Correlation Study Between Cryptocurrency Prices and Reddit Comment Sentiment

Abstract

For our project we analyzed if comment data from the online forum Reddit can be used to predict the price development of cryptocurrencies. We first tested several sentiment analysis models and found that own trained models perform better on cryptocurrency comment data than pre-trained libraries. After using the best performing sentiment model to classify the sentiment of all reddit comments available to us, we performed a correlation analysis between the sentiment and cryptocurrency prices. According to our result, there is only a weak correlation. Finally, we build cryptocurrency price prediction models. We chose ARIMA as the baseline model and compared its performance to an ARMAX and VAR model for which we added the sentiment as an additional feature. Our comparison shows that adding the sentiment as a feature does not increase the predictive power of the models.

1. Introduction

1.1 Background

Cryptocurrencies are digital assets that utilize cryptographic technology to secure transaction records. These assets are typically decentralized using blockchain technology. Bitcoin was the first established decentralized cryptocurrency introduced in 2009. Many alternative cryptocurrencies have been created since then to rival Bitcoin. As of January 2021, there were more than 4000 cryptocurrencies in the market. The emergence of an alternative to fiat currencies in the form of cryptocurrencies coupled with the volatile nature of cryptocurrencies' values has attracted a lot of speculation and discussions on social media and online forums such as Reddit.

According to a study by Pulsar (2018), the price of Bitcoin is correlated with the volume and sentiment of comments on social media. The study found that a rise in online conversation volumes on Bitcoin preceded spikes in its price by about 2 days. Our project builds upon these initial findings and analyzes whether online sentiment has a similar effect on other alternative cryptocurrency prices. For our work we chose 5 cryptocurrencies for a comparison to Bitcoin: Ethereum, Monero, XRP, IOTA and Neo.

1.2 Approach and methodology

We first scraped unstructured comment data from the forum Reddit (https://www.reddit.com). Next, we performed a sentiment analysis on the comment data. As cryptocurrency comments on Reddit may not conform to standard English, contain many new words and discuss very specific topics, it is likely that sentiment analysis by standard pre-trained NLP libraries (i.e., NLKT, Textblob & Stanford CoreNLP) may not be able to predict the sentiment of Reddit comments well. To evaluate the performance of each model, we manually classified 5000 comments into negative, neutral and positive sentiment. Since we expected a limited performance by standard pre-trained NLP libraries, we also build our own sentiment models using the 5000 labeled comments.

Hypothesis 1: Models from standard pre-trained NLP libraries perform poorly on cryptocurrency comments on Reddit compared to models specifically trained on cryptocurrency comment data.

After evaluating all models, we chose the best performing model to classify all our scraped data into negative, neutral, and positive sentiment. We used the new sentiment feature to calculate the correlation between the sentiment and the price movement of the selected cryptocurrencies.

Hypothesis 2: Reddit sentiment correlates with prices strongly.

With the continued rise in popularity of cryptocurrencies, many cryptocurrencies caught the attention of institutional investors and established companies. In early 2021 news media (Kovach, 2021) reported that Tesla invested USD 1.5bn in Bitcoin and started accepting Bitcoin as payment for its products. With the entry of institutional investors and companies into the cryptocurrency market, we believe that social media sentiment generated by retail investors becomes less indicative of price movements. Hence, for our analysis we compared the sentiment price correlation for two periods. The first period covers the cryptocurrency price frenzy and subsequent crash in 2017 to 2018 during which cryptocurrencies were not yet popular amongst institutional investors and the second period covers 2019 to early 2021.

Hypothesis 3: Correlation of reddit sentiment with prices will be stronger from 2017 to 2018 as it is less mainstream compared to 2020 to 2021.

In the current market, institutional investors primarily focus on established mainstream cryptocurrencies such as Bitcoin and Ethereum. Hence, we expect a difference between mainstream and non-mainstream cryptocurrencies in terms of their correlation with our sentiment analysis. Therefore, we compared the correlation results amongst cryptocurrencies to identify differences.

Hypothesis 4: Reddit sentiment correlates more strongly with prices of non-mainstream cryptocurrency than mainstream cryptocurrency.

To better understand the results of our sentiment analysis and the correlation, we performed a deep dive analysis of the results for two coins. We chose Bitcoin as a mainstream coin and Neo as a non-mainstream coin.

Finally, we analyzed if the Reddit sentiment has predictive power when used in a forecasting model. For this analysis, we compare the performance of ARIMA, a pure time series model, with the performance of ARIMAX and VAR, two models that allow us to use the sentiment feature on top of price time series information. For our analysis we again focused on the two coins Bitcoin and Neo.

Hypothesis 5: Reddit sentiment performs as a strong predictor over cryptocurrency prices forecasting.

2. Data collection

2.1 Reddit comments data

Reddit has the official PRAW API (https://praw.readthedocs.io) for data scraping. However, the PRAW API only provides access to recent commentary data. Hence, we had to rely on the PushShift API (https://github.com/pushshift/api) that is currently in active development to scrape historical commentary data. The PushShift API takes in arguments in a base URL for the title, search term, time range of posting etc. An example of the URL to search for submissions[1] is as follows:

https://api.pushshift.io/reddit/search/submission/?q=bitcoin&after=180d&before=179d&sort=asc

Using the URL as explained above, the commentary data can be retrieved from Reddit in a JSON format.

