A Logistic Regression Based Sentiment Analysis of Cryptocurrency Tweets for Smart Portfolio Management

Victor Plesco

11-05-2021

Contents

1. Introduction

By April 2021, the two largest cryptocurrencies, measures in terms of market capitalization, had a combined market value of 1623.8 billion dollars. Bitcoin alone made up nearly $1065 billion of this value. Given the significant value of these currencies, some people see value in them through use as actual currencies, while others view them as an investment opportunities. The result has been large swings in value of both currencies over short periods of time. During 2017 the value of a single Bitcoin increased 2000% going from $863 on January 9, 2017 to a high of $17,550 on December 11, 2017. The growth and interest were primarily caused by news stories which reported the unprecedented returns of cryptocurrencies, that subsequently attracted a type of gold rush. By eight weeks later, on February 5, 2018, the price of a single Bitcoin had been more than halved with a value of $7,964. Today, May 2, 2021, you can buy a single Bitcoin for a modest price of $56,846. The promising technology behind cryptocurrencies, the blockchain, makes its likely that cryptocurrencies will continue to be used in some capacity, and that their use will grow. Simultaneously, current global regulations on cryptocurrencies are very limited, as cryptocurrencies are not yet acknowledged as a mature asset class. This regulatory void, in combination with the high popularity and lack of an institutional guarantor, makes the cryptocurrency market so volatile that it has even been called a "*wild west*".

2. Problem Statement

The volatility in the value of cryptocurrencies is strongly fueled by news messages and posts on social media and means uncertainty for both investors, and people who wish to use them as a currency rather than an investment. This effect is further reinforced, as investors struggle to discover whether the posted information is true or false. Due to the relatively young age of the cryptocurrency market (Bitcoin was created in 2009) relative to fiat currencies such as the U.S. dollar (USD) or the Japanese Yen, traditional news outlets do not always timely report events, what has led to social media being a primary source of information for cryptocurrency investors. Specifically, micro-blogging website Twitter is a widely used source for cryptocurrency information. Not only does Twitter provide live updates on cryptocurrencies, it is also a rich source of emotional intelligence, as investors frequently express their sentiment. Behavioral economics tells us that sentiment and emotions can profoundly affect individual behavior and decision-making. With the vast amount of easily available data from Twitter containing the emotional intelligence of cryptocurrency users and investors, it is the main goal of this study to research to what extent public Twitter sentiment can be used to forecast the price fluctuations of cryptocurrencies.

3. Methodology

This study focuses on the prediction of price returns of fifteen randomly chosen cryptocurrencies. Specifically, in descending order of market capitalisation Bitcoin (*BTC*), Ethereum (*ETH*), Cardano (*ADA*), Ripple (*XRP*), Polkadot (*DOT*), Litecoin (*LTC*), Uniswap (*UNI*), Stellar (*XLM*), TRON (*TRX*), EOS (*EOS*), Monero (*XMR*), Neo (*NEO*), PancakeSwap (*CAKE*), IOTA (*MIOTA*) and Dash (*DASH*) are researched.

3.1. Twitter Data

Tweets were obtained separately for each cryptocurrency between the period of 6 April 2021 and 20 April 2021, resulting in fifteen datasets with a total of 274,950 public Tweets. Due to the recent upgrade of the standard v1.1 Twitter API to v2 (a.t.m. in early access), the accessibility of the full-archive endpoint has been negated to most of the existing libraries. This pushed me into an undesired adventure, lasted for uncountable weeks, which in the end has brought to life racademic: an implementation of calls designed to access Tweets via the recent and full-archive search REST endpoints of the Twitter API v2 fully written in R (first and last time, I swear). Going back to the gathering process, it is common for the Twitter community to use cashtags ($) as a prefix to communicate about financial products such as cryptocurrencies or stocks. Therefore, with the purpose of increasing the precision of the Twitter querying process (i.e. imagine requesting tweets for a cryptocurrency called *CAKE*), cashtags alongside cryptocurrency symbols have been

used as query parameters. Non-English Tweets, as well as *retweets*, *replies* and *quoted* Tweets were filtered out and numerous (user) variables were collected to be used for detecting bots. More details of the Twitter datasets can be found in **Table 1** and **Table 2**.

