A Sentiment and Emotion Annotated Dataset for Bitcoin Price Forecasting Based on Reddit Posts

Pavlo Seroyizhko* , **Zhanel Zhexenova*** , **Muhammad Zohaib Shafiq*** , **Fabio Merizzi*** , **Andrea Galassi†** and **Federico Ruggeri†**
University of Bologna, Bologna, Italy
{pavlo.seroyizhko,zhanel.zhexenova,muhammad.shafiq6,fabio.merizzi}@studio.unibo.it,
{a.galassi,federico.ruggeri6}@unibo.it

Abstract

Cryptocurrencies have gained enormous momentum in finance and are nowadays commonly adopted as a medium of exchange for online payments. After recent events during which GameStop's stocks were believed to be influenced by WallStreetBets subReddit, Reddit has become a very hot topic on the cryptocurrency market. The influence of public opinions on cryptocurrency price trends has inspired researchers on exploring solutions that integrate such information in crypto price change forecasting. A popular integration technique regards representing social media opinions via sentiment features. However, this research direction is still in its infancy, where a limited number of publicly available datasets with sentiment annotations exists. We propose a novel Bitcoin Reddit Sentiment Dataset, a ready-to-use dataset annotated with state-of-the-art sentiment and emotion recognition. The dataset contains preprocessed Reddit posts and comments about Bitcoin from several domain-related subReddits along with Bitcoin's financial data. We evaluate several widely adopted neural architectures for crypto price change forecasting. Our results show controversial benefits of sentiment and emotion features advocating for more sophisticated social media integration techniques. We make our dataset publicly available for research.

1 Introduction

Cryptocurrencies, often referred to as crypto-currency or cryptos, are any type of currencies that exist virtually and are secured by encryption or cryptography to safeguard transactions. These currencies are decentralized and do not have any regulating or governing authorities to track them or to create new units. Bitcoin (BTC) is one of the most dominant cryptocurrency that is driven by investor expectations, and its demand is becoming increasingly appealing [Foley *et al.*, 2019], with investors adopting it to diversify their portfolios. Due to the similarity of its features with speculative stocks and decentralised nature, perception and sentiments of investors are likely to drive the price of bitcoin [Kraaijeveld and De Smedt, 2020].

Among the many challenges presented by economics and finance businesses there is indeed the modeling of customers' sentiment, including their polarity and diversity, and the intentions that may be associated to them. For this purpose, social media constitute rich and useful sources of information that can be analysed to obtain useful insights. Additionally, it is worth noting that social media may contain misinformation and biases that can potentially exacerbate the information extraction process, eventually making it unreliable [Cao, 2022].

Reddit is a social media platform that has recently gained a lot of attention due to its influence in affecting cryptocurrencies trend. For instance, the subReddit *r/wallstreetbets* allegedly played a role in influencing the GameStop's stocks in 2021.[1] Reddit is structured in communities, i.e., subReddits, with a user base of more than 50 milions and more than 1.5 bilions of monthly visitors and has become one of the most popular social media in the US.[2] Recently, the gained momentum of cryptocurrencies has been observed and enhanced by the increasing number of dedicated subReddits. In particular, almost every influential cryptocurrency has its dedicated subReddit. The popularity of some of them, such as r/wallstreebets (12 milion subscribers), r/CryptoCurrency (4.8 milion subscribers), and r/Bitcoin (4.2 milion subscribers), highlight the widespread interest of cryptocurrencies in social media.[3]

One peculiar aspect of cryptocurrencies is that they are not regulated by governments or other international institutions, but rather public opinions represents the main cause of crypto price changes. Thus, we commonly observe the rise and fall of popular cryptocurrencies due to their high dependence on people's opinion. In this perspective, the role of public communities and social media platforms like Reddit is highly influential in determining the trend of a cryptocurrency.

