Financial Innovation

RESEARCH Open Access

Bitcoin price change and trend prediction through twitter sentiment and data volume

Jacques Vella Critien[1*], Albert Gatt[2] and Joshua Ellul[1]

*Correspondence: jacques.vella-critien.18@um.edu.mt
[1] Centre for DLT, University of Malta, Msida MSD 2080, Malta
Full list of author information is available at the end of the article

Abstract

Twitter sentiment has been shown to be useful in predicting whether Bitcoin's price will increase or decrease. Yet the state-of-the-art is limited to predicting the price direction and not the magnitude of increase/decrease. In this paper, we seek to build on the state-of-the-art to not only predict the direction yet to also predict the magnitude of increase/decrease. We utilise not only sentiment extracted from tweets, but also the volume of tweets. We present results from experiments exploring the relation between sentiment and future price at different temporal granularities, with the goal of discovering the optimal time interval at which the sentiment expressed becomes a reliable indicator of price change. Two different neural network models are explored and evaluated, one based on recurrent nets and one based on convolutional networks. An additional model is presented to predict the magnitude of change, which is framed as a multi-class classification problem. It is shown that this model yields more reliable predictions when used alongside a price trend prediction model. The main research contribution from this paper is that we demonstrate that not only can price direction prediction be made but the magnitude in price change can be predicted with relative accuracy (63%).

Keywords: Bitcoin, Sentiment analysis, Prediction methods, Cryptocurrencies

Introduction

Bitcoin (Nakamoto 2009) has, since its introduction in 2008/9, attracted extensive attention for various reasons from different stakeholders with varying opinions regarding its utility and potential. Despite the fact that it and its underlying blockchain technology have been declared 'dead' on a number of occations,[1] recent all-time high prices[2] and research indicates otherwise (Ellul 2021). Given that such opinions may and do sway potential investors' interest, it is desirable to investigate whether better methods could be developed to support investors. In fact, not only could public opinion sway potential investors' interest, the same public sentiment could be used to empower investors to make better informed decisions regarding future price predictions.

There is no doubt that sentiment affects an asset's price—and as put by Baker and Wurgler: "the question is no longer, as it was a few decades ago, whether investor sentiment

[1] https://markets.businessinsider.com/currencies/news1030156713.

[2] This article was written in April 2021.

affects stock prices, but rather how to measure investor sentiment and quantify its effects" (Baker and Wurgler 2007). Whilst this statement is referring to a well-founded body of literature on applying sentiment analysis to traditional markets (Gunter et al. 2014; Rao and Srivastava 2012; Li et al. 2014; Mittal and Goel 2012), sentiment analysis can similarly be used for cryptocurrency price prediction as demonstrated extensively in recent work (Valencia et al. 2019; Kraaijeveld and De Smedt 2020; Abraham et al. 2018; Stenqvist and Lönnö 2017; Pant 2018; Galeshchuk et al. 2018; Kilimci 2020; Naeem et al. 2020; Serafini et al. 2020; Wołk 2019; Balfagih and Keselj 2019; Mohapatra et al. 2020).

Twitter[3] is widely-used as a source of sentiment data, and has become a popular social media platform amongst crypto-communities (Kraaijeveld and De Smedt 2020; Abraham et al. 2018; Stenqvist and Lönnö 2017; Pant 2018; Galeshchuk et al. 2018; Kilimci 2020; Naeem et al. 2020; Balfagih and Keselj 2019; Mohapatra et al. 2020). Whilst the current state-of-the-art has achieved encouraging results, yet further research effort is required to overcome a number of issues. Many of these following issues are frequently encountered in price prediction models based on sentiment analysis of Twitter data: (i) evaluation is typically based on minimal historical data (Pant 2018; Valencia et al. 2019; Stenqvist and Lönnö 2017; Kilimci 2020); (ii) predictions tend to lag behind when the predicted prices are actually seen on the market (Serafini et al. 2020; iii) models are frequently limited to the prediction of the direction (up or down) (Kilimci 2020; Valencia et al. 2019; Galeshchuk et al. 2018), though some studies have proposed to make predictions of exact prices with limited success (Pant 2018; Li and Dai 2020; Serafini et al. 2020).

Some of the issues described above may be due to the following particular challenges. The direction of price change and magnitude is often non-linear, and therefore not straightforward to solve (Kimoto et al. 1990). Furthermore, tweets are often duplicated for marketing purposes and may also be automated by tweet bots (Valencia et al. 2019). Tweets also typically contain features that result in noise (when it comes to sentiment analysis) including hashtags, profile mentions and URLs (Kraaijeveld and De Smedt 2020). At the same time, the use of sarcasm in tweets could skew sentiment predictions (Rosenthal et al. 2014). Therefore, prior to tackling the general problem of extracting sentiment, pre-processing of tweets should be undertaken to reduce such noise.

In this paper, we investigate predicting the magnitude of price change (beyond just the direction) and to the best of our knowledge, this is the first paper proposed to do so. We present exhaustive evaluation and conclusive results for a number of models. Furthermore, the models proposed overcome the late prediction problem see in the state-of-the-art.

Furthermore, we investigate the predictive relationship between Twitter sentiment and associated price changes as a function of different time lags. Through this we thereby address the question of which temporal interval between expressed sentiment and price change provides the best results.

An in-depth study was undertaken to determine how different types of neural networks and features used may affect accuracy, in which each model investigated was

[3] https://twitter.com.

evaluated against different combinations of features used as well as against different time lags introduced between sentiment and price change.

The rest of this paper is organised as follows. The following section gives an overview of sentinent analysis, followed by recent work on Bitcoin and other cryptocurrency price prediction. Then the classification problems addressed, and the methods used for data preprocessing, feature extraction, and the neural models we propose are presented. After which results are presented and discussed. Finally, then conclude with some directions for future work.

Background on sentiment analysis

In its raw form, natural language text is meaningless to a computer—nothing more than encoded bytes. Over the past decade strides have been made within the field of Natural Language Processing (NLP) with the aim of enabling computational systems to reason better about natural language. Sentiment analysis, as its name implies, analyses and extracts sentiment, opinion, subjectivity and polarity from text. Use-cases for sentiment analysis are plenty—including but not limited to product market analysis and automated flagging of positive/negative/potentially-harmful comments on websites and social media platforms.

Since the introduction of Sentiment Analysis to the NLP community (Pang and Lee 2008), several techniques have been proposed to associate polarity with a piece of text. It can be framed as a classification problem, whereby a text segment is classified as positive, negative or neutral. Some approaches also assign a value reflecting the degree of confidence associated with the respective polarity. Lexicon-based approaches make use of a lexicon (a collection of words) and associated sentiment scores to compare with the text being classified to determine a final polarity score. One widely-used lexicon-based implementation, VADER (Valence Aware Dictionary and Sentiment Reasoner) (Hutto and Gilbert 2015), further makes use of rule-matching, which attempts to identify polarity based on the input text using linguistic patterns. VADER combines selected features from three validated lexicons[4] along with tweet intensity rules that were extracted from analysis of terms of syntax, grammar and valence values of 800 tweets (Stenqvist and Lönnö 2017). Sentiment analysis has over the past decade been extensively applied to Twitter data which indeed 'poses newer and different challenges' (Agarwal et al. 2011) beyond those associated with more traditional sentiment analysis applications (Hussein 2018) and the wide range of approaches (Medhat et al. 2014). Furthermore, the SemEval international workshop[5] helped facilitate further research by making available a set of shared challenges for the community. Particularly related to the context of this work, since the 2013 workshop (Nakov et al. 2013) a shared task focused on Twitter sentiment analysis has been published every year.