Table 1: Symbol, Name and Number of Tweets per researched cryptocurrency.

	Symbol	Name	Tweets
1.	TRX	TRON	14,155
2.	XMR	Monero	3,087
3.	EOS	EOS	4,351
4.	CAKE	PancakeSwap	8,649
5.	DASH	Dash	2,656
6.	BTC	Bitcoin	118,096
7.	ETH	Ethereum	39,579
8.	XRP	Ripple	41,386
9.	ADA	Cardano	19,411
10.	DOT	Polkadot	4,642
11.	LTC	Litecoin	9,871
12.	UNI	Uniswap	2,734
13.	XLM	Stellar	3,327
14.	NEO	Neo	2,533
15.	MIOTA	IOTA	473

3.2. Financial Data & Custom Labelling

Financial data for the nine researched cryptocurrencies was sourced from CoinMarketCap between 6 April 2021 and 20 April 2021. The rationale is that the prices of cryptocurrencies can vary substantially across various exchanges. CoinMarketCap is a widely used proxy for cryptocurrency prices as it combines prices from a large number of exchanges, thereby offering a more accurate and general value representation that is independent of any exchange price bias. Once gathered the daily prices, Tweets have been labeled with *0s* or *1s* according to the positive or negative change in daily price of each cryptocurrency (e.g. a Tweet on Bitcoin from 7 April is labeled as *1* if the closing price of Bitcoin on 8 April is higher than the closing price on 7 April).

Table 2: Overview of Twitter dataset.

text	created_at	username	target
How did I end up holding a large bag of $TRX?	2021-04-20 11:56:00	yy_crypto	0
Bought $TRX on the dip @ .123, LFG	2021-04-20 10:51:44	Third_iQ	0
Just add some more $DOGE & $TRX	2021-04-20 06:03:16	DScatts	0
Keep going $XRP!	2021-04-13 05:51:13	SirHodlsLong	1
$XRP just might be taking off.	2021-04-13 05:51:21	BLOCKCHAIN_DNA	1
$XRP breaking out violently!!	2021-04-13 05:51:02	Cryptozillaa	1
I bought more $ADA today. #Cardano	2021-04-16 19:40:26	LatCrypto	0
Next coin going to fly is $ada	2021-04-16 19:16:07	smpltrade	0
$ADA I don't like people who talk to say nothing	2021-04-16 18:16:33	Kryptokoko1	0
Longed $ADA, any opinions?	2021-04-12 13:03:18	ExwhereAngel1	1

3.3. Data Pre-Processing and Feature Selection

Twitter data is known for its lack of structure and its high levels of noise. As a result, the collected Twitter data requires extensive pre-processing to make it useful in sentiment analysis. An array of 11 pre-processing techniques is applied, in combination with specifically designed techniques to filter out noise elements from Tweet texts. First, the Python `BeautifulSoup` is applied for HTML and UTF-8 BOM (Byte Order Mark) decoding. Next normalization is applied by removing URLs, excess (white) spaces, and user mentions (e.g. *@account*) from the Tweets. Following this, the contractions are expanded (e.g. *"we're"* into *"we are"*), both CashTag symbols (e.g. *"$BTC"*), numerical characters (e.g. *"2nd"*) and HashTags containing cryptocurrency symbols (e.g. *"#BTC"*) are removed and negations are handled (e.g. *"haven't"* into *"have not"*). Note that CashTag symbols were used to obtain Tweets but are removed as they are noise in the context of sentiment analysis. Emojis are pre-processed by first identifying part of them with the Python library `emoji` and subsequently using a series of RegEx functions to identify the remaining symbols, to finally decode them into textual data. Punctuation is removed to further reduce noise and finally tokenization and lemmatisation are applied by using the Python library `SpaCy`. An example of how the above techniques have impacted the number of characters per Tweet is shown in **Figure 1**.

