Information Processing and Management 58 (2021) 102708

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Global cryptocurrency trend prediction using social media

Poongodi M. [a], Tu N. Nguyen [b], Mounir Hamdi [a], Korhan Cengiz [c,*]

[a] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[b] Department of Computer Science, Kennesaw State University, Marietta, GA 30060, USA
[c] Department of Electrical-Electronics Engineering, Trakya University, Edirne 22030, Turkey

ARTICLE INFO

Keywords:
Cryptocurrency
Social media
Machine learning

ABSTRACT

This paper aims to investigate the global crypto-currency price movement trends with respect to the social media communication data. The idea is to analyze the topical trends in the online communities and social media platforms to understand and extract insights that could be used to predict the price fluctuations in crypto-currencies. The hypothesis rests in finding the empirical evidence to exploit the relationship between price variations and social media activities. Such models and insights will help us better understand the crypto currency ecosystems in context of social media behavior which can be used for real-time trading systems.

1. Introduction

Bitcoin is cryptographic currency that is utilized to make online installments and has turned out to be exceptionally prominent these days. Thus, it turned out to be imperative these days to see how the bitcoin cost changes and what sorts of variables impact its vacillation. The ascent of digital currencies has changed the method for monetary exchanges significantly. Other than the bitcoin, a few other digital forms of money have also appeared. The expansion of the bitcoin brought many clients into social media and online discussion forums to impart their insights and responses about this digital currency.

Individuals who have regular interests will make posts about certain points. In addition, as bitcoin is generally exchanged on the Internet, they settle on their choices about purchasing or offering the bitcoins for the most part dependent on the data that they acquire from the equivalent Web, as a rule from online discussions and social media. Furthermore, the relationship of such discussion dialogs and bitcoin value changes is not very much concentrated today. This paper intends to find such connections by making opinion and point examination of Bitcoin exchange posts from bitcointalk.org and develop a prediction model. To design the proposed model, we have downloaded information from bitcoin talk, cleaned the information and utilized machine learning to figure out how to make sense of connection between blog entries and bitcoin cost. We have scratched the information from the specified source (Bitcoin discussion), then cleaned that to keep just the helpful data by expelling incorrect words and additionally, HTML labels, stemmed the words. After that we developed a subject model with Latent Dirichlet designation (LDA) (Blei, Ng, & Jordan, 2003). At that point, we made a slant investigation of our information. Later, by having that data we dissected which sort of effect does that have on bitcoin value, by building a deep neural network with numerous layers.

Contributions

1. **Data Collection:** Data is accumulated from original observations. Information gathering plays a fundamental job in each task, as the viability and the tidiness of the data straightforwardly influences the results.

* Corresponding author.
E-mail address: korhancengiz@trakya.edu.tr (K. Cengiz).

https://doi.org/10.1016/j.ipm.2021.102708
Received 25 March 2021; Received in revised form 11 July 2021; Accepted 27 July 2021
Available online 16 August 2021

2. **Bitcoin Talk Forums:** Examining the diverse gatherings of online life and discovered bitcoin related information on Reddit, Twitter and Bitcoin Talk.
3. **LDA Topic Model:** Our main goal is to find relations between forum talks and bitcoin price changes. We also considered topic modeling in order to include topic distributions of the posts as an additional feature in our model to have better results.
4. **Construct Document Term Matrix:** We constructed a list for each of our original documents, which means we grouped all posts for each subject into one list (Ghahramani, Jordan, & Adams, 2010). In order to build an LDA model we need to get the frequencies of the terms within the documents. To do that, we constructed a document-term matrix, which will contain the frequencies of the words for each of the documents.
5. **Applying LDA Model:**The model has many functions that provide us with the needed information. For instance, we can review our topics with the print_topic and print_topics methods. Also we can get the topics for each of our documents. Hence, we can understand which topics were most popular in particular periods of time and what kind of impact does that have on the performance of the neural network predictions through this model.

Our Proposed paper notes the inefficiency of existing methods to predict price variations of cryptocurrencies. It proposes the use of simple data points such as user comments and replies in online cryptocurrency communities and forums to predict the price fluctuations. It focuses on the three cryptocurrencies to show that such a simple method is indeed useful in predicting fluctuations.

