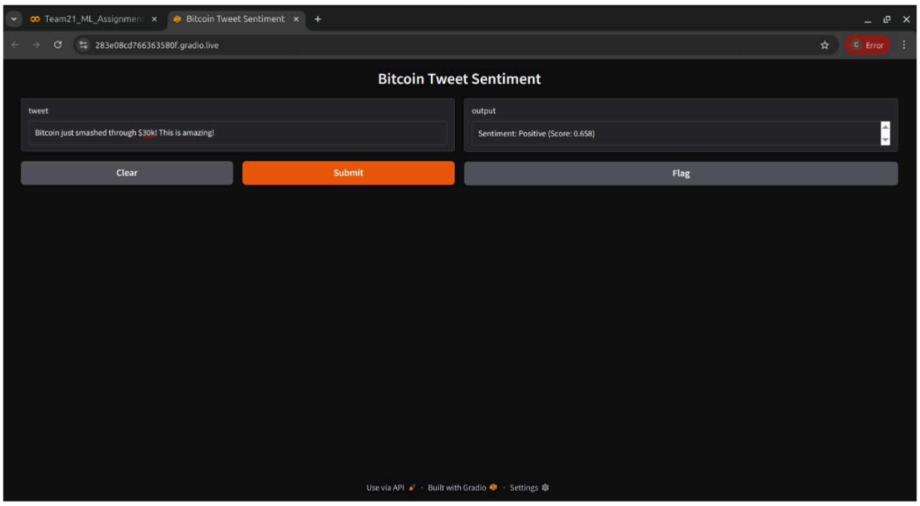
However, the small test set size and challenges in predicting price decreases highlight the need for further research with larger datasets and potentially more complex models incorporating additional features. Sentiment appears to be a promising signal, but not a sole predictor, for Bitcoin price movements.

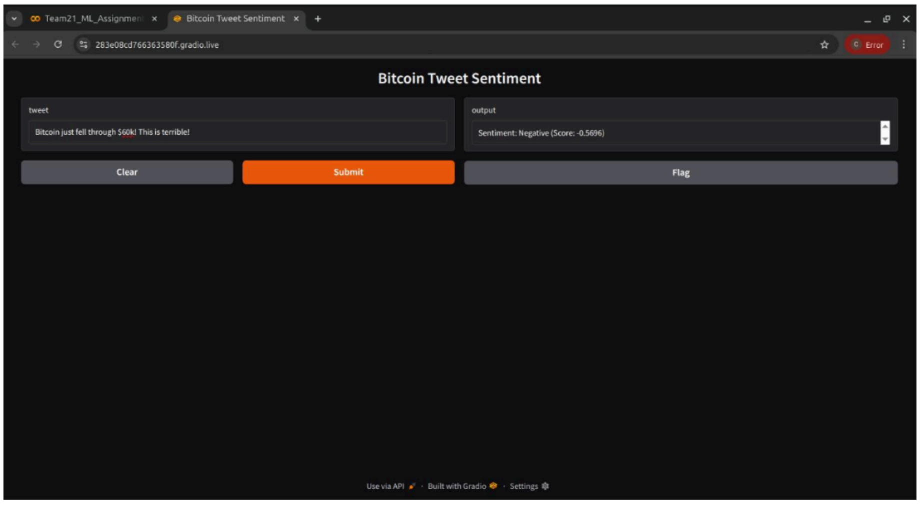
Some screenshots:

