

# Predicting the Price of a “Small” Crypto-Currency Using Twitter-Based Sentiment Indicators

## The Case of Optimism

Alessandro Ciancetta, Alessandro Tenderini, Nikita Baklazhenko

March 3, 2023

## 1 Introduction

The efficient market hypothesis has been widely confirmed in financial markets. These markets are indeed characterized by their players devoting many resources in collecting and processing the information that regards price formation, as this is relevant to evaluating their portfolios and taking advantage of possible trading opportunities. The efficient market hypothesis has been studied also for the market cryptocurrencies. For instance, many studies tried to forecast the price movements of Bitcoins based on sentiment indexes that collect the opinions that professional investors posted on Twitter, but none of them showed conclusive results about the predictability of Bitcoin price. Since the Bitcoin is by far the most widespread and well-known cryptocurrency, it can be argued that its price may depend on many factors other than investors’ sentiment, as it is for almost any traditional financial instrument, and that sentiment indexes do not convey additional information. On the contrary, smaller and newer cryptocurrencies could be much more exposed to the “animal spirits” of the investors, precisely because they have no known history and the market perceives them distrustfully. According to this interpretation, new and small cryptocurrencies could react much more hysterically to fluctuations in investors’ perceptions of the currency’s soundness, with sudden rises and falls in the market price of these assets that are not justified by their fundamentals. Then, contrary to the efficient market hypothesis, the price of these currencies could be affected by metrics that are not publicly available and that can actually be gamed. Examples are Twitter posts or Telegram/Reddit accounts that are easily manipulated by creation of fake accounts or purchase of engagement on social media.

The aim of this work is to test the efficient market hypothesis on the market of new and small cryptocurrencies. We study the case of Optimism, a relatively small cryptocurrency traded since May 31st, 2022. We build a sentiment index about this cryptocurrency using all the tweets posted on Twitter containing the string ‘\$OP’ since the first day on the market of Optimism, i.e. we cover the whole lifetime of the currency with 223,899 tweets. We then study whether including the information about the sentiment improves the predictive performances of a baseline model. We find that the sentiment index does not improve the forecasting performances of the models, thus supporting the hypothesis that this market is efficient.

## 2 Fundamental analysis for cryptocurrencies

Fundamental analysis is a financial and investment approach to establish the “intrinsic value” of an asset or business by looking at its internal and external factors. This method is particularly useful for cryptocurrencies as the blockchain technology and its decentralized and transparent nature create a scenario of perfect information with many available information about every crypto asset. The fundamental analysis of cryptocurrencies typically includes three main metrics: on-chain metrics, project metrics, and financial metrics. On-chain metrics are data provided by the blockchain, such as transaction volume, active addresses, and network hash rate. Project metrics involve a qualitative approach to evaluating a cryptocurrency, considering factors such as the development team and the project’s roadmap. Last, financial metrics are classic financial factors such as market capitalization and trading volume. The combination of fundamental analysis in crypto and the transparent and decentralized nature of the blockchain is an argument in favour of the efficient market hypothesis.

When considering the efficiency of the market, it is important to note that if all of the relevant information is available, then no further significant information is disclosed by Twitter users and their sentiment. Therefore, while Twitter sentiment analysis can provide interesting insights into the mood and opinions of the community, it is unlikely to be useful in predicting future price movements. While it is true that for well-established cryptocurrencies like Bitcoin, the same might not hold for some smaller and new cryptocurrencies. Indeed, some smaller cryptos heavily rely on community support and the developer community is the driving force behind such projects. These kind of “small” crypto, which are usually tokens of a new blockchain, derive their value from the interaction and functioning of the underlying protocols. They are largely decentralized and rely on a network of users and developers to maintain, improve and create new protocols on the corresponding blockchain. In these cases, Twitter is a key platform for them to share updates, discuss technical issues, and analyze and evaluate new protocols to a wider audience.

