Decoding Market Sentiments: Leveraging Machine Learning to Predict Bitcoin Trends from Social Media Data

Project Problem Statement:

This project explores the potential of using machine learning to predict Bitcoin trends. The core question it seeks to answer is whether social media data, specifically tweets, can be a reliable source for predicting the fluctuations in Bitcoin prices. This investigation is rooted in the growing interest in cryptocurrency and its volatile nature. The project adds value both in the business and societal context by offering insights into market trends based on public sentiment and discussions, potentially leading to more informed investment decisions.

Background on Subject Matter:

The intersection of financial market prediction and data science is not new; however, the use of social media data for such predictions is a relatively novel approach. Traditionally, market predictions have relied on quantitative financial data, but the advent of big data and machine learning has opened doors to qualitative data sources like social media. Previous studies have shown some correlation between public sentiment and market trends, making this a promising area for applying data science techniques.

Details on Dataset:

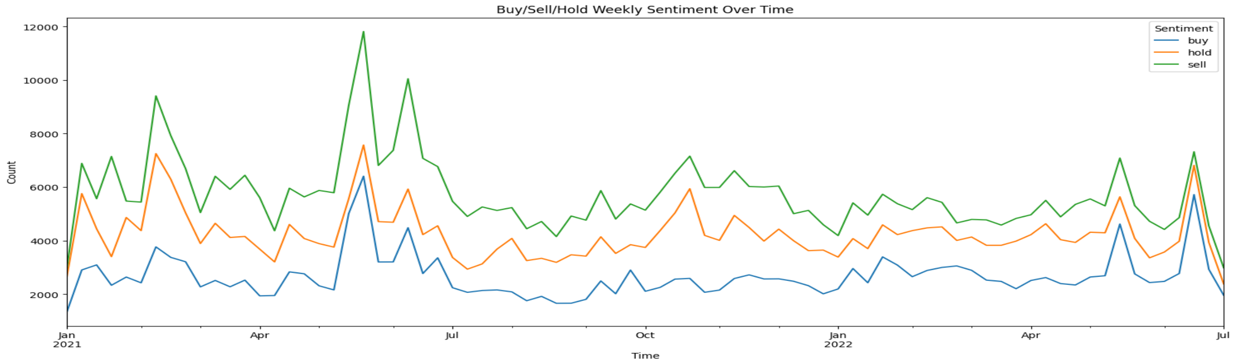
The dataset originates from Kaggle and contains over 22 million tweets related to Bitcoin, spanning 2021-2022. This large and unstructured dataset poses unique challenges and opportunities. It offers a real-time reflection of public opinion and sentiment but requires extensive preprocessing to be usable for modeling. Due to lack of resources, this report is built on a sample of 1 million tweets.

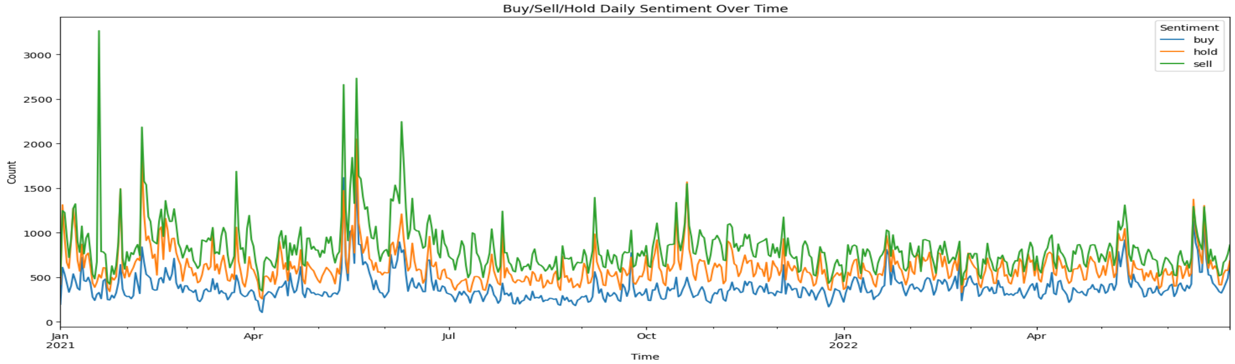
Summary of Cleaning and Preprocessing:

The preprocessing involved several steps like removing irrelevant information and standardizing text data. Extracting meaningful attributes from the tweets, like sentiment scores.

