Journal of Computer Science

Original Research Paper

Sentiment Identification on Tweets to Forecast Cryptocurrency's Volatility

[1]Rafael Calixto Ferreira de Araújo, [2]Alex Sandro Roschildt Pinto and [3]Mauri Ferrandin

[1]Department of Computer Science, Universidade Federal de Santa Catarina, Brazil
[2]Department of Informatics and Statistics, Federal University of Santa Catarina, Brazil
[3]Department of Control Engineering and Automation, Federal University of Santa Catarina, Brazil

Article history
Received: 06-10-2022
Revised: 16-02-2023
Accepted: 22-02-2023

Corresponding Author:
Rafael Calixto Ferreira de Araújo
Department of Computer Science, Universidade Federal de Santa Catarina, Brazil
Email: rafaelcf.araujo@gmail.com

Abstract: Cryptocurrencies have had a huge presence on social media since their creation. In current days, the constant increase of the mass of data produced by this environment has attracted several researchers to try to identify patterns with the potential to allow identification of the volatility in the crypto market before it happens. This approach involves the concept of the wisdom of the crowds, a popular theory in the economy field that in the current days may have the perfect tools to prove itself true. This scenario creates an opportunity to unite two new technologies, social media, and cryptocurrencies to the newest Natural Language Processing (NLP) tools, and produces a study in a rich and unexplored field. Executing a detailed sentiment analysis, this study intents to analyze the forecast of the volatility of cryptocurrencies through the detection and evaluation of several categories of sentiments on messages on twitter when it is associated with a specific cryptocurrency. To achieve this, an NLP model was trained with the GoEmotions dataset to identify and categorize emotions, and results were used to calculate the forecast of the cryptocurrency. Index terms cryptocurrency, social media, Natural Language Processing (NLP), GoEmotions.

Keywords: Cryptocurrency, Social Media, Natural Language Processing, NLP, GoEmotions

Introduction

Since the beginning of the internet, multiple discussions and debates are developed online by communities formed by members that share common interests, from old forums to modern's social media, those discussions have a strong potential to generate useful information. In the financial debates, one of the greatest issues debated is the possibility to create an autonomous and digital currency, native to the digital environment and maintained only by the users, having a structure safe enough to avoid fraud and thus gaining the trust of users (Marple, 2021). As a solution to this demand, the Bitcoin project (Nakamoto and Bitcoin, 2008) has been developed and generated all the crypto environment. This context is important to understand the intrinsic relationship between digital communities and cryptocurrencies and how, since the beginning, it became natural to express the understanding and feelings about cryptocurrencies on digital platforms. Also, the cryptocurrency environment evolved into a complex and valuable trading market that still spreading the adoption of cryptocurrency through several scenarios, including the traditional trading market (Özkaynar, 2022). Even in the traditional market, the relationship between social media and the stock market is becoming clearer. The episode of the GameStop stock market prices is an example of the current power of social media influencing even the traditional market (Gianstefani *et al*., 2022). However, if the relationship between the traditional trade market and social media only now is starting to be evident, the crypto market is already born with a natural proximity to it, having several important debates and speculations happening on social media platforms (Mai *et al*., 2015). It produces a huge mass of data daily, calling attention to an opportunity to explore the economic theories as the wisdom of the crowd (Hossain *et al*., 2022) applying it to the investment field and analyzing the potential of social media to generate accurate forecasts for the volatility of cryptocurrency. Although the technologies listed at this point are relatively new, it is already possible to find in the literature several works exploring the relationship between those technologies on the financial market forecast (Nassirtoussi *et al*., 2014). Furthermore, besides most of

those works providing interesting results, they are far to be conclusive. The techniques applied in those works found in literature, most of the time, use simple sentiment analysis techniques, or other algorithms that use texts just to extract simple measures and concentrate on other features. In works recently published it is possible to find investigations where recent techniques and technologies of Natural Language Processing (NLP) are applied (Tran, 2022), however, it is far from filling all the possibilities of the theme. This study uses a model of NLP trained with the GoEmotions (Demszky *et al*., 2020) dataset to extract 27 categories of sentiment, plus neutral, and use those sentiments as the main parameter to generate a forecast for cryptocurrency volatility. Those categories were based on the classification proposed by Ekman (1992) being an advanced classification, going away further than the simple classification as positive, negative, or neutral. With this approach, this study intends to explore social media data extracting more detailed features to investigate the capacity of prediction of social media to cryptocurrency values. The system presented in this study detects the sentiment in social media posts feeding an algorithm that generates a score that is combined with the post's metadata. All the values are inputted in a formula that generates the forecast of the volatility indicating not only the increase or decrease of the prices but the proportion of the volatility's movement. After a round of predictions was generated, the results were compared to the volatility of the cryptocurrencies in the market for 8 different time windows, analyzing the performance of the forecasts with standards metrics in different scenarios. This study is structured with an introduction section where the theme is presented followed by the sections scope delimitation, related works, proposed approach, experimental evaluation, discussion, and conclusion.

