# Forecasting Dogecoin Prices with Twitter Sentiment Analysis and Deep Learning

K P M A L Karunarathne[1] Dr. M G N A S Fernando[2]
University of Colombo School of Computing, Sri Lanka
[1]anjulakarunarathne@gmail.com [2]nas@ucsc.cmb.ac.lk

***Abstract*—Cryptocurrencies are generally considered an investment opportunity with high risk and rewards. There has been a huge uprise in interest to invest in lesser-known cryptocurrencies, as opposed to popular cryptocurrencies like Bitcoin and Ethereum. Dogecoin is one of the lesser-known cryptocurrencies with immense growth in a brief period, possibly caused by the massive attention and demand it received from social media platforms. Nevertheless, Dogecoin value swings on both short-term and long-term valuations like most cryptocurrencies. Such volatility means uncertainty for potential investors and those interested in using Dogecoin for financial transactions.**

**We propose to use deep learning to predict the future value of Dogecoin. We use sentiment indicators extracted from Twitter and historical Dogecoin prices for training and evaluation purposes. Several Recurrent Neural Network (RNN) will be experimented and evaluated on their prediction capabilities. Experiments show that artificial recurrent neural network architectures such as Long Short-Term Memory (LSTM) exhibit substantial capabilities in predicting multiple steps into the future.**

## I. Introduction

DOGECOIN is an open-source, peer-to-peer digital currency initially created as a joke to make fun of wild speculations of various cryptocurrencies at the time [1]. Over the past few months, the cryptocurrency became increasingly popular due to attention from social media, mostly from Twitter, and is slowly seeing mass adoption. It reached a market capitalization of $44,056,548,828 by June 01, 2021, which is a staggering growth of 12,000% since January 01, 2021 [2]. Dogecoin presents a unique opportunity for price forecasting with social media activity due to its comparatively young age with Bitcoin and resulting in far greater volatility than alternative cryptocurrencies. The Bitcoin market has matured to the point that even though it is generally considered secluded, it regularly interacts with stock markets, commodities, and even foreign exchange markets [3]. This makes Bitcoin more appealing to long-term investors and therefore increases the risks of market manipulation, implying that it is less likely to be affected by sentiment from online social media. In comparison, the Dogecoin market does not suffer from such risks as a smaller cryptocurrency driven by its online community.

Dogecoin has unique differentiators that set it apart from Bitcoin and other cryptocurrencies. Dogecoin offers a lower transaction fee of about $0.01 compared to about $1 per transaction with Bitcoin. Furthermore, financial transactions in Dogecoin are significantly faster than in Bitcoin [4]. Such advantages heavily promote Dogecoin's acceptance in both trading and day-to-day usage. Considering the massive popularity of Bitcoin and its limited supply of 21 million [5], its value naturally tends to increase in the long run [6]. In contrast, Dogecoin has an endless supply of digital currency, thus its value heavily depends on its demand as a financial commodity, and therefore it is extremely volatile and hard to predict.

We created a staggering 2.5 quintillion bytes of information every day in the year 2020 [7]. More than 500 million tweets are being published daily [8], and 1.84 billion people actively use Facebook each day [9]. The snippets of information published on these platforms vary from what the person had for their breakfast, to excitement over their favorite sports team winning a tournament. More specifically, Twitter has become known as a platform where such information is spread quickly in a summarized format.

Public confidence and opinion regarding a financial commodity are arguably the core bases of its value, and social media platforms have allowed the general public to express their opinion since their inception. These pieces of information become available for studies with the open APIs provided by Facebook and Twitter [10], [11]. As such, understanding the effects of public opinion on financial commodities becomes possible. Dogecoin is similar to commonly known financial assets in the sense that it is affected by socially constructed opinions, disregarding the fact of whether those opinions are factual or not. It was created in 2013 [1], but didn't gain significant public attention at the time. However, it recently became a trending topic of discussion on various social media platforms, resulting in a staggering amount of growth in a short period. Moreover, it is quite interesting to observe how the growth fluctuates with the recent social media discussions concerning the high energy requirements of cryptocurrencies [12]. As such, opinions, and news about Dogecoin are widespread throughout the social media sphere and keep affecting its growth and value.

Cryptocurrencies are a fresh store of value relative to fiat currencies such as the U.S. dollar (USD) or the Sri Lankan Rupee (LKR). What causes price changes in this new store of value is an area of debate. It has been argued that cryptocurrencies are unique assets in that their price behaves in ways like both a standard financial asset and a speculative one [13]. Given that cryptocurrency prices behave nothing like traditional currencies, the prices are extremely difficult to predict. Meanwhile, Twitter is increasingly used as a news source influencing purchase decisions by informing users of the cryptocurrency and its increasing popularity. Understanding the impact of tweets on price fluctuations can provide a purchasing and selling advantage to both cryptocurrency users and traders. This research paper will allow one to make better-informed purchase decisions and selling decisions, potentially leading to profits.

The remainder of this research paper is organized as follows. Section 2 reviews the existing related works and summarizes them. Section 3 describes the designs and implementation details of the research methodology. Experimentation results and their evaluation are illustrated in Section 4. Lastly, Section 5 presents the conclusion, findings, and future work.

## II. Literature Review

Several studies have been conducted on predicting the prices of cryptocurrencies such as Bitcoin and Ethereum. However, the lack of research studies related to Dogecoin price prediction is revealed in the comprehensive literature review. There have been numerous studies using textual data from online social media to predict Bitcoin prices as well.

Kahneman and Tversky established that decisions, even the ones revolving around financial consequences are impacted by emotions, and not just by the value of the commodity alone [14]. Dolan et. al. discussed how human emotions interact with and influences other domains of cognition such as attention, memory, and specific reasoning. He further supported that decision-making is heavily influenced by emotions [15]. Later studies also found that the information people gathered online has an impact on their purchase decisions. Panger studied the emotions expressed in social media and concluded that the sentiment of a person's tweets correlated with their general emotional state. Moreover, he found that social media like Twitter tends to have a reassuring effect on the user, instead of amplifying their emotional state [16]. It has also been shown in a study by Gartner that most people relied on social media to encourage their purchase decisions [17]. These findings open the possibility for tools such as sentiment analysis to be used for price prediction, as they indicate the demand for a commodity, and thus price impact.