Therefore, while Twitter sentiment may not be useful for predicting the price of well-established cryptocurrencies like Bitcoin, it could be useful in predicting the price movements of smaller and new cryptocurrencies, which are more exposed to “animal spirits” and rely on community sentiment and support.

## 3 Data

To collect data for our study, we used Twarc to retrieve all tweets containing the string ‘\$OP’, which is the usual way to refer to the cryptocurrency “optimism”. Further, we did not include retweets and we only considered tweets in English. We used the following code:

```
twarc2 search --archive --start-time "2022-05-30" --end-time  
"2023-02-20" "$OP lang:en -is:retweet" tweets_op.jsonl.
```

Because Optimism entered the market on May 30th we began researching tweets on that date until the 20th of February and we ended up with 220 000 total tweets. Next, to ensure that our dataset consisted of tweets relevant to cryptocurrencies, we filtered the tweets by examining the user’s description and keeping only tweets from users whose biography included one or more of the

following keywords: BTC, Bitcoin, DeFi, Crypto, Cryptocurrency, Blockchain, Finance, Analyst, Trader, Trading, Trade, NFT, web3, Optimism, OP. With this filtering step, we eliminated 30% of the tweets and we ended up with a dataset of 160 000 tweets. Regarding the price of Optimism, we used CoinGeko’s API in python to retrieve the daily price of OP from May 30, 2022 to February 20 2023.

## 4 Methodology

### 4.1 Text Preprocessing

The first step in the analysis involved tokenization of tweets, which was performed using the Natural Language Toolkit (nltk) library. The tweets were tokenized based on whitespace, resulting in a tokenized version of each tweet containing irrelevant characters, meaningless words, and emojis. The tokens were then converted to lowercase to reduce the dimensionality of the bag of words and simplify modeling. Although some uppercase abbreviations and names were initially considered to be kept, it was ultimately deemed appropriate to increase their dimensions, given the prevalence of uppercase words in crypto-specific tweets.

Several clearing steps were performed subsequently, including:

1. Removal of hyperlinks, retweet information, mentions of other users, and non-alphabetic characters.
2. Emojis were filtered out and placed in a separate column of a dataframe, as they were deemed to provide additional meaning beyond text.
3. Although bigrams and trigrams were initially explored, no improvement was observed, as unigrams were found to be the most prevalent. Also, most of the libraries for sentiment analysis were only working with unigrams, hence there was no point in using spatially structured texts and to incorporate bigrams and trigrams into sentiment analysis, unless a user-defined sentiment dictionary is defined. Thus, only basic sentiment modeling was attempted, word order was not considered crucial to the analysis.

English language stop-words, sourced from nltk, were removed, and additional user-specific stop-words were added to the list (this list could have been extended if your study tweets more in depth).

The next step in the tweet analysis process involved the application of two lemmatization methods: WordNetLemmatizer from NLTK and Pattern lemmatizer. WordNetLemmatizer works by mapping the input word to its lemma based on the part-of-speech (POS) tag of the word. Pattern lemmatizer, on the other hand, employs a rule-based approach to lemmatization, in which it applies a set of predefined rules to identify the base form of each word.

The choice to use lemmatization over stemming and trimming was made for several reasons. Firstly, lemmatization produces the base form of a word, which is more likely to be a meaningful word compared to stemming or trimming. This is important for tweet analysis because tweets often contain abbreviations, misspellings, and other forms of non-standard language, and retaining the base form of words can help ensure that their meaning is preserved.

Text	Cleaned	Cleaned_Pattern	Only_Emoji
\$BTC ltf\n\nlocal bottom formation between \$16...	btc ltf local bottom formation kk tcowtawonk	have btc ltf local bottom formation have kk ht...	
Great move. Zero-fee \$BTC trading can help to ...	great move zerofee btc trading help accelerate...	great move have zerofee have btc trade help ac...	
With the release of @CoreApp, the Avalanche B...	release coreapp avalanche bridge added support...	release have coreapp have avalanche bridge add...	👉
Some mining pool participants from Poolin just...	mining pool participant poolin sent btc binanc...	mine pool participant poolin send k have btc h...	👉
Crypto literally dumps during 100% of crypto m...	crypto literally dump crypto meetups another e...	crypto literally dump have have crypto meetup ...	