[4] LIWC (Linguistic Inquiry and Word Count), ANEW (Affective Norms for English Words) and GI (General Inquirer).

[5] ff

Table 1 Existing studies and models

Study	Date range	Data days	Model-accuracy %	Data used
Predicting the price of Bitcoin using machine learning-2018 McNally et al.	19/08/2013– 19/07/2016	~1065	LSTM-52.78% RNN-50.25%	Market
An advanced CNN-LSTM model for cryptocurrency forecasting-2021 Livieris et al.	01/01/2017–31/10/2020	~1400	(CNN-LSTM) Model1-55.03% Model2-53.64% MICDL-53.04%	Market
Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model-2019 Yan Li, Wei	30/12/2016–31/08/2018	~600	(Precision) BP-59% CNN-64% LSTM-58% CNN-LSTM-64%	Market
Bitcoin response to twitter sentiments Galenchuk et al.	01/2014–09/2017	~912	RW-46.2% ARIMA-47.2% MLP-47.5% CNN-68.6%	Twitter market
Price movement prediction of cryptocurrencies using sentiment analysis and machine learning- 2019 Valencia et al.	16/02/2018– 21/04/2018	~60	MLP-72% SVM-55% RF - 44%	Twitter market
Recurrent neural network based bitcoin price prediction by twitter sentiment analysis-2018 Pant et al.	01/01/2018–30/06/2018	~180	RNN–77.62%	Twitter market
Predicting bitcoin price fluctuation with Twitter sentiment analysis-2017 Stenqvist, Lönnö	11/05/2017–11/06/2017	~30	No machine learning. Predicting direction solely on sentiment change in tweets 1hour_shift3-83.33% 30mins_shift4-78.78% 45mins_shift3-70.59%	Twitter Market
Sentiment analysis based direction prediction in Bitcoin using deep learning algorithms and word embedding models-2020 Kilimci	01/05/2019–01/08/2019	~90	GloVe-82.01% RNN-83.77% CNN-84.3% LSTM-87.45% FastText-89.13%	Twitter market

Related work on cryptocurrency price prediction

We now provide an overview of approaches used in specifically the domain of cyrptocurrency price prediction. Further detail and a comparison with the state-of-the-art will be woven into the methodology and evaluation sections, in order to provide a direct comparison of relevant aspects.

Attempts to predict Bitcoin price using Long Short Term Memory Cells (LSTM), Convolutional Neural Network (CNN) and hybrid CNN-LSTM models without undertaking sentiment analysis were originally proposed (Li and Dai 2020; Livieris et al. 2021; Kwon et al. 2019). Initial work in investigating the use of sentiment analysis of Twitter data for price prediction was proposed by Pant (2018); Galeshchuk et al. (2018) and Serafini et al. (2020). To eliminate the vanishing gradient problem seen in Recurrent Neural Networks (RNNs), Pant (2018) proposed to make use of an RNN predictor with LSTM and Gated Recurrent Unit (GRU) variations. This work resulted in a moderate correlation between rising negative sentiment and consequent falling of prices. A comparison of LSTM and ARIMA model-based approaches were conducted in Serafini et al. (2020), in which it is stated that the ARIMA model performs better. Further comparisons with other models

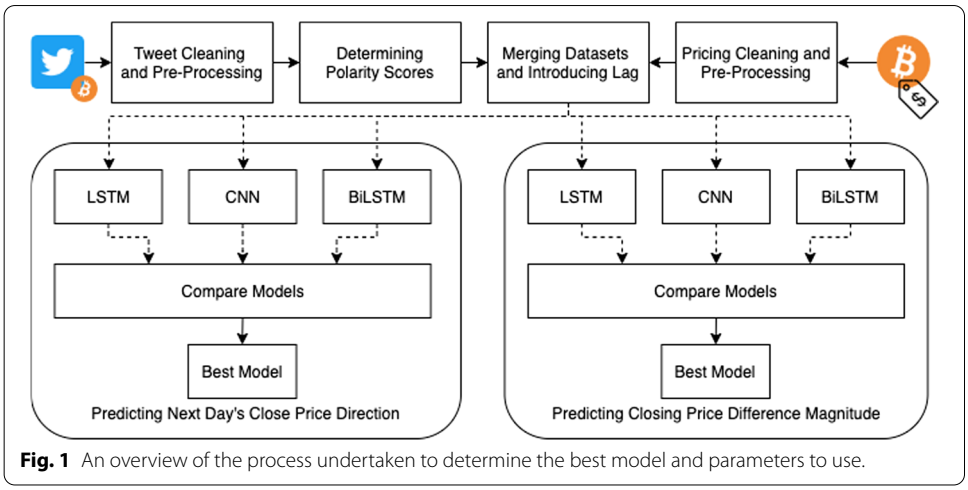

Fig. 1 An overview of the process undertaken to determine the best model and parameters to use.

and on different cryptocurrencies were further presented in Valencia et al. (2019) and Wołk (2019). According to Valencia et al. (2019), Twitter data is not sufficient to predict Bitcoin price on its own, but can help when combined with other market data.

An overview of the related studies is presented in Table 1.

Methodology

Figure 1 provides an overview of the process followed to determine the best model for predicting: (i) the next day's close price direction (i.e. whether it will increase/decrease); and (ii) the magnitude of difference in closing prices. Two main datasets are used in this study: (i) Bitcoin price data; and (ii) Twitter tweets. Historical Bitcoin price data providing a per-minute record of timestamps, opening and closing prices, high and low prices and volume of Bitcoin traded for the period between 1st January 2012 and 31st December 2020 was retrieved from Kaggle.[6] A Twitter dataset[7] (also from Kaggle) was filtered to retrieve tweets that contained either 'bitcoin' or 'btc'. The period of tweets provided in the dataset is between 1st January 2016 and 29th March 2019—including a total of over 20 million tweets. In addition to the text of each tweet, the dataset provides timestamps, tweet IDs and URLs, associated authors' usernames and full names, and the number of replies, likes and retweets that tweets received.

Data cleaning and pre-processing

The following cleaning and pre-processing tasks depicted in Fig. 2 were undertaken on the Twitter dataset:

- Removal of non-English tweets[8] and duplicate tweets made by the same user in a similar manner to Pant (2018); Valencia et al. (2019) and Ranjan et al. (2018);
- Removal of URLs from tweets, as performed in Kraaijeveld and De Smedt (2020) and Ranjan et al. (2018);

[6] https://www.kaggle.com/mczielinski/bitcoin-historical-data.

[7] https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329.

[8] Language used for tweets was detected using https://pypi.org/project/langdetect/.

