Twitter Sentiment Analysis for Cryptocurrencies Price Predict

Karina Szubert (0036544634) Patryk Grzegorek (0036544655)

1. Introduction

In this article, we will analyze the impact of emotions visible on Twitter on some of the most popular cryptocurrencies' prices. The data we collected from multiple datasets covers 105 individual days from 02/10/2021 to 01/11/2022. Together we analyzed 299,265 tweets from those 105 days. We will check how the emotions read from tweet messages influenced the prices of individual cryptocurrencies, such as Bitcoin, Ethernum, Binance coin, Cardano, Chainlink, Litecoin, Neo. For the sentiment analysis, we will use the NLP TextBlob Python library. Whereas to find the best algorithm to predict the tweets' sentiment (obtained based on the lexicon-based approach as training data) we will use the following classifiers: RandomForest, Naive Bayes, K-NN (KNearest Neighbor), MLP (Multilayer perceptron, Neural Networks), linear SVM (Support Vector Machine), AdaBoost. We will compare the obtained results and draw conclusions whether the thesis "Emotions in social network regarding cryptocurrencies have an impact on changes in their value on the market" is true.

2. Background

2.1 Cryptocurrency

Cryptocurrency, sometimes called crypto-currency or crypto, is any form of currency that exists digitally or virtually and uses cryptography to secure transactions. Cryptocurrencies don't have a central issuing or regulating authority, instead of using a decentralized system to record transactions and issue new units.

The validity of each cryptocurrency's coins is provided by a blockchain. A blockchain is a continuously growing list of records, called blocks, which are linked and secured using cryptography. Each block typically contains a hash pointer as a link to a previous block, a timestamp, and transaction data. By design, blockchains are inherently resistant to modification of the data. It is "an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way".For use as a distributed ledger, a blockchain is typically managed by a peer-to-peer network collectively adhering to a protocol for validating new blocks. Once recorded, the data in any given block cannot be altered retroactively without the alteration of all subsequent blocks, which requires the collusion of the network majority.

Blockchains are secure by design and are an example of a distributed computing system with high Byzantine fault tolerance. Decentralized consensus has therefore been achieved with a blockchain.

In cryptocurrency networks, mining is a validation of transactions. For this effort, successful miners obtain new cryptocurrency as a reward. The reward decreases transaction fees by creating a complementary incentive to contribute to the processing power of the network. The rate of generating hashes, which validate any transaction, has been increased by the use of specialized machines such as FPGAs and ASICs running complex hashing algorithms like SHA-256 and script. This arms race for cheaper-yet-efficient machines has existed since the first cryptocurrency, bitcoin was introduced in 2009.

A cryptocurrency wallet stores the public and private "keys" (address) or seed which can be used to receive or spend the cryptocurrency. With the private key, it is possible to write in the public ledger, effectively spending the associated cryptocurrency. With the public key, it is possible for others to send currency to the wallet.

There exist multiple methods of storing keys or seeds in a wallet from using paper wallets which are traditional public, private, or seed keys written on paper to using hardware wallets which are dedicated hardware to securely store your wallet information, using a digital wallet which is a computer with software hosting your wallet information, hosting your wallet using an exchange where cryptocurrency is traded. or by storing your wallet information on a digital medium such as plaintext.

Cryptocurrency markets are decentralized, which means they are not issued or backed by a central authority such as a government. Instead, they run across a network of computers. However, cryptocurrencies can be bought and sold via exchanges and stored in 'wallets'.

Crypto market capitalization is the total value of a cryptocurrency. Where stock market capitalization is calculated by multiplying share price times shares outstanding, crypto market capitalization is calculated by multiplying the price of the cryptocurrency with the number of coins in circulation.

2.2 Twitter

The term social media refers to a computer-based technology that facilitates the sharing of ideas, thoughts, and information through virtual networks and communities. Social media is internet-based and gives users quick electronic communication of content, such as personal information, documents, videos, and photos.

Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read those that are publicly available.

3. Data

3.1 Datasets

We used several datasets in our research. First "Bitcoin Tweets" dataset from kaggle.com, which contained 2,126,407 tweets with the hashtag #Bitcoin and #btc. However, we reduced the number of tweets to a maximum of 3,000 per day, thus receiving 299,265 tweets.
The remaining datasets are from cryptodatadownload.com and contain information about the daily changes in the cryptocurrency market.