This phenomenon has inspired researchers to leverage publicly available social media information, i.e., people's opin-

*Equal contribution
†Contact Authors

[1]https://edition.cnn.com/2021/01/27/investing/gamestop-reddit-stock/index.html
[2]https://www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/
[3]https://frontpagemetrics.com/

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2022), pages 50-56, Vienna, Austria, July 24th, 2022

ions, to evaluate their effect on cryptocurrency price forecasting. To do so, as a common conveyor of people's opinion, researchers have extracted sentiment information from textual data to enrich the set of input features for a forecasting model [Wooley *et al.*, 2019]. In particular, the statistical analysis of recent work has shown that a correlation between extracted social media information and crypto price trends does exist [Kraaijeveld and De Smedt, 2020]. It is worth noting that the integration of public opinion is not anew in the field of finance. Other similar domains like stock market forecasting have shown promising results when integrating heterogeneous information from social media [Kearney and Liu, 2014]. Nonetheless, the existing work on this matter is still preliminary under several perspectives. First, few available datasets providing scraped social media data exist nowadays and the number decreases drastically when considering the subset with sentiment annotations [Loginova *et al.*, 2021]. Second, the integration of social media information is carried out by a limited subset of sentiment features [Carmezim, 2018; SocialGrep, 2021], while a wide variety of sentiment and emotion tools for efficient and accurate extraction can be leveraged on this matter.

In this work, we propose a novel dataset for Bitcoin price forecasting, called Bitcoin Reddit Sentiment Dataset (BRSD). We create this dataset by pre-processing and integrating an existing dataset of Reddit posts and comments with crypto price values. We employ a wide suite of sentiment and emotion recognition techniques to automatically annotate Reddit textual data in addition to the existing comment-level sentiment annotations. We formulate the task of crypto price forecasting with both price, sentiment, and emotion features with different widely adopted architectures. In particular, we carry out an ablation study to evaluate the impact of extracted social media information by considering three different input configurations: (i) price data only, (ii) sentiment and emotion data only, (iii) all available data. We make our dataset publicly available for research.[4] Our experimental setting shows controversial results on the impact of sentiment and emotion data. In particular, the best performing models performance is deteriorated when adding social media features. In contrast, the less performing model shows trend learning capabilities only when considering the full set of features.

In Section 2 we analyze similar datasets in the field of crypto currency forecasting. Section 3 describes the dataset creation process. In Section 4, we introduce our experimental setting. Section 5 concludes.

2 Related Work

The automatic forecasting of stock market and cryptocurrency prices has gained a lot of interest in the past years [Yenidoğan *et al.*, 2018; Lahmiri and Bekiros, 2019; Pang *et al.*, 2020]. Indeed, cryptocurrencies are currently widely adopted in finance due to their popularity [Foley *et al.*, 2019]. Nonetheless, they are subject to public influence since they rely on decentralization and are not managed or regularized by government institutions. In particular, public opinion and crypto prices are strongly related [Wooley *et al.*, 2019; Phillips and Gorse, 2018a]. Recent works on this topic has shown that social media can affect price changes and stand as reasonable indicators of the current economics trends [Kraaijeveld and De Smedt, 2020; Phillips and Gorse, 2018b].

Such a characteristic allowed the introduction of datasets for crypto price forecasting with users opinions extracted from online social media like Reddit [Loginova *et al.*, 2021; Prajapati, 2020; Leukipp, 2022] and Twitter [Pant *et al.*, 2018]. However, there are few available datasets regarding Reddit posts and comments with sentiment information [Loginova *et al.*, 2021; SocialGrep, 2021]. Loginova et al. introduced a large dataset collected over 768 days (from February 2017 to April 2019) regarding several social media like Reddit, Bitcointalk and CryptoCompare. Social media data is collected by leveraging Pushshift APIs,[5] a widely adopted tool for extracting such information [Leukipp, 2022; SocialGrep, 2021; Reinerink, 2022], and later merged with financial data of five popular cryptocurrencies, namely BTC, ETH, LTC, XPR and XMR, from the website *coinmarketcap.com*.