Figure 1: Example of the final cleaned dataset for BTC tweets.

These pre-processing steps were performed for both BTC and OP tweets separately

The word clouds displayed below highlight a significant disparity between the most frequently used terms in BTC and OP tweets. Nearly half of the common terms are specific to cryptocurrencies, such as OP, eth, btc, doge, sol, and lun. Additionally, certain terms may potentially provide market-relevant information regarding the market state, recommended market actions, and public perception (e.g., buy, market, price, bull, short, and high). These findings demonstrate that relying solely on unigrams is insufficient, as it may result in ambiguous sentiments (e.g., buy/don't buy). Instead, incorporating additional bigrams and even trigrams may be advantageous. Moreover, to conduct high-quality sentiment analysis, it is imperative to create a specific dictionary based not only on the crypto-market-related vocabulary but maybe on some crypto specific vocabulary.

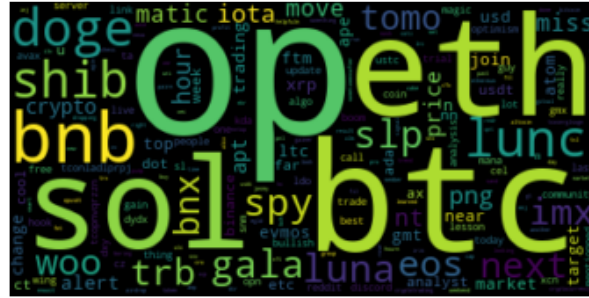


Figure 2: Word-cloud for OP tweets



Figure 3: Word-cloud for BTC tweets

Moreover, some of the terms were used way more often than the others. In accordance to Zipf's law in natural language's most frequent words occur at a rate approximately ten times greater than the next most frequent set of words, twenty times greater than the third most frequent set of words, and so on, hence we do have to filter out the most common irrelevant words to significantly reduce the size of corpus to consider.

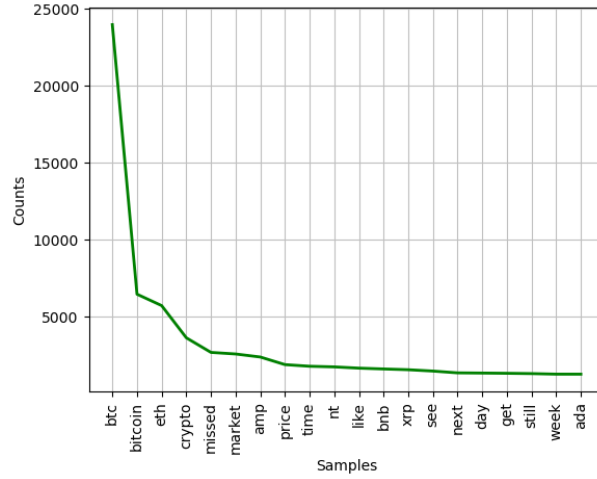


Figure 4: Most common words for BTC tweets

Emojis were also very frequently used over the tweets, and many of them can provide a lot of information on market perception (for example the rocket, market chart, alarm sign etc). However, their usage must be carefully considered, and their meaning is often contextual, requiring thorough analysis.











Emoji										
Count	1352	1255	1227	979	712	647	466	390	386	342

Figure 5: Most commonly used Emoji

To conclude, some modifications to the text processing can be done, but given method will provide the stable results for the most of the text, and open for additional fine-tuning regarding the kind of data.