The Section 2 gives details related works followed by Section 3 discusses completely on our proposed research claim, Section 4, gives the evidential proof of our proposed model, Section 5, give the complete mathematical background of the design, Section 6, shows the evaluation of results, Section 7, discusses the analysis of the results attained and conclusively Section 8 overall work research framework and future directions

2. Related work

The research work by Phillips and Gorse (2018) aims to find the relationship between cryptocurrency price fluctuations and social media topic discussions to extract topics that indicate the future price variations. As cryptocurrencies undergo a new wave of price volatility, there has been a similar rise in the activity regarding bitcoin in the social media. Since our knowledge of factors that drive the price of cryptocurrency is still primitive, this paper intends to apply dynamic topic-modeling on social media communication data to gain insights into the temporal occurrence of various topics. To find the interactions between the topics and price movements, a Hawkes model is applied. This is shown to work successfully and hence; such know-how of topic relationships can be built into a real-time trade signaling system. Recent discovery (Xie, Chen, & Hu, 2019) of the predictive ability of social media data in stock markets has caused a flurry of activity in this domain. This paper investigates the social media discussions to further separate the insights and information from noise. The Bitcoin data shows that less cohesive social media discussion networks are better future price return predictors. This is verified using several trading simulations. The major contribution of the paper is to affirm that trading strategies based on sentiments, weighed by discussion network cohesion result in two to three times gain in returns.

Bitcoin has started (Mai, Shan, Bai, Wang, & Chiang, 2018) to affect financial markets and government policy decisions. This paper aims to determine the value of Bitcoin from a social media information perspective. It extensively studies the relation between social media and the monetary value of Bitcoin. Textual analysis of social media data and vector error correction models are the methods used by the authors to perform the experiments. The results of these experiments show that more bullish forum posts are associated with higher future values of bitcoin. One important insight is that social media effects are driven by the silent majority of users who are significantly less active (Poongodi, Hamdi, Sharma, Ma, & Singh, 2019; Poongodi, Vijayakumar, Al-Turjman, Hamdi, & Ma, 2019). Another important insight is that the messages on internet forums have a stronger impact than tweets. The paper concludes that social media sentiment is indeed a good predictor of bitcoin value while also asserting that not all social media forums generate equal impact. The paper aims (Garcia & Schweitzer, 2015) to improve the design of algorithmic traders by integrating various data-sources from the social media domain. Even though there is high availability of digital trace data from online social media communities, extracting actionable insights to inform trading strategies is still a challenge. For computational finance, the digital traces of human behavior which exists in high volume shows great promise. The approach is illustrated through the analysis of Bitcoin. The data sources blended into the design of the algorithmic trader include not only the canonical Bitcoin data but also various factors such as information search, word of mouth volume, emotional valence and opinion polarization from past 3 years' Twitter data. The high profitability of the proposed trading strategy is then verified with robust statistical methods. This work tries (Lamon, Nielsen, & Redondo, 2017) to research the effective ability of news and social media data to predict price fluctuations for three cryptocurrencies: Bitcoin, Litecoin and Ethereum. The methods used in the paper were the supervised learning algorithms for text-based sentiment classification. However, instead of labeling positive and negative sentiments, the labels were based on the actual price changes of a day. Hence, the model was able to directly predict price fluctuations. As a result, the model was able to correctly predict the days with largest fluctuations in price for Bitcoin and Ethereum.

The work focuses (Kim et al., 2017) on using text-mining methods on Bitcoin online forum data to analytically predict the price and transaction fluidity of the currency. The paper notes that although sentiment analysis of user communication data from Bitcoin online forums has been done before, significant attention has not been paid to the textual tokens from a natural language perspective. The major contribution of this paper is to extract key-words from noteworthy user comments from the forums to enable price prediction. The effectiveness of this method is then validated on the historical Bitcoin data.