Scope Delimitation

This study analyzes the messages related to cryptocurrencies published on the internet. Considering social media an environment that currently concentrates most of the messages with opinions about many things, a selection to elect the best social media to be used as the source of data for this study has been done considering the following criteria: Number of users, the volume of debates about crypto assets, accessibility to the data and available metadata about the post. Considering the last criterion, the forums were excluded from the selection because of the lack of metadata related to users' profiles. To evaluate the volume of users in each social media was used the raking provided by the portal Statista (2022) where are listed the platforms with the highest number of users, removing from the list the platforms for direct messages and the social media focused on images or videos. Thus, the final list of social media was formed by facebook, sina weibo, twitter, reddit, and quora. Through the second criterion, facebook is excluded, presenting a lack of relevant debates about cryptocurrencies on the platform. Finally, the third criterion excludes Sina Weibo for most of the content is in Chinese languages, making the content of the texts inaccessible to this study. With a final list with twitter, reddit, and quora, this study elects to use only Twitter as a source of data for being the option with the highest number of Monthly Active Users (MAUs), including reddit and quora in opportunities for future works. The selection of the cryptocurrencies was done based on the list of the cryptocurrencies with the highest market cap provided by Cap (2023). The market cap is a strong indication of the adoption of cryptocurrency and how valuable it is for the market. This list provided the top five cryptocurrencies with the highest market cap and used as the unique excluding criteria for those crypto assets that belong to the category of the stablecoin. Thus, the initial list of cryptocurrencies started with Bitcoin, Ethereum, Tether, Binance coin (BNB), and USD coin. Two of those assets are classified as a stablecoin, Tether, and USD coin, remaining in the list Bitcoin, Ethereum, and Binance coin (BNB). All three remaining cryptocurrencies were elected to be analyzed in this study. To measure the performance of the forecast score it will be compared to the quote of each cryptocurrency obtained from the Binance exchange in the format acronymous for Open, High, Low, and Close (OHLC), the measures used in each time window in a candle chart). From those values in OHLC format have been considered only the values of opening for each time window. The Binance exchange has been elected because it's the only exchange with the quote for the BNB. One particularity of this exchange is that it doesn't have the quotes in US dollars, so this study used the quote of the cryptocurrencies in Tether, a stablecoin backed in US dollars. After the extraction process of the data from Twitter to identify the sentiments in each message, an experimental equation to generate the score for the forecast was applied to generate a group of forecasts where standards market metrics were applied to measure the performance of the forecasts. The metrics applied were a correlation, accuracy, precision, recall, and F1-score, and the forecasts were compared to eight-time windows starting from the immediate time to two days of interval. This study will not analyze the causation relation in the forecast, letting this theme be explored in future works.

Materials and Methods

The review of the literature shows that the research around the use of content from forums and social media to forecast the financial market is a field already been explored since the beginning of this century. This study used a systematic review as a base for the literature review and spread it by updating it with more recent works (Nassirtoussi *et al*., 2014).

Table 1: Related works

Reference	Text type	Source	Dataset
Tran (2022)	Tweets and posts	Twitter and reddit	NA
Özkaynar (2022)	News	News	NA
Raheman *et al*. (2022)	Tweets and posts	Twitter and reddit	100,000
Kane *et al*. (2022)	Posts	Reddit	58,000
Aslam *et al*. (2022)	Tweets	Twitter	40,000
Seong and Nam (2021)	News	News	1,397,800
Groß-Klußmann *et al*. (2019)	Tweets	Twitter	102,737,864
Burnie and Yilmaz (2019)	Submissions	Reddit	338,415
Zhang *et al*. (2016)	Post	Weibo	139,855
Liu *et al*. (2015)	Tweets	Twitter	NA
Chatrath *et al*. (2014)	News	Bloomberg	NA
Jin *et al*. (2013)	General news	Bloomberg	361,782
Yu *et al*. (2013)	social media and other	Blogs, news, and twitter	52,746
Vu *et al*. (2012)	Tweets	Twitter	5,001,460