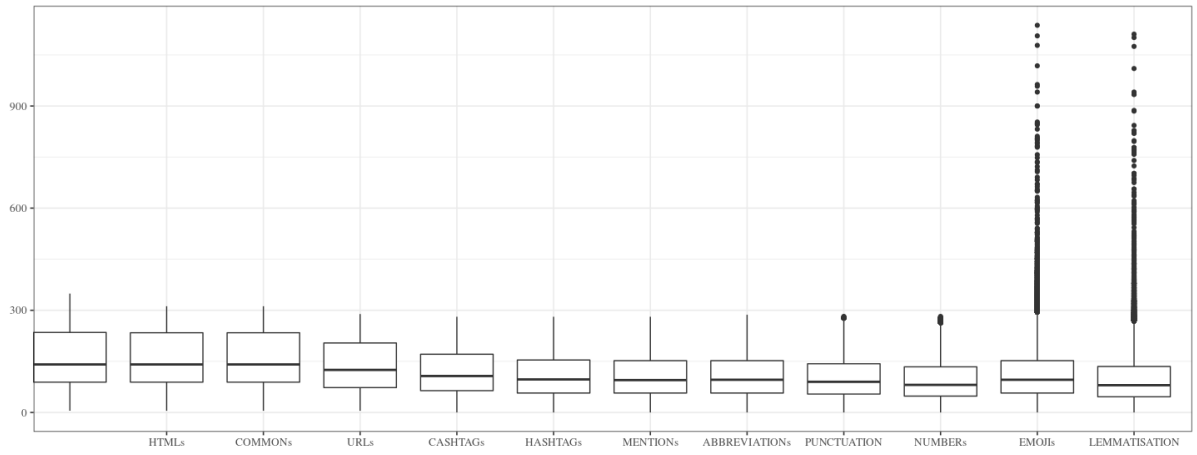

Figure 1: Box-Plot for Length of Characters after Each Pre-processing Phase

3.4. Exploratory Data Analysis

The hard pre-processing has heavily reduced the variance while increasing the character length of each Tweet. This because converting emojis to text has added to Tweets long strings of bytes used for the encoding of the latter. As can be seen from Word Cloud in **Figure 2**, most of the frequent words appear to be emojis such as **rocket_rocket**, identifying typically written phrases as *to the moon*, **backhand_index_pointing_right** and **join channel**, mostly used by twitter bots in order to spam Telegram cryptocurrency channels, and **spouting_whale** and **whale**, usually used in contexts where a steady drop or increase in price is associated to the entrance of a big player in the market. That said, a further improvement within the pre-processing phase could be the reduction of duplicated emojis in order still maintain their information while reducing the redundancy.