This paper introduces (Li, Chamrajnagar, Fong, Rizik, & Fu, 2019) the idea of using data from social media platforms to predict price fluctuations in the highly speculative alternative cryptocurrencies. The experiments use Twitter data to model user

sentiments on an hourly basis for a period of 3.5 weeks. This is combined with the actual pricing data of the low-cap alternative cryptocurrency, ZClassic. An Extreme Gradient Boosting Regression Tree Model is trained on this ground truth. The resulting predictions are shown to have 0.81 correlation with the historical testing data of price fluctuations. The results are also shown to be statistically significant at a very low p-value as well. This paper examines (Xie, Wu, & Wu, 2017) the relationship between Bitcoin price variations and social media sentiment. It bases it off on the foundation that the data from social media platforms do provide value-relevant information. The experiments are run across different information channels and user groups. The findings suggest that while speculative information from social media can be used for both short-term and long-term predictions, fundamental related information is only good for long-term returns. The results also reiterate the point that the predictive power or accuracy is significantly higher for less active users especially for long-term predictions. The Bitcoin network (Somin, Gordon, & Altshuler, 2018) serves a trading ledger for exchanging tokens of value. However, this paper aims to model it as a social network in order to understand the ecosystem in tandem with the external and internal forces. The ERC20 protocol compliant crypto coins trading data is used for the study of network properties. The results in the paper demonstrate that the network indeed displays strong power-law properties which coincide with the current network theory expectations. Even with all the hype surrounding cryptocurrencies especially Bitcoin, (Li & Wang, 2017) there is limited theoretical understanding of the value and drivers of these currencies. This paper conducts a theory driven empirical study of Bitcoin exchange network to understand the technological and economic factors behind it. The results show that in a short-term period, the Bitcoin exchange rate does indeed correlate and adjust to variations in economic market conditions. However, the long-term exchange rate is more sensitive to basic economic fundamentals than technology. Significant impact of mining technology is also identified in Bitcoin price determination (Kim et al., 2016).

Our Proposed paper notes the inefficiency of existing methods to predict price variations of cryptocurrencies. It proposes the use of simple data points such as user comments and replies in online cryptocurrency communities and forums to predict the price fluctuations. It focuses on the three cryptocurrencies to show that such a simple method is indeed useful in predicting fluctuations.

3. Proposed work

3.1. Data collection

Data is an accumulated data that originates from original observations. Information gathering plays a fundamental job in each task, as the viability and the tidiness of the data straightforwardly influences the results. For this task, the data is primarily assembled from two assets; from the bitcoin talk forums and from the verifiable bitcoin value trade information.

3.2. Bitcoin talk forums

Examining the diverse gatherings of online life, and also for the most part discovered bitcoin related information on Reddit, Twitter and Bitcoin Talk. We attempted to assemble information from Reddit and do examination on, be that as it may Bitcoin talk gathering was more identified with our venture as Reddit contained more summed up posts.

In spite of the fact that there are numerous undertakings that did opinion investigation on Twitter information, yet we chose to utilize Bitcoin Talk information to enhance the general research and venture. We assembled information from https://bitcointalk.org/discussion, which comprises of subjects and every theme contains answers. Utilizing the library Scrapy in python (Doc.scrapy.org, 2019), we scratched the HTML information of each page from bitcoin talk gatherings since April 23, 2011 6:24:16 PM until May 05, 2018 01:15:00 AM, which took very nearly 5 h to finish. At that point we isolated every subject with its answers. We pre-processed the information by cleaning the message part and isolating the quoteheader for the message.

3.3. Topic models LDA

As our main goal is to find relations between forum talks and bitcoin price changes, we also considered topic modeling in order to include topic distributions of the posts as an additional feature in our model to have better results. Topic modeling is described as a way of finding a group of words (topics) from a set of documents, which can be a collection of sentences, that best shows the information of each of the documents. It can also be considered as a method of text mining — a way of obtaining recurring patterns of words in textual context (Garcia & Schweitzer, 2015). There are many techniques of such modelings, however, the widely used one is the Latent Dirichlet Allocation (LDA), which we used in our model to separate our posts into topics (Ghahramani et al., 2010). The LDA model discovers the different topics that are contained in a particular document and the portion of the topic that is present in a document. We took bitcoin price value per 5 min as labels of our model; hence as input we took the set of all the posts that were made during every 5 min. So the LDA model was constructed on the set of documents, which are represented as bags of posts per 5 min. We built a model with 15 topics. Then for each document we got the weights of the topics, and added those weights as additional features to our input data for the neural network.

3.4. Data preparation and transformation

Data preparation is very important in order to build a meaningful topic model, because documents may contain a lot of nonsensical words which will interfere generating useful topics. So the following transformations were performed to prepare the data.