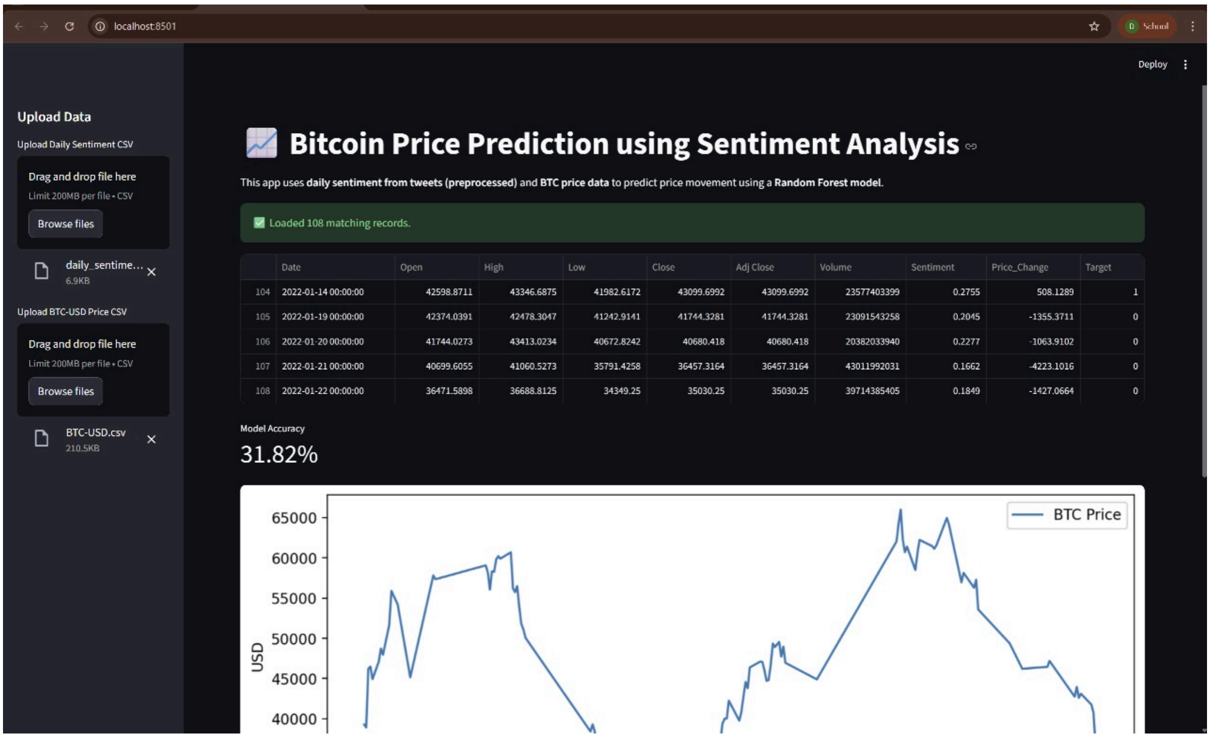


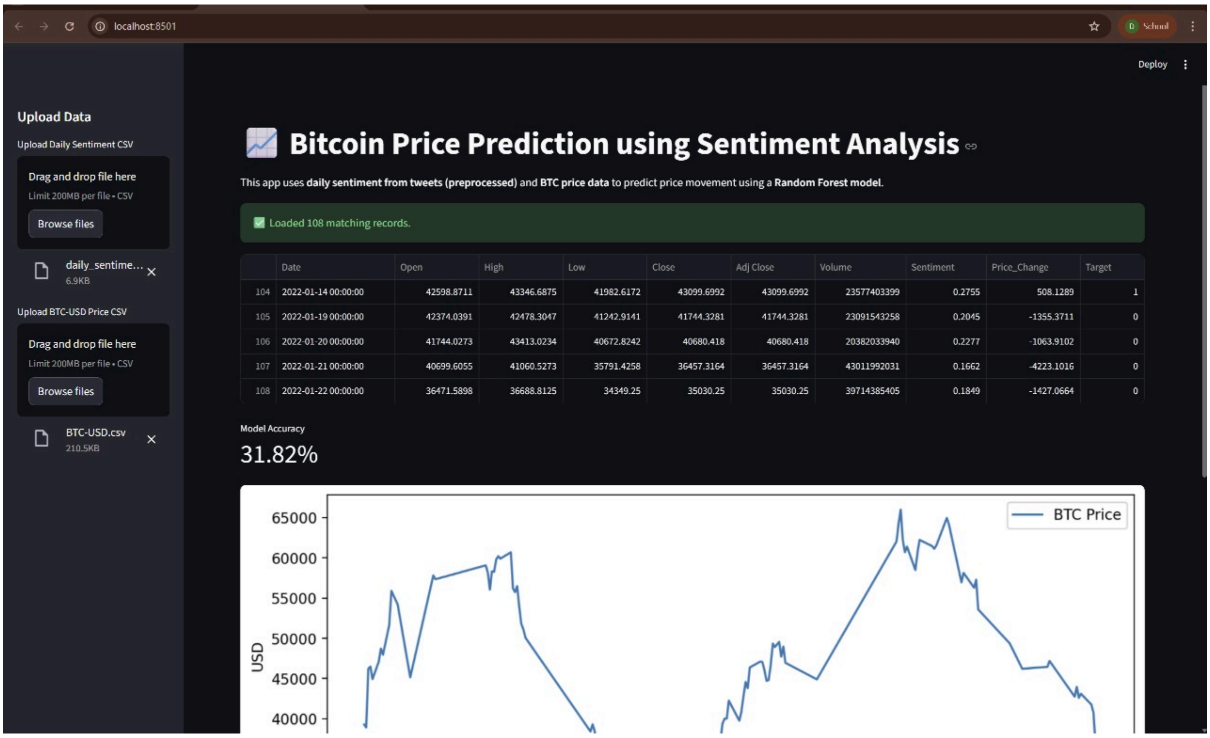
Model Accuracy Comparison
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Random Forest
Logistic Regression
LSTM
BERT-based (Logistic Regression)
BERT-based (Random Forest)
Model F1 Score Comparison
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
Random Forest
Logistic Regression
LSTM
BERT-based (Random Forest)