The first dataset before data preprocessing:

	date	text
0	2021-02-10	Blue Ridge Bank shares halted by NYSE after #b...
1	2021-02-10	😎 Today, that's this #Thursday, we will do a "...
2	2021-02-10	Guys evening, I have read this article about B...
3	2021-02-10	$BTC A big chance in a billion! Price: \487264...
4	2021-02-10	This network is secured by 9 508 nodes as of t...
...	...	...
2126402	2022-01-13	I simply can't understand how they can't list ...
2126403	2022-01-13	#BTC is now at $43306.96
2126404	2022-01-13	@feiprotocol \n\nSell #tribe $tribe , now \n...
2126405	2022-01-13	#XMR - long alert 💰💵\n\nExchange : BINANCE...
2126406	2022-01-13	With great ideas come great changes. #BTC is t...

3.2 Data pre-processing

Before analyzing the sentiment of the tweet, we had to properly prepare the data.

We have carried out 6 stages of data pre-processing.

- Noise cleaning: Noise removal is about removing characters digits and pieces of text that can interfere with your text analysis such as usernames, etc.

- Tokenizing: Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units.

- Lower-case conversion: In this step, all the tokens are converted to the lower cases.

- Remove stop words: In this step, all the meaningless words are removed.

- Lemmatization: Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

- Delete empty tweets.

Sample of the first dataset after preprocessing the data:

	date	text	extracted_text
798932	2021-08-08	Buy #Gold they said.\n\n@PeterSplifff \n\nBut ...	[buy, said, smart, money, nt, listen, bought, ...
94254	2021-04-20	#ARB SP did remarkably well in yesterday's #Bi...	[sp, remarkably, well, yesterday, drop, short,...
67642	2021-04-07	Learn how to earn from home , it's easy.... I'...	[learn, earn, home, easy, dm, away, earn, much...
1647343	2021-11-05	@CryptoSavy1 #bitcoin is bullish the bears who...	[bullish, bear, block, hiding]
148394	2021-06-23	Is this guy still alive #Bitcoin https://t.c...	[guy, still, alive]
669156	2021-07-27	@BigImpactHumans Hey check out \n@G_dogetoken\...	[hey, check, telegram, website, amazing, profe...
55256	2021-04-10	Talking about #Bitcoin with Bobby Lee, CEO of ...	[talking, bobby, lee, ceo, btc, china, la, vega]
1290646	2021-10-19	https://t.co/Qos91fUosr Ethereum, Solana, and ...	[ethereum, solana, elrond, eyeing, massive, su...
905388	2021-08-19	westbeachmusic found #bitcoin in a User vault ...	[westbeachmusic, found, user, vault, location,...
1795024	2021-11-19	Let's take a look under the hood today at (IMO...	[let, take, look, hood, today, imo, huge, step...
113653	2021-05-28	$BTC went down to my buy zone as expected. Let...	[went, buy, zone, expected, let, see, push, br...
1942524	2021-11-24	@minipekkatoken Wow, projectin Future , Good ...	[wow, projectin, future, good, futer, whoever,...
808384	2021-08-07	Who the fxck do you think you are, pleb.\nIf y...	[fxck, think, pleb, ca, nt, afford, lose, ever...
1570454	2021-10-27	@shibafloki_army @gate_io @Shibaflokitoken #sh...	[next]
79714	2021-04-23	#Bitcoin #BTC current price (GBP): £35,358\nLi...	[current, price, gbp, like, update, tip, 3l9dz...

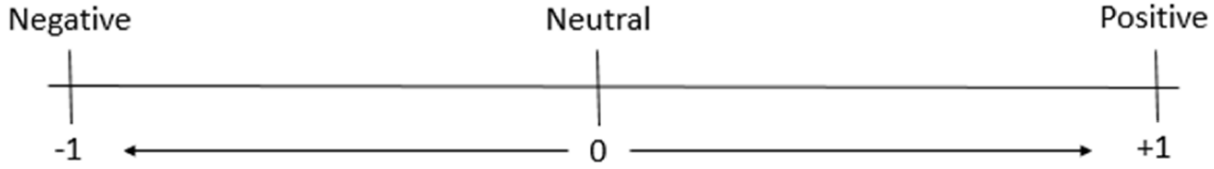
4. Sentiment Analysis

Sentiment Analysis (also known as opinion mining or emotion AI) is a sub-field of NLP that tries to identify and extract opinions within a given text across blogs, reviews, social media, forums, news, etc.

After preparing the data, we used BlobText which returns the polarization and subjectivity of the sentence. The polarity lies between [-1,1], polarity below 0 defines negative sentiment, equal to 0 - neutral, and the value above 0 defines positive sentiment. Words of negation reverse the polarity. TextBlob has semantic labels to aid in detailed analysis.