- **Tokenizing:** converting a document to its atomic elements. In our case, we are interested in tokenizing to words (Blei et al., 2003).
- **Stopping:** removing meaningless words. Certain parts of English speech, like conjunctions (“for”, “or”) or the word “the” are meaningless to a topic model. These terms are called stop words and were removed from our token list (Blei et al., 2003).
- **Stemming:** bringing words, that have equivalent meaning, to the same term. For example, the words “stemming”, “stemmer”, “stemmed”, should be interpreted as the same; hence stemming reduces those words to “stem”. So this is very important, because otherwise our model would view those words as separate entities and will reduce the frequency of that model which will influence the decision of the topics. For this purpose the python implementation of Porter stemming algorithm was used to stem the words, which performs different operations to remove unnecessary parts from the words by having lists of English suffixes and prefixes in accordance with their usage rules (Willett, 2006).

3.5. Constructing a document-term matrix

Now at this stage, we have a tokenized, stopped and stemmed list of words. Then we constructed a list of lists, one list for each of our original documents, which means we grouped all posts for each subject into one list (Ghahramani et al., 2010). As was already mentioned, in order to build an LDA model we need to get the frequencies of the terms within the documents. To do that we constructed a document-term matrix, which will contain the frequencies of the words for each of the documents. First, we assigned a unique integer id to each unique token then collected word counts and relevant statistics. As a result we got an object called corpus, which represents a list of vectors equal to the number of documents. In each document vector is a sequence of tuples. For example one of the vectors can be: [(0, 2), (1, 1), (2, 2), (3, 2), (4, 1), (5, 1)], which represents one of the documents. The tuples are (term ID, term frequency) pairs.

3.6. Applying the LDA model

Now we have a document-term matrix and we generate an LDA model using gensim library (Ghahramani et al., 2010). Our LDA model is then stored. This model has a lot of functions that provide us with the needed information. For instance, we can review our topics with the print_topic and print_topics methods. Also we can get the topics for each of our documents. Hence by having this model, we can understand which topics were most popular in particular periods of time and which kind of impact does that have on the performance of the neural network predictions.

Based on these topic representations as given in Fig. 1 we can see that indeed there are meaningful topics in the documents. We can notice that the discussions of the forum are about important topics which indeed can influence the price fluctuations. We can even label these topics. For instance, if we look at topic number 2, based on the represented words, we can conclude that this topic is mainly about stock investments in the market. As another example, we can take topic number 5. In this case we can even make a complete meaningful sentence with the provided words. An example of such sentence can be: ”It is good time to invest money and buy coins to make a profit”. Another interesting topic is the 10th. We can see that this topic is about future price rise. To sum up the results of LDA topic modeling, we can say that, the topics of the forum discussions are indeed related to bitcoin. Hence, it is logical to make a hypothesis that the discussions around the relevant topics may play a significant role in bitcoin price fluctuations. Therefore as we became confident about the propriety of the topics, we decided to include the weights of the 15 topics for each document as an auxiliary feature to our neural network

4. Implementation

As our primary objective is to discover relations between discussion talks and bitcoin value transforms, we too considered theme demonstrating with the end goal to incorporate subject dispersions of the posts as an extra highlight in our model to have better outcomes as shown in Fig. 1. Topic modeling is depicted as a method for finding a gathering of words (points) from an arrangement of reports, which can be an accumulation of sentences, that best demonstrates the data of every one of the archives. It can likewise be considered as a technique for content mining a method for acquiring repeating examples of words in literary setting as shown in Fig. 2. There are numerous methods of such modeling, be that as it may, the generally utilized one is the Latent Dirichlet Distribution (LDA), which we utilized in our model to isolate our posts into themes (Ghahramani et al., 2010). The LDA demonstrate finds the diverse topics that are contained in a specific record and the part of the subject that is available in an archive as shown in Fig. 3. We took bitcoin value esteem per 5 min as marks of our model; thus, as information we took the arrangement of the considerable number of posts that were made amid at regular intervals.

