Bitcoin Price Prediction based on Sentiment Analysis of Relevant Tweets

J. Brook, I. Postavaru, P. Ngare

Amsterdam University College

900307SCIY: Text Mining

Jelke Bloem

03/06/2022

Abstract

This project employs sentiment analysis on tweets related to Bitcoin, in order to create a model that can correlate how the price performance might change based on public sentiment. We decided to do this to see whether there is a correlation between the behaviour of Bitcoin and people's sentiments about them since they are based on intangible services and as such could be considered volatile based on the availability of these services. We have taken a sentiment-tagged dataset of Bitcoin (BTC) related tweets, and used that to train a machine learning model in order to sentiment-tag a curated dataset which we have pulled from the Twitter API. We then compare the change in sentiment per day to the change in price of BTC per day, in order to analyse the correlation. Finally, we attempt to determine whether the sentiment influences the price or vice versa.

Keywords: Bitcoin, sentiment analysis, machine learning, natural language processing,

1. Introduction

There's no doubt that Bitcoin ($BTC), and cryptocurrencies in general have a big implied volatility. The sentiment of online discourse tends to also fluctuate, seemingly roughly in tandem with the price changes. This research aims to answer how correlated the sentiment of BTC-related tweets is with the actual fluctuation in its price. We are especially interested to see whether big dips in price cause a positive sentiment, due to the capacity to buy more for less, or whether the general sentiment will reduce, due to a loss of value of currently owned BTC.

1.1. Research questions

Is the sentiment of online discourse about Bitcoin representative of the change of its market price? If there is a correlation, which one affects the other more?

2. Related Work

Twitter Sentiment Analysis for Bitcoin Price Prediction (Abdali & Hoskins, 2021) provided us with some targets and baselines for understanding how feasible this research is and what sort of results we could expect. Pano & Kashef's (2020) paper, A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19, has also proven a great resource of information and methodology for accurately predicting sentiment and comparing it to BTC price data.

3. Dataset Description

We made use of the Bitcoin Sentiments (Alexandrayuliu, 2021) dataset from Kaggle, in order to train our sentiment-tagging model. This dataset contains 1.5 million BTC-related tweets which have been sentiment-tagged by a *Valence Aware Dictionary and sEntiment Reasoner* (VADER). VADER is a lexicon- and rule-based sentiment analysis tool built to handle words, abbreviations, slang, and emojis, which are all commonly found in tweets. This data will be useful for creating our own learned sentiment tagging model, as it can be used to train a model through regression in order to tag our curated dataset.We also web-crawled our own dataset of Bitcoin-related discourse by using the Twitter Developer API to pull relevant tweets from the week of the 17th to 23rd of May, and subsequently have stored them in a CSV file.

We theorised that a dataset only containing data after 2020 would be much better at indicating the sentiment of current twitter data. This is mainly because of the increase in retail investors at the beginning of 2021. Bitcoin's market cap increased almost 3 times from late 2020 to its peak in 2021 (Statista, 2022), and had an increase in non-zero addresses of 12% in the first quarter alone (Graffeo, 2021).

We theorise that this wave of new investors brought a different approach of investing, based on hype, influencers and trends rather than fundamentals. This influenced the vocabulary used in the circumstances of social media discussion about $BTC. To test this theory, we trained a model on a smaller dataset from 2018, also tagged with sentiment, both in polarity and intensity. This was a much smaller sized dataset with around 50.000 tweets, and we split it in a 80-20% train/test ratio. We got its native MSE which was 0.0078, and then we tested the 2021 model on the same test data, which yielded an MSE of 0.1035, which is an increase of roughly 1300%. When we tried to test the 2018 model on the 2021 data, we got a dimension error. One of the possible reasons for this could be because of the vocabulary difference between the two. It would make sense that the bigger model would be able to predict the 2018 one and not vice versa, because of size but also because of the added vocabulary. We tested this by training another model on 50.000 tweets from the 2021 dataset ,which was then able to predict the test sample designated for the big model, with an increase of around 30% in the MSE.

4. Method

First, we preprocess all of our data, which allows for more accurate training as well as sentiment tagging and collocation/frequency analysis. Next, we train a machine learning (ML) model on our preprocessed tagged data in order to apply it to our Twitter API dataset. Finally, we compare the average sentiment computed per day with the change in price of BTC, in order to determine the correlation.

4.1 Preprocessing

The preprocessing pipeline involves filtering, tokenization, part of speech (POS) tagging, and lemmatization. Firstly, links, user @s, and non-letter characters (except for spaces and sentence delimiters) are removed. Next, tokenization is achieved through the *regular expression* tokenizer provided by the Python Natural Language ToolKit (NLTK), as this model allows for high customizability. NLTK also provides part of speech (POS) tagging, which is used to improve the accuracy of the lemmatization.

4.2 Training the Model

TF-IDF (term frequency inverse document frequency) is used as a baseline for our word representation, and is given by:

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

Where the Term Frequency (TF), which is the number of times a term appears in a document, is multiplied by the Inverse Document Frequency (IDF), the log of the number of documents over the document frequency of the term.