Several researchers have explicitly studied the sentiment analysis of tweets. According to Kouloumpis et. al., applying traditional natural language processing techniques such as document-level sentiment scoring and sentence-level sentiment scoring is ineffective in the context of tweets, due to the limited number of characters (originally 140-characters, which was increased to 280-characters in 2018 [18]) and uniqueness of language used [19]. Pak et. al. built a classifier that categorizes tweets into positive, negative, and neutral classes which provided higher performance than the previously proposed models [20].

Gilbert and Hutto presented an effective yet simple rule-based model for sentiment analysis named VADER (for Valence Aware Dictionary for sEntiment Reasoning) which uses a "gold-standard" sentiment lexicon that is precisely accustomed to the context of micro-blogs such as Twitter. Their model considers many grammatical and syntactical conventions that humans use to express sentiment intensity, which greatly improves the accuracy across various domains such as social media text, movie reviews, product reviews, etc. According to their correlation coefficient measurements, it was concluded that VADER performs as well as human raters at matching the ground truth, and in most cases outperforms individual human raters at classifying tweets into positive, negative, or neutral categories. VADER is fast and requires less computational power compared to complex models like Support Vector Machine (SVM). Moreover, both the lexicon and rules used by VADER are easily accessible for inspection. Consequently, they can be easily understood, modified, or extended as necessary [21], [22].

Connor, Balasubramanyan et. al. discovered that the sentiments shown in tweets were reflective of the public opinion on various topics in national polling [23]. The potential of using text streams i.e., tweets as a replacement and supplement for traditional polling was highlighted in the study which in turn can save costs compared to national polling, as tweets are publicly available via open APIs. If the sentiment indicators from tweets do accurately represent the public opinion on some financial commodity like Dogecoin, it implies the possibility of using the sentiment scores to predict the demand for that commodity, and accordingly the fluctuations of its price.

Several pieces of research have studied the efficacy of specifically using the above approach of using sentiment analysis to predict Bitcoin prices. Stenqvist and Lonno describe their tweets sanitizing approach where they collected tweets related to Bitcoin from May 11th to June 11th of 2017. During the period, a total of 2, 271, 815 tweets were collected. Even though it is a substantial amount, their clean-up approach removed 44.8% of tweets from the dataset. Non-alphanumeric symbols (i.e., "#" and "@") were removed from the tweets during the clean-up process. Tweets that were duplicates, not relevant, or too influential were also removed from the analysis. The authors used VADER to analyze the sentiment of each tweet and classified them as either positive, negative, or neutral. Only the tweets that were classified as positive and negative were kept in the final analysis. Their prediction model aggregated tweet sentiments over 30 minutes with 4 shifts forward and resulted in 79% accuracy with a sentiment change threshold of 2.2% [24].

Lamon et. al. analyzed the sentiment of tweets combined with news headlines over 67 days to predict price fluctuations in Bitcoin, Litecoin, and Ethereum. (Two of the popular alternative cryptocurrencies in the market.) However, their approach is unique from other methods in the sense that, rather than labeling texts with positive or negative sentiment first, they label them according to the actual price changes of one day in the future for each of the cryptocurrencies. Their model is generally able to predict the largest price increases and decreases correctly according to the magnitude, where they correctly predicted 43.9% of price increases and 61.9% of price decreases accurately. Similarly, Colianni et. al. reported a price fluctuation prediction accuracy exceeding 90% using supervised learning algorithms on tweets; however, their dataset was labeled using an online text sentiment service [25]. Thus, their accuracy measurement corresponds to how well their model matched the online text sentiment service, as opposed to the accuracy of predicting price fluctuations.

Georgoula et. al. investigated sentiment analysis on tweets with an SVM-based approach coupled with the frequency of Wikipedia views and the network hash rate [26]. Tweet quantity has also significantly shown predictability for future Realized Volatility (RV) and trade volume of Bitcoin [27]. Matta et. al. uncovered significant correlations between Google Trends statistics and Bitcoin prices [28]. Matta, Lunesu et. al. discovered a similar correlation except with Google Trends statistics and Bitcoin trade volumes [29].

Kaminski studied the interconnection between Bitcoin market indicators and Twitter posts of sentimental significance regarding Bitcoin. He considered about 160,000 tweets extracted from the hashtag “bitcoin” in his analysis. During the period from November 23rd of 2013 to March 7th of 2014, he found significant positive correlations between emotional tweets, daily closing prices, trading volume, and intra-day price spread of Bitcoin when considering the Pearson correlation. For example, whenever the trading volume of Bitcoin was higher, emotions expressed on Twitter were also of higher intensity. He further interpreted Twitter as the virtual trading floor of Bitcoin, and that it reflects Bitcoin’s trading dynamics with emotions [30].

Pant, Neupane et. al. built their own sentiment analyzer which classifies tweets into positive and negative classes with an accuracy of 81.39%, and further developed an RNN model that can predict next-day Bitcoin price with 77.62% of accuracy. They also reported a significant Pearson correlation of 0.41 between the negative sentiment of tweets and the reduction in Bitcoin price. However, their approach is limited to a comparatively smaller dataset of just 7,454 tweets for 3 years (From January 1st of 2015 to December 31st of 2017) from a few Twitter accounts that supply news related to Bitcoin and other cryptocurrencies [30]. Even though these Twitter accounts might be able to influence people’s purchase decisions to a certain degree, it is also possible that the price fluctuations are causing the tweets in the first place, which is the case with news in general.

The major limitation of the above publications is often the smaller sample size as many studies consider daily closing prices of the cryptocurrency. Recent speculations suggest that the effects of social media attention can become visible within mere hours [31]. Therefore, it motivates researchers to study how immediate such effects are by studying a larger sample size considering price fluctuations in shorter periods. This observation is reassured by the work of Sattarov, Jeon et. al. where they achieved an accuracy of 62.48% in predicting Bitcoin price fluctuations on a minute-by-minute basis with Bitcoin-related sentiment scores from Twitter and historical Bitcoin prices [32]. Moreover, due to the limited liquidity in the cryptocurrency exchanges, the market suffers from a high risk of manipulation. This causes misinformation to spread throughout various social media channels such as Twitter and message boards such as Reddit, which might cause artificial price inflation and deflation.