## 4.2 Sentiment Analysis

Sentiment analysis can be performed using various approaches such as rule-based, lexicon-based, machine learning-based, and advanced methods. Lexicon-based analysis is one of the most commonly used approaches due to its simplicity, interpretability, and widespread use. Therefore, in this work, we considered two different lexicon-based sentiment score models: TextBlob and VADER.

TextBlob computes the polarity score by averaging the sentiment scores of individual words in the text. The sentiment scores are obtained from a pre-labeled lexicon of words and phrases associated with a polarity score ranging from -1 (negative) to 1 (positive), which are then normalized to a range of -1 to 1. On the other hand, VADER uses a pre-labeled lexicon of words and phrases with sentiment scores ranging from -4 to 4, along with sentiment-intensity calculation and rules to compute the polarity score.

To compare the performance of VADER and TextBlob, we conducted sentiment analysis on a tweet corpus and individual tweets. The sentiment scores for the entire corpus were 0.11 for TextBlob and 0.05 for VADER, which give us a baseline for the typical sentiment in the text. The average sentiment scores for individual tweets were 0.09 for TextBlob and 0.05 for VADER. However, these scores do not account for the influence of the author, such as retweets, followers, and comments.

To obtain a more representative score, we removed tweets with a sentiment score of 0, resulting in a sentiment score of 0.14 for TextBlob, indicating that the text became more positive on average. We also conducted the same analysis using a different lemmatizer (Pattern Lemmatizer) and obtained similar scores (less than 5

Overall, the differences in the computation of polarity scores and the presence of additional features in VADER such as sentiment-intensity calculation and rule application can result in different sentiment scores for the same piece of text compared to TextBlob. The specific choice of sentiment analysis tool should be made based on the characteristics of the text data and the requirements of the analysis.

For our future modeling, we opted to utilize Vader sentiment scores with the exclusion of 0 sentiment tweets. To obtain the daily sentiment score, we computed the average sentiment score for each day, with the time period being adjustable based on preference.

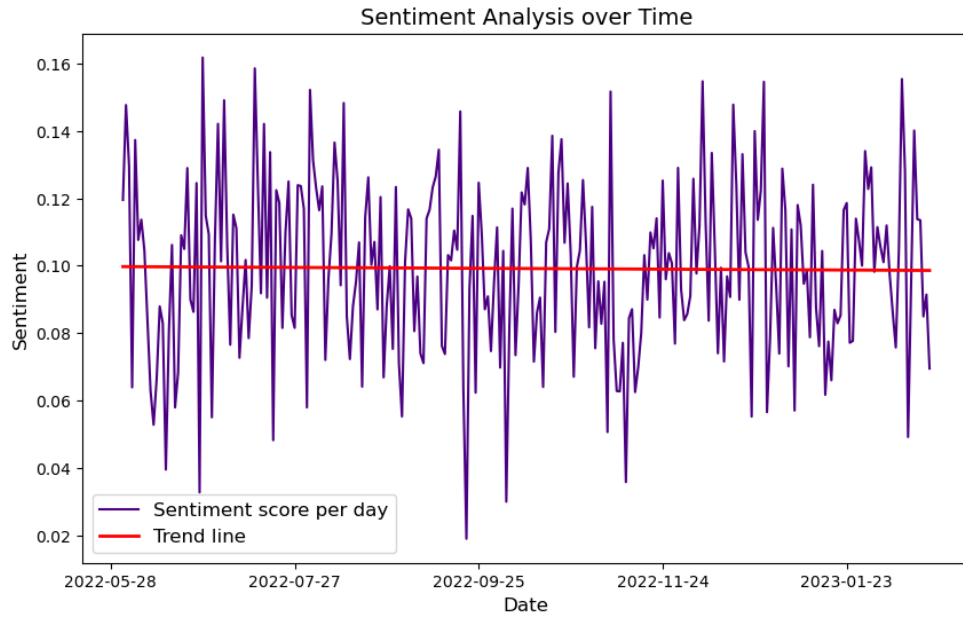


Figure 6: Daily Sentiment Scores

Upon analyzing the resulting charts, it is apparent that both the rolling mean and daily sentiment scores exhibit a high degree of inconsistency and susceptibility to fluctuations. However, upon comparing the sentiment charts with the overall market trends, some similarities and patterns become evident, such as a sharp fall in market prices corresponding to a decrease in sentiment scores and vice versa.