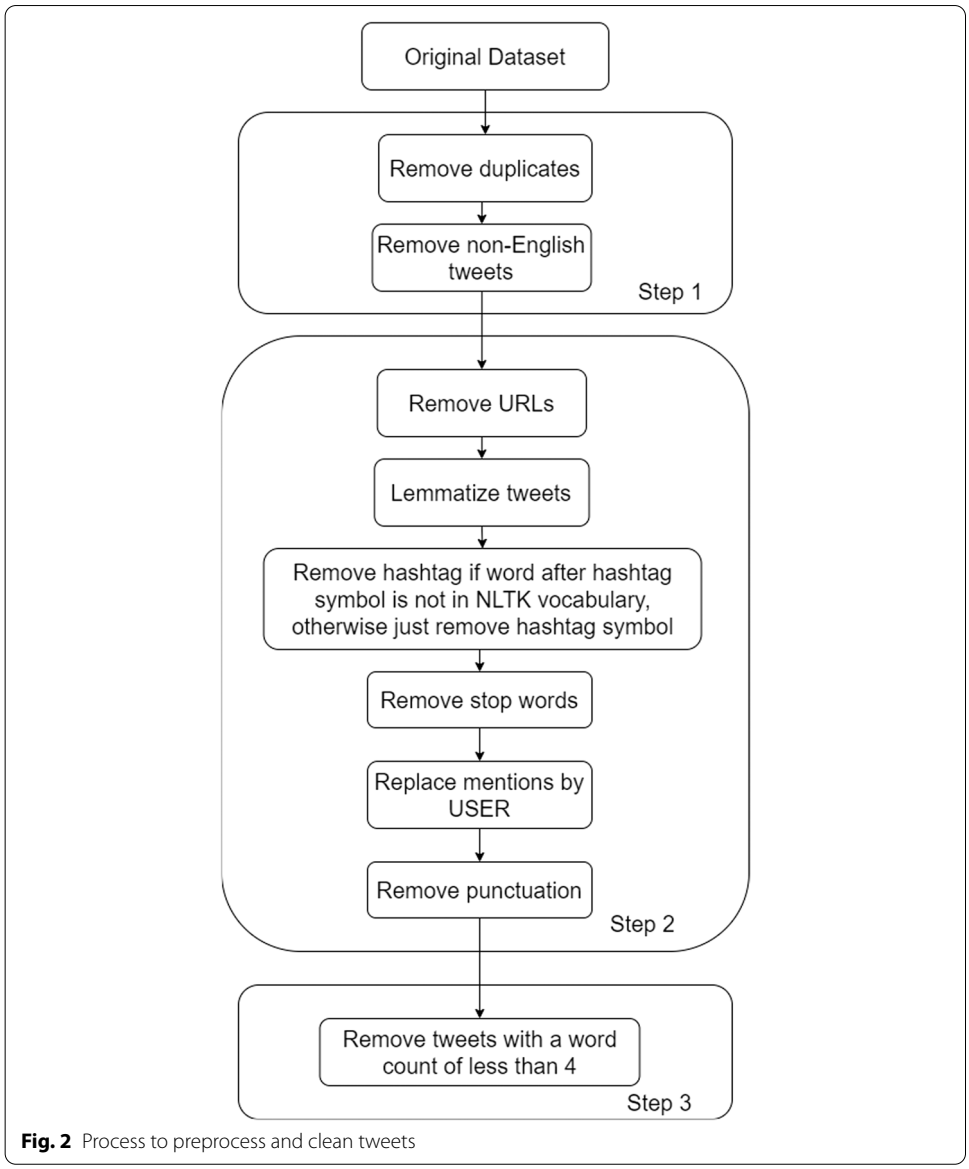

Fig. 2 Process to preprocess and clean tweets

- Tokenization and lemmatization in a similar manner to Pagolu et al. (2016) (that is, mapping each token to its morphological base form, so as to enable reasoning about words in a form-agnostic manner);[9]
- Removal of stop words (e.g. 'a', 'the', etc.), similarly to what was done by Pagolu et al. (2016);
- Replacement of user mentions (akin to tagging, which takes the form of '@' followed by the username) with the text 'USER', again following Pagolu et al. (2016);
- Removal of all punctuation (Abraham et al. 2018);

[9] Tokenization was carried out using the `TweetTokenizer` model designed for tweets and incorporated in the Natural Language Toolkit (NLTK) (Bird et al. 2009). Lemmatization was carried out using the `WordNetLemmatizer` in NLTK. See https://www.nltk.org/api/nltk.tokenize.html.

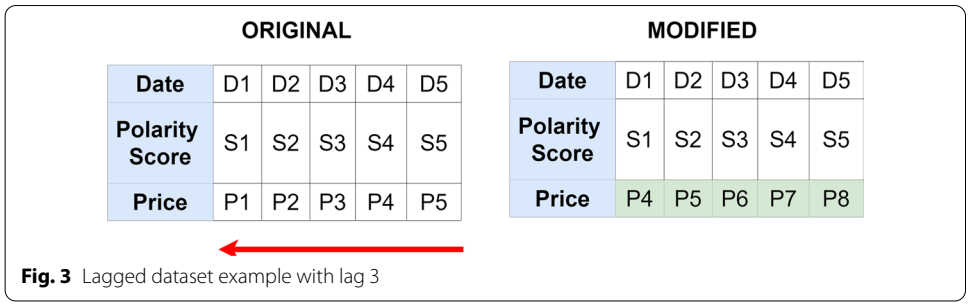

Fig. 3 Lagged dataset example with lag 3

- Processing of hashtags: if the word following the hash sign was not found in a precompiled English wordlist[10], it was removed; otherwise, the '#' sign was dropped and the word retained as was done by Kraaijeveld and De Smedt (2020);
- Removal of tweets containing fewer than 4 words, similarly to Kraaijeveld and De Smedt (2020);
- The normalised datasets are available from the repository we have made available to download from https://github.com/jacquesvcritien/fyp.

In relation to the Bitcoin pricing dataset, the high and low prices were removed from the feature list so as to only keep the average price per minute[11]. After the the cleaning and pre-processing steps, this study ended up with tweets and prices ranging between 30th August 2018 and 23rd November 2019.

A desirable property of the resulting dataset is that Bitcoin prices within the specified date range evinced both downward ($6500 to $3300 = -49%) and upward trends ($3300 to $11500 = +248%).

Determining polarity scores

Following preprocessing, VADER (Hutto and Gilbert 2015) is used to assign sentiment scores to tweets. Similar approaches are used by Valencia et al. (2019); Abraham et al. (2018) and Kraaijeveld and De Smedt (2020)[12]. VADER scores each tweet with a negative, positive, neutral and compound polarity score. The compound score is a sum of the individual sentiment scores, adjusted according to a set of rules and normalised to fall within the $[-1, +1]$ range. However, for the purposes of this study, only positive and negative polarity scores are included in the training and evaluation data sets. VADER was widely used in related work (Valencia et al. 2019; Abraham et al. 2018; Kraaijeveld and De Smedt 2020; Mohapatra et al. 2020; Serafini et al. 2020) and provides advantages including the following: it is open source and free; it is human validated and tuned for Twitter content (Valencia et al. 2019); and it has also been shown to perform competitively with human annotators and has outperformed several benchmarks, especially on social media content (Hutto and Gilbert 2015).

[10] We use the Wordlist provided in the `nltk.corpus.words` library in Python.

[11] and the provided UNIX timestamps were changed to UTC format (so as to match the Twitter dataset).

[12] https://github.com/cjhutto/vaderSentiment.

Merging datasets and introducing lag

One of the research questions which this work aims to address is the optimal lag to consider that would enable the discovery of a relationship between Bitcoin-related tweets (and in particular the sentiment they express) and actual price change. Indeed, it is not certain that such tweets are the cause of the change in price. However in this work we investigate whether a potential correlation can be seen, and if so what the optimal time lag is (between tweets and the price being affected). This approach was similarly followed by (Stenqvist and Lönnö 2017) and (Balfagih and Keselj 2019), who explored lags ranging from minutes to hours. In contrast to these approaches, in this paper we investigate lag intervals of a number of days—to be exact, 1, 3 or 7 days. To illustrate, Fig. 3 depicts the effect of introducing a lag of 3 days on a dataset—where the original dataset is on the left, and the dataset with a lag introduced is on the right. From the lagged dataset (on the right) note how, as an example, Day 1's score is associated with Day 4's price—i.e. tweets from day 1 are being assumed to affect prices 3 days later (in this example). The reason for choosing to investigate lags of 1, 3 and 7 days is that since this study focuses on making a daily prediction, the minimum lag to be observed should be of at least 1 day. Thereafter it was decided to observe a granularity of a week. The choice of a 3-day lag represents an interval between these two extremes.