In contrast, our dataset leverages multiple state-of-the-art sentiment and emotion recognition tools like lexicons, rule-based, and contextual techniques rather than employing aspect-based sentiment analysis. For what concern the time period, our dataset covers only one month (August 2021) and focuses only on the BTC cryptocurrency. Nonetheless, its size is comparable to previous works and it is publicly available. Another difference with respect to [Loginova *et al.*, 2021] is the time range on which we base our predictions. Due to the length of our dataset, we predict the price in 15 minutes based on the last hour, whereas, they predict the price in 1 day based on features collected in 7 days averaged with a rolling window of 1 day.

Table 1 provides a summary of existing datasets highlighting their characteristics.

3 Dataset Creation

We rely on a publicly available dataset of Reddit content and on a well-known finance platform to build our dataset. We firstly collect and process data for each of the two datasets. Subsequently, we derive our dataset by merging the information extracted from these two datasets and report statistics.

3.1 Kaggle Dataset

We selected a publicly available dataset of Reddit posts and comments from Kaggle [SocialGrep, 2021]. The dataset contains 250,569 posts and 3,756,097 comments collected from Reddit in August 2021. In particular, the following subReddits were considered during data collection:

- /r/cryptocurrency
- /r/cryptocurrencyclassic
- /r/cryptocurrencyico
- /r/cryptomars
- /r/cryptomoon

[4] https://www.kaggle.com/datasets/paulsero/bitcoin-reddit-sentiment-dataset

[5] https://github.com/pushshift/api

Name	Date	Social media	Data	Source	Sentiments	Sentiment type
[Carmezim, 2018]	2018	Twitter	1,578,627 tweets	unspecified, general	Yes	Word polarity
[Leukipp, 2022]	Jan 2022- May 2022	Reddit	518,610 posts	51 subreddit	No	
[Loginova *et al.*, 2021]	Feb 2017- Apr 2019	CryptoCompare, Reddit, and Bitcointalk	Respectively 78,902, 2,635,046, and 1,643,705 texts	r\cryptocurrency, news headlines	Yes	Vader, TextBlob, JST, and TS-LDA
[Reinerink, 2022]	Nov 2017- Mar 2018	Reddit	2,161,000 comments	r\cryptocurrency	No	
Kaggle dateset [SocialGrep, 2021]	August 2021	Reddit	3,756,097 comments and 250,569 posts	7 subreddits	Yes	Kaggle-Sentiment
[Pano and Kashef, 2020]	May 2022 - Jun 2022	Twitter	4,169,709 tweets	BTC related tweets	No	
BRSD (our dataset)	August 2021	Reddit	2,755,329 comments and 55,002 posts	7 subreddits (BTC only)	Yes	Vader, TextBlob, BERT, RoBERTa, Flair, Kaggle-Sentiment

Table 1: A comparison between the Bitcoin Reddit Sentiment Dataset and similar cryptocurrency datasets that integrate social media information.

Dataset → Property ↓	Original	Deleted	Bots/Ads	Cleaned
Posts	250,569	127,891	67,676	55,002
Avg Posts/Hour	336.96	171.99	89.68	73.95
Avg Posts/Minute	5.71	2.9	1.52	1.25
Comments	3,756,097	998,568	2,200	2,755,329
Avg Comments/Hour	3,708.91	1,109.90	2.96	3,705.89
Avg Comments/Minute	62.84	20.13	0.05	62.75

Table 2: Kaggle Reddit dataset statistics throughout our preliminary pre-processing pipeline. We report information about filtered posts and comments after each pre-processing phase.

- /r/cryptomoonshots
- /r/satoshistreetbets

The Kaggle dataset does not come with additional pre-processing steps concerning posts filtering and text cleaning. To this end, we devised a preliminary pre-processing phase to only select posts and comments that could potentially correlate with crypto price changes. We removed posts that were tagged as *'deleted'* or *'removed'* since they do not have any relevant textual content. In particular, these posts constituted around 49% of the Kaggle dataset. This is a known phenomena of popular and controversial subReddits like the ones related to cryptocurrencies. Additionally, we remove every comment that contained blacklisted words regarding scam/phishing or advertisements (e.g. *'giveaway'* or *'pump join'*. We use the custom blacklist suggested by [Kraaijeveld and De Smedt, 2020].