There have also been studies that use historical prices alone, and sometimes statistics derived from the cryptocurrency blockchain to predict future price variations. Shah et. al. utilized a latent source model to predict Bitcoin prices using 10-second historical price and Bitcoin limit order book features, claiming 89% returns in 50 days with a Sharpe ratio of 4.10 using a simulated trading strategy [33]. Greaves et. al. used the Bitcoin transaction graph combined with historical time deltas of 30, 60, and 120 minutes to predict the price of Bitcoin [34]. They used SVM and Artificial Neural Network (ANN) resulting in an accuracy of 55% for predicting the direction of price fluctuations. Their findings indicated limited success, indicating the possible effects of external factors such as the ones discussed above. There have also been attempts of using wavelets to predict Bitcoin prices [35]. Kristoufek et. al. identified positive correlations between search engine views, blockchain hash rate, and mining difficulty with Bitcoin price [13].

Predicting the price of Dogecoin is comparable to traditional time series prediction tasks such as forex and stock market prediction. Mao et. al. have shown that Twitter sentiment can predict daily up and down changes in stock with an accuracy of 86.7% [36],. Kimoto et. al. discusses a buying and selling timing prediction system for stocks on the Tokyo Stock Exchange which achieved accurate predictions. Furthermore, their simulation exercise on stocks trading using the timing prediction system showed excellent profits [37].

Zhu et. al. used stock trading volumes as input to a neural network and achieved modest improvements in prediction performance over medium and long terms [38]. Several studies have used the MultiLayer Perceptron (MLP) to predict stock prices [39], [40]. According to Koskela et. al., the main limitation of an MLP-based price prediction approach is that it only analyses one observation at a time [41]. RNN have the advantage that the output of each layer is stored in a context layer, which is looped back in with the output from the next layer. This approach gives the RNN some sort of memory as opposed to the MLP. Giles, Lawrence et. al. identified as such the temporal relationship of the series is explicitly modeled by the internal states, which heavily contributes to the model effectiveness [42]. Rather, Agarwal et. al. continued this approach and combined an RNN with a network optimization genetic algorithm for predicting stock returns [43]. LSTM networks are another form of RNN networks. LSTM networks can choose which data to remember, and which data to forget based on the weight and importance of that attribute, which is the main differentiator between them‘ and Elman RNN networks. Gers, Felix A. et. al. implemented an LSTM network for a time series prediction task, where the LSTM network performed as well as the RNN network [44]. It was suggested to use LSTM only when simpler traditional approaches fail. Yang et. al. proposed an LSTM model to predict stock closing prices with sentiment analysis. Their reasoning for selecting an LSTM model is its memory function, which is rather useful in analyzing relationships among time-series data. The model outperformed in predicting closer closing prices, higher rise and fall classification accuracy, and lower time offset [45].

This research paper also builds upon the RNN, more specifically LSTM networks due to their relatively high performance shown in similar tasks. Dogecoin is distinctive from other cryptocurrencies in the sense that it is highly volatile and extremely sensitive to sentiment from social media such as Twitter. Because existing solutions are attuned to specific use cases such as daily closing price prediction of Bitcoin, they cannot be used directly for short-term Dogecoin price prediction, providing the opportunity for exploring a different solution. Furthermore, most of the existing solutions for Bitcoin price prediction focus on predicting a single value based on one or more variables. As explained in earlier sections, this is not ideal nor helpful for cryptocurrency investors. This research paper focus on forecasting several hours into the future using multi-step prediction capabilities of deep neural networks makes more sense from the investors' perspective.

## III. Methodology

This section presents the overall methodology of the research paper. The research process is broken down into several stages for better understanding, which are explained in their relevant subsections.

The high-level research design is presented in Figure 1.

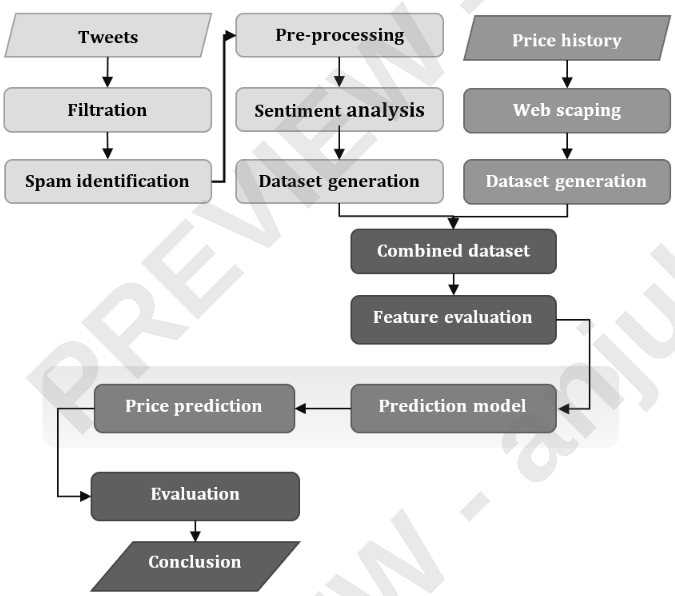


Fig. 1: High-level architecture diagram of the research methodology

### A. Price History

CoinGecko provides hourly historical prices of Dogecoin in US dollars for any given day individually [46]. However, there is no public API to obtain hourly historical prices for custom periods longer than one day. Using a custom Python script, the CoinGecko website is scraped to obtain the hourly price information by sending one request per each day from January 1st to April 30th of 2021. The resulting fragments of data are combined into a single dataset.

### B. Tweets Filtration and Spam Identification

The tweets are available on Kaggle, [47] containing tweets referencing "DOGE" and "DOGECOIN" hashtags posted in the previously mentioned. The dataset is filtered where non-English tweets and retweets are removed. Authors who posted two identical tweets in a row are flagged as autonomous accounts (i.e., bots), and tweets authored from such accounts are removed from the final analysis.

### C. Pre-processing the Tweets

During the pre-processing stage before sentiment analysis, multiple things happen. Twitter return handles are removed from the text content. People use return handles at the beginning of the tweets (i.e., RT @username) to indicate that they are re-posting someone else's content. Twitter mentioning handles are also removed from the text content. These are used to indicate that they are mentioning someone else in their tweets. Removing such elements from tweets will ensure that usernames will not be mistaken for words that contribute to the overall sentiment of the tweets.

