An analysis of social media sentiment using transformer language model to predict cryptocurrency market trends

Candidate no : 245414

Introduction

The mysterious release of a whitepaper named - “Bitcoin: A Peer-to-Peer Electronic Cash System” by an unknown person/s under the pseudonym Satoshi Nakamoto in 2008 marked the introduction of the world's first established cryptocurrency- Bitcoin.
According to coinmarketcap.com there are over 19,900 cryptocurrencies that exist today. Bitcoin is the most popular with a market cap of around $400billion.According to Forbes, there are over 600 cryptocurrency exchanges worldwide facilitating the trading of cryptocurrencies. According to coinmarketcap, the largest cryptocurrency exchanges measured by daily trading volume are Binance, FTX, and Coinbase.
Unlike traditional markets which open and close during specific time periods, cryptocurrency markets run 24 hours a day, 365 days a year nonstop without pause. Due to the fact that the cryptocurrency markets never close, it becomes increasingly hard to track price movements at all times, diversify and reduce risk, and act accordingly to sudden fluctuations and volatility, something which is very prevalent in crypto markets. This has led to the emergence of trading bots.
Trading bots are a piece of automated software that essentially plugs into a cryptocurrency exchange’s API over the internet and performs trades and tasks efficiently based on certain pre-established conditions set by the user.
An algorithmic bot is a piece of code that executes trading orders such as buying and selling based on pre-established conditions set by the user. These pre-established conditions must be thoroughly thought out based on extensive research and they must also follow a strict set of rules to be as effective as possible in taking advantage of the market conditions.
The amount of information out there relating to cryptocurrency trading is extensive and different social media play a huge role in different cryptocurrencies' price actions. These include but are not limited to: Twitter, Telegram, Discord, and YouTube. Finding reliable information relating to cryptocurrency is getting increasingly difficult and to make an effective piece of automated trading that relies on this is a challenge. According to time.com in 2021, $2.8billion was taken in from cryptocurrency pumps and dumps. Due to the vast amount of disinformation in the cryptocurrency space, creating an algorithmic bot that relies more on technical analysis and technical indicators may be more effective in creating a more successful trading bot.
Twitter/Reddit are the types of social media where anyone can express their thoughts, reviews, memes, or daily life events. These tweets and feeds can affect the cryptocurrency markets due to the large number of people who are deeply into the cryptocurrency markets and publish technical

analyses and thoughts on the markets. Therefore, they become ‘reference’ sources of thoughts/analyses which leads to a majority of people following them. With this information, it is clear to say the feedback/ thoughts from social media are very important and can help create a better-involving prediction of the price movements.

Literature review

In its raw form, natural language text is meaningless to a computer—nothing more than encoded bytes. Over the past decade, strides have been made within the field of Natural Language Processing (NLP) with the aim of enabling computational systems to reason better about natural language. Sentiment analysis, as its name implies, analyses and extracts sentiment, opinion, subjectivity, and polarity from the text. Use-cases for sentiment analysis are plenty—including but not limited to product market analysis and automated flagging of positive, negative, and potentially-harmful comments on websites and social media platforms.

Sentiment Analysis (SA), also known as opinion analysis or emotion AI, can be defined as the process of calculating emotions, opinions, and attitudes scores. This score can be used for further analysis and usually, the sentiment scores are 'Positive', 'Negative', and 'Neutral'.

Sentiment analysis was first used in the 1950s and the field has been continuously evolving ever since. There is no doubt that sentiment affects an asset’s price—and as put by Baker and Wurgler: “the question is no longer, as it was a few decades ago, whether investor sentiment affects stock prices, but rather how to measure investor sentiment and quantify its effects”.

Social media provides diverse exposure to businesses and the various ways to connect to their customers. Consumers can use the product or service and can provide feedback (reviews) on said service/product. Sentiment analysis is widely used to extract valuable insights from the received feedback, which can help improve or evolve the service/product for future customers.

The rise of Bitcoin and altcoins has produced a deluge of data on social media platforms, blogs, forums, and countless other online mediums. There have been quite a few researchers trying to predict Cryptocurrency prices' behavior based on its emotions on social media platforms, such as Twitter, Reddit, YouTube comments, discord and etc. using various algorithms. Several attempts have been made that use sentiment analysis to predict the early market movement of cryptocurrencies using tweets sentiment. Now we provide an overview of approaches used in specifically the domain of cryptocurrency price prediction.

Ali Raheman et al. (8) explored different methods to calculate the sentiment metrics from a text and they come up with interpretable artificial intelligence and natural language processing methods. For their text data, the authors focused on Twitter and Reddit sources and collected about six months of data for their experiments. Their main purpose was to compare different machine learning (ML) models as well as models based on lexicons and “n-grams” and analyze their performance. Ali Raheman et al. tried experiments evaluating the sentiment from raw textual

data on a total of 21 different models including Afinn, Vader, TextBlob, GoogleNLP, AWS, Aigents, and 15 BERT-based models. A great deal of challenges that were encountered during these experiments was sarcasm in their posts or conversation, idioms used in the texts, Negations, and Non-text data. Ali Raheman et al. found a model for social media sentiment analysis in the cryptocurrency domain. They had explored the potential causal connection between social media sentiment and the price movements as an increase of expression of particular sentiment metrics two or one day before corresponding changes in price.