Subsequently, we applied a series of traditional text normalization operations for the remaining posts. These operations included (i) merging the title and the body of a post; (ii) cleaning URLs and special symbols and (iii) removing stopwords. This preliminary pre-processing phase reduced the number of posts to 121,593 and the number of comments to 2,755,329.

Lastly, on Reddit the presence of bots is common. We detected and removed spam or bots sentences from Reddit posts by relying on the set of heuristics proposed by [Kraaijeveld and De Smedt, 2020]. This approach assumes that bot and spam sentences are short length sentences that are frequently repeated throughout a document. The spam and bot filtering procedure further reduced the number of posts to 55,002. Table 2 summarizes the described pre-processing process.

3.2 Extracting Sentiment and Emotion from Reddit texts

The use of Reddit posts and comments for crypto price forecasting relies on the extraction on sentiment features that could potentially act as indicators for price changes. The Kaggle dataset comes with comment-level sentiment annotations, which we refer to as **Kaggle-Sentiment**. Nonetheless, no information about how these annotations have been produced is reported.

We have therefore decided to enrich our data through additional unsupervised labels. Indeed, a wide suite of off-of-the-shelf tools for accurate and efficient sentiment extraction can be found in the literature. In this work, we consider multiple state-of-the-art tools to extract sentiment and emotion features from Reddit posts and comments. We employed the following tools:

- **Flair** [Akbik *et al.*, 2018]: A multilingual library that comprises of several state-of-the-art contextual text embedding methods like BERT. Flair methods are pre-trained on the well-known IMDB movie review dataset. We adopted the English embedding models of this library. In particular, each method extract sentiments scores in the $[0, 1]$ range and polarity scores as either being `negative` or `positive`. In particular, the two scores are combined to provide a single sentiment score.
- **TextBlob** [Loria, 2018]: A high-level library built on top of NLTK [Bird *et al.*, 2009]. TextBlob extracts sentiment scores in the $[-1, 1]$ range. Additionally, the library extracts subjectivity scores ($[0, 1]$ range) and related subjectivity intensity scores. Subjectivity determines if a text is subjective or factual, whereas the intensity score of a word quantifies to what extent the word modifies the meaning of the next word. We use the extracted sentiment, subjectivity and intensity scores as features.
- **VADER** [Hutto and Gilbert, 2014]: A rule-based model based on a fixed list of lexical features. VADER does not take into account the contextual information of a word unlike Flair. VADER extracts classifies the sentiment of a text as `positive`, `negative` or `neutral` and return their probabilities which we use a features.
- **BERT** [Devlin *et al.*, 2019]: We employ a BERT model pre-trained on multilingual product reviews for senti-

Tool	Feature	Range
TextBlob	Polarity	[-1, 1]
	Subjectivity	[-1, 1]
Kaggle	Kaggle-Sentiment	[-1, 1]
RoBERTA	Anger	[0, 1]
	Joy	[0, 1]
	Optimism	[0, 1]
	Sadness	[0, 1]
VADER	Positive	[0, 1]
	Negative	[0, 1]
	Neutral	[0, 1]
	Compound	[-1, 1]
Flair	Flair-Sentiment	[-1, 1]
BERT	BERT-Sentiment	[1, 5]

Table 3: Score ranges of sentiment analysis tools

ment analysis.[6] In particular, the extracted sentiment of a text ranges from 1 (negative) to 5 (positive).

- **RoBERTA** [Liu *et al.*, 2019]: We employ a RoBERTa model pre-trained on the TweetEval benchmark [Barbieri *et al.*, 2020] for the emotion recognition task.[7] The model is trained on English tweets and identifies the following emotion categories: *anger*, *joy*, *optimism*, *sadness*. We use the prediction probabilities of emotion classes as features.