Each article found in the literature explores different sources of data (news from relevant portals, forums, or social media), the financial market segment (traditional assets or cryptocurrencies), and techniques to generate forecasts of the assets. Most of the approaches use a combination of techniques since algorithms developed specifically for this task to Machine Learning (ML) models trained to calculate the forecast. The literature also reviews approaches to generate the forecasts where was applied artificial intelligence techniques, such as the use of unsupervised ML models to classify the assets in groups (Liu *et al*., 2015), training of supervised models with different ML techniques, feeding them with features extracted from the text and analyzing the performance of each model (Seong and Nam, 2021) and the application of techniques of sentiment analysis through different ML techniques (Yang *et al*., 2018). Some works investigate more deeply the financial market of new technologies, such as the crypto market (Özkaynar, 2022). However, in recent works is possible to find surveys with a scope similar to this study, applying new techniques of NLP, but with significant differences, such as the use of other datasets than the GoEmotions (Tran, 2022). Table 1 compares features of the systematic review used as a reference to this study. With this comparison is possible to observe the great diversity of setting in this field are and, thereby, the large range of possibilities that still unexplored.

Proposed Approach

To extract and process the data from twitter a pipeline has been developed using the python language and the library tweepy that provides functions to easily connect with twitter's API and extract the data according to the input parameters. This library provides a function to search in tweets containing a specific term where was used the name of the cryptocurrencies, loading the messages where those were mentioned. Python is an open-source programming language largely applied in projects of data processing, being elected to this project for the large set of available libraries for ML and NLP. The election of the library tweepy intends to simplify the process of data extraction, abstracting the complexity of communication with Twitter's API. The data extracted from twitter shows a few challenges in text processing because of the range of possibilities in free text. To pre-process the text a function was built to normalize the data tokenizing each word and then excluding punctuations, stop words, emoticons, and other characters that could produce noise to the analysis as URLs, retweet marks, and so on. The spacy library was elected because it provides functions that simplify this process, being elected to be used to build this part of the algorithm. This library is a toolkit for NLP processing providing intuitive functions and is one of the most popular on the market for NLP processing (Aswathy *et al*., 2022). However, to handle specifically the emoticons were applied the library emoji, this library was selected because provides functions unavailable on Spacy. With the clean text, the texts were ready to be processed by the ML model to identify all the sentiments in the tweets. The model was trained with the Bidirectional Encoder Representations from Transformers (BERT) (Liu *et al*., 2021) algorithm applying the transformers architecture (Kane *et al*., 2022). This algorithm was selected because it is considered the state of art for NLP processing, having a high performance in the training of a model for sentiment analysis. The dataset GoEmotions

(Demszky *et al*., 2020) is loaded from the TensorFlow repository and it already provides it split into the train, test, and validation datasets. The tokenizer and the pre-trained model squeeze BERT from transformers were applied. At the end of the training process, the Receiver Operating Characteristic Curve (ROC) curve in Fig. 1 shows that only the emotions of pride and realization had an accuracy under 80%. This dataset was selected because of its data classification over social media content, having very close proximity to the data extracted from twitter. A combination of libraries was applied to the machine learning processing, using the framework Torch as the base of the process, the Tez library to provide the model function (Zeydan and Mangues-Bafalluy, 2022), the scikit learn library to generate the metrics during the training process and the Transformers library to provide the architecture, the optimizer and the BERT framework to NLP machine learning training (Devlin *et al*., 2018). Those libraries were selected to simplify the process of deep learning, bringing functions that turn the process easier than using straight TensorFlow.

This dataset contains messages extracted from the social media Reddit and classified into 27 categories, plus neutral. When exploring the amount of representation of each emotion, it's clear that the neutral category was super represented as Fig. 2 shows, where the category neutral is presented by the number 27. However, it's expected to doesn't affect the result in this study considering that both sources of messages were social media platforms and the frequency of neutral messages on Twitter is expected to be similar to reddit.

In the process of training, the model inputs the hyperparameters with 0.3 of dropout and a linear transformation with a value of 768 for the length of the sample, that is the number of labels (28) and the size of the output. Also, it was applied as an optimizer of the function AdamW, an optimizer for its best performance in batch processing of large scale, being useful when applied to neural networks such as BERT (You *et al*., 2019). The loss function was defined as a binary function of cross entropy for each neuron and a forward function was defined to orchestrate the training process for each epoch. All the training process was executed on a CPU processing, with 8 epochs, batch size of 64, and 10 jobs, leaving approximately 30 h to complete the full training of the model.