Ahmed Ibrahim et al. (9) have attempted to demonstrate this concept(predict early market movements of cryptocurrency) by assessing Tweets' collection, manipulation, and interpretation. More specifically, sentiment analysis and text mining methods, including Logistic Regressions, Binary Classified Vector Prediction, Support Vector Mechanism, and Naive Bayes, were considered. Each model was evaluated on its ability to predict public mood states as measured by 'tweets' from Twitter during the era of covid-19. An XGBoost-Composite ensemble model is constructed, which achieved higher performance than the state-of-the-art prediction models. According to the authors, XGBoost is an ensemble classifier that provides benefits such as no need for normalized data, scalability to larger data sets, and rule-based behavior that is easier for people to interpret. Thus, this paper aims to propose a Composite Ensemble Prediction Model (CEPM) using the notion of sentiment analysis. The CEPM framework is comprised of five stages, 1) text preprocessing, 2) Sentiment Scoring, 3) individual XGBoost classifications, 4) composite ensemble aggregation, and 5) model validation. Ahmed Ibrahim et al. started with Text Data Preprocessing (Text Stemming, Stop Word Removal) to save computational time and increase the data manipulation's overall accuracy. Then for categorizing tweets the words assigned a positive or negative relative to cryptocurrency markets. Authors have done Sentiment Analysis using Vader scoring which is a lexicon and rule-based sentiment analysis tool. For categorizing the data into positive or negative reflections for cryptocurrencies in the market they used Twitter Sentiment analysis tools including Vector Support Machines and Naive Bayes.

Ahmed Ibrahim et al. built a composite of the Extreme Gradient Boosting (XGBoost) using a majority vote over multiple cross-validations iterations which are used to achieve a better overall prediction accuracy than baseline classifiers and individual boosting algorithms.

Jacques Vella et al.(10) seek not only to predict the direction yet to also predict the magnitude of increase/decrease by utilizing sentiment extracted from tweets, and also the volume of tweets.

The authors explored and evaluated two different neural network models, one based on recurrent nets and one based on convolutional networks for predicting the direction and one additional model to predict the magnitude of change, which is framed as a multi-class classification problem. It is shown that this model yields more reliable predictions when used alongside a price trend prediction model.

They used Two main datasets in this study: Bitcoin price data and Twitter tweets and started with Data cleaning and pre-processing and Determined polarity scores using Vader Similar approaches are used by Valencia et al., Abraham et al., then they trained classifiers to predict a fluctuation in price: Change, Close, Positive polarity, Negative polarity, and Tweet Volume. Three different models, (i) using an LSTM, (ii) CNN, and (iii) Bidirectional Long Short Term Memory Cells (BiLSTM), were implemented for predicting whether the following day's closing price will increase or decrease. Jacques Vella et al. reached this conclusion whilst the BiLSTM overall outperformed the CNN and LSTM implementations for price direction prediction, the CNN outperformed the others

for the change in magnitude prediction. The experiments presented in this paper show that competitive results can be achieved with a 2-layer BiLSTM model trained on a dataset with a 1-day time lag and using seven different lagged features, meaning that each instance consists of features from tweets from the seven previous days. This model achieves a maximum accuracy of 64.18%.

Yanzhao Zou et al. (11) worked on a multimodal model that used finBERT embeddings for Bitcoin price prediction. This method takes a variety of technical indicators and correlated assets as input, the author used about 10 million tweets including text, emoji, and images, … that contained the keyword 'bitcoin' (5000 tweets per day from 2015 to 2020) this dataset is available online called 'preBit'. The author proposed a hybrid model that combined FinBERT embedding with a Convolutional Neural Network, to build a predictive model for significant market movement. The final model includes an NLP model together with a model based on candlestick data, technical indicators, and correlated asset prices author used this model to build a profitable trading strategy with a reduced risk over a 'hold' or moving average strategy. For evaluation, the author created a comprehensive measure that accounts for both precision and recall rate for the prediction output as it is their harmonic mean. This means that both false positives, as well as false negatives, are considered.

Method

Data:

We are using textual social media data from sources like Twitter, Reddit, and others if we can find relevant channels to provide related data. These data will be gathered with standard APIs for each source so we can filter the data with proper keywords. The typical data format would be like the following:

Text data, Timestamp, Source

The data collection process will be based on official Reddit or Twitter APIs and will be performed exclusively on public posts in public feeds.

Each tweet contains the following attributes: username, timestamp in DateTime format (minutes), number of retweets and favourites, tweet content, mentions (user names), hashtags, unique ID, and permalink. To protect user privacy, all information related to user identity will discard.

We first need to do preprocessing to clean the data and make it less noisy. This preprocessing step is a common practice in NLP models to ensure that the remaining word tokens are meaningful. For instance, Converting all English alphabet characters to lower case, Removing the symbols '@', '#' and all the URLs, etc.