If not explicitly stated, we consider raw probability scores for categorical variables (e.g. VADER, BERT and RoBERTA). The features extracted using each tool are first transformed to obtain a uniform set of value ranges (e.g. Flair and TextBlob sentiment scores) and later normalized. In total, we extract a set of 12 sentiment and emotion features for each input Reddit post or comment.

The described suite of sentiment and emotion features is employed to encode each Reddit post and comment. In particular, we aggregate individual comments sentiment and emotion feature set by summing them. This information is added to the provided sentiment annotations of the original Kaggle dataset. Table 3 provides a summary.

3.3 Finance Bitstamp Dataset

We extracted raw crypto price values from the exchanger Bitstamp platform.[8] We considered the time period of the posts and comments in the Reddit Finance dataset to extract time-aligned raw crypto price values. We observed that the Reddit Finance dataset covered the period of August 2021 and downloaded corresponding finance data from Bitstamp. We denote this dataset as the Finance Bitstamp dataset. The collected dataset contains crypto price values on a minute base and has 530,324 entries. Each entry contains metadata (*id*, *date*, *crypto currency*, *open*), trend-related values (*high*, *low*, *close*) and price information (*Volume BTC*, *Volume USD*).

3.4 BTC Reddit Sentiment Dataset

We integrate the extracted sentiment features from Reddit posts and comments to the Finance Bitstamp dataset. To do that, Reddit posts and comments are first temporally aligned with crypto price values. We leverage the reported timestamp metadata of both datasets to perform this operation. The alignment of Reddit posts and comments with crypto price values is at minute-level. We opted for a minute-based alignment motivated by two main observations. First, the Reddit Finance dataset reports textual data at minute-level. Thus, the integration of crypto price values based on this granularity is straightforward. Second, this dataset contains information extracted from the month of August 2021. A more coarse-grained granularity would significantly reduce the amount of samples for forecasting. Note that it is still possible to operate with more coarse-grained granularities (e.g. hours) to further evaluate the impact of social media information. Indeed, the influence of social media opinions works at higher scales (e.g. hours, days) and cause-effect delays have to be considered as well.

We collect Reddit posts published in each minute and aggregate them by summing the set of sentiment features, obtaining a new layer of sentiment and emotion annotations. The built dataset, which we denote as BTC Reddit Sentiment dataset, contains 44.639 entries on a minute-base with 19 features concerning sentiment, emotion and price values.

4 Experiments

We address the task of crypto value forecasting by jointly leveraging price and sentiment features of domain-related Reddit posts. Formally, a forecasting model receives a sequence of price values $\{v_1, v_2, \ldots, v_T\}$ regarding a time-window of size T and a sequence of sentiment features $\{s_1, s_2, \ldots, s_T\}$. Each sentiment feature s_t is a collection of sentiment values as described in Section 3The two sequences are concatenated temporally-wise and fed as input to a model. The forecasting model then outputs a price values v_{T+W} where W is the forecasting window size.

Our main objective is to study the impact of social media data on crypto price changes. Therefore, we define an ablation study by considering three experimental input configurations for a forecasting model: (**P**) only price values $\{v_1, v_2, \ldots, v_T\}$ are considered; (**S**) only sentiment features $\{s_1, s_2, \ldots, s_T\}$ are considered; (**P + S**) both sentiment and price information are considered.

We generally consider the forecasting task of predicting a time-window of future crypto price values given a past time-window of price, sentiment and emotion features. In this work, we set $T = 60$ minutes and $W = 15$ minutes.

To quantitatively evaluate the selected set of forecasting models, we split our dataset into train (25,025), validation (10,698) and test (8,916) set splits. Splits follow the sequential flow of crypto price time-series and, thus, no data shuffling is involved. We devise a preliminary hyper-parameter calibration phase based on the validation set. Models are

[6] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[8] https://www.bitstamp.net/

trained to minimize the mean squared error loss on the train set and are later evaluated on the test set. We apply a L^2 regularization to all the described models. Furthermore, we consider early stopping regularization with patience set to 10 epochs. Models are trained with Adam optimizer [Kingma and Ba, 2015].