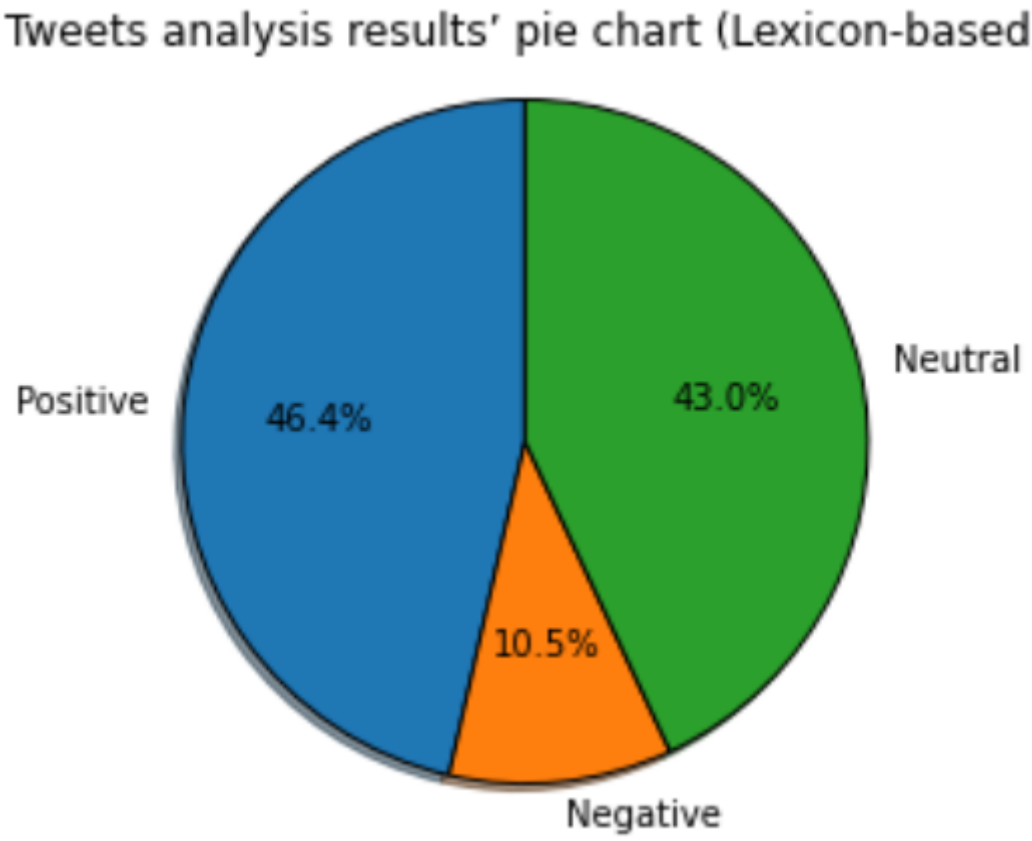
	date	text	extracted_text	sentiment
301069	2021-07-04	$Rally is 15bn token supply ? Are you fucking ...	[15bn, token, supply, fucking, shutting]	-0.600000
1834646	2021-11-18	#Dogecoin #memes #crypto #cryptocurrency #doge...	[forex, signal, available, forexworld843, gmai...	0.400000
2708	2021-02-10	#Centralbanks always use the same lies.\n\nPeo...	[always, use, lie, people, care, le, le, use]	0.000000
478524	2021-07-20	Prominent #Bitcoin maximalist Max Keiser belie...	[prominent, maximalist, max, keiser, belief, l...	0.250000
41456	2021-02-22	Absolute chaos today. Hope all you cryptonites...	[absolute, chaos, today, hope, cryptonites, ke...	0.316667
116769	2021-05-27	@alexisolinart Happiness is... A Warm NFT Coll...	[happiness, warm, nft, collection, http, tcouq...	0.650000
111600	2021-05-28	Bitcoin: $35,024.27\n - 10.153955.77)\nHig...	[bitcoin, high, low, volume]	0.080000
18074	2021-02-06	FORGET #BITCOIN, #ALTCOINS are MOVING! With ht...	[forget, moving, analyze, volumeprice, movin]	0.000000
12133	2021-02-08	"This is $TSLA and @elonmusk diving into the d...	[diving, deep, end, pool, crypto, something, r...	0.000000
1419450	2021-10-21	What a load of crap! Are they still trying to ...	[load, crap, still, trying, fool, u, instituti...	-0.750000
542819	2021-07-18	Very good project, I like it ❤️\nHopefully thi...	[good, project, like, hopefully, project, deve...	0.266667
1438725	2021-10-21	@lisafriedrich_ Alts can take us for a ride on...	[alt, take, u, ride, highway, get, faster, wan...	0.000000
1823323	2021-11-18	BTC Healthy Pullback for BITSTAMP:BTCUSD by ni...	[btc, healthy, pullback, bitstamp, btcusd, nic...	0.500000
80966	2021-04-23	Stop panicking people. The government and medi...	[stop, panicking, people, government, medium, ...	0.000000
784486	2021-08-08	The #Bitcoin supply shock is going to send the...	[supply, shock, going, send, price, one, highe...	0.225000