Some screenshots of gradio based GUI






localhost:8501
Deploy
Upload Data
Upload Daily Sentiment CSV
Drag and drop file here
Limit 200MB per file • CSV
Browse files
daily_sentime...
6.9KB
Upload BTC-USD Price CSV
Drag and drop file here
Limit 200MB per file • CSV
Browse files
BTC-USD.csv
210.5KB
Bitcoin Price Prediction using Sentiment Analysis
This app uses daily sentiment from tweets (preprocessed) and BTC price data to predict price movement using a Random Forest model.
Loaded 108 matching records.
Date Open High Low Close Adj Close Volume Sentiment Price_Change Target
104 2022-01-14 00:00:00 42598.8711 43346.6875 41982.6172 43099.6992 43099.6992 23577403399 0.2755 508.1289 1
105 2022-01-19 00:00:00 42374.0391 42478.3047 41242.9141 41744.3281 41744.3281 23091543258 0.2045 -1355.3711 0
106 2022-01-20 00:00:00 41744.0273 43413.0234 40672.8242 40680.418 40680.418 20382033940 0.2277 -1063.9102 0
107 2022-01-21 00:00:00 40699.6055 41060.5273 35791.4258 36457.3164 36457.3164 43011992031 0.1662 -4223.1016 0
108 2022-01-22 00:00:00 36471.5898 36688.8125 34349.25 35030.25 35030.25 39714385405 0.1849 -1427.0664 0
Model Accuracy
31.82%
BTC Price
USD
65000
60000
55000
50000
45000
40000


localhost:8501
Deploy
Upload Data
Upload Daily Sentiment CSV
Drag and drop file here
Limit 200MB per file • CSV
Browse files
daily_sentime...
6.9KB
Upload BTC-USD Price CSV
Drag and drop file here
Limit 200MB per file • CSV
Browse files
BTC-USD.csv
210.5KB
Bitcoin Price Prediction using Sentiment Analysis
This app uses daily sentiment from tweets (preprocessed) and BTC price data to predict price movement using a Random Forest model.
Loaded 108 matching records.
Date Open High Low Close Adj Close Volume Sentiment Price_Change Target
104 2022-01-14 00:00:00 42598.8711 43346.6875 41982.6172 43099.6992 43099.6992 23577403399 0.2755 508.1289 1
105 2022-01-19 00:00:00 42374.0391 42478.3047 41242.9141 41744.3281 41744.3281 23091543258 0.2045 -1355.3711 0
106 2022-01-20 00:00:00 41744.0273 43413.0234 40672.8242 40680.418 40680.418 20382033940 0.2277 -1063.9102 0
107 2022-01-21 00:00:00 40699.6055 41060.5273 35791.4258 36457.3164 36457.3164 43011992031 0.1662 -4223.1016 0
108 2022-01-22 00:00:00 36471.5898 36688.8125 34349.25 35030.25 35030.25 39714385405 0.1849 -1427.0664 0
Model Accuracy
31.82%
BTC Price
USD
65000
60000
55000
50000
45000
40000