Ridge Regression from SKLearn excels for this type of regression, as it estimates coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. In this model, the loss function is a *linear least squares* function and regularisation is given by the *l2-norm*.We ran some hyperparameter search with RandomizedSearchCV from SKLearn, to determine the optimal value for alpha, which is the constant that multiplies the L2 term, controlling regularisation strength. This model differs from normal Linear Regression due to its penalty factor, which is represented by alpha in the formula.

$$\min_w ||Xw - y||_2^2$$

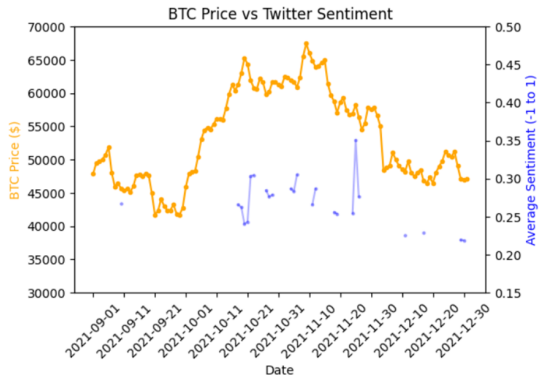
Linear regression formula

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2$$

Ridge regression formula

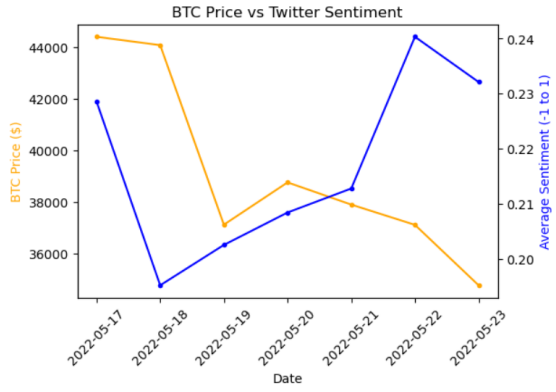
4.3 Graphing and Analysis

Fig 4.1 BTC price vs Average sentiment for September to December 2021



As expected, our results were not comprehensive. The dataset chosen between the 1st of September 2021 and 31st of December 2021 was too sparse with only 819,160 data points which does not provide enough information considering the volume of tweets made in a day is over a million .

Fig 4.2 BTC price vs Average sentiment for 17th to 23rd May 2021



On the other hand, testing the model on the larger Api tagged dataset from the shorter time period showed a better correlation.

We performed a correlation analysis on the model and the data and obtained the following results (Manu (CodingFun), 2020)

Fig 4.3 Correlation Analysis

```
x = df["BTC Price"].to_frame()
y = pd.Series(df["Sentiment"], name="BTC Price").to_frame()

corr = [x.corrwith(y, method=method)[0] for method in ["pearson", "kendall", "spearman"]]
corr
```

```
[0.4943735195451249, 0.2318840579710145, 0.3321739130434782]
```

5.Conclusion

Our goal was to look at the relationship between $BTC price performance and retail sentiment through Twitter. In our research, we found out that there are strong indicators for a shift in vocabulary related to $BTC since 2021, and that a sentiment analysis model trained on newer data would outperform one before 2021. Even though we didn't get enough data for a bigger period of trading, we found correlation between retail sentiment and price action, as well odd trends such as an increase in sentiment when the price is falling, which could indicate a general threshold for retail when $BTC is under-valued. Overall, the sentiment of Bitcoin related tweets seem to have the potential for good indicators of $BTC's market.

References

Abdali, S., & Hoskins, B. (2021). *Twitter Sentiment Analysis for Bitcoin Price Prediction*. Stanford CS229. http://cs229.stanford.edu/proj2021spr/report2/81988764.pdf

Alexandrayuliu. (2021). *Bitcoin tweets sentiment analysis dataset*. Kaggle. https://www.kaggle.com/code/alexandrayuliu/bitcoin-tweets-sentiment-analysis/data

Manu (CodingFun), J. (2020, March 3). *Moving Average Technical Analysis with Python*. Medium. https://towardsdatascience.com/moving-average-technical-analysis-with-python-2e77633929cb

Pano, T., & Kashef, R. (2020). A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing, 4*(4). https://doi.org/10.3390/bdcc4040033

Graffeo, E. (2021). *Participation in the bitcoin market grew 12% in the 1st quarter, the 3rd-highest quarterly jump ever, according to CoinDesk*. (2021, April 7). Markets Insider.

Statista. (2022, May 19). *Daily Bitcoin (BTC) market cap history up until May 18, 2022*. https://www.statista.com/statistics/377382/bitcoin-market-capitalization/

Pocs, M. (2022). *Hyperparameters in Lasso and Ridge. Towards Data Science*. Medium. https://towardsdatascience.com/hyperparameter-tuning-in-lasso-and-ridge-regressions-70a4b158ae6d