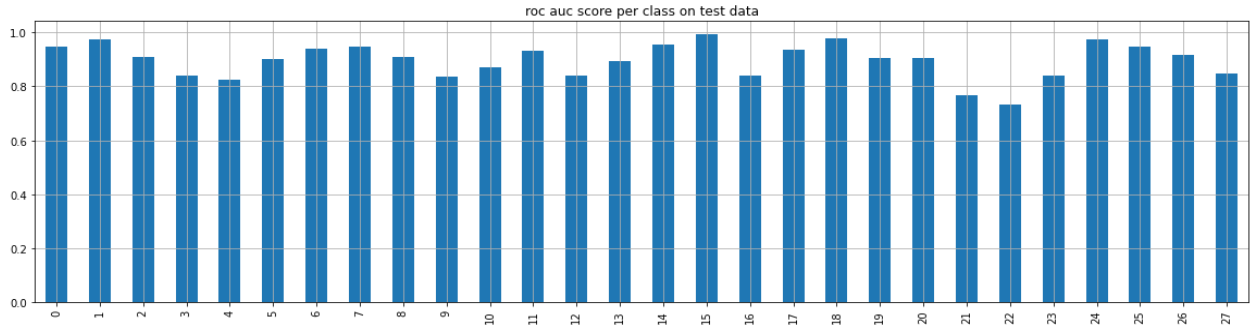

Fig. 1: ROC curve in the training

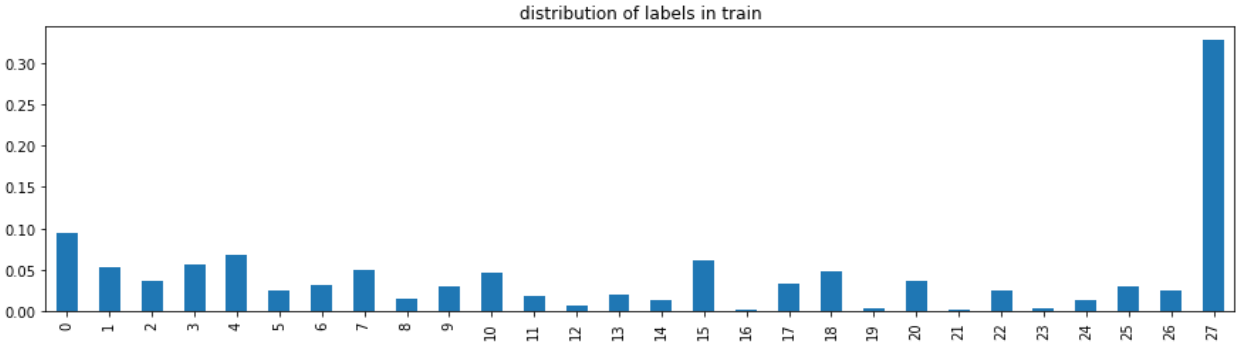

Fig. 2: Distribution between classes

With the model ready and the data pre-processed, the ML classifies the sentiment of each tweet extracted from social media and uses it to calculate the score to forecast the volatility of each cryptocurrency. The data extracted from twitter's API also contains the metadata of each tweet, so the algorithm creates a python dictionary with the most relevant metadata about the tweet's author, including the user's id, number of author's followers, current timestamp subtracted by the date of creation of the account in the timestamp. Other metadata are related to the tweet's text, such as the number of retweets and the number of likes. Using those metadata and the score of sentiment classification, extracted by the weight relation presented in Table 2, the algorithm calculates the forecast based on each message processed. All the steps of the algorithm are represented in Fig. 3. The weights presented in Table 2 were filled with an initial value without systematic criteria and it will be changed in future works, testing different combinations of values aiming to improve the performance of the forecast. In this way, in this study, those weights were applied generally in the process to generate all the scores. The criteria to generate the initial configuration of weights was to set positive values for positive emotions and negative values for negative emotions (Demszky *et al*., 2020). The ambiguous emotions received either positive value, except the sentiment of confusion, which received a negative value. The range of values goes from 10-10, where 0 represents neutral and the other values were set based on the understanding of its relevance to the crypto market, being an initial guess to this stage of the project and letting open the opportunity to explore in future works. With the sentiment's weight, the algorithm can execute the forecast for cryptocurrency volatility. Thus, a formula was been developed applying the values of the meta-data to potentiate tweets according to their interaction and the range of the author. Furthermore, the value has smooth based on how much time it was tweeted, considering the date in timestamp (seconds). It's important to highlight that this formula isn't related to any market measure being just an adjustment to the value of the sentiment according to the metadata of the tweet:

$$P = \frac{\left(\left(1 + Nf * 0{,}001\right) + \left(Nrt * 0{,}01\right)\right) * \left(S * 0{,}1\right)}{1 + \left(\left(Dtc - Dtn\right)\right) * 0{,}001}$$

Table 3 indicates the meaning of each variable. The investigation in this study is to measure the sentiment captured in social media adjusting it with the metadata and analyzing if the crypto market follows the sentiment of social media with the same intensity or, at least, in the same direction (positive or negative).