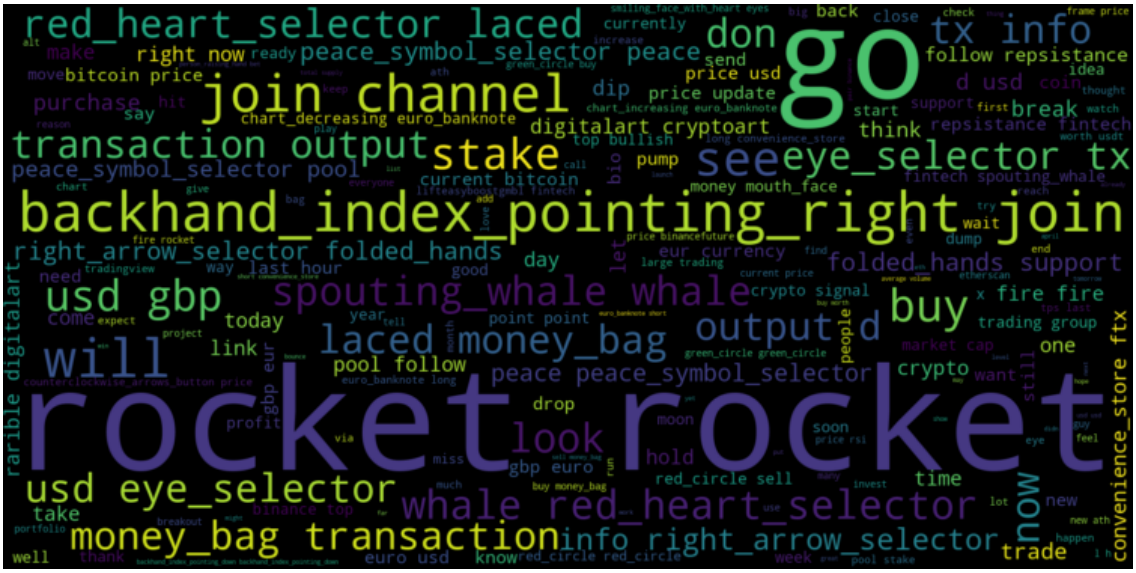

Figure 2: Wordcloud for Tokens after Pre-processing Phase

A further curiosity that was decided to be explored is the Zipf's Law. The latter has been introduced by the french stenographer Jean-Baptiste Estoup and later named after the american linguist George Kingslev Zipf and states that a small number of words are used all the time, while the vast majority are used very rarely. What's interesting about it is that given some corpus of natural language processing utterances, the frequency of any word is inversely proportional to its rank in the frequency table. The corpus used within this research appears to have passed the Zipf's Law as can be seen from the plot in **Figure 3**.

Figure 3: Frequency Barplot for Top 250 Tokens (Testing Zipf's Law)

As final step of the exploratory data analysis, a visualization of the frequency distribution of words within each class appeared to be more than necessary. For this purpose, it was decided to represent this information through a scatter plot having on the x-axis the frequency of words for the *negative* class and on the y-axis the *positive* one. Conceptually, having different distributions across the two target classes would certainly improve the accuracy of any classification model, while in the opposite case, having similar frequencies in

both classes would just add noise to the model. As can be seen in **Figure 4**, words from our corpus appear to lay on the diagonal, representing the worst case for a classification model.

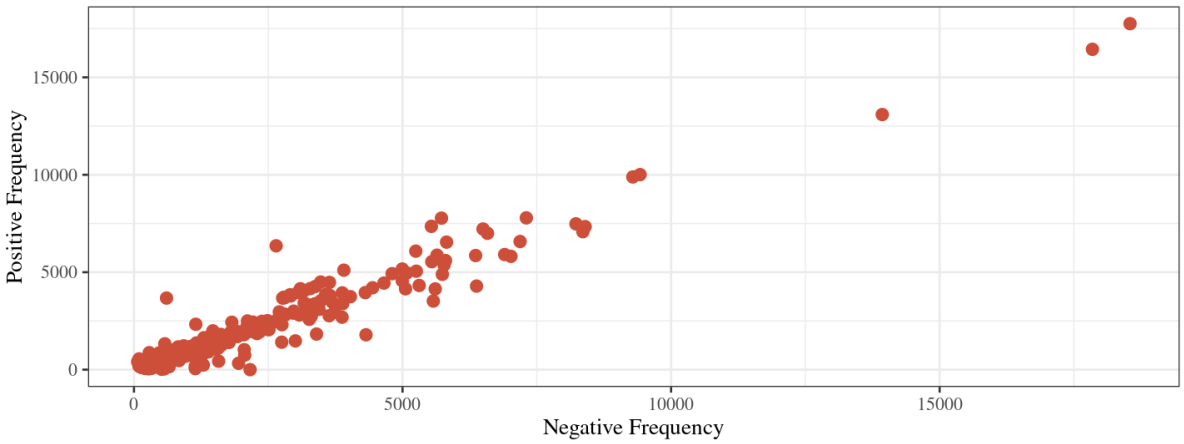

Figure 4: Positive and Negative Frequency Scatter Plot

4. Results

4.1. Model Selection

The classification problem that we decided to face will be handled throughout the implementation of a Logistic Regression on four different word vectorizations, namely a count and a tf-idf vectorization. For both of the latter we'll be considering two different types of tokenizations: the first with *ngrams*$(1, 3)$ and the second with *nchars*$(2, 5)$. In both cases *english* stopwords are deleted. For each of the corpuses we'll proceed with a chi-squared feature selection and output the results as mean of a 5-fold cross validation. The training set is considered to be the corpus created by the Tweets gathered during the first week of the data collection. The remaining Tweets will serve as a validation set for the conclusion of this research.

