Cryptocurrency Curated Database – Two Pager Summary

Abstract

Purpose – The study aims to help informed decisions by investors and cryptocurrency holders by providing news tagged with relevance (relationship with cryptocurrency), sentiment (positive / neutral / negative) and strength (normal/abnormal return) scores. The study also presents the pipeline for reproducibility and recreation of relevance, sentiment and strength scores. The study explores changes in the cryptocurrency world, it does it by using the news relevance, sentiment and strength scores trends and assess its relationship with cryptocurrency prices.

Design/methodology/approach –

Event studies in general rely on having a high quality curated database of events, which is especially important in a volatile markets like cryptocurrencies. The curated data is an extraction of GDELT's quarter billion georeferenced records covering the entire world over 30 years and over 100 languages.

This study presents the pipeline for the filtering, curation and enrichment of relevant cryptocurrency news events extracted from GDELT containing crypto news from April 6, 2022 to May 1, 2022 with relevance, sentiment and strength scores.

Data for the development of the relevance score was extracted from Google (www.google.com) containing crypto news from April 6, 2022 to May 1, 2022 with 18,439 crypto related news ¥ non-crypto related news. The data was used for creating a NLP classification model to identify if news was related or not to cryptocurrency.

Data for retraining of the sentiment analysis algorithm FinBERT was extracted from https://coinmarketcal.com/en/news containing 2683 news from July 2018 to January 2022 with 1366 positive news, 1134 neutral news and 396 negative news. The data was used for retraining FinBERT classification model to assess if a news is positive, neutral or negative. The best set of parameters attained was batch_size = 16, max_len = 128 (tokeniser), epochs = 10, learning_rate = 1e-05. Data was split into 2146 news (80%) for training and 537 news (20%) for test data. The retrained FinBERT achieved the following accuracy, precision, recall, f1-score 99.487%, 99.487%, 99.487%, 99.487% in the train set and 70.764%, 70.836%, 70.764%, 70.765% in the test set.

Data for the strength score was acquired from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ website containing 22 days from 2022-04-06 to 2022-05-02 with 3 days with abnormal positive returns, 5 days with abnormal negative returns days and 14 days with normal returns. Classes were combined into 3+5 abnormal returns days and 14 normal returns

Originality/value –

Our work is original, because on top what was requested we did new columns for FinBERT sentiment analysis in order to fine tune the model. What the FinBERT model does is that for each news title, it gives you the probability of the sentiment being positive, neutral, or negative. Thus, you would choose the column with the higher probability to define the sentiment. To complement this, we decided to create a new column that would return a number greater than 1 if the sentiment is positive, and a number smaller than 1 if the sentiment is negative. Finally, we created a new column that combined the positive, negative, and neutral columns that were previously created for FinBERT into one column that ranges from -1 to 1 like the other sentiment analysis algorithms. We assigned values ranging -1 to -0.05 for negative sentiment, .05 to 1 for positive sentiment, and -0,05 to .05 for neutral sentiment`