We then grouped the data into individual days. We calculated how many tweets on a given day were positive, negative, and neutral, we presented the result in%. We averaged the sentiments arithmetically and we counted its median. We classified sentiments into 3 class
-1 - negative
0 - neutral
1 - positive



The total amount of negative/positive/neutral tweets for the whole dataset

Sentiments class	Quantity	Percent
Positive	137363	46.4
Negative	31199	10.5
Neutral	127413	43.1



Each day the sentiments average was greater than 0, so each day the sentiments class was 1.

date	sentiment_mean	sentiment_median	sentiment_class	positive_tweets_percent	negative_tweets_percent	neutral_tweets_percent
2021-07-04	0.178581	0.012121	1	0.5018	0.1020	0.3961
2021-04-19	0.083677	0.000000	1	0.3991	0.1058	0.4951
2021-07-21	0.135492	0.000000	1	0.4572	0.1188	0.4240
2021-02-19	0.089602	0.000000	1	0.3711	0.0969	0.5321
2021-06-23	0.127629	0.000000	1	0.4315	0.1129	0.4556
2021-11-04	0.169846	0.033333	1	0.5077	0.0908	0.4015
2021-04-18	0.085812	0.000000	1	0.4037	0.1191	0.4772
2021-10-18	0.142802	0.000000	1	0.4822	0.1201	0.3977
2021-10-22	0.173432	0.033333	1	0.5069	0.1015	0.3916
2021-04-24	0.099253	0.000000	1	0.4598	0.1243	0.4159
2021-08-16	0.163038	0.012917	1	0.5020	0.0945	0.4034
2021-02-13	0.086875	0.000000	1	0.3303	0.0792	0.5905
2021-02-07	0.105510	0.000000	1	0.3774	0.0682	0.5544
2021-08-24	0.164298	0.025476	1	0.5086	0.0979	0.3935
2021-03-11	0.078426	0.000000	1	0.3832	0.0841	0.5327

5. Cryptocurrencies Price Predict

5.1 Algorithms

Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets. It can be used for Binary as well as Multi-class Classifications. It performs well in Multi-class predictions as compared to the other Algorithms. It is the most popular choice for text classification problems.

Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time. For regression tasks, the mean or average prediction of the individual trees is returned.

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classify a data point based on how its neighbours are classified.

A multilayer perceptron (MLP) is a class of feedforward artificial neural networks (ANN). MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

The results of the classifier validation will be presented separately along with the accuracy criterion (the percentage of correct forecasts or the relative number of correctly classified examples). Additionally, the average accuracy of each classifier is presented in:

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} * \ 100$$

[True positive (TP), true negative (TN), false positive (FP), false negative (FN)]

5.2 Results

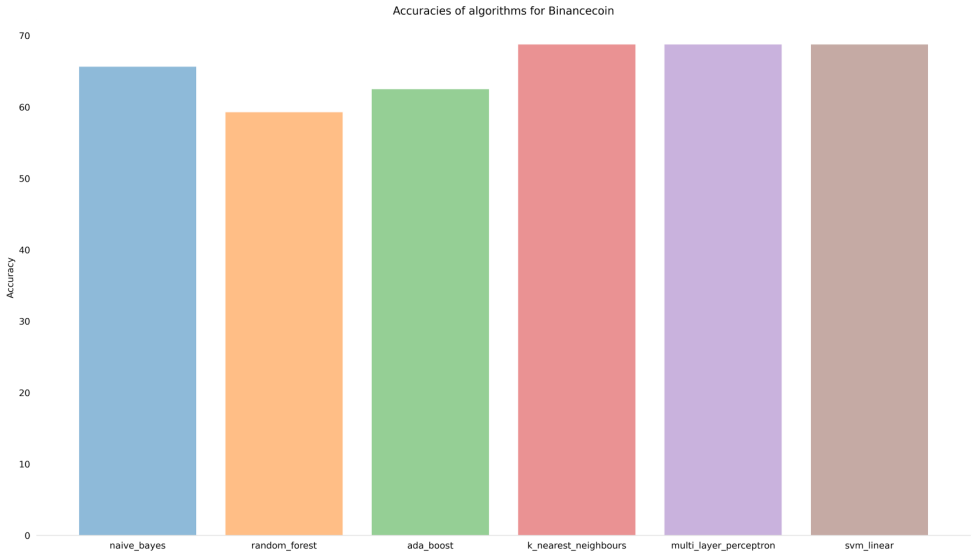
Based on the cryptocurrency price datasets, we calculated whether the price went up or down the next day. If the price rose, the column value was 1, if it fell, it was 0. We then linked this information to the sentiment data frame.