Table 2: Sentiments

Index	Sentiment	Weight
0	Admiration	8
1	Amusement	4
2	Anger	-7
3	Annoyance	-9
4	Approval	7
5	Caring	-8
6	Confusion	-5
7	Curiosity	5
8	Desire	9
9	Disappointment	-10
10	Disapproval	-9
11	Disgust	-5
12	Embarrassment	-6
13	Excitement	10
14	Fear	-8
15	Gratitude	6
16	Grief	-2
17	Joy	1
18	Love	2
19	Nervousness	-4
20	Optimism	10
21	Pride	8
22	Realization	3
23	Relief	4
24	Remorse	-10
25	Sadness	-3
26	Surprise	1
27	Neutral	0

Table 3: Variables

Letter	Description
Nf	Number of followers
Nrt	Number of retweets
S	Sentiment
Dtc	Creation date
Dtn	Current date

To highlight some limitations of this approach, the source of data can be used as an example. Although social media provides a source of data never seen before, it still being just a small fraction of the data related to the crypto market. Other sources as news platforms and expert analysis have the potential to improve the algorithm. Furthermore, other limitations are highlighted throughout this study as opportunities for future works, as the use of general data to train the ML model instead uses a mass of data related to the crypto market and the simple system of weights applied that can be improved with an investigation to find the values that represent the weight of each sentiment more accurately.

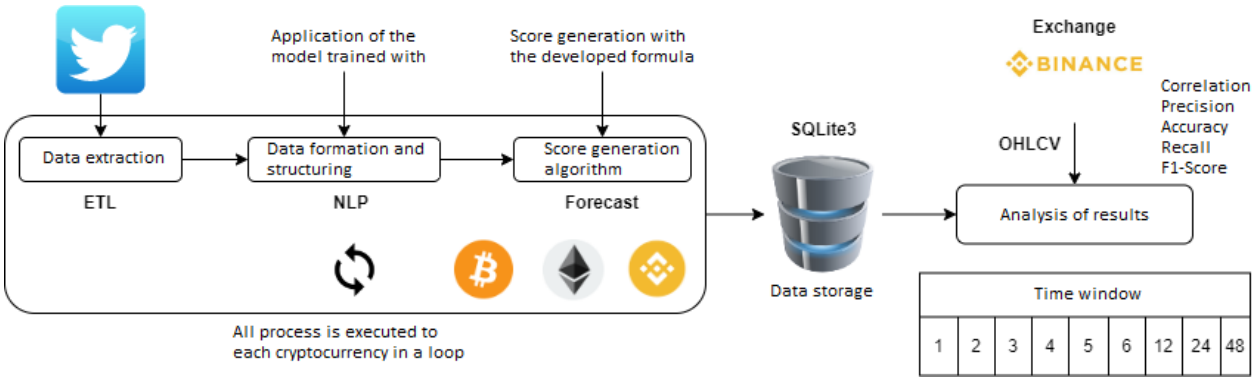

Fig. 3: Algorithm steps

Experiment Evaluation

The algorithm executed properly, taking around 40 sec to execute the process to each cryptocurrency, from the extraction of data on Twitter to generate the forecast score. During this process around 200 tweets are processed, generating the forecast with the sum of all the tweets captured. The slowest step of the process is the sentiment classification taking around 30 sec to be executed. This latency in the processing of data is a limitation to trades of high frequency such as operations with less than one minute. After running 57 inferences to generate the score for each one of the cryptocurrencies, the results were analyzed to measure the performance of the forecast. To execute the analysis an algorithm was developed using the Jupyter notebook, starting with the load of the forecast saved in the SQLite3 database and then splitting the data into datasets organized by cryptocurrency and time interval as in Table 4. This database was selected to store the data because of its simplicity and liability to handle the data of this process. The results reveal a small correlation between the scores of the forecast and the values of the quote, but the most significant is that the correlation to the 1 h for Bitcoin is relatively more significant than to the other cryptocurrencies. Besides the values of correlation pointing to a lack of a straight link between the score value and the cryptocurrencies' quote value, it doesn't mean that the score fails as a tool to forecast volatility. It can be used in several ways and combined with other indexes to increase its capacity for prediction. To evaluate the forecast performance in a bigger picture was considered just the direction of the forecast and the direction of the volatility in 8 different time windows, meaning, for example, that if the forecast indicates an increase in the price and the volatility moves in the same direction, it will be considered a correct forecast. Thus, the results were classified as shown the Table 5.