As discussed before, we focused on 5 cryptocurrency alternatives to Bitcoin. For scraping purposes, we not only used the names of the cryptocurrencies, but also the abbreviations.

Table 1. Search terms for crypto currencies

CRYPTO CURRENCY	ABBREVIATION
Bitcoin	BTC
Ethereum	ETH
Monero	XMR
Ripple[2]	XRP
IOTA	n/a
Neo	n/a

Reddit has several sub-reddits[3] on which people discuss specific topics. For cryptocurrencies, there are a wide range of different sub-reddits which can be classified into two types: currency unspecific and currency specific sub-reddits. For our project, we used the following currency unspecific and currency specific sub-reddits as presented in table 2.

Table 2. Sub-reddits

CURRENCY UNSPECIFIC	CURRENCY SPECIFIC
CryptoCurrency	Bitcoin
CryptoMarkets	btc
binance	ethereum
altcoin	XRP
IOTA	Monero
SatoshiStreetBets	IOTA
	NEO

Reddit users that are active on a coin specific sub-reddit preselected themselves to a specific coin. Hence, it is reasonable to expect that coin specific sub-reddits could be more biased in their sentiment. Hence, we chose to scrape comments from different sub-reddits in order to generate a more diverse dataset.

Using the described approach, we scraped in total 1.9 million comments from Reddit with 5 features.

Table 3 – Reddit data feature set

ATTRIBUTE	DATA TYPE
coin_name	String
created_time	Integer

[1] Reddit distinguishes between submissions and comments. Submissions are the initial post and comments are follow-up comments. For the purpose of this project, we do not distinguish between the two.

[2] Ripple refers to the company that created XRP and not the currency itself. However, on Reddit the terms are frequently used interchangeably.

[3] Sub-reddits are topics specific sub-forums of Reddit

message_body	String
score	Float
total_awards	Integer

The score relates to the number of upvotes a comment received from the reddit community. Reddit users can upvote comments if they deem the comment content to be relevant. The total awards relate to the number of awards a comment received from the Reddit community. Reddit users can give topic specific awards to comments if they deem their content to be exceptionally relevant for specific topics.

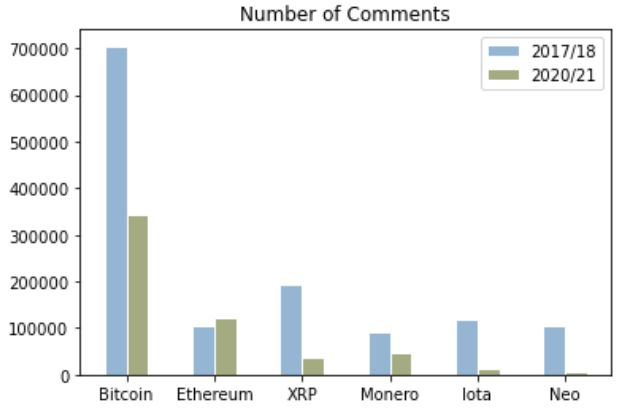
2.2 Crypto currency price data

For the recent period covering 2019 to 2021, we scraped the crypto currency price data from the Alpha Vantage API (https://www.alphavantage.co).
For the period covering 2017 to 2018, we relied on a historical price dataset published on Kaggle (SRK, 2021) since most APIs do not cover an extensive historical period. The historical price data from Kaggle was originally collected from coinmarketcap.com, a site that reports daily prices and market capitalization of all cryptocurrencies.

2.3 Data exploration

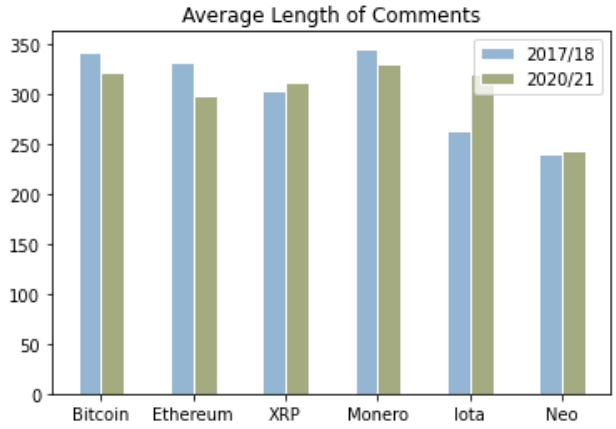
The number of comments and the average length of comments has been analyzed during the data exploration.

Figure 1 – Number of comments per cryptocurrency



One observation is that Bitcoin counts more comments (1,048k) than the other five coins combined (831k). Another observation is that there are much more comments (1,314k) for the 2017/18 period before the crypto-crash than for 2020/21 (566k). The largest decrease of comments are seen for Neo (-92%), Iota (-90%), and XRP (-81%). However, this does not mean that less comments indicate less popularity because the market capitalization of these coins has actually increased from Feb-2018 to Feb-2021: Neo (+28%), Iota (+61%), and XRP (+73%) based on https://coinmarketcap.com/.

Figure 2 – Average length of comments per cryptocurrency



The average length of comments has been analyzed as well but there is no significant change between both timeframes.