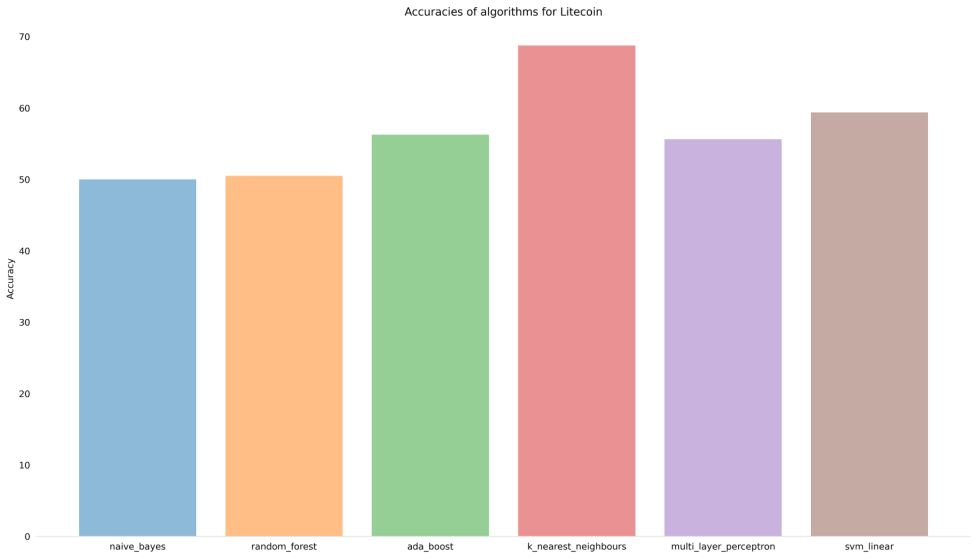
Sample data frame part

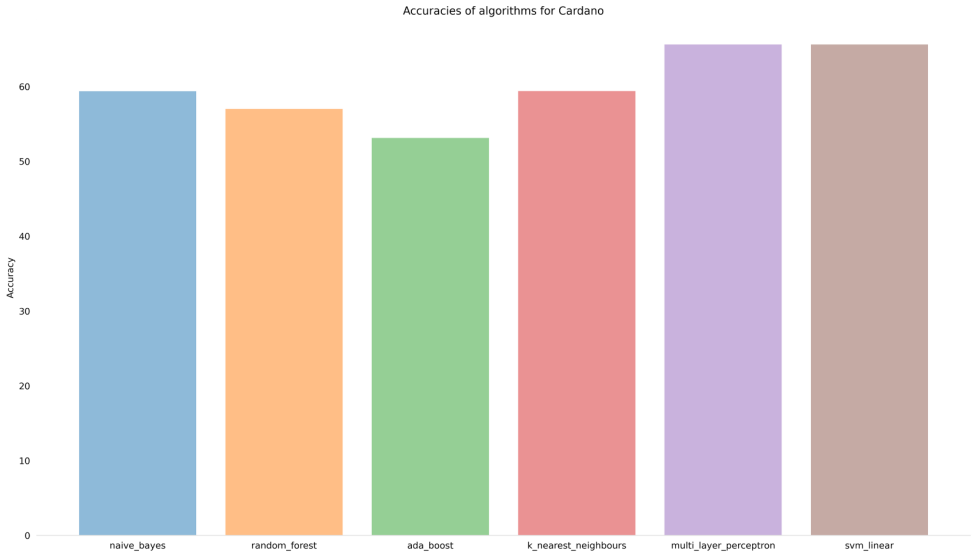
date	sentiment_mean	sentiment_median	sentiment_class	positive_tweets_percent	negative_tweets_percent	neutral_tweets_percent	bitcoin_price_change
2021-08-08	0.151952	0.000000	1	0.4851	0.1082	0.4068	0
2021-08-23	0.137805	0.000000	1	0.4628	0.1071	0.4301	1
2021-02-10	0.092868	0.000000	1	0.3728	0.0983	0.5289	0
2021-07-30	0.132719	0.000000	1	0.4525	0.1126	0.4349	1
2021-07-20	0.128066	0.000000	1	0.4595	0.1173	0.4233	0
2021-02-06	0.090390	0.000000	1	0.3496	0.0744	0.5760	1
2021-07-02	0.154256	0.000000	1	0.4540	0.0984	0.4476	1
2021-07-26	0.083673	0.000000	1	0.3324	0.1023	0.5652	1
2021-04-06	0.099984	0.000000	1	0.3961	0.1418	0.4621	0
2021-04-22	0.096787	0.000000	1	0.4524	0.1266	0.4210	0
2022-01-12	0.172582	0.107273	1	0.5682	0.0821	0.3497	1
2021-07-04	0.178581	0.012121	1	0.5018	0.1020	0.3961	1
2021-06-20	0.156114	0.000000	1	0.4758	0.1208	0.4034	1
2021-10-28	0.158357	0.000000	1	0.4929	0.1038	0.4033	1
2021-06-23	0.127629	0.000000	1	0.4315	0.1129	0.4556	1