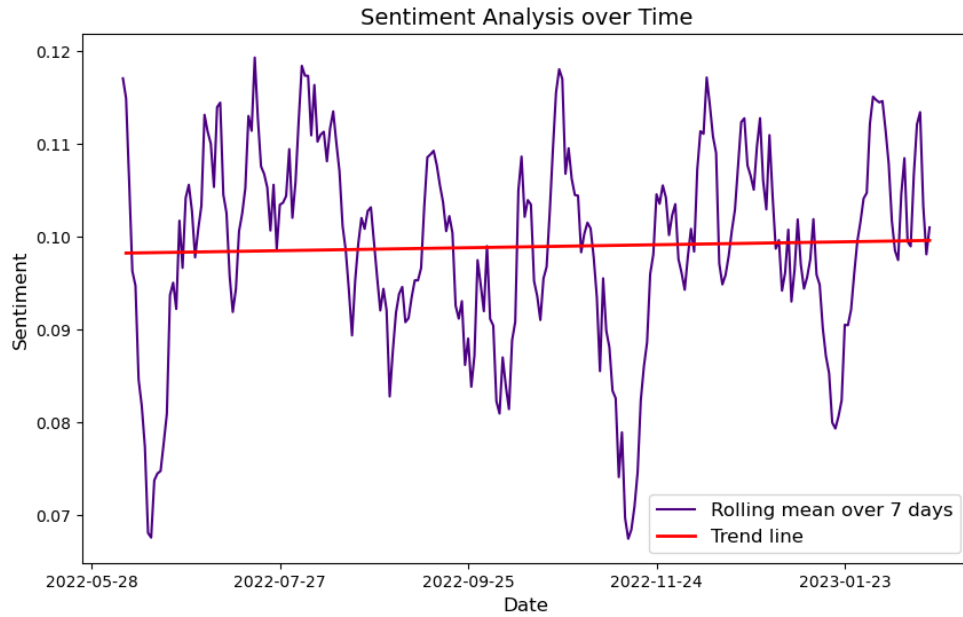


Figure 7: Rolling Mean Sentiment scores



Figure 8: Market Prices of the Coin over observed period

Notably, the most significant drops in sentiment scores are observed during prolonged periods of market decline, as opposed to brief fluctuations. Moreover, once the market stabilizes, we observe an increase in sentiment scores to previous levels, even in instances where the currency prices remain below their previous levels. Overall, the sentiment scores appear to remain stable over time, despite the overall decline in currency prices.



### 4.3 Prediction Models

In this section, we will evaluate the predictive ability of the Twitter sentiment data in predicting the price of Optimism. To this end, we will compare a simple Distributed Lag Model (DLM) in two versions. The first specification does not include the sentiment indicators, while the second one is unrestricted and includes both indicators as predictors. Let  $y_t$  denote the first-difference of the price of Optimism at time  $t$ ,  $z_t$  and let  $s_{1t}, s_{2t}$  denote the two sentiment indicators at time  $t$ . Let also  $z_t$  denote the market capitalization at day  $t$ . We consider the following models:

$$\begin{aligned} y_t &= \beta_0 + \sum_{k=1}^p \beta_k y_{t-k} + u_t \\ y_t &= \beta_0 + \sum_{k=1}^p \beta_k y_{t-k} + \sum_{k=1}^s \beta_{p+k} s_{1,t-k} + \sum_{k=1}^s \beta_{p+s+k} s_{2,t-k} + v_t, \end{aligned} \tag{1}$$

where  $p$  denotes the autoregressive order and  $s$  is the number of lags considered for the sentiment index. In this application, we will consider  $p = 14$  and  $s = 7$ , so the model will explicitly account for the price of Optimism over the past two weeks and the sentiment over the last week. We choose these lag orders as they led to the best forecasting performances in the experiments. Moreover, we consider the same autoregressive order  $p$  for both models in order to make them more comparable.