4.1 Models

We select several widely adopted architectures for price forecasting to evaluate our dataset. In particular, we consider two recurrent neural network models which have been proved to achieve state-of-the-art performance for the stock price prediction task [Gao *et al.*, 2021]. Additionally, we consider a transformer model due to its widespread popularity [Vaswani *et al.*, 2017]. More precisely, the multi-head attention [Galassi *et al.*, 2021] of the transformer is used to capture high-level interactions between the heterogeneous set of features.

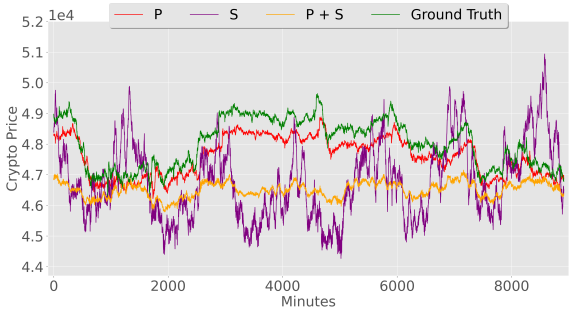
- **RNN**: A 2-layer recurrent neural network with a linear regression layer on top. The first recurrent layer is a simple RNN with 64 units followed by a GRU layer with 64 units.
- **LSTM-GRU**: A 2-layer recurrent neural network with a linear regression layer on top. The first recurrent layer is a LSTM with 64 units followed by a GRU layer with 64 units and L2 regularization.
- **Transformer**: A 6 heads attention model with 4 transformer-encoder blocks followed by a linear layer on top with 256 units.

4.2 Results

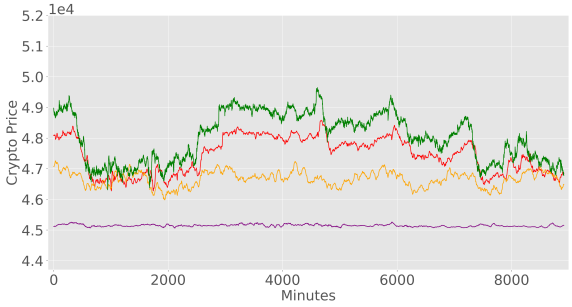
We evaluate the selected set of forecasting models on our dataset by computing the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) metrics. Table 4 reports the achieved performance on our dataset. Additionally, we report each model forecast on the test set in Figure 1. We distinguish between each input configuration for forecasting. We observe that RNN-GRU is the best performing model in terms of regression metrics with a RMSE of 468, and a MAE of 445,98. Additionally, the LSTM-GRU model achieves comparable performance while the transformer model falls behind. We notice that sentiment features have controversial effects on selected models. In particular, recurrent models performance is significantly deteriorated when considering the (P + S) input configuration compared to the (P) one (Figures 1a and 1b). In contrast, the transformer model achieves higher regression performance in the (P + S) setting while it fails to learn the trend of crypto price changes when relying on price values only (P) (Figure 1c). We speculate that this result could be motivated by the limited time period considered in our dataset.

As expected, the (S) setting leads to the worst performing results for all models. This is mainly motivated by the fact the social media information captures trend changes and provides general opinions on the cryptocurrency status rather than discussing exact crypto price forecasts.