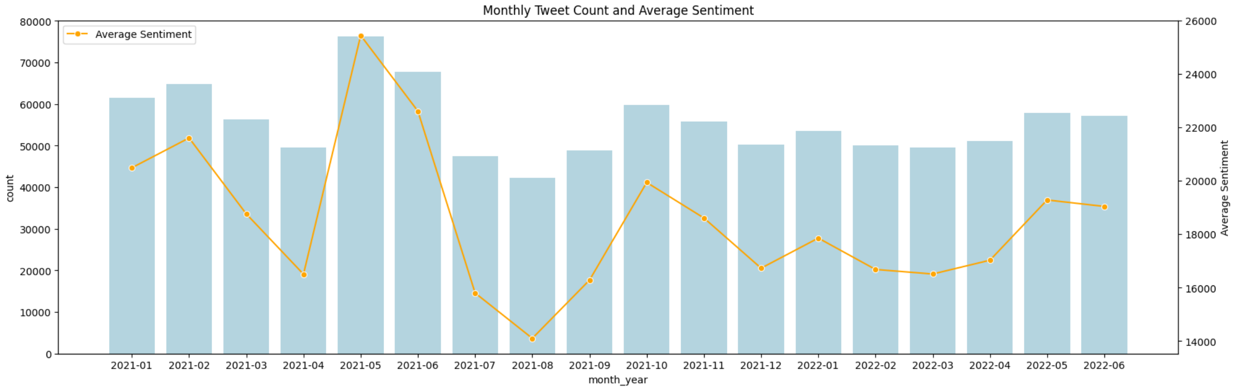
Exploratory Data Analysis (EDA): Conducting initial analysis to understand data patterns and distributions.

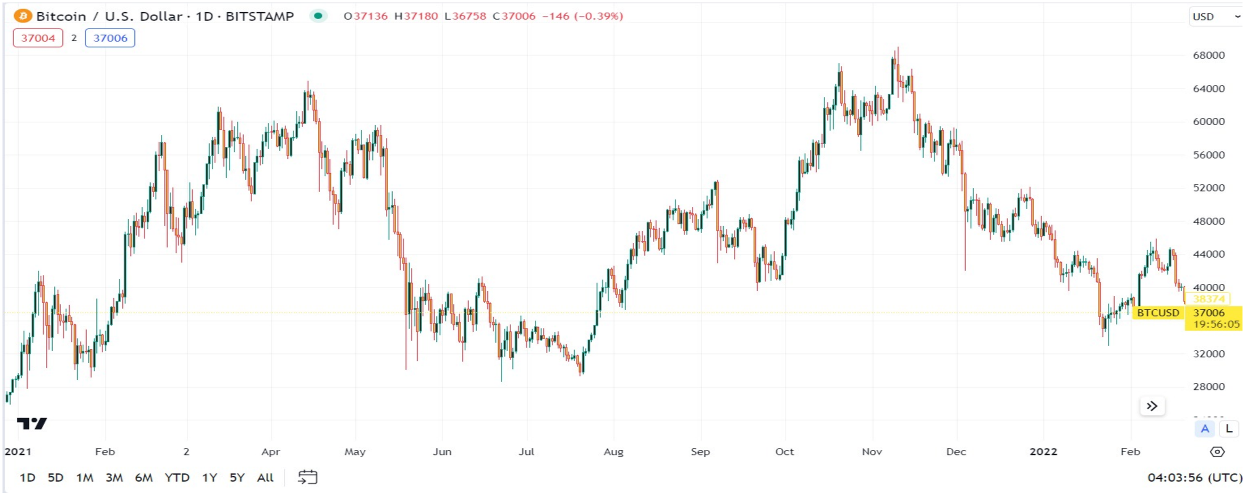
- Weekly and Daily sentiment help us understand the pulse of the market.