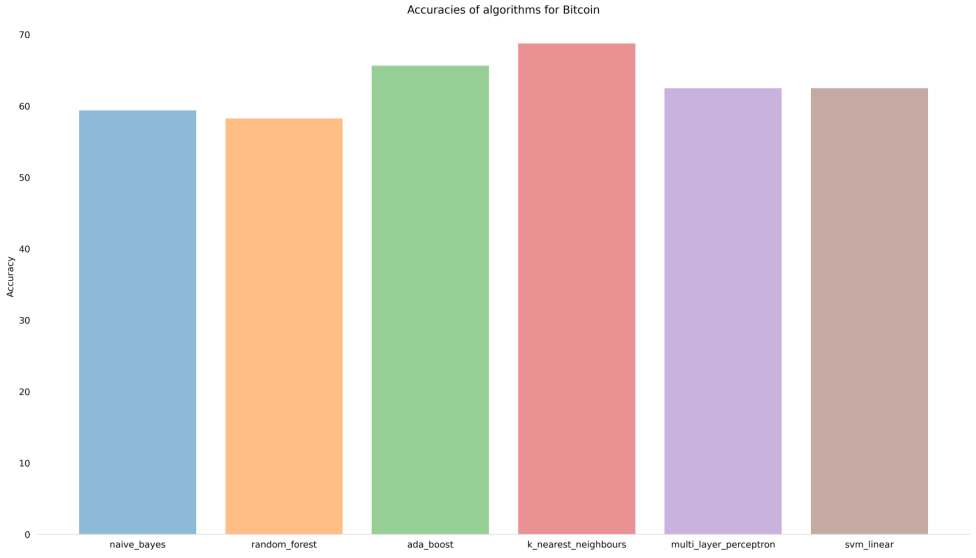
We analyzed each data frame 50 times with each algorithm with x_price_change as the target column. We averaged the results arithmetically.

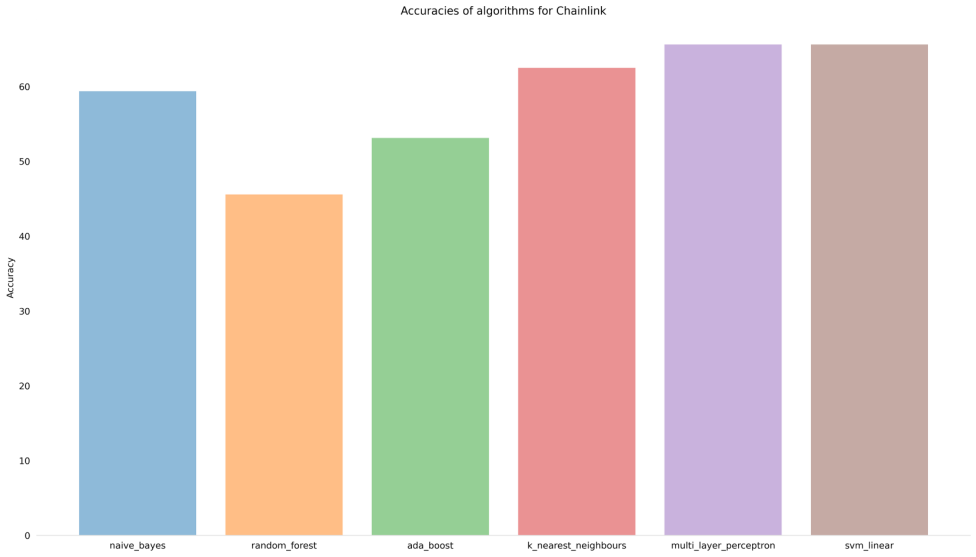
	naive_bayes	random_forest	ada_boost	k_nearest_neighbours	multi_layer_perceptron	svm_linear	Average
Bitcoin	59.375	58.25	65.62	68.75	62.5	62.5	62.83
Ethernum	59.375	45.62	46.88	62.5	65.62	65.625	57.60
Binancecoin	65.625	59.25	62.5	68.75	68.75	68.75	65.60
Cardano	59.375	57.0	53.12	59.38	65.62	65.625	60.02
Chainlink	59.375	45.56	53.12	62.5	65.62	65.625	58.63
Litecoin	50.0	50.5	56.25	68.75	55.62	59.375	56.75
Neo	62.5	56.06	59.38	65.62	68.75	65.625	62.99
Average	59.38	53.18	56.7	65.18	64.64	64.73	60.64