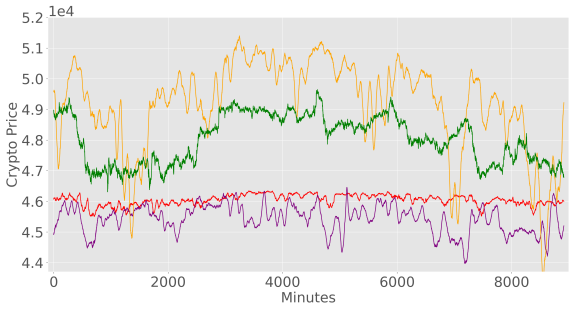
The proposed method for social media information integration has shown promising results in similar settings [Prajapati, 2020]. Nonetheless, our experimental results suggest that more sophisticated methodologies for social media information integration could be explored. In particular, we identify two major challenges. First, textual information encoding should scale to large sets of sources For instance, in our experimental setup, 3,705 comments and about 74 posts are reported hourly (Table 2). Properly encoding and aggregating such a large amount of textual data still remains an open research direction. We show that summing sentiment and emotion scores for aggregating Reddit posts and comments is not sufficient to achieve satisfying results.



(a) RNN.



(b) LSTM-GRU.



(c) Transformer.

Figure 1: Minute-based models forecasting performance with different input configurations.

5 Conclusions

We have introduced Bitcoin Reddit Sentiment, a novel dataset for crypto price values forecasting. Our dataset is built upon

Model	RMSE ↓	Δ RMSE	MAE ↓	Δ MAE
RNN (P)	**468,39**	-	**445,98**	-
RNN (S)	1961,30	-1492,91	1634,07	-1188,09
RNN (P + S)	1685,18	-1216,79	1518,82	-1072,84
LSTM-GRU (P)	623,70	-	589,75	-
LSTM-GRU (S)	2984,22	-2360,52	2883,50	-2293,75
LSTM-GRU (P + S)	1532,76	-909,06	1365,96	-776,21
Transformer (P)	2099,18	-	1992,60	-
Transformer (S)	2708,79	-609,61	2592,62	-600,02
Transformer (P + S)	1693,90	**405,28**	1515,48	**477,12**

Table 4: Model forecasting regression performance on our dataset. We report performance for each model input configuration. Additionally, we report the performance delta (Δ columns) between the (P) configuration and the remaining ones for each model.

an existing Kaggle dataset with Reddit posts and comments taken from a wide set of domain-related subReddits. We enrich such a dataset by leveraging several state-of-the-art sentiment and emotion extraction tools to encode textual data. This approach is motivated by the assumption that public platforms like social media can affect the current trend of cryptocurrencies. Thus, sentiment and emotion features stand as valuable indicators of the opinions reported in those platforms. The collected dataset is challenging due to the high number of textual data that has to be digested. Our experimental results show that well-known neural architectures like recurrent neural models and transformers can reach satisfying to modest forecasting performance when using price information only. In contrast, the high amount of encoded textual data deteriorates their successful integration by leveraging sentiment and emotion features and no significant benefit is shown regarding the task. Our results suggest that the integration of social media information is still an open research direction. We advocate for novel techniques that adapt easily to large and heterogeneous sources of information like Reddit, Twitter, Facebook and Google News. In future works, a critical point of investigation would be the evaluation of the contribution of each set of sentiment features. Another possible research direction would be to analyze the argumentative content of the social media [Lytos *et al.*, 2019] to obtain a score that can be used as feature in the forecasting [Lippi *et al.*, 2022].

Acknowledgments

We would like to thank Paolo Torroni for his help and supervision. This work was partially funded by EU Horizon 2020 project StairwAI, grant agreement number 101017142.

References

[Akbik *et al.*, 2018] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *COLING*, pages 1638–1649. Association for Computational Linguistics, 2018.

[Barbieri *et al.*, 2020] Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 1644–1650. Association for Computational Linguistics, 2020.

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[Cao, 2022] Longbing Cao. Ai in finance: Challenges, techniques, and opportunities. *ACM Comput. Surv.*, 55(3), feb 2022.

[Carmezim, 2018] Adriano Carmezim. Sentiment analysis on crypto tweets. https://github.com/Carmezim/crypto-twitter-sentiment-analysis, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Foley *et al.*, 2019] Sean Foley, Jonathan R Karlsen, and Tālis J Putniņš. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies*, 32(5):1798–1853, 2019.