The three different lagged datasets (for 1-, 3- and 7-day lags) were created by shifting the price data (of the cleaned and merged dataset) back by the respective number of days in the lag being tested.

Grouping lagged datasets

Lagged datasets consist of preprocessed tweets coupled with their Bitcoin price at the minute the tweet was posted. Subsequently, these are grouped by day in order to allow a model to make daily predictions. Grouping is done in the following manner:

- Timestamps of tweets are floored to the hour or day when the tweet was posted;
- Tweets are grouped by their floored timestamp;
- For a given group, the polarity scores are averaged;
- The tweet *volume* is added as an additional feature, where the volume is the number of tweets in a given day;
- The closing Bitcoin price for the day is then identified as the price for the last record for the given day.

Features and labels

The classifiers described below are trained to predict a fluctuation in price based on the following features:

1. **Change**: Bitcoin price change direction of that day (binary, indicating whether the price rises or falls);
2. **Close**: Bitcoin's closing price for that day;
3. **Positive polarity**: The positive polarity score obtained from VADER;

Table 2 Hyperparameters for the best models for predicting price direction DBMLA - Difference between Maximum and Lowest Accuracies

	LSTM	CNN	BiLSTM
# Layers	1	3	2
Layer Size	32	32	64
Batch Size	80	50	80
Dataset	1 day lag	1 day lag	1 day lag
Lagged Features	7	7	7
Train-Test Split	85:15	85:15	85:15
Loss Function	Categorical Crossentropy	Categorical Crossentropy	Categorical Crossentropy
Early Stopping Parameter	Validation Loss	Validation Loss	Validation Loss
Early Stopping Patience	20	20	20
Maximum Accuracy	67.16%	64.18%	64.18%
Mean Accuracy	59.10%	58.51%	60.90%
DBMLA	14.93%	11.94%	7.46%

ORIGINAL

Date	f1	f2	f3
D1	0.1	0.2	0.3
D2	0.4	0.5	0.6
D3	0.7	0.8	0.9
D4	0.1	0.11	0.12

MODIFIED

Date	f1(t-2)	f2(t-2)	f3(t-2)	f1(t-1)	f2(t-1)	f3(t-1)
D3	0.1	0.2	0.3	0.4	0.5	0.6
D4	0.4	0.5	0.6	0.7	0.8	0.9

Fig. 4 Dataset with lagging features of 2

4 **Negative polarity**: The negative polarity score obtained from VADER;
5 **Tweet Volume**: The volume of tweets in the relevant interval. This was also investigated in Abraham et al. (2018) which demonstrated that prince changes were highly correlated with tweet volume.

Note that lagged datasets also include the above features for the previous days. For example, as shown in Fig. 4, if the lag is 2, a training instance would include data from the last 2 days. Finally, the label of that instance would be the price change direction of the day following the last lagged feature.

Data split and resampling

In order to train and test the data, the dataset was split using a train-test ratio of 85:15. The reason for this is because of the small number of records available for training and testing after grouping and averaging the original datasets per day. Therefore, such a split allows the model to have a good percentage of the available data to train on while also having a fair number of records to test.

When testing different models and parameters, each set of parameters is tested on 3 different shuffled datasets. However, when shuffling the dataset, the same seed is set to ensure that each model and set of parameters is tested on the same three sets of datasets, allowing for a fair comparison. The training and test sets are prepared

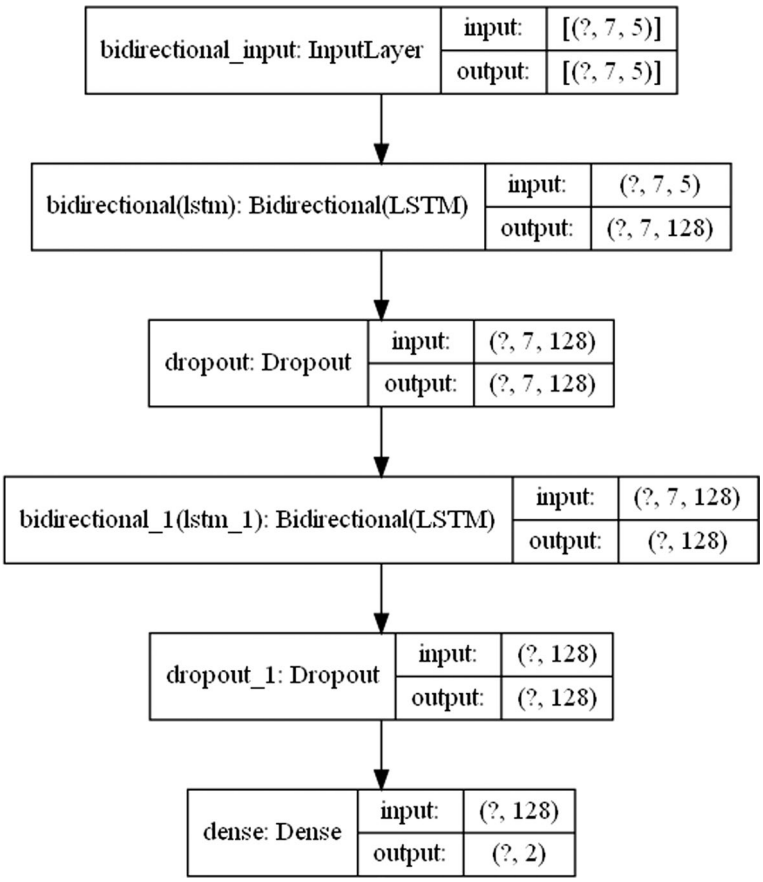

Fig. 5 Chosen BiLSTM model

by first shuffling the original datasets and then using the first 85% as the training set and the last 15% as the test set.

Predicting next day's close price direction

The direction of the closing price can be framed as a binary classification problem where, given the input corresponding to features extracted from tweets, the task is to predict whether the price will go up or down.

Three different models, (i) using an LSTM, (ii) CNN and (iii) Bidirectional Long Short Term Memory Cells (BiLSTM), were implemented for predicting whether the following day's closing price will increase or decrease. These are hereafter referred to as Direction-LSTM, Direction-CNN and Direction-BiLSTM. Table 2 outlines the hyperparameters used for each model. The table also gives the accuracy statistics for each model. These are further discussed in the evaluation section below. It is however evident that the best performing model, in terms of mean accuracy, is Direction-BiLSTM. The architecture of this model is depicted in Fig. 5.

Table 3 Bin ranges for each class

Class	Range
1	Less than $1320
2	−$1320 to −$989
3	−$990 to −$659
4	−$660 to −$329
5	−$330 to −$1
6	$0 to $330
7	$330 to $660
8	$660 to $990
9	$990 to $1320
10	Greater than $1320

Daily price change magnitude prediction

Another prediction model tries to predict the magnitude of the change of closing day prices as a multi-class classification problem. This is done by predicting which interval the closing day price changes would fall into.

Closing day prices were categorised into ten different bins/classes. An average of the maximum positive ($1563) and maximum negative price ($1746) changes was calculated (and rounded to $1650) to define the lower and upper bins/classes (that were extended to included any greater price change), and then equal steps (of $330) calculated for each bin in between, as can be seen in Table 3.