3. Sentiment analysis

3.1 Data preprocessing

Before we performed our sentiment analysis, we first applied the following preprocessing steps to the comment data:

- Remove non alphanumeric characters
- Remove all punctuation
- Remove dashes and concatenate words
- Remove underscores
- Remove digits
- Remove words with three consecutive letters
- Remove stopwords
- Lemmatize

These preprocessing steps helped us to clean the data and improve the performance of the sentiment models. However, at the same time we see that many comments only included very short text that would be removed during pre-processing. For example, our 5000 labeled comments were reduced to 3958 usable datapoints after pre-processing.

3.2 Pre-trained libraries

We chose 3 pre-trained libraries for our evaluation: NLTK, Stanford CoreNLP and TextBlob. We set-up each model so that it returns either a negative, neutral or positive

sentiment class for every comment. Additionally, we build an ensemble model using all three libraries.

3.2.1 NLTK:

NLTK (Natural Language Toolkit, https://www.nltk.org) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Table 4 – NLTK performance

Accuracy	Macro F1	Weighted F1
0.43	0.39	0.43

3.2.2 Stanford CoreNLP:

Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/) provides a set of natural language analysis tools written in Java. It can take raw human language text input and give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize and interpret dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases or word dependencies, and indicate which noun phrases refer to the same entities.

Table 5 – Stanford CoreNLP performance

Accuracy	Macro F1	Weighted F1
0.36	0.35	0.36

3.2.3 TextBlob:

TextBlob (https://textblob.readthedocs.io/en/dev/) provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Table 6 – TextBlob performance

Accuracy	Macro F1	Weighted F1
0.41	0.35	0.40

3.2.4 Compound (ensemble):

Lastly, we build an ensemble model out of the results generated by the pre-trained libraries. We chose to aggregate the scoring using the following weighting as the Stanford CoreNLP model tends to return discrete values of 1, 0, -1 which may have a larger impact than the other two models if a simple average is taken:

$$compound = 0.4 \times NLTK + 0.4 \times TextBlob + 0.2 \times Stanford\ CoreNLP$$

Table 6 – Compound performance

Accuracy	Macro F1	Weighted F1
0.41	0.38	0.41

3.3 Self-trained models

In addition to the pre-trained libraries, we also build our own trained models using the 5000 labeled comments. Since the 5000 comments got reduced to 3958 usable datapoints after pre-processing, we split the 3958 comments into a training set of 3562 comments and a test set of 395 comments. 3958 labelled comments are a limited dataset for training a sentiment classifier. In order to have as much training data as possible we decided to keep the test set small. However, we are aware that a small test set reduces the robustness of our performance evaluation.

We chose three models for our evaluation: Multinomial Naïve Bayes (MNB), Linear SVC (LSVC) and one-to-rest XG Boost (XGB). For each model, we used TF-IDF as the vectorizer and either simple random over-sampling or SMOTE for the minority classes depending on the model performance.

Table 7 – Trained models performance

Model	Accuracy	Macro F1	Weighted F1
MNB	0.49	0.47	0.50
LSVC	0.50	0.48	0.51
XGB	0.51	0.47	0.51

Finally, we experimented with majority vote ensemble structures to further tune the model performance. After experimenting with several structures, the best preforming structure consisted of combining the linear SVC model, the on-to-rest XG Boost model and the compound model into a majority voting ensemble structure. Since XGB performed the best on a stand-alone basis, we chose XGB to be the tiebreaker.

Table 8 – Ensemble model performance

Accuracy	Macro F1	Weighted F1
0.54	0.51	0.54

3.4 Performance evaluation

The performance evaluation shows that our ensemble model performed best across all models. Hence, we used our ensemble model to predict the sentiment of our entire dataset of 1.9 million comments. The model classified 61% of all comments as neutral, 29% as positive and 10% as negative sentiment.

4. Correlation analysis

For comparison, VADER is used as an unsupervised model for sentiment analysis as it is a tool that is specifically attuned to sentiments expressed in social media (Luis et al., 2020). With the sentiments predicted from both supervised and unsupervised model, daily sentiment rate is calculated by:

$$\frac{No.\,of\ positive\ comments}{No.\,of\ positive\ comments + No.\,of\ negative\ comments}$$

Daily sentiment rate above 0.5 will be considered positive, while 0.5 is considered neutral and less than 0.5 is considered negative.

For null and nan values, the daily sentiment is assumed to be neutral (0.5).

4.1 Spearman Correlation Coefficient

As the relationship between close price and sentiment rate is non-linear, Spearman is used to determine monotonic relationship and dependency of the variables (Jason, 2018).

