SENTIMENT ANALYSIS OF CRYPTOCURRENCY TWEETS: PREDICTING BITCOIN MARKET SENTIMENT AND PRICE MOVEMENTS USING MACHINE LEARNING

CIND820: Big Data Analytics Project

Amarpreet Kaur

St No.: 501213603

Toronto Metropolitan University

Supervisor: Tamer Abdou

Date of Submission: Jun 29, 2023

Abstract

Sentiment analysis of tweets on cryptocurrency, particularly Bitcoin, provides valuable insights into market sentiment and public perception, helping investors and enthusiasts make informed decisions. By analyzing the emotional pulse of the community, we can uncover trends, gauge optimism or skepticism, and stay attuned to the ever-changing landscape of digital currencies.

I utilized Machine Learning (ML) and Sentiment Analysis (SA) techniques to predict stock market movements, specifically focusing on cryptocurrencies, with a primary emphasis on Bitcoin. My approach involved gathering data from social media sites, specifically Twitter, and combining it with cryptocurrency data from both Kaggle and the CryptoCompare API for real-time information.

To begin, I collected a dataset of tweets from Kaggle, which served as my source of Twitter data. Additionally, I obtained cryptocurrency data from both Kaggle and the CryptoCompare API, enabling me to access live data for my analysis.

I applied Sentiment Analysis to the tweets I gathered, allowing me to determine the sentiment (extreme positive, positive, extreme negative, negative, and neutral) associated with each tweet. This sentiment data was then used as a feature in my ML models.

For sentiment prediction, I implemented several ML classification models, including Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic

Regression (LR), and Gradient Boosting (GB). These models were trained using the sentiment data from the tweets to predict the next sentiment.

To predict cryptocurrency prices and changes in price (delta), I employed ML regression models such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, Gradient Boosting Regressor, and Neural Network Regressor. These models were trained to predict bitcoin price and change in price based on sentiment.

Furthermore, I utilized Time Series Models to predict the prices of five cryptocurrencies (BTC, ETH, XRP, LTC, USDC) based on their historical data.

The key novelty of my work lies in the integration of multiple SA and ML methods, with a specific focus on extracting additional features from social media, particularly public sentiment from Twitter. By incorporating sentiment analysis and leveraging various ML algorithms, I aimed to enhance the accuracy of cryptocurrency price prediction.

Overall, my project demonstrates the effectiveness of combining SA, ML, and historical data analysis to predict cryptocurrency movements, with an emphasis on Bitcoin, by leveraging social media sentiment and other relevant features.

The research questions delve into the effectiveness of sentiment analysis, compare different machine learning algorithms, assess regression models for cryptocurrency price prediction, and

evaluate the impact of incorporating social media sentiment on prediction accuracy. They address the core aspects of your project and provide opportunities for in-depth analysis and exploration.

1. How do different machine learning algorithms, such as Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting, perform in sentiment prediction for cryptocurrency-related tweets on Twitter, and which algorithm(s) yield the highest prediction accuracy?

2. What is the comparative effectiveness of various machine learning regression models, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, Gradient Boosting Regressor, and Neural Network Regressor, in predicting cryptocurrency prices and changes in price (delta) based on sentiments from tweets?

3. How does the utilization of Time Series Models improve the accuracy of cryptocurrency price prediction for Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Litecoin (LTC), and USD Coin (USDC) based on their historical data? What are the potential limitations and challenges associated with using Time Series Models for cryptocurrency price prediction?

4. To what extent does sentiment analysis of Twitter data improve the accuracy of cryptocurrency price prediction models, specifically focusing on Bitcoin?

Keywords: sentiment analysis, cryptocurrency, Bitcoin, machine learning, classification, regression, Twitter, social media, historical data, prediction.

Datasets:

Tweets from Twitter: https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

Bitcoin Data for Sentiment Analysis: https://min-api.cryptocompare.com/data/v2/histohour

Historical Crypto Data:

https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

Github Link: https://github.com/Amarpreet3/CIND-820-CAPSTONE