So, the LDA show was developed on the arrangement of archives, which are spoken to as packs of posts per 5 min. We constructed a model with 15 topics. At that point for each report we got the weights of the subjects, and included those weights as extra highlights to our info information for the neural system.

```
(0, '0.044*"country" + 0.033*"ban" + 0.024*"china" + 0.022*"government" + 0.020*"news" + 0.018*"economy" + 0.01
4*"economic" + 0.014*"world" + 0.010*"crisis" + 0.010*"state"')
----------
(1, '0.098*"gold" + 0.037*"fork" + 0.022*"silver" + 0.020*"hard" + 0.014*"network" + 0.014*"mate" + 0.014*"ligh
tning" + 0.012*"proper" + 0.012*"skill" + 0.011*"segwit"')
----------
(2, '0.162*"market" + 0.043*"stock" + 0.037*"crypto" + 0.018*"cap" + 0.018*"cryptocurrency" + 0.015*"trade" + 0
.014*"investor" + 0.014*"asset" + 0.013*"billion" + 0.012*"value"')
----------
(3, '0.020*"know" + 0.019*"people" + 0.019*"say" + 0.017*"like" + 0.016*"one" + 0.014*"go" + 0.013*"don" + 0.01
2*"thing" + 0.011*"even" + 0.011*"really"')
----------
(4, '0.219*"usd" + 0.088*"position" + 0.034*"bch" + 0.029*"co" + 0.019*"positively" + 0.014*"ratio" + 0.013*"cl
ose" + 0.011*"toâ" + 0.010*"pessimistic" + 0.010*"trade"')
----------
(5, '0.026*"invest" + 0.026*"money" + 0.023*"good" + 0.020*"buy" + 0.019*"make" + 0.017*"investment" + 0.017*"c
oin" + 0.017*"time" + 0.015*"profit" + 0.015*"people"')
----------
(6, '0.030*"business" + 0.026*"work" + 0.015*"blockchain" + 0.015*"job" + 0.013*"family" + 0.012*"wallet" + 0.0
11*"success" + 0.009*"key" + 0.009*"help" + 0.009*"social"')
----------
(7, '0.047*"mine" + 0.038*"supply" + 0.032*"transaction" + 0.029*"demand" + 0.026*"fee" + 0.023*"exchange" + 0.
020*"coin" + 0.017*"cost" + 0.015*"rise" + 0.013*"high"')
----------
(8, '0.024*"entry" + 0.022*"assurance" + 0.021*"image" + 0.015*"theâ" + 0.013*"click" + 0.013*"bitcoin" + 0.011
*"nope" + 0.010*"consistent" + 0.010*"ofâ" + 0.009*"lift"')
----------
(9, '0.031*"price" + 0.029*"market" + 0.020*"go" + 0.014*"see" + 0.012*"time" + 0.010*"back" + 0.010*"sell" + 0
.009*"buy" + 0.008*"panic" + 0.008*"news"')
----------
(10, '0.081*"price" + 0.033*"year" + 0.032*"rise" + 0.019*"go" + 0.018*"time" + 0.015*"happen" + 0.015*"high" +
0.014*"reach" + 0.013*"see" + 0.011*"value"')
----------
(11, '0.073*"campaign" + 0.025*"signature" + 0.022*"management" + 0.016*"discipline" + 0.016*"join" + 0.012*"pr
imary" + 0.012*"divide" + 0.011*"vote" + 0.011*"register" + 0.010*"withdrawal"')
----------
(12, '0.096*"ico" + 0.050*"project" + 0.034*"icos" + 0.024*"scam" + 0.024*"facebook" + 0.012*"twitter" + 0.012*
"watch" + 0.010*"game" + 0.010*"youre" + 0.010*"kyc"')
----------
(13, '0.034*"currency" + 0.021*"people" + 0.021*"crypto" + 0.021*"bank" + 0.018*"money" + 0.014*"country" + 0.0
14*"world" + 0.013*"government" + 0.010*"make" + 0.009*"cryptocurrency"')
----------
(14, '0.025*"index" + 0.024*"org" + 0.024*"topic" + 0.023*"budget" + 0.023*"bitcointalk" + 0.020*"emotion" + 0.
017*"charity" + 0.014*"content" + 0.014*"merit" + 0.012*"pray"')
----------
```

Fig. 1. LDA Topics.