Table 9 – Spearman Correlation Coefficient (p-value)

	Model	2017	2018
BTC	Supervised	0.723 (0.000)	0.454 (0.000)
	Unsupervised	-0.249 (0.000)	0.191 (0.000)
ETH	Supervised	-0.207 (0.000)	0.558 (0.000)
	Unsupervised	0.066 (0.209)	0.120 (0.022)
XMR	Supervised	0.193 (0.000)	0.312 (0.000)
	Unsupervised	-0.041 (0.430)	-0.133 (0.011)
Iota	Supervised	0.217 (0.001)	0.157 (0.003)
	Unsupervised	-0.095 (0.147)	-0.004 (0.945)
XRP	Supervised	-0.160 (0.002)	0.049 (0.349)
	Unsupervised	0.013 (0.807)	0.126 (0.016)
Neo	Supervised	0.362 (0.000)	0.125 (0.016)
	Unsupervised	0.427 (0.000)	-0.088 (0.092)

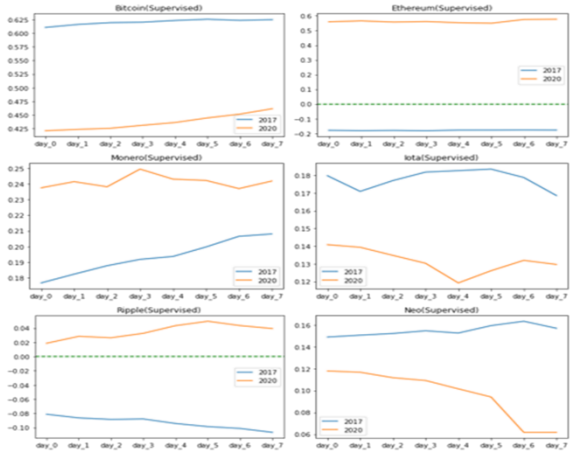
Table 9 shows results at 95% confidence interval, values highlighted in red implies that close price is independent to sentiment rate. The results show that the supervised ensemble model could extract more dependency between sentiment rating and close price than the unsupervised VADER model. In addition, except for ETH and XMR, the correlation coefficient is larger in 2017 than 2020, confirming Hypothesis 2 and 3.

However, it was noted that the mainstream coins (BTC and ETH) have larger correlation coefficients than non-mainstream coins. This result does not align with Hypothesis 4.

4.2 Lagging sentiment rate

Since supervised models gives better correlation, its sentiment rates are lagged by 7 to explore time effect of sentiments on close price.

Figure 3 – Correlation between sentiment rate lags and close price



From Figure 3, the 2017(blue line) data corresponds to the results of Spearman that correlation is stronger in 2017 than that of 2020 (orange line) except for ETH and XMR. Lag data does not show a significant trend except for XMR which increases with lag. This might be due to the speculative nature of cryptocurrency market in 2017, making investors more reactive to immediate online sentiments. In 2020, generally, the correlation appears to increase after about 2 days. The lag and weaker correlation in general could be due to the participation of more institutional investors (Olga, 2018) which are likely to adopt more assessment tools, thus less reactive to immediate online sentiments.

5. Deep Dive Investigation

In this section, Bitcoin will be selected as the representative of the mainstream cryptocurrency while Neo will be the representative of the non-mainstream cryptocurrency based on their market capitalization where Neo market capitalization is only at $5 billion while bitcoin market capitalization is at $923 billion (https://coinmarketcap.com/). This section serves to provide a deeper analysis to give an understanding on what kind of topics of interest are being discussed in general and

possibly give insight on the performance of the sentiment prediction model.

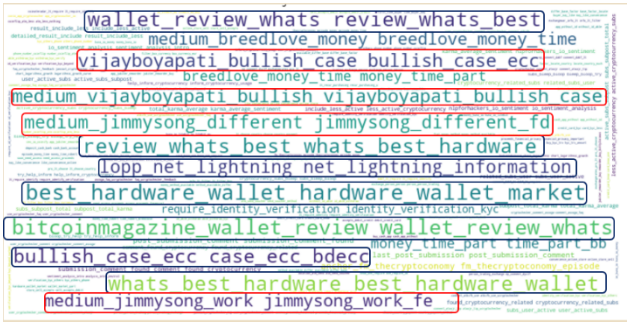
5.1 Word Cloud Analysis

Word clouds is a visual representation on the frequency of words that appear within a given text which would give general insight on what are the commonly discussed topics by redditors on Bitcoin and Neo. The generated unigram and bigram word clouds contain mostly frequently used verbs and words without context thus does not produce any interesting findings. Thus, the trigram word clouds will be discussed in this section.

5.1.1 Trigram Word Cloud – Bitcoin

As seen in Figure 4, the trigram word cloud generated from all comments on Bitcoin by redditors shows two large themes of discussion. Firstly, as highlighted in red, redditors are sharing their opinions of articles written by popular bitcoin advocates Vijay (2018) & Jimmy Song (2021). Another theme that redditors are discussing about are technical aspects of Bitcoin such as the lightning network upgrade to speed up transactions (2021) and reviews on hardware wallets for bitcoin.

Figure 4 – Trigram Word Cloud for Bitcoin



These discussions may be crucial for Bitcoin's future mainstream adoption by the general public but they are not likely to correlate directly to short term Bitcoin price changes thus would be manually labelled as neutral comments. However, by using pre-trained sentiment packages, these discussion comments may be labelled as positive or negative comments thus might be one contributing cause to the low prediction accuracy of our sentiment prediction model.

5.1.2 Trigram Word Cloud – Neo

The trigram word cloud for Neo shows that redditors are predominantly discussing about: (1) founder and Neo-related news, (2) investment topics, and (3) technical aspects. In order to understand what specific topics are being discussed, topic modelling will be done next.