Furthermore, URLs beginning with "http://" and "https://" are also removed to stop misleading the sentiment classifier with the words amongst the website domains and URLs.

Special characters, numbers, and punctuation except for "#" are also discarded. The resulting tweets are subjected to sentiment analysis with VADER.

### D. Sentiment Analysis and Summarising

VADER analyzes the sentiment of each tweet and returns a positive score, a negative score, and a neutral score. All values are in the range of [0.0, +1.0], where the number implies the intensity of the emotion. Based on these values, a compound score will be calculated in the range of [-1.0, +1.0] in which "-1.0" means the sentiment is completely negative, "0.0" means neutral, and "+1.0" means completely positive.

It is required that we re-sample the sentiment scores into hourly observations to map them into relevant Dogecoin price data points. Tweets contain the relevant date and time they were published, which can be used to group them into hourly bins. We can summarize each bin using by taking the sum of compound scores mapped into the individual bins.

### E. Prediction Models

We use the recent Dogecoin prices and sentiment indicators as input to forecast the expected Dogecoin prices for the next six hours. The observations from January 1st to March 31st of 2021 are used for training the prediction models, while the remaining observations from April 1st to April 30th of 2021 are used for evaluation.

This kind of problem framing is generally known as the multi-step time series forecasting problem, which is a type of sequence-to-sequence (also known as *seq2seq*) prediction problem. We explore several ANN models intended for producing multi-step sequence predictions in this research paper.

*1) Vanilla LSTM Model:* Recurrent Neural Networks such as LSTMs exploit the recurrent connection of the outputs from previous timesteps. LSTMs have sort of an internal memory which allows them to carry a state over the input sequences. They are also intended to take sequences as input data. It is one of the biggest advantages of LSTMs compared to other artificial neural networks where lag observations should be provided to them as input features.

We implement a simple LSTM model which reads in a sequence of hours of historical Dogecoin prices and Twitter sentiment indicators and predicts a vector output of Dogecoin price for the next six hours. We use an artificial neural network with a single hidden LSTM layer with 50 units. The hidden LSTM layer is followed by a fully-connected layer with 20 nodes that will understand the features learned by the previous layer. Lastly, the output layer will produce a six-element vector, one for each hour in the output sequence.

Figure 2 graphically represents the architecture of the artificial neural network.

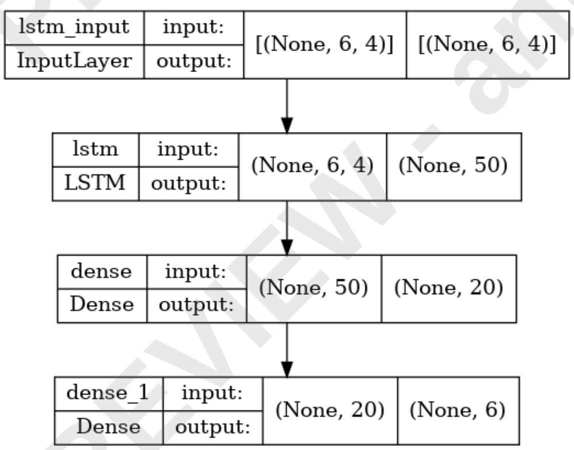


Fig. 2: Architecture of the vanilla LSTM model

We use the MSE loss function to calculate training loss because it matches the preferred error measurement of Root Mean Square Error (RMSE), and train the prediction model for 70 epochs with a batch size of 24 using the efficient Adam implementation of stochastic gradient descent.

*2) Encoder-Decoder LSTM Model:* The Encoder-Decoder LSTM architecture is another RNN designed for *seq2seq* problems. It consists of two sub-models: the encoder which reads the input sequences and compresses them to a fixed-length internal representation, and the decoder which interprets the internal representation and uses it to predict the output sequence.

We define the encoder using a hidden layer consisting of 50 LSTM units. It will read the input sequence and produce a 50-element vector that interprets features from the input sequence. Consequentially, the internal representation of the input sequence is repeated several times, once for each timestep in the output sequence. It is then passed down to the decoder layer, which is defined as an LSTM hidden layer with 50 units. Each of these units will output the entire sequence, containing values for each of the six hours. These sequences are the basis for what to predict for each hour in the output sequence.

We then use a fully connected layer to interpret each timestep in the output sequence before the final output layer. The output layer produces a single step in the output sequence, not all six hours at once. Therefore, we use the same layers applied to each timestep in the output sequence. To achieve this, we wrap the interpretation layer and the output layer in a TimeDistributed wrapper from Keras that allows the wrapped layers to be used for each timestep from the decoder [48]. This allows the decoder sub-model to figure out the context required for each timestep in the output sequence, and the wrapped dense layers to interpret each timestep separately, but reuse the same weights to perform the interpretation.

Figure 3 graphically represents the architecture of the artificial neural network.

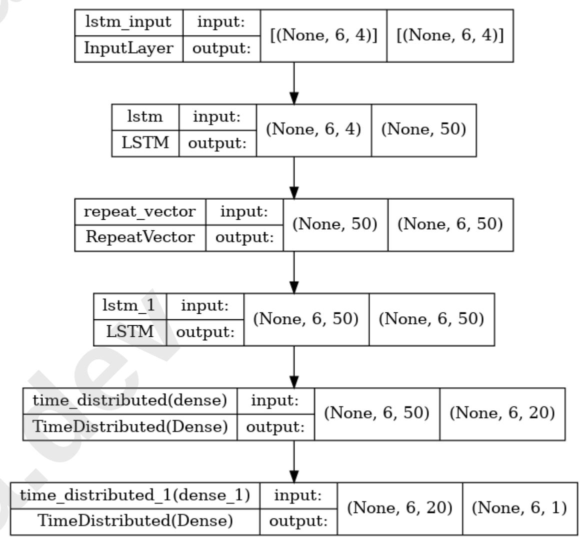


Fig. 3: Architecture of the Encoder-Decoder LSTM model

We train the prediction model for 20 epochs with a batch size of 16 using the Adam implementation of stochastic gradient descent.

*3) CNN-LSTM Encoder-Decoder Model:* A Convolutional Neural Network (CNN) could be used as the encoder in the Encoder-Decoder architecture. Even though CNN does not directly support sequential input, we can easily use a 1-D CNN to read across sequential input and automatically learn the salient features. These features can be understood by an LSTM decoder layer. This hybrid architecture is usually referred to as the CNN-LSTM Encoder-Decoder architecture.