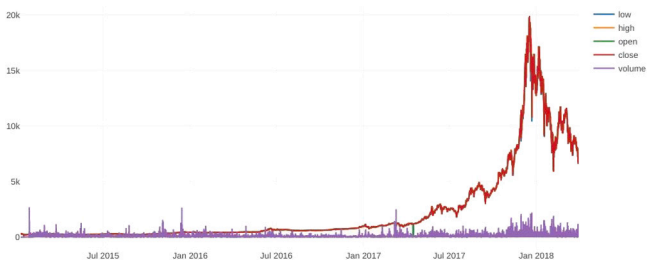

Fig. 2. The plot of bitcoin price historical data.

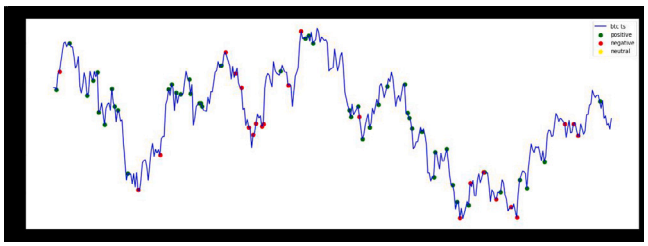

Fig. 3. Sentiment analysis and price change relation.

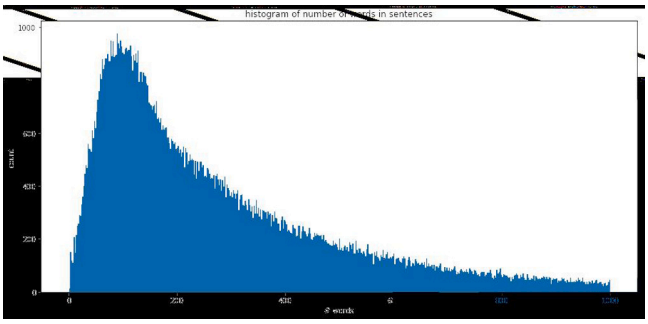

Fig. 4. Histogram of number of words in a sentence.

5. System architecture

Next, we began to develop a neural system that predicts bitcoin cost dependent on our information, which is an arrangement of client posts with some assistant highlights. We constructed Keras neural system (Lee, Chen, & Lee, 2012). As the system required the contributions to be vectors, we chose to utilize StarSpace word inserting in request to outline to a vector in a space that has high measurements. These word embeddings give semantic and syntactic data of the words. For example, the words, which are comparable, are near one another in this space and the words which are divergent, are far separated (Linton, Teo, Bommes, Chen, & Härdle, 2017). Before preparing the model, we marked the information. We attempted diverse strategies for naming. First, we figured the marks as the logarithm of the proportion of the normal cost of bitcoin in next 5 min to the normal cost of past 5 min. Be that as it may, in the wake of preparing our model we got awful outcomes with Eq. (1), our model was not getting the hang of anything and the misfortune was not diminishing. At that point we accepting the name as the standardized estimations of bitcoin cost for every 5-minute information and showed signs of improvement results. Also, we attempted distinctive methods for normalizing and the outcomes for every technique are spoken to in Section 7. The Histogram of number of words in the sentences is shown in Fig. 4. The Bitcoin price peaks is clearly illustrated in Fig. 5

$$s_t = \varphi(W_{s_{t-1}} + U_{x1} + b) \tag{1}$$

Where

- $\varphi - ActivationFunction$
- $s_t \in R^n - CurrentState$
- $s_{t-1} \in R^n - PriorState$
- $x_t \in R^n - CurrentInput$
- $n - State$
- $m - inputsize$

The principle contribution to the model was the gathering of the posts for like clockwork, spoke to as a succession of words which we got from StarSpace in a vector shape. Yet, to zest things up, we too added assistant contributions to the model, for example, extremity and subjectivity, whose connection to the information was 3%. At that point we incorporated the volume of exchanges inside every 5 min as another highlight, whose connection was 12%. Lastly, we included the 15 weights of the themes of our LDA demonstrate as new assistant highlights. By having such measure of highlights, we analyzed their diverse mixes in the neural system and spoken to all acquired outcomes in Section 7. To join the fundamental information (the posts) with the helper highlights we previously passed our vectors to the LSTM layer. LSTM changed the arrangement of the vectors into a solitary vector, containing data about the whole arrangement. At that point we sustained into the model our assistant info information by linking it with the LSTM yield. Next, we stacked three profound thickly associated systems with various measures of neurons lastly connected the principle calculated relapse layer with sigmoid enactment as our names run from 0 to 1.