Figure 5 – Trigram Word Cloud for Neo

5.2 Topic Modelling – Latent Dirichlet Allocation (LDA)

The positive & negative comments for both Bitcoin & Neo will be analyzed separately to further investigate the topics discussed that are of positive & negative sentiment, respectively. The following section will detail the implementation of Latent Dirichlet Allocation (LDA) to identify topics for each analysis group.

As part of text pre-processing before LDA is implemented, comments that are might be generated by moderator bots or malicious spam bots are removed by filtering by the send_replies flag and whether the comment contains the words moderators. Several common words that are used by spammers and links are also removed to improve data quality for better topic modelling results. To obtain the optimal k topics for each analysis group of each coin, a grid search is conducted for values of k from 2 to 15 and the c_v measure is used to calculate the coherence score for each k value. The above grid search is conducted with a subset of 5000 comments of each analysis group if the original dataset is over 5000 comments due to high computational cost of performing a grid search.

From Figure 6, the optimal k value for positive comments & negative comments for Bitcoin is 3 and 6 respectively and from Figure 7, the optimal k value for positive comments & negative comments for Neo is 6 and 13, respectively.

Figure 6 – Coherence score for k topics (Bitcoin)

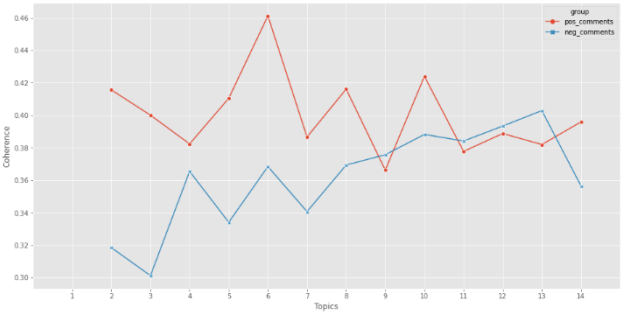
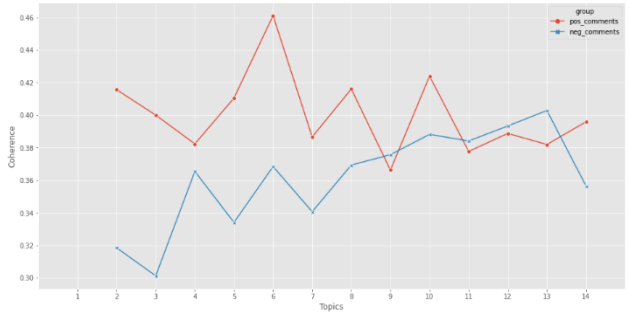


Figure 7 – Coherence score for k topics (Neo)



With the optimal k values, LDA is implemented for all comments in each analysis group to obtain the top 30 most salient words of each modelled topic.

5.2.1 Analysis of LDA topics - Bitcoin

From the top 30 most salient words of each topic given by LDA, relevant words that give context are picked out to obtain the postulated topic.

The LDA topics from positive comments for Bitcoin are as shown in Table 10. With investment related words like 'stock market', 'invest', long term', 'bull run' & 'store_vlaue' in topic 1, these are likely to be comments from mainstream retail investors who are taking an interest into Bitcoin due to its growing popularity as a viable investment tool.

For topic 2, there are a mix of common investment terms such as 'time_high', 'worth' & 'value' and more technical investment terms such as 'market_cap', 'world','sentiment'. This may be from retail investors discussing Bitcoin's technical investment metrics and the world's sentiment on Bitcoin. Bitcoin specific words such as 'wallet', 'platform_like' and 'global_reserve' which indicates that redditors are engaging in a more technical discussion on wallets & trading platforms for Bitcoin and positive discussion on Bitcoin being touted as a global digital reserve. Thus, topic 2 is likely to be a more technical & researched driven discussion on Bitcoin trading and usage.

Topic 3 contains mostly internet slang that is specific to cryptocurrency discussion such as 'pump', 'boom', 'moon', 'dyor','tldr','minerd'. It also contains recent news that are perpetuated by mainstream media such as 'gold' where Bitcoin was compared to gold and 'trading' where more retail investors are moving into trading bitcoin. Topic 3 is likely to be short hyped-up comments without backed up research from redditors who are Bitcoin loyalists.

The 3 topics of the positive comments from Bitcoin generally are investment related & serves as a good indicator of short term Bitcoin prices.

Table 10 – LDA topics for Bitcoin positive comments

No.	Relevant words in Topic	Postulated Topic
1	stock_market, long_term, future, bull_run, invest, see_potential, store_value, gain	Discussion between mainstream investors
2	worth, value, wallet, time_high, sentiment, market_cap, world, platform_like, global_reserve	Technical discussion between hardcore bitcoin supporters
3	dyor, tldr, minerd, doge, moon, gold, lol, boom, trading, pump, million	Hyped up messages from redditors

The LDA topics from negative comments for Bitcoin are as shown in Table 11. Topic 1 and 5 are likely to be either comparison of bitcoin to other alt-coins or discussion of alt-coins or other hyped up news such dogecoin bull run and whether hex is a scam (Turner; Terence 2020). These topics do not have Bitcoin as the subject of discussion thus should be labelled as a neutral comment but are labelled as negative instead thus might be contributing source of error to the sentiment prediction.