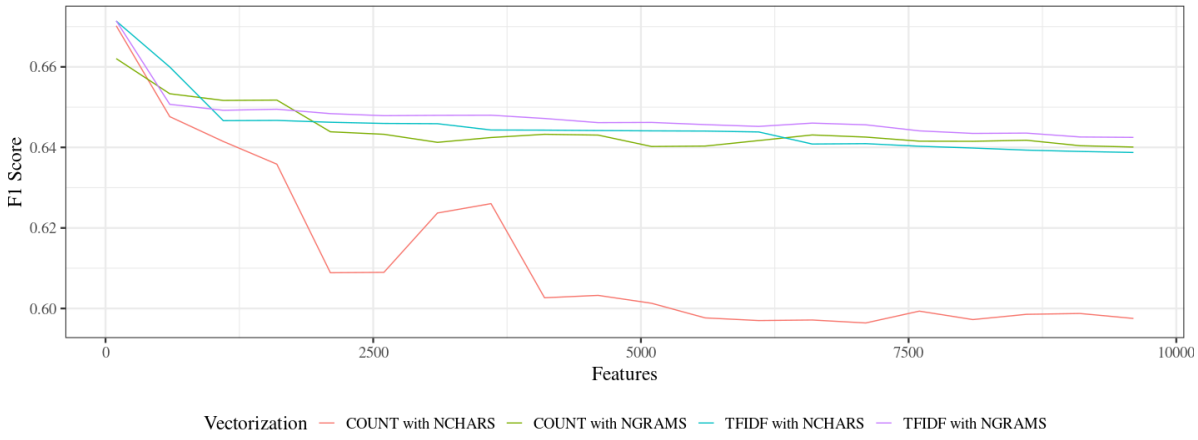

Figure 5: Logistic Regression 5-Fold Cross Validation Results

Figure 5 presents the results of the Logistic Regression applied on the four previously introduced corpuses. The model has proven interest towards smaller sets of features, in fact behaving the best with less than one hundred predictors. Count vectorizer with *nchars* tokenization has had a steady decrease in performance as the features increased. One reason explaining this may be the exponential increase in the length

of tweets subsequently to the decoding of emojis. This may have introduced unneeded noise that resulted in a strong drop of accuracy. On the overall, the remaining models have behaved similarly.