The encoder layer consists of two convolutional layers, followed by a max-pooling layer. The output of the max-pooling layer is flattened. The first convolutional layer reads across the input sequence and projects the results onto feature maps. The second performs the same operation, but on the feature maps created by the first layer, attempting to amplify any salient features. We use 64 feature maps per convolutional layer and read the input sequences with a kernel size of three timesteps.

The max-pooling layer minimizes the feature maps by keeping $\frac{1}{4}$ of the values with the largest (max) signal. The distilled feature maps after the pooling layer are then flattened into one long vector that can then be used as input to the decoding process. We use the same decoder layer explained in the previous section.

Figure 4 graphically represents the architecture of the artificial neural network.

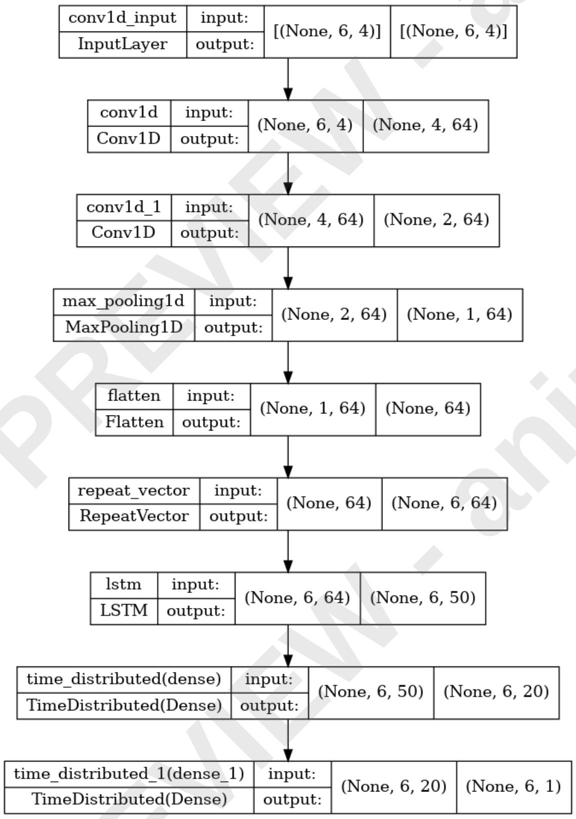


Fig. 4: Architecture of the CNN-LSTM Encoder-Decoder model

We train the prediction model for 20 epochs with a batch size of 16 using the Adam implementation of stochastic gradient descent.

## IV. Evaluation

The prediction models produce vectors consisting of six individual forecasts, one for each hour ahead. It is useful for multi-step forecasting problems to evaluate each and every timestep independently and evaluate the skill for a specific lead time. For instance, what is the performance of the prediction model when predicting the next hour, compared to predicting the third hour from the present?

Having an error metric in the same unit as the predicted values can also be useful in certain scenarios. Both RMSE (Equation 1) and Mean Absolute Error (MAE) (Equation 2) are suitable for the situation. However, RMSE will be adopted in this research paper due to its relatively high weight to large errors. Mean Absolute Percentage Error (MAPE) (Equation 3) is also calculated because it is an intuitive interpretation in terms of relative error.

The primary performance metric for this research study is the RMSE for each leading timestep from the first hour to

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(predicted_i - actual_i)^2} \tag{1}$$

1: Root Mean Square Error

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|predicted_i - acctual_i| \tag{2}$$

2: Mean Absolute Error

$$MAPE = \frac{100\%}{n}\Sigma_{i=1}^{n}\ |\frac{predicted_i - actual_i}{actual_i}| \tag{3}$$

3: Mean Absolute Percentage Error

the sixth hour. However, it might also be useful to summarize the overall performance of the prediction model using a single score. One possible score is the RMSE across all forecasted timesteps.

The prediction model is evaluated using the walk-forward validation approach. The model is required to make the initial six-hour window prediction. Following, the correct data for the predicted six-hour window is made available to the model so that it can be used for predicting subsequent windows. This is both realistic considering how the model might be used in practice and advantageous to the models because it allows them to make use of the best available data.

Another parameter to consider is the length of the window that the prediction model takes as input to produce the subsequent output window. We explore different window lengths and evaluate the results on every prediction model.

Table I demonstrates the separation of input and predicted data. Recall that one window contains six timesteps each representing one hour. Assume that [n] represents the nth six-hour window. To predict the $n^{th}$ window, we can train the model to use either the last [n-1] window, or both [n-2, n-1] windows, and likewise. Different window lengths can significantly affect the performance of the prediction model.

| Input windows | | | Predicted window |
|---|---|---|---|
| Last one | Last two | Last three | |
| [n-1] | [n-2, n-1] | [n-3, n-2, n-1] | [n] |

TABLE I: Demonstration of the input and output windows

More than three million tweets were processed in the initial stage. Approximately 16.04% of tweets were removed from non-English languages. Furthermore, 72.92% were removed from the remaining tweets during the spam removal process, resulting in a total reduction of 77.25% of tweets.

A random sample of tweets identified as spam and published by autonomous accounts is presented in Table II.

Consequentially, the prediction models are implemented according to the architectures presented in Section III Subsection E, and trained with the first three months of data from January 1st to March 31st of 2021.

Walk-forward validation is used during the evaluation process. The model predicts the first six hours from 01.00 AM to 6.00 AM on April 1st of 2021 initially. Afterward, the

| Author | Tweet content |
|---|---|
| zayd00208042 | 100$ #doge giveaway if this post reach 1000 retweet Must like + follow #dogecoin #dogearmy #safemoon |
| traderwisdom | Unusual #volume spike in $ANTBTC 6 times the average volume in $ANT - $BTC binance bsc crypto $crypto #BTC $BTC ETH $ETH #LTC $LTC DOGE $DOGE Don't have Binance? Create a new account (5% off on fee) [Link removed] |
| zeroenvysxs | It's time to start investing in the #StockMarket if you haven't already! PLUS if you click the link you'll receive a FREE Stock JUST FOR SIGN-ING UP! NO FEES, 100% FREE & you can also trade #Crypto for FREE! #Robinhood #DOGE-COIN #DOGE #Bitcoin #GME #AMC #FB [Link removed] |
| yekturus | Strong #Uptrends in 30/60/120 mins (Binance $BTC pairs): $DOGE $PERP $PPT $TVK $WNXM #crypto #coin #RSI |
| zealdorn | Free Doge every hour! [Link removed] #doge #do-gecoin #cryptocurrency #crypto |

TABLE II: Sample of tweets identified as spam

actual values are made available to the model. Subsequently, the model predicts the next six hours from 7.00 AM to 12.00 PM using the newly available history.