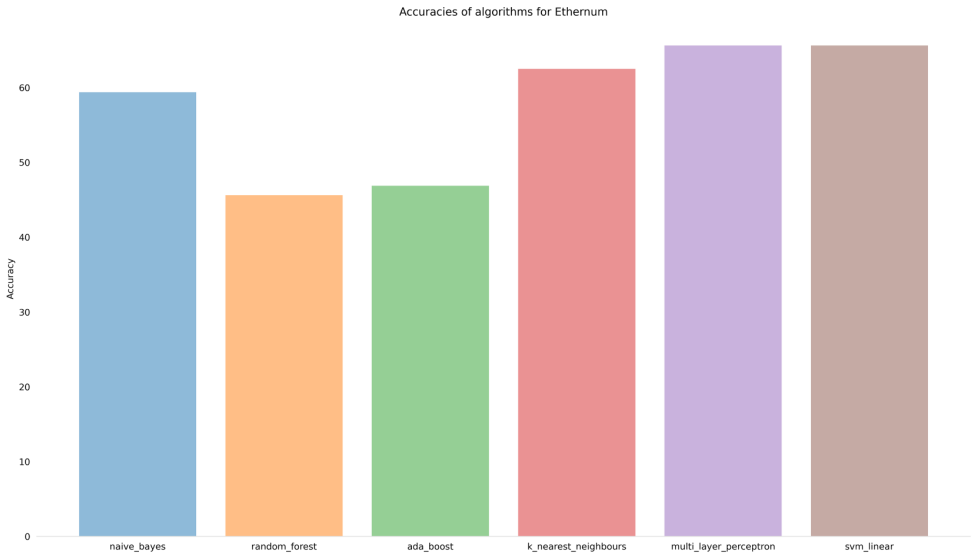

Accuracies of algorithms for Binancecoin
Accuracy
0
10
20
30
40
50
60
70
naive_bayes
random_forest
ada_boost
k_nearest_neighbours
multi_layer_perceptron
svm_linear

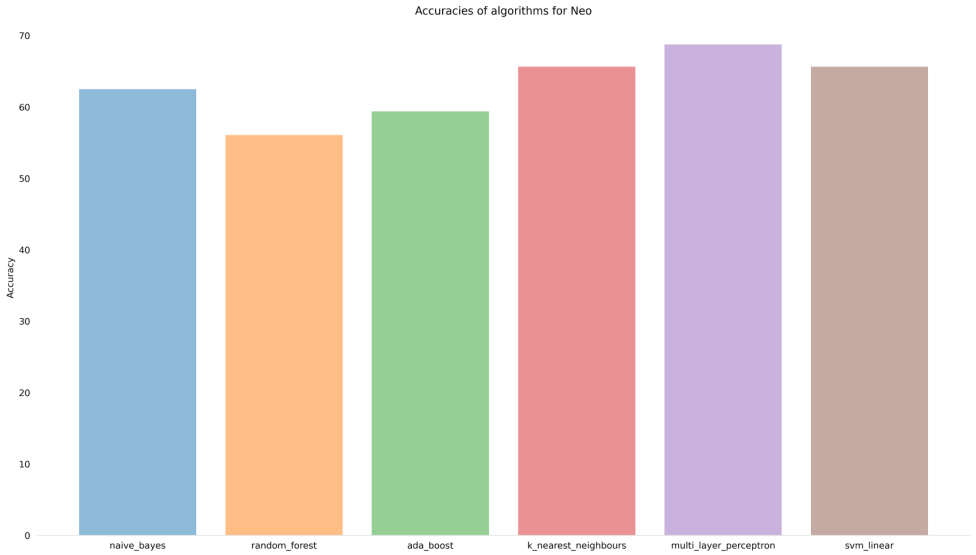

Accuracies of algorithms for Litecoin
Accuracy
0
10
20
30
40
50
60
70
naive_bayes
random_forest
ada_boost
k_nearest_neighbours
multi_layer_perceptron
svm_linear