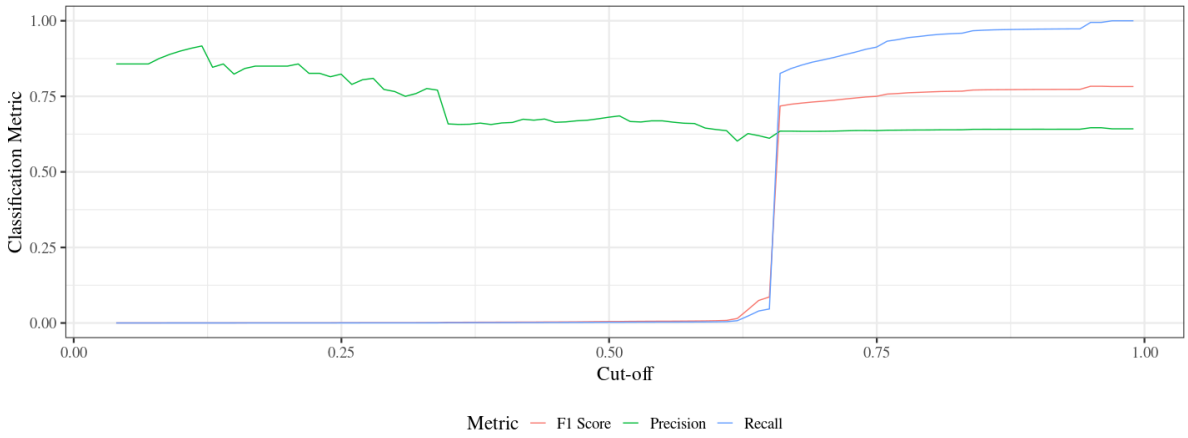

Figure 6: Choosing the Optimal Cut-off

It was decided, at this point, to proceed with the tf-idf vectorization with *ngrams* tokenization model as the best among the considered one. Feature selection for the latter has been set to one hundred and the optimal cut-off, considering *F1 Score* as our optimization metric, has been chosen by iterating the predicted probabilities through all the possible cut-offs lying between the minimum and maximum predicted probabilities. The precision of this range of values has been set to 0.001 (i.e. third decimal number). From **Figure 6** it is possible to denote the point of intersection between *Precision*, *Recall* and *F1 Score*, which is 0.66.

4.2. Baseline Portfolio Management

In order to assess the performance of the Logistic Regression it was decided to consider as metric the profit gained during the second week of data collection, namely the validation set previously cited. Considering that the model doesn't take into consideration the intensity with which prices increase or decrease on a daily basis, but simply try to binary classify the price movement for the next day, we'll proceed with a simple assumption making the comparison, between the baseline and our model, possible.

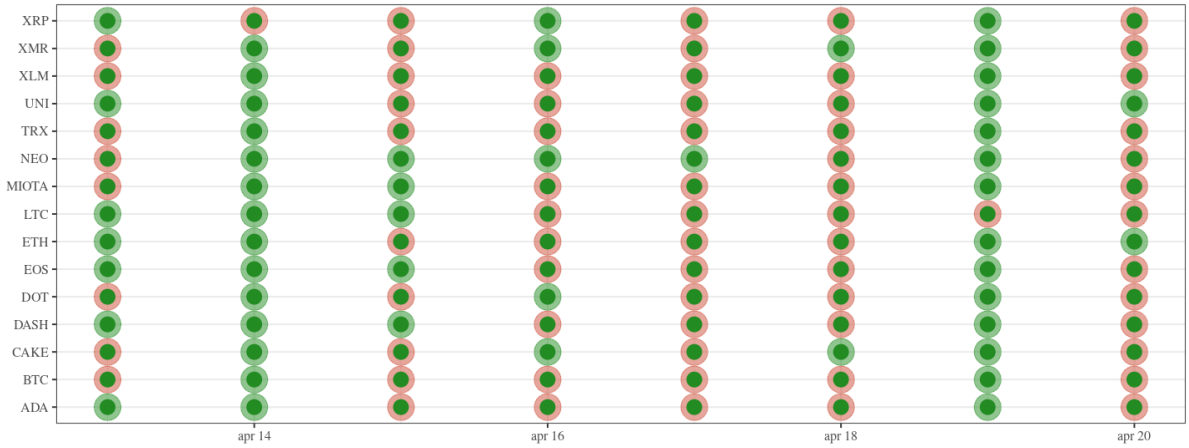

Figure 7: Choosing the Optimal Cut-off

We'll assume the baseline model to be a simple investor interested into the long term investments (i.e. in

our case weekly) who splits equaly its portfolio across all the considered cryptocurrencies. This investor buys on mondays and sells on sundays, without modifying its open positions (neither if the market is collapsing). For every daily change in price we'll consider it having a constant impact on the portfolio. In particular for each change in price from one day to another, if the price has increased the portfolio will as well increase by a 10%, while in case of a drop in price, the portfolio will lose a 10% of the invested money. **Figure 7** depicts the investment decisions undertaken by the baseline strategy. Outer circles represent the actual change in price from day to day, *green* if positive and *red* if negative. The inner circles depict the decisions taken by the simple investor which appears to have found himself in a week characterized by continuous decreases in prices for most of the cryptocurrencies. Considering, therefore, a weekly investment of 100$, the profit at last day for the investor with the baseline strategy is -14.70$ with a final portfolio value of 85.30$.