Table III summarizes the overall performances of previously discussed prediction models. Experimentation results of individual models are also presented in respective subsections.

| Model | MAPE | RMSE | MAE |
|---|---|---|---|
| Vanilla LSTM | 8.7477% | 0.080482 | 0.049931 |
| Encoder-Decoder LSTM | 12.7120% | 0.105823 | 0.070515 |
| CNN-LSTM Encoder-Decoder | 16.2562% | 0.165237 | 0.106038 |

TABLE III: Performance summary of the prediction models

*1) Vanilla LSTM Model:* Figure 5 demonstrates the individual RMSE values when predicting Dogecoin value in US dollars for six hourly timesteps ahead.

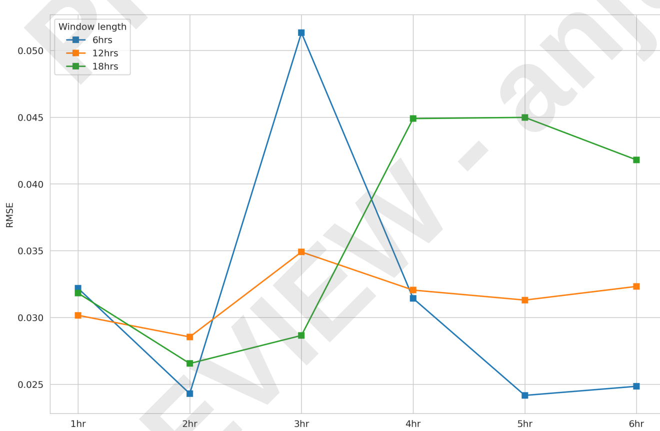


Fig. 5: RMSE of individual timesteps obtained from the vanilla LSTM model with different window lengths

Using a six-hour window length, (equivalent to one window) the model has the lowest RMSE when predicting the second, fifth, and sixth hours. Interestingly, using a twelve-hour window gives the lowest error when predicting the first hour.

Considering the performance of overall predictions, the model acquired 0.08 RMSE and 8.74% MAPE when trained and evaluated with a six-hour window length. The predictions and actual values can be compared in Figure 6.

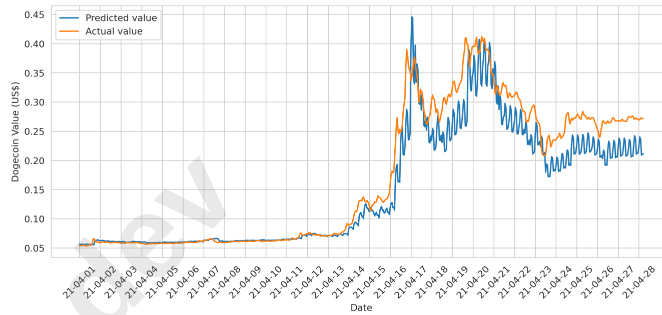


Fig. 6: Dogecoin value predictions obtained from the vanilla LSTM model compared to actual values

*2) Encoder-Decoder LSTM Model:* Figure 7 demonstrates the individual RMSE values when predicting Dogecoin value in US dollars for six hourly timesteps ahead.

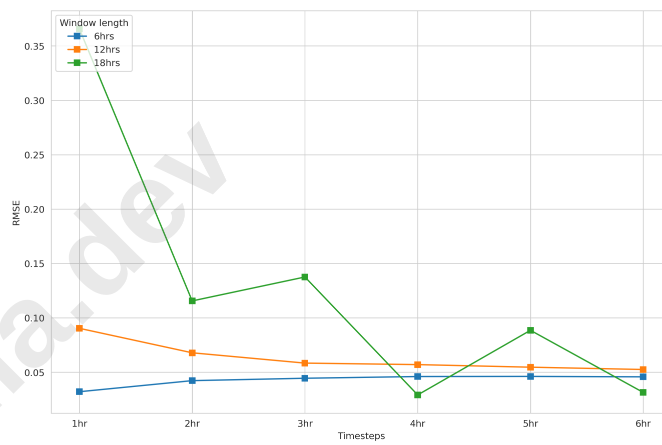


Fig. 7: RMSE of individual timesteps obtained from Encoder-Decoder LSTM model with different window lengths

Encoder-Decoder LSTM model has average prediction performance compared to other models. It behaves quite similarly when using either one or two six-hour windows, except on the first and second-hour timesteps. However, the model can achieve the highest performance when forecasting the fourth and sixth hours using three six-hour windows,

The model acquired 0.10 RMSE and 12.71% MAPE when trained and evaluated with a six-hour window length. The predictions and actual values are presented in Figure 8.

*3) CNN-LSTM Encoder-Decoder Model:* Figure 9 demonstrates the individual RMSE values when predicting Dogecoin value in US dollars using CNN-LSTM Encoder-Decoder Model for six hourly timesteps ahead.

Compared to previous models, CNN-LSTM Encoder-Decoder model has the worst prediction performance. Nonetheless, the model seems to prefer longer window lengths according to results presented in Figure 9. The model obtained the highest performance using three windows of eighteen-hours for every timestep except the first hour.