Blockchain technology specific words such as 'block_chain', 'transaction',' decentralize', 'developer', 'smart_contract', 'network', 'transaction_fee', ' block_reward', 'miner' appears in topics 2 & 6. Redditors are concerned about current technical issues that bitcoin faces like high transaction fees (Colin, 2021), slow transaction speed or skeptical about the advantages of smart contracts and how blockchain being decentralized. While these topics are indicative of the low confidence for current Bitcoin widespread adoption, but it also provides a platform of active discussion for such issues to be resolved which might be positive in the long run. Additionally, short term price changes are less likely to be affected by negative comments on technical aspects of Bitcoin since retail investors are likely to more concerned about factors that might cause short term price changes. Thus, comments of such topics are mislabeled as negative, leading to further errors in sentiment prediction.

For topics 3 & 4, it is likely to be about recent popularized political, social and investment news related to Bitcoin. Words such as 'government',' bank', 'china' might be indicative of negative news such as government or banks denouncing Bitcoin (Aftab & Nupur, 2021) or China's dominance in Bitcoin mining (Shawn, 2021). Redditors might also be concerned on the large fluctuations in prices due to rampant trading of bitcoin from words such as 'fiat', 'money','profit','usd','wrong' & 'pump' indicative of the issue of pumping Bitcoin prices up by hype for a profit. There is also likely to be disagreement on Bitcoin having store value or being a legitimate currency despite strong

proponents from well-known advocates such as Elon Musk. These topics are likely correlate well with short term Bitcoin price thus are likely to be labelled correctly as negative.

Table 11 – LDA topics for Bitcoin negative comments

No.	Relevant words in Topic	Postulated Topic
1	nano, alt, coin, high, invest, gain, investment, litecoin	Comparison with Alt-Coins
2	smart_contract, blockchain, defi, transaction, nano, network, user, wallet, exchange, mine, binance, decentralize, platform, developer	General discussion on blockchain technology
3	sell, run, bank, government, pump, money, usd, profit, wrong, data, china	Political & social issues regarding bitcoin
4	store_value, currency, tehter, gold, value, inflation, use_case, argument, asset, utility, fiat, future	Bitcoin as a commodity & its future adoption
5	doge, mining, hex, news	Hyped news related to bitcoin
6	fee, transaction_fee, trader, gamble, miner, moon, block_reward	Technical issues related to bitcoin

5.2.2 Analysis of LDA topics – Neo

The k values for Neo are 6 and 13, which represents the number of positive (6) and negative (13) sentiment topics. Table 12 and 13 show the identified topics for Neo.

Compared to Bitcoin, the positive topics are not only investment-related but also deal with technological developments (no. 2, 4) and with Neo's potential value to China's society (no. 3) as largest Chinese cryptocurrency.

Table 12 – LDA topics for Neo positive comments

No.	Relevant words in Topic	Postulated Topic
1	time, price, hold, pump, value, high, sell, day, another	Growth speculation
2	flamingo, new, how, long, transaction, maybe, cloud, would	Flamingo protocol updates
3	community, china, chinese, research, experiment, usage, life	Impact to Chinese society
4	team, building, release, ecosystem, update, ngd, platform, foundation	Neo Global Development (NGD) updates
5	trading, hit, sell, antshares, king, nano, nneo, go_moon, want_buy, trx, litecoin	General cryptocurrency hype
6	gas, coinbase, swap, cheap, low_sell, atl, loss, high	Buying discussion during all-time-low

The negative topics are more diverse but may be summarized into investment-related topics (no. 2, 4, 8, 9), technology issues (no. 1, 5, 7, 10, 11, 12), and founder & market-related topics (no. 3, 6, 13).

Table 13 – LDA topics for Neo negative comments

No.	Relevant words in Topic	Postulated Topic
1	blockchain, really, works, work, smart_contract	Technology doubts
2	sell, exchange, hold, buy, binance, swap, low, bought	Buy or sell discussions
3	market, investor, people, value, hype, marketing, company	Value perception by market
4	maybe, play, may, invest, enter, return	Investment experiments
5	token, need, require, give, receive, usd, innovation, account, transaction, introduce	Transaction improvement areas
6	protocol, incentive, network, polkadot	Polkadot competition
7	gas, ledger, error, send, fuck, transfer, generate_gas	Transaction errors
8	people, invest, emotional_invest, scam, lot_people	Emotional investment discussions
9	sell, profit, ath, try_get	Selling discussion during all-time-high
10	send, string, address, mint, try_send, can_not	Technical issues during MINT rush
11	liquidity, defi, liquidity_pool, defi_project, time, mess, delay	Delay of DeFi liquidity pools
12	wallet, transaction, shitty	Wallet transaction issues
13	communication, da_hongfei, resolve, news, conversation, speak, wrong	Founder's comments to challenges

Similar to Bitcoin, some topics are rather about the long-term adoption of the cryptocurrency and hence may not be well correlated to the short-term crypto-price development. The investment-related topics however of both, negative and positive comments, may be a good predictor of price increase or decrease.

6. Price prediction

6.1 Data Process

Given the result of correlation analysis in section 4, Bitcoin and NEO are also chosen for time series forecasting of close prices. This section will show in detail whether the sentiment of reddit comments can help perform better on price prediction.