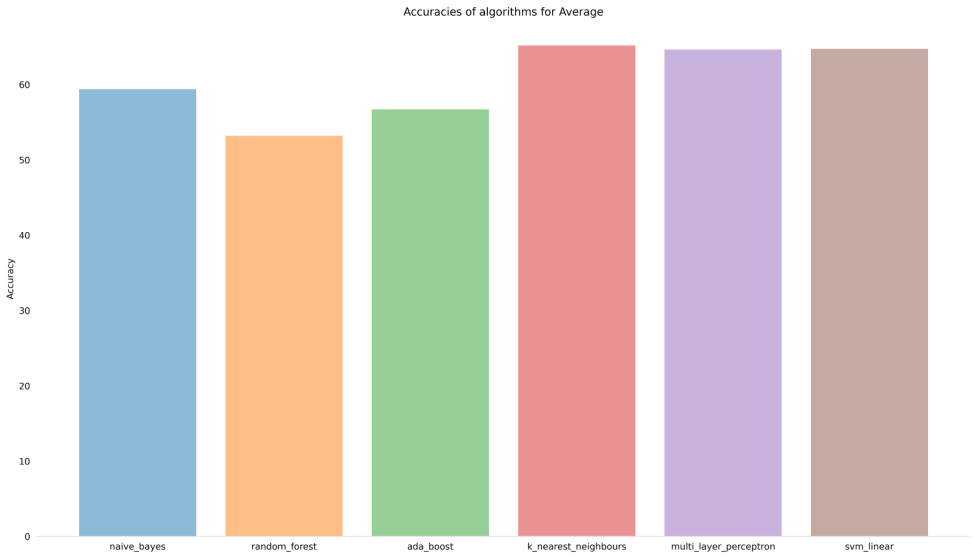

Accuracies of algorithms for Cardano
Accuracy
0
10
20
30
40
50
60
naive_bayes
random_forest
ada_boost
k_nearest_neighbours
multi_layer_perceptron
svm_linear


Accuracies of algorithms for Bitcoin
Accuracy
0
10
20
30
40
50
60
70
naive_bayes
random_forest
ada_boost
k_nearest_neighbours
multi_layer_perceptron
svm_linear


Accuracies of algorithms for Chainlink
Accuracy
0
10
20
30
40
50
60
naive_bayes
random_forest
ada_boost
k_nearest_neighbours
multi_layer_perceptron
svm_linear


Accuracies of algorithms for Ethernum
Accuracy
0
10
20
30
40
50
60
naive_bayes
random_forest
ada_boost
k_nearest_neighbours
multi_layer_perceptron
svm_linear





5.3 Conclusion

A study was conducted to find out if Tweets had an impact on cryptocurrency prices. With all the above techniques, we obtained the arithmetic mean value above 50%. On average, we got the best results for KNearest Neighbor 65.18%. The results also depend on the cryptocurrency. All cryptocurrencies scored higher than 50%. The best result was obtained by Binance coin 65.60%. The arithmetic mean of predictions for all cryptocurrencies was 60.64%. From the above results, it can be concluded that the sentiments on Twitter have an impact on cryptocurrency prices, but they are not strictly dependent on them.

We believe that the above results can be improved. We would especially suggest using a different database for gathering tweets, which would have collected tweets from more than just 105 days. Another thing to improve could be tuning hyperparameters for the above algorithms, for example using GridSearch.

6. References

Dataset of tweets about Bitcoin:
https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets

Datasets of cryptocurrencies' prices:
https://www.cryptodatadownload.com/data/binance/?fbclid=IwAR2hn6pLxAKOzcCf4EwMYpXiRozayDJDtBqCMvgwyL6oClxVwXNMgcMXQzE#google_vignette

https://scikit-learn.org/stable/

https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

https://en.wikipedia.org/wiki/Cryptocurrency

https://en.wikipedia.org/wiki/Twitter