As before, to predict the magnitude of change in price on the following day, three models were implemented using an LSTM, CNN and BiLSTM, which we'll refer to as Magnitude-LSTM, Magnitude-CNN and Magnitude-BiLSTM for the remainder of this paper. Table 4 summarises the hyperparameters and training settings used for these models, together with the evaluation results (see the evaluation section for this discussion). The Magnitude-CNN model outperforms the other two for this task, as is evident from the mean accuracy and F1 scores. Figure 6 depicts the architecture of this model.

Voting classifier

The best performing models from each of the aforementioned predictive tasks, more specifically, the Direction-BiLSTM and Magnitude-CNN models, were merged together to create a voting classifier model which takes into consideration the outputs from the two models. As Fig. 7 shows, the voting classifier works by first predicting the next day's closing price direction and then, the magnitude of the next day's closing price using the second model. Then, it checks whether the next day's closing price direction matches the direction of the predicted change magnitude. In other words, a match happens: (i) if the first model outputs a 0, which means a decrease in price, and the second model outputs a class from 1 to 5 (negative magnitude of price change); or (ii) if the first model outputs a 1, which means an increase in price, and the second model outputs a class from 6 to 10 (positive magnitude of price change). The prediction of the next day's closing price direction is kept if there is a match in the output of the two classifiers. Moreover, the voting classifier is evaluated on 50 different runs with 50 differently shuffled datasets.

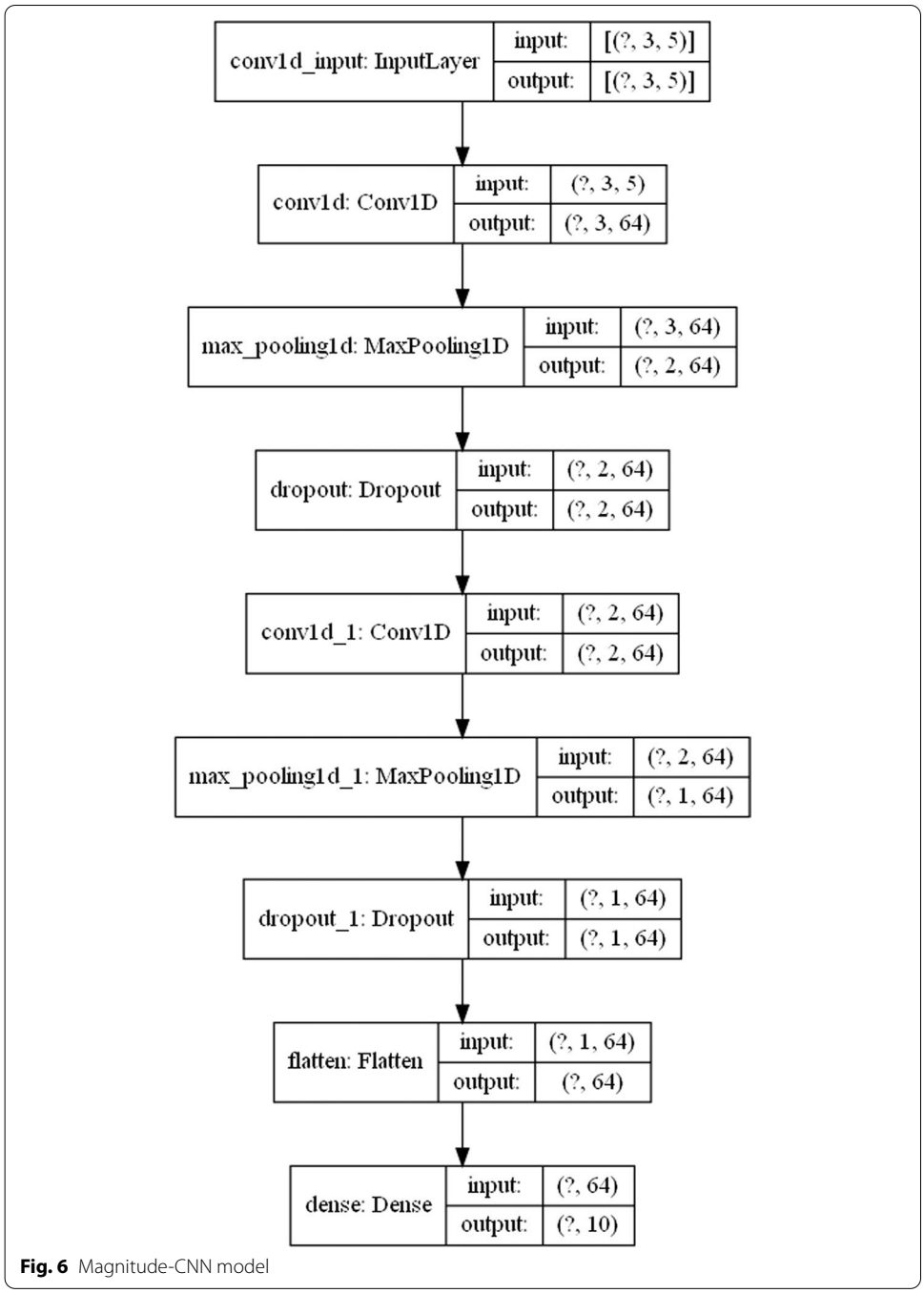

Fig. 6 Magnitude-CNN model

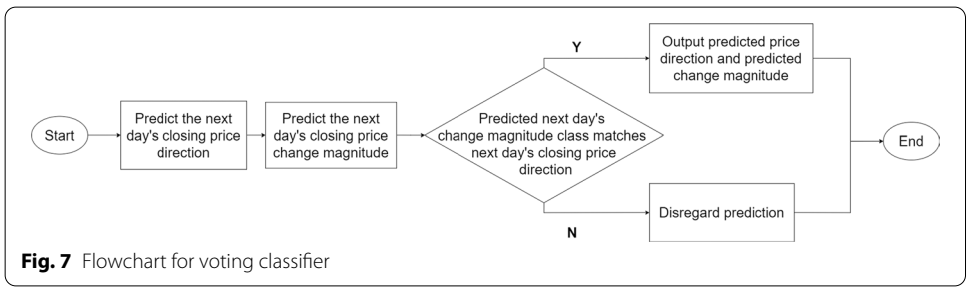

Fig. 7 Flowchart for voting classifier

Table 4 Hyperparameters for the best models for predicting magnitude of change Acc. F1 Score - The sum of the individual F1 Scores of each label/class

	LSTM	CNN	BiLSTM
# Layers	3	2	2
Layer Size	256	128	256
Batch Size	50	80	80
Dataset	1 day lag	3 day lag	1 day lag
Lagged Features	7	3	7
Train-Test Split	85:15	85:15	85:15
Loss Function	Categorical Crossentropy	Categorical Crossentropy	Categorical Crossentropy
Early Stopping Parameter	Validation Loss	Validation Loss	Validation Loss
Early Stopping Patience	20	20	20
Maximum Accuracy	58.21%	57.35%	59.09%
Mean Accuracy	46.76%	51.47%	46.67%
Acc. F1 Score	12.33%	14.21%	12.88%

Evaluation

Tweet analysis

The tweets in the dataset are analysed to see how many of the tweets are positive, negative or neutral according to the sentiment scores assigned by VADER. As can be seen, there are more positive and neutral tweets than negative ones. Moreover, the same query was made on the daily grouped data. This resulted in all days being either positive or neutral on average with no days being negative at all. When analysing this metric and the graphs in Fig. 11, one can come to the conclusion that the dataset is imbalanced. However, this dataset contains all tweets with '#btc' and '#bitcoin' and therefore, the dataset reflects the reality of tweets. In addition, it could very well be the case that people tend to tweet more positively than negatively as seen in Pantano et al. (2018), Zhou et al. (2013), Abraham et al. (2018) and Kraaijeveld and De Smedt (2020).