6.1.1 Price Data

In accordance with the sentiment analysis period, same 2 periods for each coin are picked as the price dataset: close prices between 2017 and 2018 and close prices between 2020 and 2021 are the target variables.

6.1.2 Sentiment Feature

To further discover the deeper connection between price and daily sentiment, some simple feature engineering is done. Besides from considering only the daily sentiment comment rate, other features, including daily compound sentiment score are generated out based on the sentiment prediction for the whole reddit comment dataset.

Table 14 – Exogenous Sentiment Features

Feature	Definition
pos_rate	Daily sentiment rate (Section 4)
y_pred_ensemble	Mean of daily sentiment prediction
sen_sum	Sum of daily sentiment prediction
compound_score	Mean of the daily ensemble sentiment score of different models

6.2 Initial Forecasting with Lagging

6.2.1 Model Choosing and Evaluation Metrics

ARIMA, ARIMAX and VAR are used to predict prices:

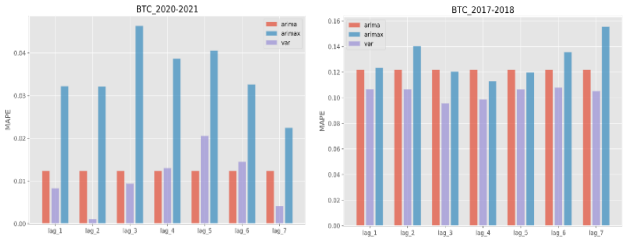
- ARIMA: pure time series according to past price.
- ARIMAX: exogenous variables to ARIMA.
- VAR: multiple time series influencing each other.

The Mean Absolute Percentage Error (MAPE) will be used as the evaluation metric between the models. Normal ARIMA is the benchmark model of price forecasting, the impact of sentiment is summarized by comparing the other 2 models using exogenous variables.

At the initial stage, 7 lags of sentiment features are added to the price prediction models.

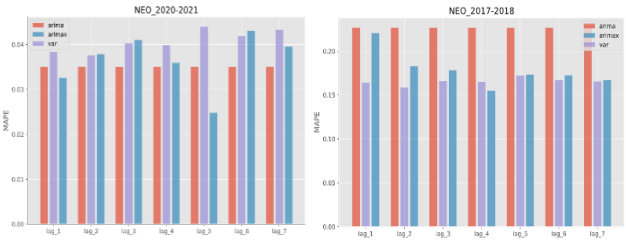
6.2.2 Initial Result Analysis – Bitcoin

Figure 8 – Price Forecast - Different Models (BTC)



6.2.3 Initial Result Analysis – Neo

Figure 9 – Price Forecast - Different Models (NEO)



The results above show further indicate that, in most cases, there might not be a strong correlation between sentiment and prices, especially from 2020 to 2021. For example, by adding exogenous sentiment features into ARIMAX for Bitcoin price forecasting seems to get worse in prediction. Same period for NEO cannot clearly distinguish the difference between 3 models.

However, a clear pattern can be observed from the comparison of NEO prediction in 2017 to 2018: by adding sentiment features, MAPEs go down significantly, both for ARIMAX and VAR. Similar but slighter reductions of error rate happen in bitcoin for VAR model.

To further explore the pattern, rolling window prediction is done in period 2017 - 2018 only in next stage.

6.3 Rolling Window Prediction

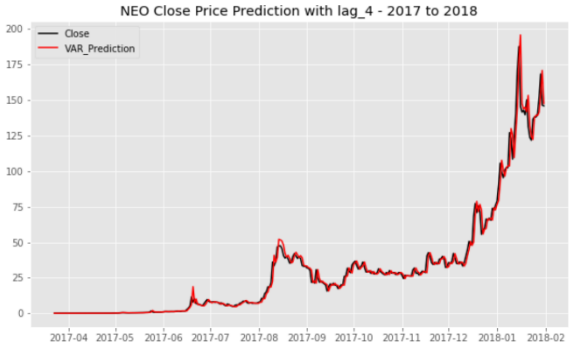
6.3.1 Introduction for Rolling Prediction

Rolling prediction is commonly used in time-series models, which is closer to the real-world situation, constructing a set of persistent time series models.

In this section, the first 50 days of period from 2017 to 2018 is considered as the initial prediction window with data-point size of 50: 49 training data and 1 test data (as the prediction period is 1 day). As time goes by, the prediction window size expands by constantly adding new data point by day until the window size equals to the whole period size (365 days). Therefore, we can obtain 315 samples of

test error for each set of time-series models with different lags. Below shows one example of rolling prediction:

Figure 10 – VAR Rolling Prediction with lag_4 Sentiment



6.3.2 Two Sample T-Test

By using the test error samples, test the hypothesizes:

H0: MAPE of ARIMA <= MAPE of VAR / ARIMAX

H1: MAPE of ARIMA > MAPE of VAR / ARIMAX

Table 15 – T-Test Result for NEO Rolling Prediction

LAG	ARIMA_VAR (P-VALUE - 0.1)	ARIMA_ARIMAX (P-VALUE - 0.1)
1	0.1058	0.5689
2	0.0996	0.3057
3	0.1422	0.3813
4	0.0962	0.1949
5	0.1207	0.3584
6	0.1234	0.536
7	0.148	0.3719

Table 16 – T-Test Result for BTC Rolling Prediction

LAG	ARIMA_VAR (P-VALUE - 0.1)	ARIMA_ARIMAX (P-VALUE - 0.1)
1	0.4436	0.7125
2	0.4160	0.8280
3	0.3687	0.7599
4	0.3663	0.7152
5	0.4188	0.8453
6	0.4072	0.7441
7	0.4364	0.6126

As observed from the tables:

For NEO: only 2 test samples of VAR models (Lag_2 sentiment and Lag_4 sentiment) reject H0, indicating a smaller test error rate when adding sentiment.