Sentiment analysis is a text analysis method that detects polarity (e.g. a positive or negative opinion) within the text, whether a whole document, paragraph, sentence, or clause. Sentiment analysis aims to measure the attitude, sentiments, evaluations, attitudes, and emotions of a speaker/writer based on the computational treatment of subjectivity in a text.

To categorize tweets, the words will be assigned a positive or negative relative to cryptocurrency markets.VADER is a lexicon and rule-based sentiment analysis tool that can handle words, abbreviations, slang, emoticons, and emojis commonly found in social media. It is typically much faster than machine learning algorithms as it requires no training. VADER scores each tweet with a negative, positive, neutral, and compound polarity score. The compound score is a sum of the individual sentiment scores, adjusted according to a set of rules and normalised to fall within the [−1,+1][−1,+1] range. However, for the purposes of this study, only positive and negative polarity scores are included in the training and evaluation data sets.

Model:

The DeBERTa[1] model will be used in this study. This is an enhanced model of the famous BERT model and RoBERTa model which outperform most of the current language models to this date as shown in Figure 1. DeBERTa is a Transformer-based neural language model trained on large amounts of raw text corpora using self-supervised learning. Like other PLMs, DeBERTa is intended to learn universal language representations that can be adapted to various downstream NLU tasks. DeBERTa is represented using two vectors that encode its content and position, respectively, and the attention weights among words are computed using disentangled matrices based on their contents and relative positions, respectively.
Given the small number of parameters (1,5B), this model will be accurate as fast so it can be used in a fast-moving market like cryptocurrency at an affordable cost.
We will use DeBERTa model download from Huggingface with its pre-trained weights.
By fine-tuning this pre-trained model with our dataset, we will investigate the results and compare them with the previous state-of-the-art studies.

Questions that we are trying to answer is:
1- is it reliable both from accuracy and speed perspective to use AI models to read data and predict the market?
2- how relatively small models can perform compare to other studied models in this area?

[1] Decoding-enhanced BERT with Disentangled Attention

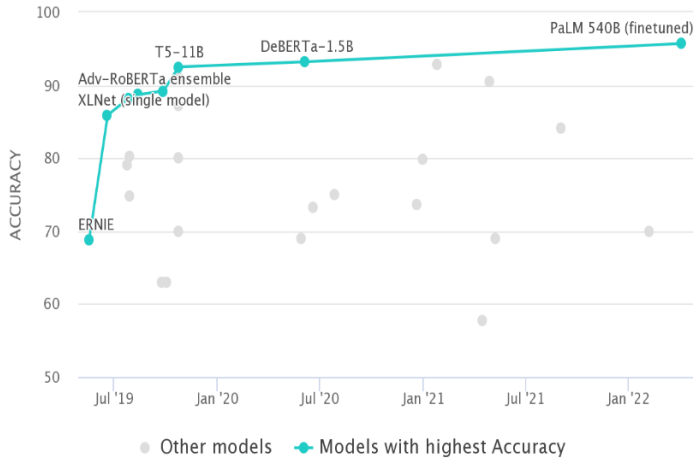

Figure 1 Natural Language Inference on RTE [12]

References

[1] (https://www.forbes.com/sites/bernardmarr/2017/12/06/a-short-history-of-bitcoin-and-crypto-currency-everyone-should-read/)
[2] (https://cybercrew.uk/blog/cryptocurrency-statistics-uk/#:~:text=in%20the%20UK-,Over%203.3%20million%20people%20in%20the%20UK%20currently%20own%20cryptocurrency,total%20population%20currently%20owns%20crypto)
[3] (https://www.statista.com/statistics/1209654/most-popular-cryptocurrency-wallets-uk/)
[4] (https://time.com/nextadvisor/investing/cryptocurrency/protect-yourself-from-crypto-pump-and-dump/)
[5] (https://coinmarketcap.com/)
[6] (https://paperswithcode.com/sota/natural-language-inference-on-rte?p=spanbert-improving-pre-training-by)
[7] (https://www.microsoft.com/en-us/research/blog/microsoft-deberta-surpasses-human-performance-on-the-superglue-benchmark/)
[8] Raheman, Ali, et al. "Social Media Sentiment Analysis for Cryptocurrency Market Prediction." *arXiv preprint arXiv:2204.10185* (2022).
[9] Ibrahim, Ahmed. "Forecasting the early market movement in bitcoin using twitter's sentiment analysis: An ensemble-based prediction model." *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021.
[10] Critien, Jacques Vella, Albert Gatt, and Joshua Ellul. "Bitcoin price change and trend prediction through twitter sentiment and data volume." *Financial Innovation* 8.1 (2022): 1-20.
[11] Zou, Yanzhao, and Dorien Herremans. "A multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin." *arXiv preprint arXiv:2206.00648* (2022).
[12] (https://paperswithcode.com/sota/natural-language-inference-on-rte?p=spanbert-improving-pre-training-by)