Daily price trend prediction

For the three price direction prediction models (Direction-LSTM, Direction-CNN and Direction-BiLSTM) at different time-lags of 1, 3 and 7 days, Figs. 8, 9 and 10 display the mean accuracy, the maximum accuracy and the range—the difference between maximum and minimum accuracies, or Difference between Maximum and Lowest Accuracies (DBMLA) respectively. See Table 2 for descriptive statistics corresponding to these plots (Fig. 11).

The mean and maximum are computed over 5 runs for each model, with data randomly shuffled for each run. These results give rise to the following observations.

First, mean accuracy is highest for a single day lag, with the 7-day lag in second place. Despite the highest mean accuracy being obtained in the shortest time lag, we cannot definitively conclude that the price fluctuation is strongest over shorter time periods since we also observe that a 7-day lag achieves a higher mean accuracy than a 3-day lag. At the same time, the relationship is not linear. Whilst a 3-day lag results in the lowest accuracies on average, when considering the maximum accuracy obtained by a model,

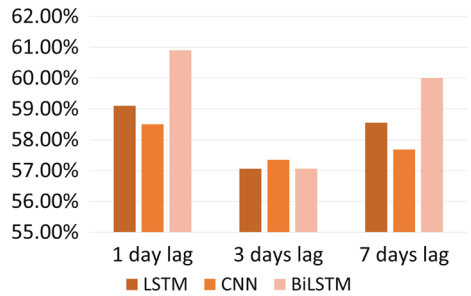

Fig. 8 Mean accuracy

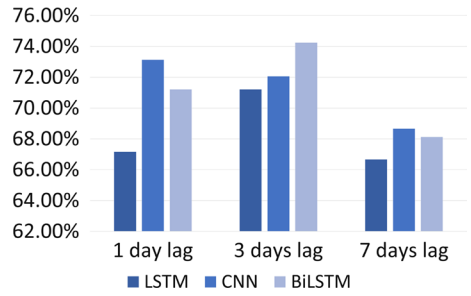

Fig. 9 Maximum accuracy

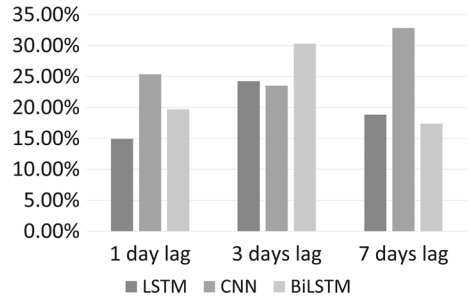

Fig. 10 Difference between maximum and lowest accuracy (DBMLA)

higher maximum scores are observed for a 3-day lag (with the exception of the CNN model). This implies that the results suggest that mean accuracies are subject to considerable variation, leading to an effect whereby short time lags of 1 day may benefit from an 'immediacy effect' (fluctuations which are closer in time can be better predicted on average). While longer time lags may result in lower variance in the data overall (so that averages over 7-day lags are better than those over 3 days).

Figure 12 shows a direct comparison of the best CNN, BiLSTM and LSTM models. The BiLSTM architecture (Fig. 5) achieves overall higher mean accuracies coupled with lower variance, as evidenced by its lower DBMLA. Hence, it is considered the best

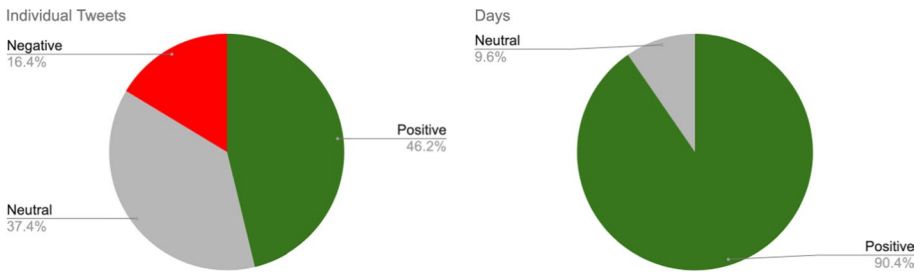

Fig. 11 Tweet sentiment polarity analysis

model overall. It is worth noting that this model achieves a maximum accuracy that outperforms those reported in Galeshchuk et al. (2018), Li and Dai (2020) and Livieris et al. (2021). The main reason for why LSTMs seem well-suited for this specific problem is that they are inherently sequential models and thus, expected to do reasonably well at predicting a trend over time. Furthermore, we can also see that if we introduce bidirectionality and allow the model to look both forwards and backwards around a given time, we can also achieve better results. The daily price trend prediction algorithm in this study results in a 64.2% maximum accuracy which is less than that of the models proposed by Pant (2018) and Valencia et al. (2019) which resulted in 77.6% and 72% respectively. This is likely due to the data periods used. Our study spans across around 450 days, whilst their studies were based on around 180 and 60 days respectively—and looking more closely at the the data for the given periods[13] it is clear that the periods used in their studies were times of rather low volatility (by looking at the standard deviation of daily returns[14]). Whilst in our study (which makes use of a substantially larger window) volatility is seen to fluctuate much more over the whole period.

Daily price change magnitude prediction

Figure 13 displays F1 scores for the three types of models (Magnitude-LSTM, Magnitude-CNN and Magnitude-BiLSTM) for price change magnitude prediction. The corresponding descriptive statistics can be found in Table 4.

Once again, performance is generally worse with a 7-day lag in nearly all cases, whereas the shorter time lag of 1 day results in the best F1 scores. However, the CNN model outperforms the other models on the 3-day lag dataset. This is confirmed by the per-class F1 scores in Fig. 14, where the CNN model outperforms both the LSTM and BiLSTM models in nearly all classes apart from class 4, in which the difference is in any case small.

While sequence information was found to be useful when predicting the direction of the change, here we can see that sequence information is less important when predicting the actual size of the change. Based on these results, the CNN model can be identified as the best model to predict the magnitude in price change. However, it is worth noting

[13] Our study includes data spanning across 30/08/2018 to 23/11/2019; whilst Pant et al's across 01/01/2018 to 30/06/2018; and Valenica's across 16/02/2018 to 21/04/2018

[14] E.g. one can use https://www.buybitcoinworldwide.com/volatility-index/.

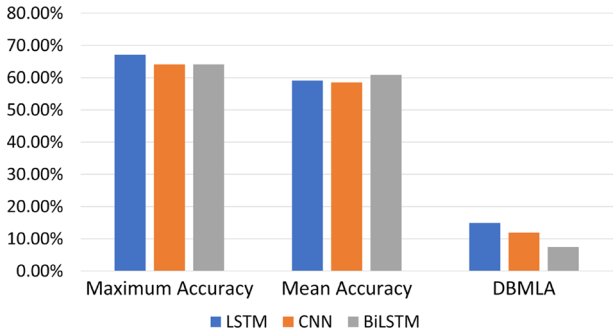

Fig. 12 Prediction accuracy: Comparison of LSTM, CNN and BiLSTM models

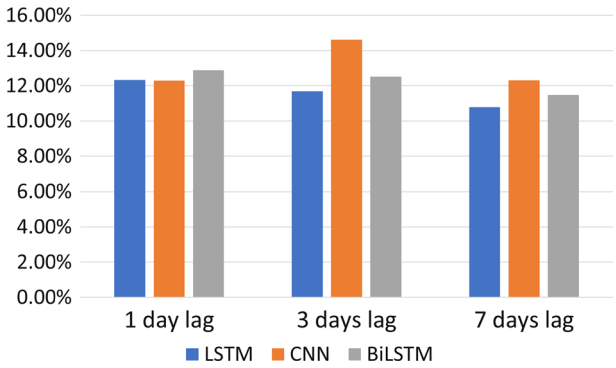

Fig. 13 F1 scores of different models on price change magnitude prediction across time lags