6. Model evaluation

Our last final data contains 351666 gathering answers, which we changed into 5-minute sacks, which made a sum of 141981, and partitioned haphazardly to three sets: train, test and approval with the particular proportions: 80%, 10%, 10%. To quantify the execution of our model we utilized the estimations of misfortune work estimated in mean squared blunder, and in addition the R squared score which measures how the information indicates are close the fitted qualities. Neural Network Architecture with Long Short-Term Memory is shown in Fig. 6. Price label histograms Price label min max vs price label uniform min max is shown in Fig. 7.

Accepting that the best naive model without preparing it on the data is to foresee the mean estimation of the underlying name, we can compare at our model to the base model to measure its performance. From the outcomes we can presume that the model

Fig. 5. Bitcoin Price Peaks.

Table 1
Model results 1.

	Model 1	Model 2	Model 3	Model 4
Word padding	1500	750	750	150
Price label	Min Max	Min Max	Min Max	Min Max
LDA topics	False	False	True	False
Polarity	True	True	True	True
Subjectivity	True	True	True	True
Volume	True	True	True	True
epochs	8	8	8	15
RMSE train set	0.044	0.044	0.041	0.028
RMSE test set	0.045	0.046	0.042	0.029
RMSE validation set	0.043	0.0436	0.040	0.030
R squared train set	0.163	0.161	0.224	0.470
R squared test set	0.172	0.167	0.225	0.446
R squared validation set	0.165	0.165	0.225	0.444
# hours to train on CPU	2.8	0.33	1	0.33

Table 2
Model results 2.

Model 5	Model 6	Model 7	Model 8	Model 9	
Word padding	1500	750	750	750	750
Price label	Uniform min max	Uniform min max	Uniform min max	Uniform min max	Uniform min max
LDA topics	False	False	True	False	False
Polarity	True	True	True	True	False
Subjectivity	True	True	True	True	False
Volume	True	True	True	False	True
epochs	8	8	8	8	8
Loss train set	0.072	0.722	0.065	0.085	0.072
Loss validation set	0.072	0.071	0.653	0.086	0.072
RMSE train set	0.072	0.071	0.065	0.086	0.072
RMSE test set	0.072	0.072	0.066	0.087	0.073
RMSE validation set	0.072	0.071	0.065	0.086	0.072
R squared train set	0.208	0.212	0.283	0.056	0.204
R squared test set	0.214	0.218	0.282	0.056	0.211
R squared validation set	0.211	0.214	0.283	0.049	0.206
# hours to train on CPU	1.25	0.83	0.66	0.53	1.2

works as given in Tables 1 and 2 entirely well considering that the information was less and was taken just from one gathering. We can see that the most elevated R squared score was acquired in the event of having just 150 words in each report, nonetheless, this is not a decent expectation in light of the fact that a lot of posts were excluded for this situation. Next, we can see that changing that esteem from 750 to 1500 does not have much effect, on the grounds that there are not all that numerous records with a length bigger than 750. Min max cost naming gives better R squared score, and less misfortune, in any case, the dispersion of the cost is correct skewed, which implies we have more opportunity to be close to the genuine incentive by foreseeing the incentive close to one side pinnacle. Albeit uniform min max gives more misfortune and less R squared, however measurably it is more exact to do the test on. Utilizing uniform min max marking, with 750 we have R squared score about 0.22. Notwithstanding, when we included likewise LDA themes' weights as an element the model worked better and we got R squared score 0.28. By this outcome we can state that our speculation made in Section 4 was valid; in fact, LDA point weights enhanced the R squared score. Consequently, we can state that theme dissemination of the posts impact value changes. What is more, we ought to likewise see that if there should be an occurrence of including LDA themes the uniform standardization gave a higher score than the other one. In addition, the consequences of including the assistant factors of extremity, subjectivity and volume additionally increment the health of the

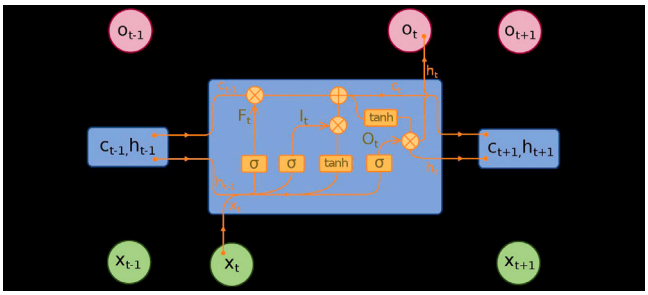

Fig. 6. Neural Network Architecture with Long Short-Term Memory.