The main reason to opt for a simple linear model is the limited sample size. Indeed, as we already pointed out, Optimism is a new crypto-currency. If, on the one hand, this fact has the advantage of making us able to manage the tweets about Optimism spanning the entire its life until now, on the other hand we have access to few examples for training and assessing the model. Our final dataset, indeed, counts 264 daily observations. This number is too little for training highly non-linear models without a serious risk of overfitting, especially because part of the observations has to be devoted for testing the performances of the model.

The procedure that we use for assessing the forecasting performances is based on comparing the RMSE obtained in an expanding-window forecast evaluation. In particular, we fix the first 220 observations in the dataset (spanning the period 2022-06-01 to 2023-01-06), we fit the model using the data from this period, we use the fitted model to predict the price of Optimism next day, and then we append the true observed price to the training data and repeat the procedure. The evaluation ends at the last available observation in the dataset, and the result is a vector of 1-step-ahead predictions that at each period mimic the prediction that the model could have made using the information available at that period.

It is worth spending some lines to introduce the data. Fig.9 reports the three series in the dataset, while Fig. 10 reports the autocorrelograms of the prices in first difference and of the sentiment index. From the autocorrelogram of the differenced prices, it is pretty clear that no evident autocorrelation exists in the day-by-day price changes, so that there are little hopes that purely autoregressive model can achieve good predictive results. That further justify the idea of testing whether additional data can prove useful in predicting the price fluctuations. Whereas the correlation between the price levels and the first sentiment index is around 0.68, the correlation between the price changes and the first sentiment index is only 14.5%. This suggests that it may

exist a spurious correlation between the price in levels and sentiment indexes that requires the price series to be made stationary by first-differencing.

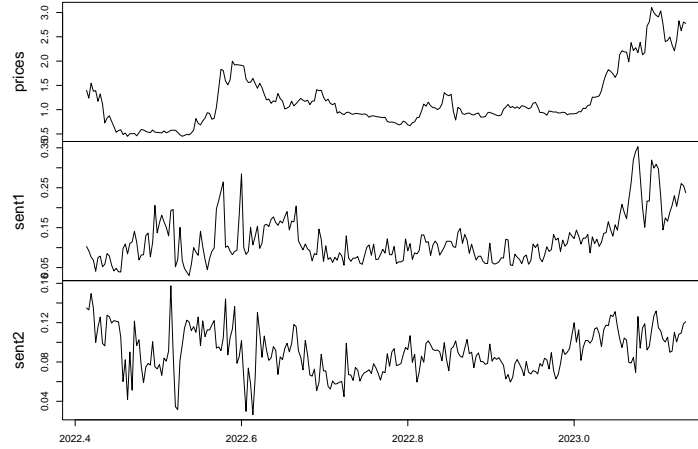


Figure 9: Final dataset. The plot represents the price (in U.S. dollars) of Optimism and two sentiment indexes obtained from the tweets about Optimism in the period between June 1st, 2022 and February 19, 2023.