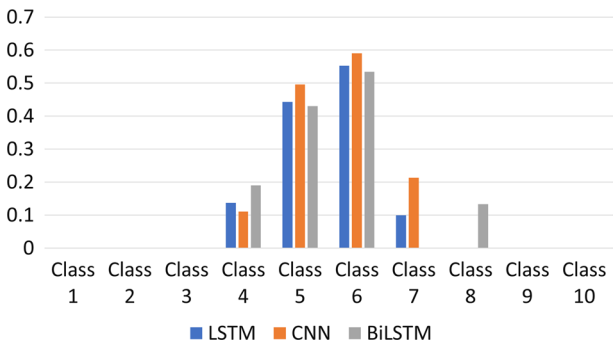

Fig. 14 F1 scores per class for the chosen Magnitude-LSTM, Magnitude-CNN and Magnitude-BiLSTM models

Table 5 Accuracies of Direction-BiLSTM, Magnitude-CNN and Voting Classifier

Model	Maximum direction accuracy (%)	Mean direction accuracy (%)
Direction-BiLSTM	64.2	60.9
Magnitude-CNN	63.3	58.3
Voting Classifier	77.2	68.4

Table 6 Comparison with other studies

Study	Maximum accuracy obtained	Days in dataset
Present (Direction-BiLSTM)	64.2%	450~
Present (Voting Classifier)	77.2%	450~
Predicting the Price of Bitcoin Using Machine Learning-2018 McNally et al.	52.78%	1065~
An Advanced CNN-LSTM Model for Cryptocurrency Forecasting-2021 Livieris et al.	55.03%	1400~
Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model-2019 Yan Li, Wei Dai	64% (Precision)	600~
Bitcoin Response to Twitter Sentiments Galenchuk et al.	68.6%	912~
Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning-2019 Valencia et al.	72%	60~
Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis-2018 Pant et al.	77.62%	180~
Predicting Bitcoin price fluctuation with Twitter sentiment analysis-2017 Stenqvist, Lönnö	83%	30~
Sentiment Analysis Based Direction Prediction in Bitcoin using Deep Learning Algorithms and Word Embedding Models-2020 Kilimci	89.13%	90~

that F1 scores could not be reliably computed for all classes. This is likely due to data sparseness, with few instances of a given class in the test set.

Voting classifier

Table 5 shows that when evaluating the voting classifier, the mean accuracy was increased by around 8%, achieving a 68.4% accuracy while the maximum accuracy achieved increased by 13%, achieving a maximum accuracy of 77.2%. This significant increase in accuracy might suggest that implementing a modular approach which first identifies the direction and then predicts the actual bin for the price change, achieves higher accuracy levels.

Conclusions

This paper compared the performance of a number of different neural models for predicting fluctuations in cryptocurrency prices from Twitter tweet data. The underlying hypothesis of this work is that opinions expressed in social media can function as useful predictors of such fluctuations, especially insofar as they incorporate features such as sentiment and opinion.

One important question is whether the predictive value of features gleaned from social media depends on the time lag between their publication and the time of prediction. The experiments presented in this paper show that competitive results can be achieved with a 2-layer BiLSTM model trained on a dataset with a 1-day time lag and using seven different lagged features, meaning that each instance consists of features from tweets from the seven previous days. This model achieves a maximum accuracy of 64.18%. It must be highlighted, that whilst this configuration heeded the best results, this does not necessarily imply that a 1-day lag always results in better predictions. In fact, from the results presented herein, other temporal lags perform better

under other configurations. Therefore, future work should be undertaken to further investigate the impact of temporal lags in more detail. Furthermore, whilst the BiLSTM overall outperformed the CNN and LSTM implementations for price direction prediction, the CNN outperformed the others for the change in magnitude prediction. Future work should be undertaken to further analyse with varying datasets to determine whether this is due to particular features in the data and/or why the different algorithms perform better/worse for varying parameters.

As can be seen in Table 6, the Direction-BiLSTM model presented here, which predicts the direction of the next day's closing price, outperforms some previously proposed models, including those reported by Li and Dai (2020), McNally et al. (2018) and Livieris et al. (2021). In addition, the voting classifier also managed to outperform an additional two studies, Galeshchuk et al. (2018) and Valencia et al. (2019). On the other hand, better results have been reported on a similar task in Pant (2018), Stenqvist and Lönnö (2017) and Kilimci (2020). However, while these studies were based on data made up of 180, 30 and 90 days respectively, in this work we sought to train and generalise over a dataset of around 1 year and 3 months' worth of data. Therefore, one would need to reevaluate the accuracy obtained for these studies on larger datasets.

A somewhat different classification task was also proposed, namely, one where rather than predicting the direction of a price fluctuation, the goal is to determine the magnitude of the next day's closing price. The performance of this task cannot be compared to other studies' because this is the first study that tries to predict the size of the next day's closing price change. Here, it was shown that a model based on a 2-layer CNN with a 2-day time lag and 3 lagged features performs reasonably well, with a 57% accuracy over 10 classes. In addition, this model manages to achieve a direction accuracy of 63.3%, which is calculated by checking the predicted class, where classes 1-5 indicate a downtrend prediction while classes 6-10 indicate an uptrend prediction.

With regards to how lag affects price, it was evident that in nearly all cases the dataset with 7 days lag performed worst, suggesting that a 7-day lag is too long to capture a predictive relationship between social media content and price. In general, a 3-day lag results in higher maximum accuracies, though at the expense of lower overall means and a higher variation (as reflected in the DBMLA, the difference between minimum and maximum accuracy). Thus, a 1-day lag might yield more reliable predictions, since tweets are closer to the affected price. On the other hand, a slightly longer time period of 3 days might allow for possible 'ripple effects', whereby sentiment and opinion accumulate over time in response to ongoing trends, yielding good predictions for future trends in some cases. The precise nature of the relationship between time and predictive power of sentiment in economic matters is an important avenue for future work.

The main obstacle singled out in relation to achieving better accuracy results is the data used to train and test the implemented model because since the data is grouped daily, it causes the dataset to shrink to only a record per day, making the dataset small and hence, more difficult for the models to generalise over. Therefore, collecting more tweets and building a bigger dataset could prove vital in following up on this research.

Furthermore, the models presented here could also be tested on several time windows with varying lengths to see whether they perform better in specific ranges of time. Moreover, when it comes to the prediction of the magnitude of price change, changing the number of classes to be predicted might also result in better accuracy results. Finally, investigating a similar approach using hourly grouped data and carrying out sentiment analysis by using tweets' raw text as input to an embedding layer are other areas which can explored in order for this study to be advanced and furthered upon.

Whilst this paper focuses on sentiment analysis for Bitcoin price prediction, many different investment and trading strategies exist, and it would be opportune to investigate in future how this and other sentiment based strategies compare with other investment and trading strategies. Given that public sentiment can provide an indication towards Bitcoin (and other cryptocurrency) price change, it would be beneficial to investigate how: (i) investor attention and sentiment dispersion (investigated by Suardi et al. (2022)) may be used to augment the work presented herein; and (ii) how to predict change in public sentiment—a recent study demonstrates the effect that popular figures (in this case Elon Musk) may have on public sentiment (Zaman et al. 2022).

