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)

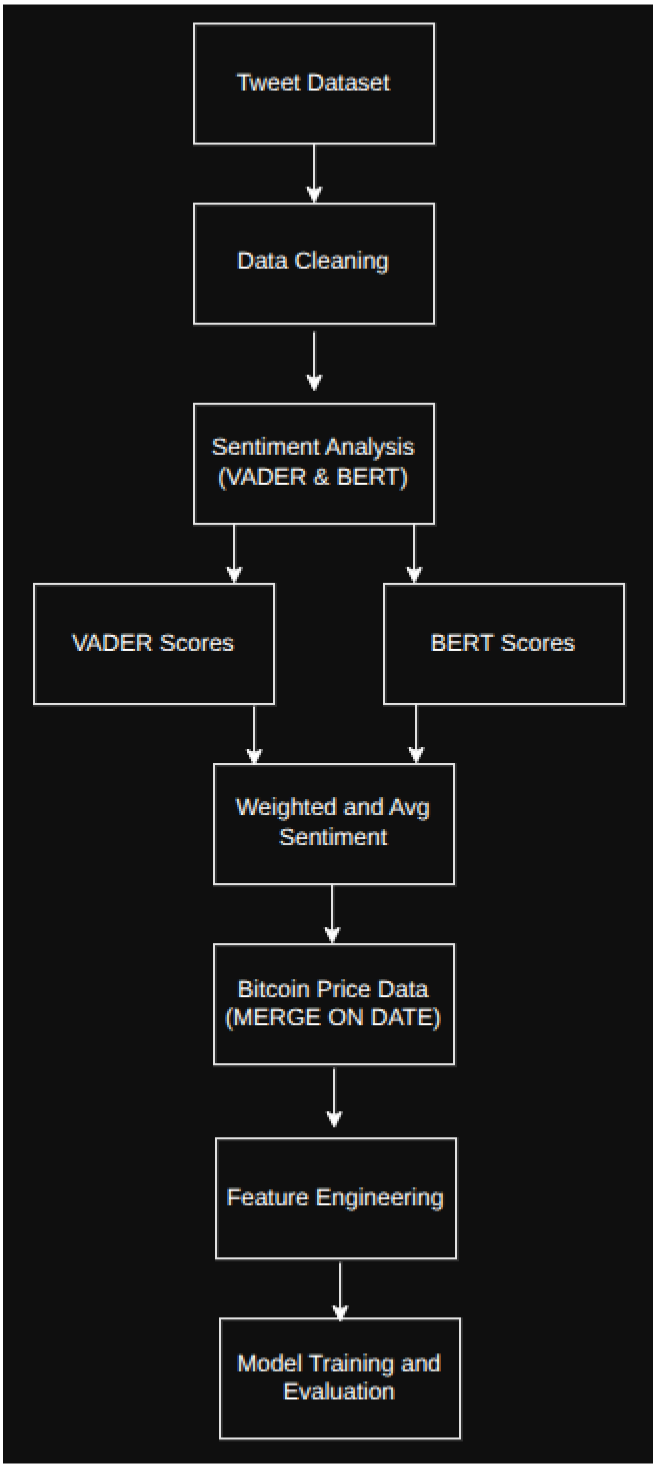
  (https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets)
- **Size:** Tweets: 1 Lakh+ samples, 13 features;

  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:

- Random Forest (VADER-based): Achieved an accuracy of 0.714 and an F1 score of 0.0. The classification report showed a precision of 1.00 and recall of 0.71 for class 0 (price down) and precision, recall, and F1-score of 0.00 for class 1 (price up). This indicates the model is good at predicting when the price goes down but completely fails to predict when it goes up.
- Logistic Regression (VADER-based): Achieved an accuracy of 0.286 and an F1 score of 0.0. The classification report showed poor performance for both classes, with a precision of 1.00 and recall of 0.29 for class 0 and 0.00 for class 1.
- LSTM (VADER-based): Achieved an accuracy of 0.0 and an F1 score of 0.0. The classification report showed a complete failure to predict either class correctly.
- BERT-based (Logistic Regression): Achieved an accuracy of 0.364 and an F1 score of 0.0. The classification report showed a precision of 0.36 and recall of 1.00 for class 0 and 0.00 for class 1. Similar to the VADER-based models, it struggles with predicting price increases.

NOTE (Explanation of low performance): The warnings about "UndefinedMetricWarning: Recall is ill-defined and being set to 0.0 in labels with no true samples" and "UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples." These warnings, particularly for class 1 (price up), indicate that in the test set, there might be very few or no instances where the price actually went up, or the models are simply not predicting any instances of the price going up. This severely impacts the F1 score for that class and suggests a significant class imbalance issue in the test set, or that the models are not capturing the signals for upward price movement and hence explains the low performance.

Conclusion

Based on the analysis, the models achieved limited success in predicting Bitcoin price direction based solely on sentiment features (both VADER and BERT).

What was learned:

- Sentiment alone is likely not sufficient for accurate price prediction: The low F1 scores across all models, particularly for predicting upward price movement, suggest that sentiment from tweets, as captured by VADER or

BERT, might not be a strong enough indicator on its own to accurately predict Bitcoin price direction. Other factors likely play a significant role.

- Class imbalance is a potential issue: The warnings in the classification reports and the consistently zero F1 scores for the positive class (price up) suggest that the dataset might have a significant imbalance between instances of price increases and decreases in the test set, or that the models are biased towards predicting the majority class (price down).
- BERT sentiment, while potentially more nuanced, didn't drastically improve performance in this context: While BERT is generally considered more powerful for sentiment analysis than VADER, using the average BERT sentiment logits as features in a simple Logistic Regression model did not lead to a significant improvement in prediction performance compared to the VADER-based models in this specific analysis, especially considering the F1 scores.

Summary: While the notebook successfully implemented and evaluated different sentiment-based models for Bitcoin price direction prediction, the results highlight the challenges of using sentiment alone for this task and point towards the need for a more comprehensive approach.