For BTC: None of the comparison shows a better performance in adding sentiment features.

Therefore, it can be concluded that the sentiment features help perform better for NEO in some cases, however, in general, these features might not be a strong predictor for CryptoCurrency prices forecasting.

7. Conclusion

Table 17 summarizes our findings.

Table 17 – Findings by hypothesis

Hypothesis	Finding
Hypothesis 1	Our sentiment analysis evaluation shows that own-trained models perform better compared to pre-trained libraries.
Hypothesis 2	Our correlation analysis shows that reddit sentiment is only slightly correlated with crypto currency prices.
Hypothesis 3	Our correlation analysis shows that the correlation between reddit sentiment on crypto currency prices was stronger during the period of 2017-2018 compared to 2019-2021.
Hypothesis 4	Our sentiment analysis does not support the hypothesis that non-mainstream currency prices have a higher correlation with reddit sentiment than mainstream currency prices.
Hypothesis 5	Our analysis does not support the hypothesis that reddit sentiment can be used as a predictive feature for crypto currency price prediction.

While our work does not support the usage of reddit sentiment for cryptocurrency price prediction, we realize that our work has several limitations. Most importantly, our analysis heavily depends on the performance of our own trained sentiment models for which we only used a limited labeled dataset of 5000 labeled comments. Higher quality sentiment features could have more predictive power than what we used for our analysis.

Also, we only used ARIMA, ARIMAX and VAR models for the price prediction. More advanced machine learning models might be able to extract more relations from the given information and hence, improve the predictive power of individual features.

GitHub Code Link

https://github.com/jieqiangt/BT5153_Crypto_Sense.git

References

Pulsar, (2018, July), Understanding cryptocurrencies through the lens of social media, https://business-reporter.co.uk/2018/07/11/understanding-cryptocurrencies-through-the-lens-of-social-media/#gsc.tab=0

Kovach, S. (2021, February, 8th), Tesla buys $1.5 billion in bitcoin, plans to accept it as payment, https://www.cnbc.com/2021/02/08/tesla-buys-1point5-billion-in-bitcoin.html

SRK, (2021), Cryptocurrency Historical Prices, Retrieved from https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory?select=ripple_price.csv

Jason, B. (2018, April), How to calculate correlation between variables in python, Machine Learning Mastery, https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/

Luis, M., Felix, W., Kay, B. Thomas, S (2020, December, 12th), Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum. Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 70–81, https://www.aclweb.org/anthology/2020.smm4h-1.11.pdf

Olga, K. (2018, October, 2nd), Institutional Investors Are Using Back Door for Crypto Buys, Bloomberg, https://www.bloomberg.com/news/articles/2018-10-01/institutional-investors-are-using-back-door-for-crypto-purchases

Vijay, B. (2018), The Bullish Case for Bitcoin, Medium, https://vijayboyapati.medium.com/the-bullish-case-for-bitcoin-6ecc8bdecc1

Jimmy, S. (2021, March), Debunking the Empty Block Attack, Medium, https://jimmysong.medium.com/

Cointelegraph (2021), What Is Lightning Network and How It Works, Cointelegraph, https://cointelegraph.com/lightning-network-101/what-is-lightning-network-and-how-it-works

Frances, C. Is a Global Digital Reserve Currency on the Horizon? American Express, https://www.americanexpress.com/us/foreign-exchange/articles/is-global-digital-reserve-currency-on-horizon/

Turner, W. (2020), HEX Still Can't Shake Scam Label as Token Approaches $1B Market Cap, Cointelegraph, https://cointelegraph.com/news/hex-still-cant-shake-scam-label-as-token-approaches-1b-market-cap

Terence, Z. (2020), Andreas Antonopoulos: Hex Team Offered Me 10 BTC to Speak Well of Their Token, https://news.bitcoin.com/andreas-antonopoulos-hex-team-offered-me-10-btc-to-speak-well-of-their-token/

Colin, H. (2021, April), Bitcoin Transactions Are More Expensive Than Ever, Coindesk, https://www.coindesk.com/bitcoin-transaction-fees-more-expensive-than-ever

Aftab, A. and Nupur, A. (2021, March), India to propose cryptocurrency ban, penalising miners, traders – source, Reuters, https://www.reuters.com/article/uk-india-cryptocurrency-ban-idUSKBN2B60QP

Shawn T. (2021, April), How much Bitcoin comes from dirty coal? A flooded mine in China just spotlighted the issue, Fortune, https://fortune.com/2021/04/20/bitcoin-mining-coal-china-environment-pollution/