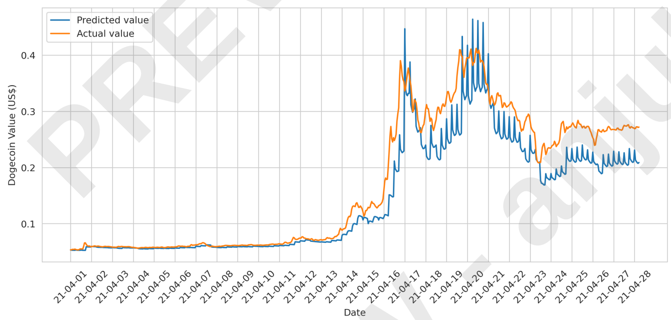


Fig. 8: Dogecoin value predictions obtained from Encoder-Decoder LSTM model compared to actual values

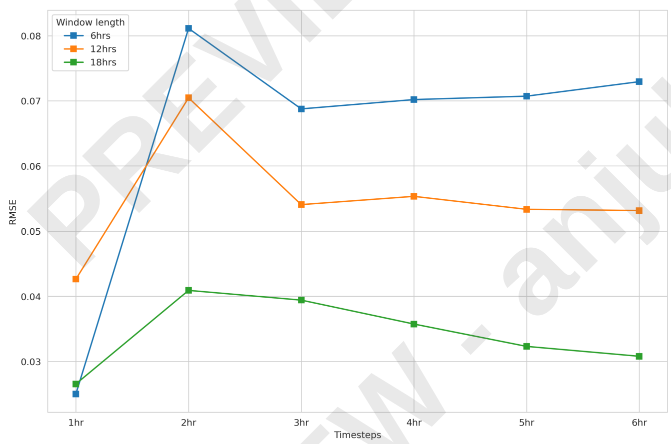


Fig. 9: RMSE of individual timesteps obtained from CNN-LSTM Encoder-Decoder model with different window lengths

The model acquired 0.16 RMSE and 16.25% MAPE when trained and evaluated with a six-hour window length. The predictions and actual values can be compared in Figure 10.

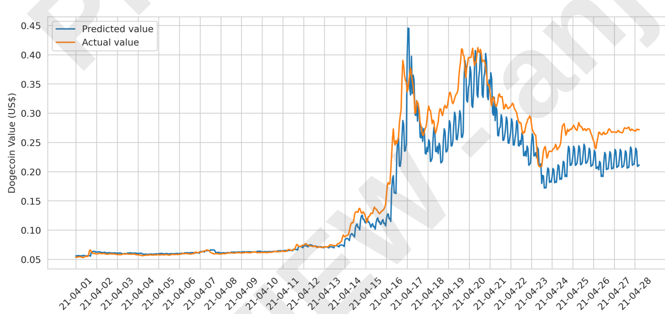


Fig. 10: Dogecoin value predictions obtained from CNN-LSTM Encoder Decoder model compared to actual values

## V. Conclusion

This research paper primarily focuses on developing an optimal deep learning model that can forecast Dogecoin value in US dollars. The results presented in Section IV suggest that the vanilla LSTM model developed in Section III can predict Dogecoin value with an error rate as low as 8.74% MAPE and 0.08 RMSE. The results also demonstrate that the prediction model can forecast multiple timesteps into the future with high performance, where it forecasts Dogecoin value five hours into the future with an even lower error rate of less than 0.025 RMSE. These results are incredibly useful for not only investors but also market analyzers and studies on alternative cryptocurrencies.

There are several aspects to consider for further research based on this study. Retraining the prediction models if the performance is inadequate is such an aspect. This is not expensive if the dataset is small and the modeling method is simple. For instance, a vanilla LSTM model can be faster to train than Encoder-Decoder LSTM and CNN-LSTM Encoder-Decoder models. Those models also require more computing resources at scale.

Furthermore, there can be many other aspects that can directly affect Dogecoin value. For instance, Reddit is also used by many users to initiate conversations on various topics including cryptocurrencies. Likewise, news articles have a much broader audience than social media platforms. Such factors can be considered in combination with Twitter or individually, which could improve the prediction performance.

## Acknowledgment

We take this opportunity to appreciate the assistance and contribution of supervisors and advisors throughout the entirety of the research.

## References

[1] U. W. Chohan, “A History of Dogecoin,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3091219, Feb. 2021. [Online]. Available: https://papers.ssrn.com/abstract=3091219

[2] N. Vega, “Dogecoin up 12,000% since January — Here’s how much money you’d have if you invested $1,000 at the beginning of 2021,” May 2021. [Online]. Available: https://www.cnbc.com/2021/05/05/how-much-a-1000-dollar-investment-in-dogecoin-is-worth.html

[3] S. M. Jang, E. Yi, W. C. Kim, and K. Ahn, “Information Flow between Bitcoin and Other Investment Assets,” *Entropy*, vol. 21, no. 11, p. 1116, Nov. 2019. [Online]. Available: https://www.mdpi.com/1099-4300/21/11/1116

[4] K. Ghaiti, “The Volatility of Bitcoin, Bitcoin Cash, Litecoin, Dogecoin and Ethereum,” Thesis, Université d’Ottawa / University of Ottawa, Apr. 2021. [Online]. Available: http://ruor.uottawa.ca/handle/10393/42016

[5] “What Happens to Bitcoin After All 21 Million Are Mined?” [Online]. Available: https://www.investopedia.com/tech/what-happens-bitcoin-after-21-million-mined/

[6] “Why Does the Price of Bitcoin Keep Going Up?” [Online]. Available: https://www.investopedia.com/why-does-bitcoin-keep-going-up-5092683

[7] J. Bulao, “How Much Data Is Created Every Day in 2022?” Mar. 2022. [Online]. Available: https://techjury.net/blog/how-much-data-is-created-every-day/

[8] Y. Lin, “10 twitter statistics every marketer should know in 2021,” Nov 2021. [Online]. Available: https://www.oberlo.com/blog/twitter-statistics

[9] M. Mohsin, “10 facebook statistics you beed to know in 2021,” Mar 2022. [Online]. Available: https://www.oberlo.com/blog/facebook-statistics

[10] “Twitter api - documentation,” Twitter Developer API. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api

[11] “Facebook graph api - documentation,” Facebook for Developers. [Online]. Available: https://developers.facebook.com/docs/graph-api

[12] G. Kay, “Elon Musk Says He’s Working with Dogecoin Developers to Make the Cryptocurrency More Efficient.” [Online]. Available: https://www.businessinsider.com/elon-musk-dogecoin-developers-environmentally-friendly-2021-5

[13] L. Kristoufek, “What are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis,” *PloS One*, vol. 10, no. 4, p. e0123923, 2015.

[14] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979. [Online]. Available: https://www.jstor.org/stable/1914185

[15] R. J. Dolan, "Emotion, Cognition, and Behavior," *Science (New York, N.Y.)*, vol. 298, no. 5596, pp. 1191–1194, Nov. 2002.