With that, a confusion matrix has been developed for each time window, allowing the calculation of the metrics of precision (Fig. 4), accuracy (Fig. 5), recall (Fig. 6), and F1-score (Fig. 7). The results of the performance review a good performance to forecast the volatility with an accuracy of up to 50% in the first 24 h to all the cryptocurrencies. The chart of recall calls attention to the high values and it was observed that the system has a low output of false negatives, meaning that it could be an interesting tool to detect decreases in the cryptocurrencies' volatility. Also, the charts show that the algorithm has a better performance in a time window between 12 and 24 h for all three cryptocurrencies with an accuracy of 90% in the forecast of Ethereum's volatility. Also, it is important to highlight that the forecast performance decreases strongly for BTC and ETH, by the same behavior isn't followed by BNB. It may happen for the different nature of the BNB in the market, being a cryptocurrency used internally by the exchange Binance, don't being negotiated on the free market. The metrics applied don't reflect the performance to forecast the intensity in each movement of the forecast, but the correlation shows that it wasn't a close fit. This indicates that the values of the weights of the sentiments need to be better defined, searching for values that result in more precise forecasts. The metrics point to a better performance in predicting the volatility of ETH than other cryptocurrencies. This may happen for the better fit of the ETH's technology with the social media platform than the other cryptocurrencies technology (Guidi, 2021). However, in the charts is possible to identify a strong similarity in the performance of both cryptocurrencies, and this similarity is also reflected in the volatility's value. This pattern observed is another indicator of the reflection of social media on the crypto market.

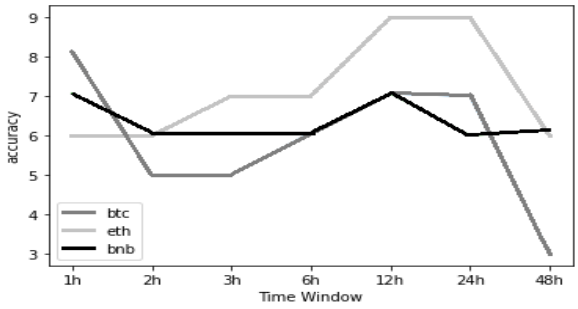

Fig. 4: Accuracy

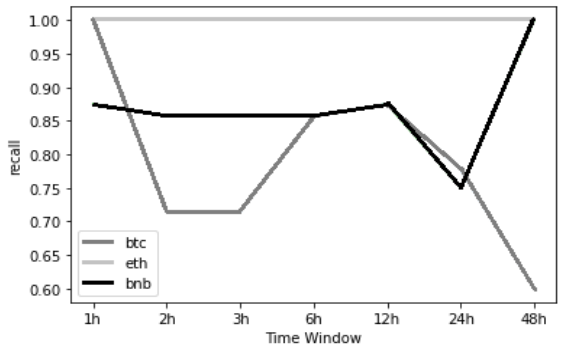

Fig. 6: Recall

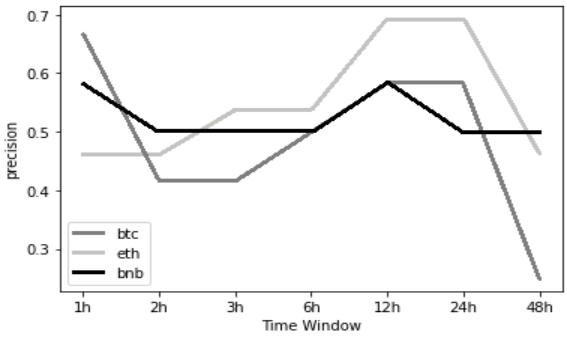

Fig. 5: Precision

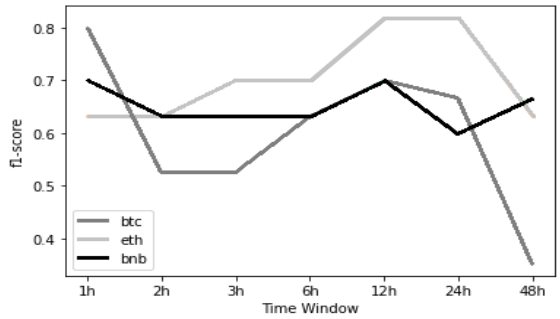

Fig. 7: F1-score

Table 4: Datasets

Cryptocurrency	Time interval	Correlation	Dataset
Bitcoin	Same time	0.2374	Test
	1 h	0.3961	Test
	2 h	0.3647	Test
	3 h	0.2190	Test
	6 h	-0.5401	Test
	12 h	-0.0154	Test
	24 h	-0.1541	Test
	48 h	0.0000	Test
Ethereum	Same time	-0.5382	Test
	1 h	-0.5484	Test
	2 h	-0.0356	Test
	3 h	0.3210	Test
	6 h	-0.0511	Test
	12 h	-0.7175	Test
	24 h	-0.7609	Test
	48 h	0.0000	Test
Binance coin	Same time	0.0514	Test
	1 h	-0.2430	Test
	2 h	-0.6223	Test
	3 h	-0.5531	Test
	6 h	0.4022	Test
	12 h	0.0982	Test
	24 h	0.1298	Test
	48 h	0.0000	Test