Code and data repository

Code developed within the scope of this paper has been open sourced at https://github.com/jacquesvcritien/fyp. This also includes all the different generated datasets, containing all the features including sentiment scores, so that all results and findings can be reproduced and validated.

Abbreviations

ANEW: Affective Norms for English Words; BiLSTM: Bidirectional Long Short Term Memory Cells; CNN: Convolutional Neural Network; DBMLA: Difference between Maximum and Lowest Accuracies; GI: General Inquirer; GRU: Gated Recurrent Unit; LIWC: Linguistic Inquiry and Word Count; LSTM: Long Short Term Memory Cells; NLP: Natural Language Processing; NLP: Natural Language Processing; NLTK: Natural Language Toolkit; RNN: Recurrent Neural Network; VADER: Valence Aware Dictionary and Sentiment Reasoner.

Author contributions

All authors made substantial contributions to the: conception of the work; analysis and interpretation of data; drafting of the work and revising it. All authors have approved the submitted version (and any substantially modified version that involves the author's contribution to the study). Furthermore, all authors agree to be personally accountable for the contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature. JCV solely created software used in the work and acquisition of data required. All authors read and approved the final manuscript.

Availability of data and materials

Data and sources are available from https://github.com/jacquesvcritien/fyp.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Centre for DLT, University of Malta, Msida MSD 2080, Malta. [2]Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands.

Received: 5 October 2021 Accepted: 29 March 2022
Published online: 05 May 2022

References

Abraham J, Higdon D, Nelson J, Ibarra J (2018) Cryptocurrency price prediction using tweet volumes and sentiment analysis

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on language in social media (LSM 2011), pp 30–38

Baker M, Wurgler J (2007) Investor sentiment in the stock market. J Econ Perspect 21(2):129–152

Balfagih AM, Keselj V (2019) Evaluating sentiment classifiers for bitcoin tweets in price prediction task. In: 2019 IEEE International conference on big data (Big Data), pp 5499–5506. https://doi.org/10.1109/BigData47090.2019.9006140

Bird S, Klein E, Loper E (2009) Natural language processing with python: analyzing text with the natural language toolkit. O'Reilly Media, USA

Ellul J (2021) Blockchain is dead! long live blockchain! J Br Blockchain Assoc, 21948

Galeshchuk S, Vasylchyshyn O, Krysovatyy A (2018) Bitcoin response to twitter sentiments. In: ICTERI workshops

Gunter B, Koteyko N, Atanasova D (2014) Sentiment analysis: a market-relevant and reliable measure of public feeling? Int J Mark Res 56(2):231–247

Hussein DME-DM (2018) A survey on sentiment analysis challenges. J King Saud Univ-Eng Sci 30(4):330–338

Hutto CJ, Gilbert E (2015) Vader: a parsimonious rule-based model for sentiment analysis of social media text

Kilimci Z (2020) Sentiment analysis based direction prediction in bitcoin using deep learning algorithms and word embedding models. Int J Intell Syst Appl Eng 8:60–65

Kimoto T, Asakawa K, Yoda M, Takeoka M (1990) Stock market prediction system with modular neural network I:1–61. https://doi.org/10.1109/IJCNN.1990.137535

Kraaijeveld O, De Smedt J (2020) The predictive power of public twitter sentiment for forecasting cryptocurrency prices. J Int Finan Markets Inst Money 65:101188. https://doi.org/10.1016/j.intfin.2020.101188

Kwon D-H, Kim J-B, Heo J-S, Kim C-M, Han Y (2019) Time series classification of cryptocurrency price trend based on a recurrent lstm neural network. J Inf Process Syst 15:694–706

Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl-Based Syst 69:14–23

Li Y, Dai W (2020) Bitcoin price forecasting method based on cnn-lstm hybrid neural network model. J Eng 2020. https://doi.org/10.1049/joe.2019.1203

Livieris IE, Kiriakidou N, Stavroyiannis S, Pintelas P (2021) An advanced cnn-lstm model for cryptocurrency forecasting. Electronics f10(3). https://doi.org/10.3390/electronics10030287

McNally S, Roche J, Caton S (2018) Predicting the price of bitcoin using machine learning, pp. 339–343. https://doi.org/10.1109/PDP2018.2018.00060

Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113

Mittal A, Goel A (2012) Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011) 15

Mohapatra S, Ahmed N, Alencar P (2020) KryptoOracle: a real-time cryptocurrency price prediction platform using twitter sentiments. arXiv:2003.04967

Naeem MA, Mbarki I, Suleman M, Vo XV, Shahzad J (2020) Does twitter happiness sentiment predict cryptocurrency? Int Rev Finance. https://doi.org/10.1111/irfi.12339

Nakamoto S (2009) Bitcoin: a peer-to-peer electronic cash system. Cryptography Mailing list at https://metzdowd.com

Nakov P, Rosenthal S, Kozareva Z, Stoyanov V, Ritter A, Wilson T (2013) SemEval-2013 task 2: sentiment analysis in Twitter. In: Second joint conference on lexical and computational semantics (*SEM), Volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), pp 312–320. Association for Computational Linguistics, Atlanta, Georgia, USA (2013). https://www.aclweb.org/anthology/S13-2052

Pagolu S, Challa K, Panda G, Majhi B (2016) Sentiment analysis of twitter data for predicting stock market movements

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 1(2):1–135. https://doi.org/10.1561/1500000001

Pant D, Neupane P, Poudel A, Pokhrel A, Lama B (2018) Recurrent neural network based bitcoin price prediction by twitter sentiment analysis, pp 128–132. https://doi.org/10.1109/CCCS.2018.8586824

Pantano E, Giglio S, Dennis C (2018) Making sense of consumers' tweets: Sentiment outcomes for fast fashion retailers through big data analytics. Int J Retail Distrib Manag 47. https://doi.org/10.1108/IJRDM-07-2018-0127

Ranjan S, Singh I, Dua S, Sood S (2018) Sentiment analysis of stock blog network communities for prediction of stock price trends. Indian J Finance 12:7. https://doi.org/10.17010/ijf/2018/v12i12/139888

Rao T, Srivastava S (2012) Analyzing stock market movements using twitter sentiment analysis. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), pp 119–123

Rosenthal S, Nakov P, Ritter A, Stoyanov V (2014) Semeval-2014 task 9: sentiment analysis in twitter

Serafini G, Yi P, Zhang Q, Brambilla M, Wang J, Hu Y, Li B (2010) Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches. In: 2020 International joint conference on neural networks, IJCNN 2020, Glasgow, United Kingdom, July 19–24, 2020, pp. 1–8. IEEE (2020). https://doi.org/10.1109/IJCNN48605.2020.9206704

Stenqvist E, Lönnö J (2017) Predicting bitcoin price fluctuation with twitter sentiment analysis

Suardi S, Rasel AR, Liu B (2022) On the predictive power of tweet sentiments and attention on bitcoin. Int Rev Econ Finance 79:289–301

Valencia F, Gómez-Espinosa A, Valdes B (2019) Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy 21:1–12. https://doi.org/10.3390/e21060589

Wołk K (2019) Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Syst 37. https://doi.org/10.1111/exsy.12493

Zaman S, Yaqub U, Saleem T (2022) Analysis of bitcoin's price spike in context of Elon Musk's twitter activity. Glob Knowl Memory Commun

Zhou X, Tao X, Yong J, Yang Z (2013) Sentiment analysis on tweets for social events, pp 557–562. https://doi.org/10.1109/CSCWD.2013.6581022