[16] "Emotion in Social Media." [Online]. Available: https://www.ischool.berkeley.edu/research/publications/2017/emotion-social-media

[17] "Gartner Says Majority of Consumers Rely on Social Networks to Guide Purchase Decisions," Jul. 2010. [Online]. Available: https://www.pressebox.com/pressrelease/gartner-uk-ltd/Gartner-Says-Majority-of-Consumers-Rely-on-Social-Networks-to-Guide-Purchase-Decisions/boxid/361969

[18] "_usGiving You More Characters to Express Yourself." [Online]. Available: https://blog.twitter.com/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself

[19] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!" *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 538–541, 2011. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14185

[20] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

[21] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[22] cjhutto, "Git Repository - cjhutto/vadersentiment," GitHub, 12 2018. [Online]. Available: https://github.com/cjhutto/vaderSentiment

[23] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," 2010.

[24] E. Stenqvist and J. Lönnö, "Predicting Bitcoin price fluctuation with Twitter sentiment analysis," *undefined*, 2017. [Online]. Available: https://www.semanticscholar.org/paper/Predicting-Bitcoin-price-fluctuation-with-Twitter-Stenqvist-L%C3%B6nn%C3%B6/0954565aebae3590e6ef654fd03410c3bdd7d15a

[25] S. G. Colianni, S. Rosales, and M. Signorotti, "Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis," *undefined*, 2015. [Online]. Available: https://www.semanticscholar.org/paper/Algorithmic-Trading-of-Cryptocurrency-Based-on-Colianni-Rosales/9b838a3177523b8179511283b9489caa0f69910d

[26] I. Georgoula, D. Pournarakis, C. Bilanakos, D. Sotiropoulos, and G. M. Giaglis, "Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2607167, May 2015. [Online]. Available: https://papers.ssrn.com/abstract=2607167

[27] D. Shen, A. Urquhart, and P. Wang, "Does Twitter predict Bitcoin?" *Economics Letters*, vol. 174, pp. 118–122, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165176518304634

[28] M. Matta, M. I. Lunesu, and M. Marchesi, "Bitcoin Spread Prediction Using Social and Web Search Media," in *UMAP Workshops*, 2015.

[29] "The Predictor Impact of Web Search Media On Bitcoin Trading Volumes - PDF Free Download." [Online]. Available: https://businessdocbox.com/Marketing/70008714-The-predictor-impact-of-web-search-media-on-bitcoin-trading-volumes.html

[30] J. Kaminski, "Nowcasting the Bitcoin Market with Twitter Signals," Jun. 2014. [Online]. Available: https://arxiv.org/abs/1406.7577v3

[31] J. Jolly, "Price of dogecoin rises by 50% following Elon Musk tweet," *The Guardian*, Feb. 2021. [Online]. Available: https://www.theguardian.com/business/2021/feb/04/price-of-dogecoin-rises-by-50-following-elon-musk-tweet

[32] O. Sattarov, H. S. Jeon, R. Oh, and J. D. Lee, "Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis," in *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, Nov. 2020, pp. 1–4.

[33] D. Shah and K. Zhang, "Bayesian regression and Bitcoin," Oct. 2014. [Online]. Available: https://arxiv.org/abs/1410.1231v1

[34] A. Greaves and B. Au, "Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin," *undefined*, 2015. [Online]. Available: https://www.semanticscholar.org/paper/Using-the-Bitcoin-Transaction-Graph-to-Predict-the-Greaves-Au/a0ce864663c100582805ffa88918910da89add47

[35] R. Delfin Vidal, "The Fractal Nature of Bitcoin: Evidence from Wavelet Power Spectra," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2547281, Dec. 2014. [Online]. Available: https://papers.ssrn.com/abstract=2547281

[36] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187775031100007X

[37] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural networks," in *1990 IJCNN International Joint Conference on Neural Networks*, Jun. 1990, pp. 1–6 vol.1.

[38] X. Zhu, H. Wang, L. Xu, and H. Li, "Predicting stock index increments by neural networks: The role of trading volume under different horizons," *Expert Systems with Applications*, vol. 34, no. 4, pp. 3043–3054, May 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417407002345

[39] White, "Economic prediction using neural networks: the case of IBM daily stock returns," in *IEEE 1988 International Conference on Neural Networks*, Jul. 1988, pp. 451–458 vol.2.

[40] Y. Yoon and G. Swales, "Predicting stock price performance: a neural network approach," in *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, vol. iv, Jan. 1991, pp. 156–162 vol.4.

[41] T. Koskela, M. Lehtokangas, J. Saarinen, and K. Kaski, "Time Series Prediction with Multilayer Perception, FIR and Elman Neural Networks," *undefined*, 1996. [Online]. Available: https://www.semanticscholar.org/paper/Time-Series-Prediction-with-Multilayer-Perception%2C-Koskela-Lehtokangas/6f506332adf9e8fcbba62fb17b1e819806b4029a

[42] C. L. Giles, S. Lawrence, and A. C. Tsoi, "Noisy Time Series Prediction using Recurrent Neural Networks and Grammatical Inference," *Machine Learning*, vol. 44, no. 1, pp. 161–183, Jul. 2001. [Online]. Available: https://doi.org/10.1023/A:1010884214864

[43] A. M. Rather, A. Agarwal, and V. N. Sastry, "Recurrent neural network and a hybrid model for prediction of stock returns," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3234–3241, Apr. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414007684

[44] F. A. Gers, D. Eck, and J. Schmidhuber, "Applying LSTM to Time Series Predictable Through Time-Window Approaches," in *Neural Nets WIRN Vietri-01*, R. Tagliaferri and M. Marinaro, Eds. London: Springer, 2002, pp. 193–200.

[45] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9713–9729, Jul. 2020. [Online]. Available: https://doi.org/10.1007/s00521-019-04504-2

[46] "Dogecoin price, DOGE chart, and market cap." [Online]. Available: https://www.coingecko.com/en/coins/dogecoin

[47] "Dogecoin Tweets 1/1/2021 - 4/30/2021." [Online]. Available: https://www.kaggle.com/tderonde/doge-tweets-1121-43021

[48] K. Team, "Keras documentation: TimeDistributed layer." [Online]. Available: https://keras.io/api/layers/recurrent_layers/time_distributed/