Table 5: Confusion matrix

	Up	Down
Up	True Positive (TP)	False Positive (FP)
Down	False Negative (FN)	True Negative (TN)

Discussion

The results presented by the experiment evaluation follow others' works being far from conclusive but revealing possible ways to develop future works. In the work developed by Aslam *et al*. (2022), the ML model is trained with data classified from tweets related to cryptocurrencies using a Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) deep learning process. Analyzing the performance, the model presented a recall of 0.97 in the identification of the sentiments positive, negative, and neutral on the validation dataset. In this study, the model trained with the dataset GoEmotions presented a recall of 0.94, considering the average of all the recalls presented in Fig. 1. Thus, the performance of both models can be considered similar. In the work of Raheman *et al*. (2022), sentiment analysis generates scores that are classified as positive, negative, contradictive, and sentiment. The analysis of the correlation between the generated score and the market prices revealed a perk correspondence with values that goes from 0.2-0.2 considering all the cases analyzed. Although in this study the correlation ranges more expressive values, they don't have a consistency to be considered relevant. Finally, the work developed by Groß-Klußmann *et al*. (2019), applied a different approach to the dataset splitting it into financial experts and non-experts, facing some issues with the small mass of data in the expert group. In this study this differentiation happens when the metadata of each tweet is used in the formula, giving a higher value weight to the tweets from sources with more followers and interactions. However, those metrics don't provide liability in the detection of knowledge on the crypto market, but just popularity. Furthermore, Groß-Klußmann *et al*. (2019) elaborated a complex and detailed strategy to generate the score, applying the others source of data and testing different ML algorithms, reaching a higher accuracy of 0.68. In this study the higher accuracy reached was 0.9 in the prediction for the ETH.

Conclusion

The algorithm performance demonstrates the viability of this approach to forecasting cryptocurrencies' volatility with the best performance to detect the direction of volatility in a time window between 12 and 24 h. The relationship between social media and the crypto market already was observed in other works (Wołk, 2020) and this relation is still sustained in the observations done in this study. Although currently, the algorithm's forecast alone isn't enough to predict the crypto market, it can be improved or can be used as an interesting solution combined with other indexes, contributing, for example, to detect when the cryptocurrencies values will decrease. The results indicate that this theme is worth being deeper developed in future works, analyzing different weights configurations to the sentiments to improve the capability to detect not only the direction of the volatility but its intensity. Another improvement to this study is the retraining of the machine learning model applying the technique of transfer learning to fine tune the model trained with the GoEmotions dataset to increase its performance in a financial context. Finally, this field will still be developed by innovations and modern tools, being revisited and explored to better understand the potential of those tools.

Acknowledgment

The author would like to thank all the professors that were part of the master degree course.

Funding Information

This article was funded with the corresponding author's own resources.

Author's Contributions

Rafael Calixto Ferreira de Araújo: Development and execution of all experiments, data analysis, and written of the manuscript.

Alex Sandro Roschildt Pinto and Mauri Ferrandin: Contributions to conception and designed of the manuscript, a general reviewed of the theme and manuscript.

Ethics

As an original article, it contains unpublished material. The corresponding author and all other authors have read and approved the manuscript and no ethical issues involved.

References

Aslam, N., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2022). Sentiment analysis and emotion detection on cryptocurrency related Tweets using ensemble LSTM-GRU Model. *IEEE Access*, *10*, 39313-39324. https://ieeexplore.ieee.org/abstract/document/9751065/

Aswathy, A., Prabha, R., Gopal, L. S., Pullarkatt, D., & Ramesh, M. V. (2022, February). An efficient twitter data collection and analytics framework for effective disaster management. In *2022 IEEE Delhi Section Conference (DELCON)* (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/9753627

Burnie, A., & Yilmaz, E. (2019, July). An analysis of the change in discussions on social media with bitcoin price. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 889-892). https://doi.org/10.1145/3331184.3331304

Cap, C. (2023). Today's cryptocurrency prices by market cap. *CoinmarketCap. https://coinmarketcap.com*

Chatrath, A., Miao, H., Ramchander, S., & Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. *Journal of International Money and Finance*, *40*, 42-62. https://doi.org/10.1016/j.jimonfin.2013.08.018

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*. https://doi.org/10.48550/arXiv.2005.00547

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169-200. https://doi.org/10.1080/02699939208411068