[Galassi *et al.*, 2021] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308, 2021.

[Gao *et al.*, 2021] Ya Gao, Rong Wang, and Enmin Zhou. Stock prediction based on optimized LSTM and GRU models. *Sci. Program.*, 2021:4055281:1–4055281:8, 2021.

[Hutto and Gilbert, 2014] Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*. The AAAI Press, 2014.

[Kearney and Liu, 2014] Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Kraaijeveld and De Smedt, 2020] Olivier Kraaijeveld and Johannes De Smedt. The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, 65:101188, 2020.

[Lahmiri and Bekiros, 2019] Salim Lahmiri and Stelios Bekiros. Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, 118:35–40, 2019.

[Leukipp, 2022] Leukipp. Reddit - crypto posts (r/cryptocurrency, r/eth...). https://www.kaggle.com/datasets/leukipp/reddit-crypto-data, 2022.

[Lippi *et al.*, 2022] Marco Lippi, Francesco Antici, Gianfranco Brambilla, Evaristo Cisbani, Andrea Galassi, Daniele Giansanti, Fabio Magurano, Antonella Rosi, Federico Ruggeri, and Paolo Torroni. Amica: An argumentative search engine for covid-19 literature. In *IJCAI*, 2022.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Loginova *et al.*, 2021] Ekaterina Loginova, Wai Kit Tsang, Guus van Heijningen, Louis-Philippe Kerkhove, and Dries F Benoit. Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. *Machine Learning*, pages 1–24, 2021.

[Loria, 2018] Steven Loria. Textblob documentation, 2018.

[Lytos *et al.*, 2019] Anastasios Lytos, Thomas Lagkas, Panagiotis G. Sarigiannidis, and Kalina Bontcheva. The evolution of argumentation mining: From models to social media and emerging tools. *Inf. Process. Manag.*, 56(6), 2019.

[Pang *et al.*, 2020] Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, and Victor Chang. An innovative neural network approach for stock market prediction. *The Journal of Supercomputing*, 76(3):2098–2118, 2020.

[Pano and Kashef, 2020] Toni Pano and Rasha Kashef. A corpus of btc tweets in the era of covid-19. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–4. IEEE, 2020.

[Pant *et al.*, 2018] Dibakar Raj Pant, Prasanga Neupane, Anuj Poudel, Anup Kumar Pokhrel, and Bishnu Kumar Lama. Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In *ICCCS*, pages 128–132. IEEE, 2018.

[Phillips and Gorse, 2018a] Ross C. Phillips and Denise Gorse. Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PLOS ONE*, 13(4):1–21, 04 2018.

[Phillips and Gorse, 2018b] Ross C. Phillips and Denise Gorse. Mutual-excitation of cryptocurrency market returns and social media topics. In *ICFET*, page 80–86, New York, NY, USA, 2018. Association for Computing Machinery.

[Prajapati, 2020] Pratikkumar Prajapati. Predictive analysis of bitcoin price considering social sentiments. *CoRR*, abs/2001.10343, 2020.

[Reinerink, 2022] Nick Reinerink. Reddit /r/cryptocurrency. https://www.kaggle.com/datasets/nickreinerink/reddit-rcryptocurrency, 2022.

[SocialGrep, 2021] SocialGrep. Reddit cryptocurrency data for august 2021. https://www.kaggle.com/pavellexyr/Reddit-cryptocurrency-data-for-august-2021, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wooley *et al.*, 2019] Stephen Wooley, Andrew Edmonds, Arunkumar Bagavathi, and Siddharth Krishnan. Extracting cryptocurrency price movements from the reddit network sentiment. In *ICMLA*, pages 500–505. IEEE, 2019.

[Yenidoğan *et al.*, 2018] Işil Yenidoğan, Aykut Çayir, Ozan Kozan, Tuğçe Dağ, and Çiğdem Arslan. Bitcoin forecasting using arima and prophet. In *UBMK*, pages 621–624. IEEE, 2018.