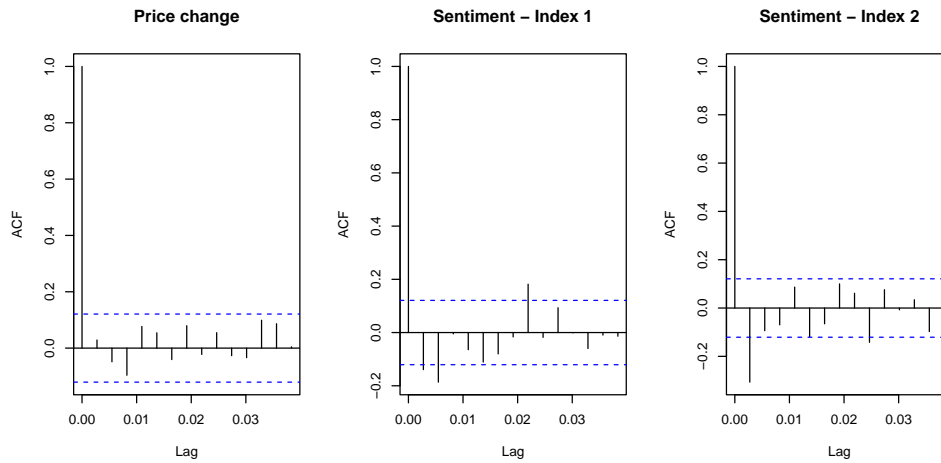
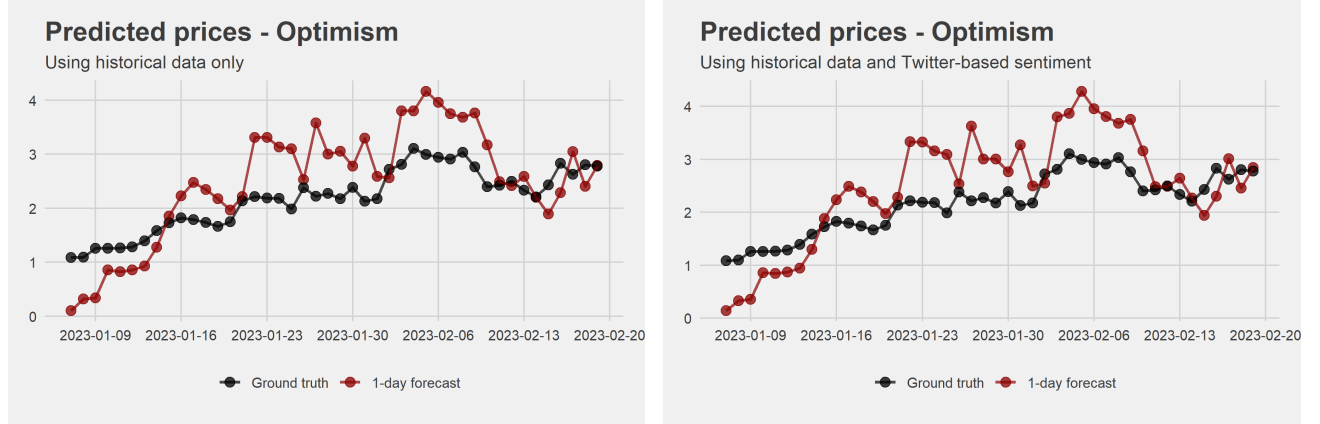


Figure 10: Autocorrelograms for the series in the dataset. The price changes do not show any apparent autocorrelation, while the sentiment index are correlated with at least their first lag.

## 5 Results

The plots below report the results of the expanding-window forecast evaluation.



As it is easily argued by looking at the plots, adding the sentiment index to the prediction does not induce any qualitative difference in the predictive performance of the model. This fact is further confirmed by inspecting the RMSE for the two models, which is almost identical with a value of about 0.70 in both cases.

## 6 Conclusions and further research

This project aimed to investigate the impact of imperfect information on the future price of a new and small crypto-currency, Optimism. We collected tweets from the active trading community on Twitter for the entire lifespan of this crypto-currency to study the daily sentiment and evaluate its potential as a predictor for future prices. Our findings suggest that the sentiment index is not a useful predictor for future price, and this result is in line with the efficient market hypothesis. Indeed, publicly available information and the fundamentals of Optimism already capture all relevant information for predicting future prices. These findings support the notion that investors should rely on publicly available information and fundamentals as indicators of future prices, while social media sentiment analysis may not provide any additional useful insights.