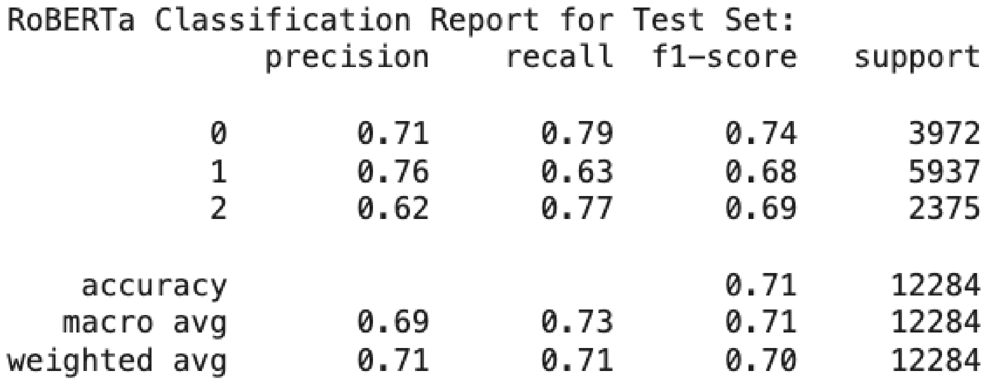
- Monthly Tweet Count and Average Sentiment viz-a-viz real world Bitcoin U.S Dollar chart





Insights, Modeling, and Results:

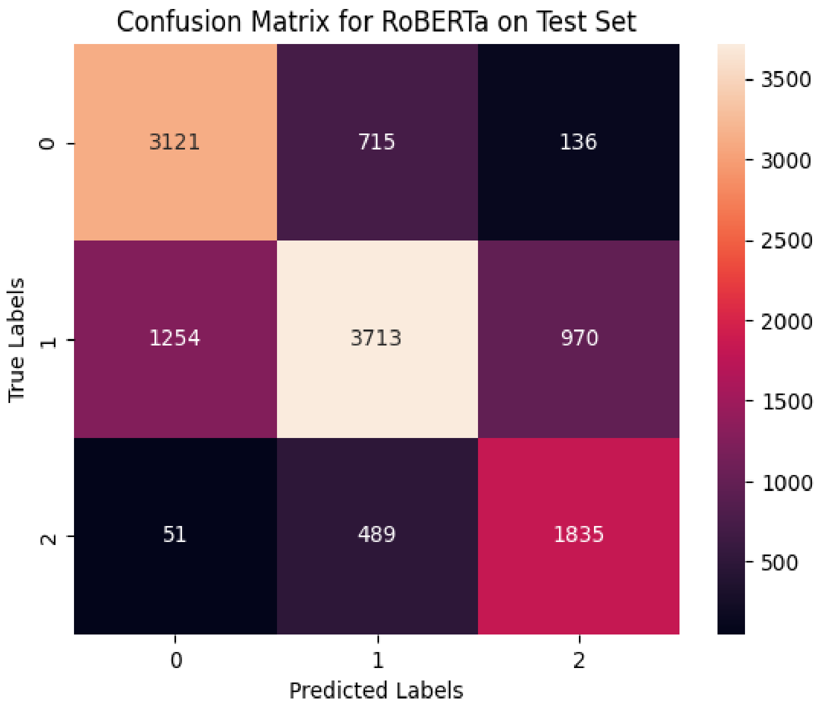
The project employed various data analysis and machine learning techniques. The initial EDA provided insights into the general sentiment. The modeling phase included the use of algorithms like Word2Vec for natural language processing and applied Logistic Regression model for prediction. Another method applied was to pre-train models on the industry benchmark; TweetEVAL. A mix of traditional and neural network models were pre-trained and the best performing model; RoBERTa, was applied for sentiment analysis.



```
RoBERTa Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.71      0.79      0.74      3972
           1       0.76      0.63      0.68      5937
           2       0.62      0.77      0.69      2375

    accuracy                           0.71     12284
   macro avg       0.69      0.73      0.71     12284
weighted avg       0.71      0.71      0.70     12284
```



Findings and Conclusions:

The project's findings were insightful, showing a notable relationship between tweet sentiments and Bitcoin price trends. However, the results also highlighted the complexity and unpredictability of financial markets. While we could capture some price movements, some nuances were missed. The practical value of the project lies in its demonstration of how social media data can be leveraged for market trend analysis. Future directions could include refining the models for greater accuracy, exploring additional data sources, data collection with a more market centric list of keywords and expanding the analysis to other fields. The project certainly stands as a testament to the potential of combining machine learning and social media data.

Data Source