Title: "Bridging Social Media and Cryptocurrency: A Deep Learning-based Twitter Sentiment Analysis for Bitcoin Market Predictions"

1 Dataset Documentation:

1.1 Project Overview:

This documentation pertains to the datasets used in the project "Bridging Social Media and Cryptocurrency," which aims to leverage deep learning techniques for sentiment analysis on Twitter data to predict Bitcoin market trends.

1.2 Bitcoin Market Dataset:

The Bitcoin market dataset comprises historical data for Bitcoin from January 2021 to January 30, 2023. The data reflects market performance and can be accessed for analysis and research purposes.

1.3 Variables:

1. Date: The date of the market data record.
2. Price: The closing price of Bitcoin on the given date.
3. Open: The price of Bitcoin at the market open on the given date.
4. High: The highest price of Bitcoin during the trading day.
5. Low: The lowest price of Bitcoin during the trading day.
6. Vol: The volume of Bitcoin traded on the given date.
7. Change %: The percentage change in the closing price from the previous trading day.

1.4 Data Source:

- Data can be downloaded from [Investing.com Historical Data for Bitcoin] (https://uk.investing.com/crypto/bitcoin/historical-data).

1.5 Twitter Dataset

The Twitter dataset contains tweets related to Bitcoin, which are used for performing sentiment analysis to understand public perception and potential market implications.

1.6 Variables:

1. User: The Twitter handle of the user posting the tweet.
2. Date: The date and time the tweet was posted.
3. Location: The geographical location of the user, if available.
4. Tweet: The content of the tweet.
5. Followers: The number of followers of the user at the time of tweet capture.
6. Likes: The number of likes the tweet received.
7. Retweets: The number of times the tweet was retweeted.

1.7 Data Source:

- The Twitter dataset is available for download at [Kaggle: Bitcoin Tweets Dataset] (https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets/).

1.8 Suggested Statistical Tests for Validation

To validate the datasets and ensure robust analysis, the following statistical tests and checks are recommended:

1. Time-Series Analysis: Analyse trends, seasonality, and irregularities in the Bitcoin market data.
2. Granger Causality Test: To test whether Twitter sentiment can predict Bitcoin prices.
3. Correlation Analysis: Determine the strength and direction of the relationship between market variables and sentiment scores.
4. Chi-Square Test for Location Distribution: Test the independence of tweet locations in relation to market movements.
5. ANOVA: Compare means of Bitcoin prices over different periods of significant tweet volume to see if there's a significant difference.
6. Sentiment Score Validation: Use a confusion matrix to validate the accuracy of sentiment classification against a pre-labeled subset of the Twitter dataset.

1.9 Accessing the Datasets

Researchers and analysts can download the datasets from the provided links for personal use, academic research, or commercial analysis. The datasets are updated regularly to provide the most recent data for ongoing studies.