Gianstefani, I., Longo, L., & Riccaboni, M. (2022). The echo chamber effect resounds on financial markets: A social media alert system for meme stocks. *arXiv preprint arXiv:2203.13790*. https://doi.org/10.48550/arXiv.2203.13790

Groß-Klußmann, A., König, S., & Ebner, M. (2019). Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market. *Expert Systems with Applications*, *136*, 171-186. https://doi.org/10.1016/j.eswa.2019.06.027

Guidi, B. (2021). An overview of blockchain online social media from the technical point of view. *Applied Sciences*, *11*(21), 9880. https://doi.org/10.3390/app11219880

Hossain, M. M., Mammadov, B., & Vakilzadeh, H. (2022). Wisdom of the crowd and stock price crash risk: Evidence from social media. *Review of Quantitative Finance and Accounting*, *58*(2), 709-742. https://doi.org/10.1007/s11156-021-01007-x

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., & Ramakrishnan, N. (2013, August). Forex-foreteller: Currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1470-1473). https://doi.org/10.1145/2487575.2487710

Kane, A., Patankar, S., Khose, S., & Kirtane, N. (2022). Transformer based ensemble for emotion detection. *arXiv preprint arXiv:2203.11899*. https://doi.org/10.48550/arXiv.2203.11899

Liu, L., Wu, J., Li, P., & Li, Q. (2015). A social-media-based approach to predicting stock movement. *Expert Systems with Applications*, *42*(8), 3893-3901. https://doi.org/10.1016/j.eswa.2014.12.049

Liu, Z., Jiang, F., Hu, Y., Shi, C., & Fung, P. (2021). Ner-BERT: A pre-trained model for low-resource entity tagging. *arXiv preprint arXiv:2112.00405*. https://doi.org/10.48550/arXiv.2112.00405

Mai, F., Bai, Q., Shan, J., Wang, X. S., & Chiang, R. H. (2015). The impacts of social media on Bitcoin performance. *SSRN Electronic Journal*. 2015. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2545957

Marple, T. (2021). Bigger than Bitcoin: A theoretical typology and research agenda for digital currencies. *Business and Politics*, *23*(4), 439-455. https://doi.org/10.1017/bap.2021.12

Nakamoto, S., & Bitcoin, A. (2008). A peer-to-peer electronic cash system. *Bitcoin. URL: https://bitcoin. org/bitcoin. pdf*, *4*(2). https://www.klausnordby.com/bitcoin/Bitcoin_Whitepaper_Document_HD.pdf

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653-7670. https://doi.org/10.1016/j.eswa.2014.06.009

Özkaynar, K. (2022). Marketing strategies of banks in the period of Metaverse, Block-chain and Cryptocurrency in the context of consumer behavior theories. *International Journal of Insurance and Finance*, *2*(1), 1-12.

Raheman, A., Kolonin, A., Fridkins, I., Ansari, I., & Vishwas, M. (2022). Social Media Sentiment Analysis for Cryptocurrency Market Prediction. *arXiv preprint arXiv:2204.10185*. https://doi.org/10.48550/arXiv.2204.10185

Seong, N., & Nam, K. (2021). Predicting stock movements based on financial news with segmentation. *Expert Systems with Applications*, *164*, 113988. https://doi.org/10.1016/j.eswa.2020.113988

Statista. (2022). Most popular social networks worldwide as of January 2022, ranked by number of monthly active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Tran, T. (2022). Predicting Digital Asset Prices using Natural Language Processing: A survey. *arXiv preprint arXiv:2212.00726*. https://doi.org/10.48550/arXiv.2212.00726

Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012, December). An experiment in integrating sentiment features for tech stock prediction in twitter. In *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data* (pp. 23-38). https://aclanthology.org/W12-5503.pdf

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*, *37*(2), e12493. https://doi.org/10.1111/exsy.12493

Yang, S. Y., Yu, Y., & Almahdi, S. (2018). An investor sentiment reward-based trading system using Gaussian inverse reinforcement learning algorithm. *Expert Systems with Applications*, *114*, 388-401. https://doi.org/10.1016/j.eswa.2018.07.056

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., ... & Hsieh, C. J. (2019). Large batch optimization for deep learning: Training BERT in 76 min. *arXiv preprint arXiv:1904.00962*. https://doi.org/10.48550/arXiv.1904.00962

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, *55*(4), 919-926. https://doi.org/10.1016/j.dss.2012.12.028

Zeydan, E., & Mangues-Bafalluy, J. (2022). Recent Advances in Data Engineering for Networking. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/9743922

Zhang, L., Zhang, L., Xiao, K., & Liu, Q. (2016, August). Forecasting price shocks with social attention and sentiment analysis. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 559-566). IEEE. https://ieeexplore.ieee.org/abstract/document/7752291