4.3. Custom Portfolio Management

Moving towards our custom portfolio management strategy, we'll proceed with specifying that, while in the baseline strategy it wasn't possible to change your ideas about the investment done, in the custom strategy we assume it to be possible (although we still keep the equal split of the portfolio across the considered cryptocurrencies). In particular the custom strategy is characterized by an investor interested in the short-term investments (i.e. daily basis). The latter buys at the closing price of the previous day and sells before the closing price of the next day. This allows the short-term investor to have more control over the high fluctuations happening within the cryptocurrency market. It is possible to denote from **Figure 8** a completely different investment strategy decided by the Logistic Regression. Almost all of the undertaken decisions move towards a prediction of a price decrease. The profit at last day for the investor with the custom strategy is +7.95$ with a final portfolio value of 107.95$.

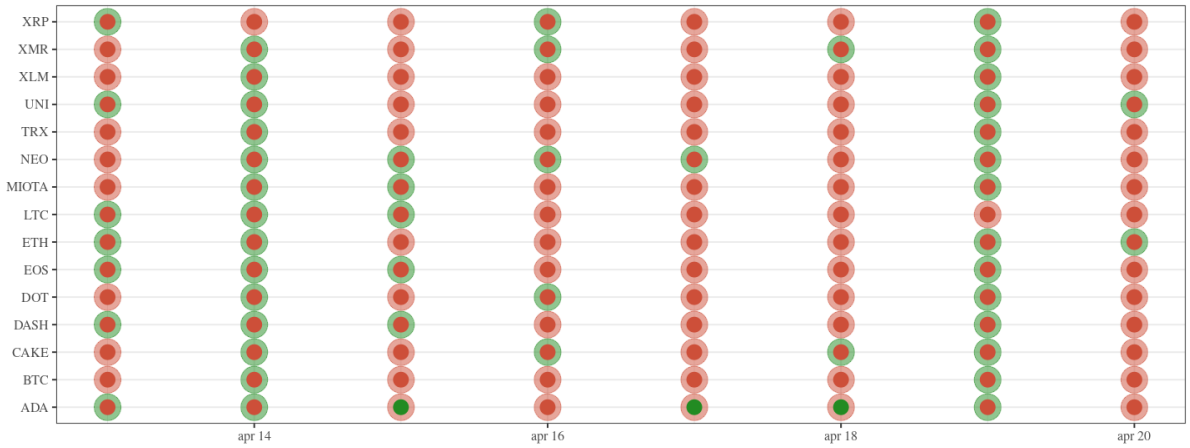

Figure 8: Choosing the Optimal Cut-off

5. Conclusion

This research has had the objective of identifying to what extent public Twitter sentiment can be used to forecast the price fluctuations of cryptocurrencies. The considered time window of Tweets appeared to have really small differences among the custom implemented labels. Further steps into the pre-processing phase may be undertaken in order to improve this issue. The Logistic Regression has been able to achieve, even if slight, a small increase of the portfolio, even though the predictions have been all, or almost, the same. Vectorization techniques appeared to behave similarly, letting us now that maybe different other approaches for word embeddings should be considered. As a further step into improving the Logistic Regression, positive and negative symbols across the Tweets should be tokenized rather than deleted. The bot presence (here almost nonexistent) should be as well considered as one of the causes of poor model performance, this because of the pre-defined Tweet's structure which do not add valuable information to the classification algorithm.

References

1. Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol. 1 : No. 3, Article 1.
2. Kristoufek L (2015) What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis. PLoS ONE 10(4): e0123923.
3. Ciaian, P., Kancs, D. and Rajcaniova, M., Virtual Relationships: Short- and Long-run Evidence from BitCoin and Altcoin Markets: JRC Working Papers in Economics and Finance - 5/2017 , Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-67442-6, doi:10.2760/133614, JRC107108.
4. Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. J. Comput. Sci. 2 (1), 1–8.