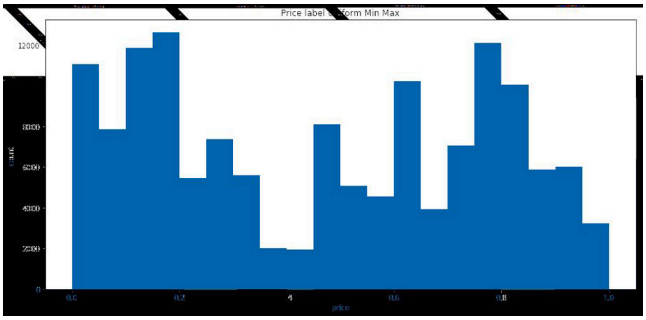

Fig. 7. Price label histograms Price label min max vs price label uniform min max.

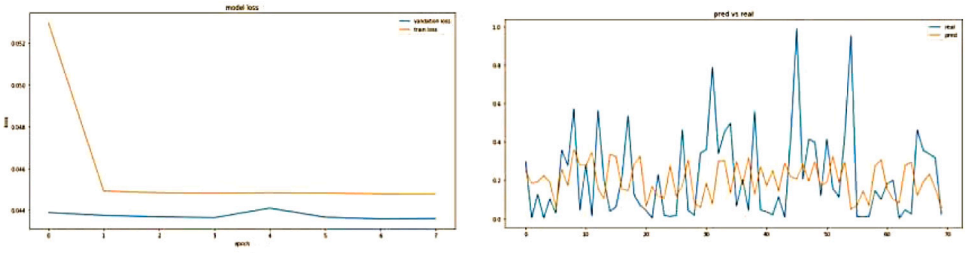

Fig. 8. Model 1.

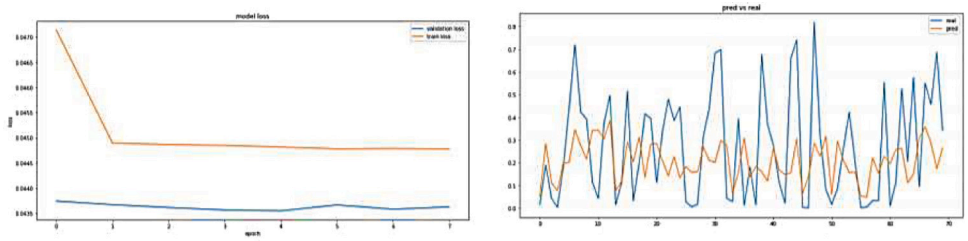

Fig. 9. Model 2.

model. All in all, from the plots of the misfortunes of our model we can see that notwithstanding not terrible exactness's of significant worth expectations, the model predicts the snapshots of increments and declines of the cost extremely well. Furthermore, for further upgrades we will attempt to test our model on bigger dataset from various gatherings and in addition online networking.

7. Results and analysis

This investigation is done to attempt diverse word paddings if there should be an occurrence of the non-uniform cost marking (min max), and one investigation utilizing LDA topics

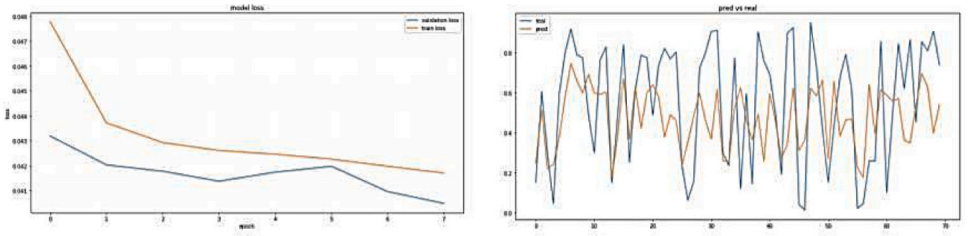

Fig. 10. Model 3.

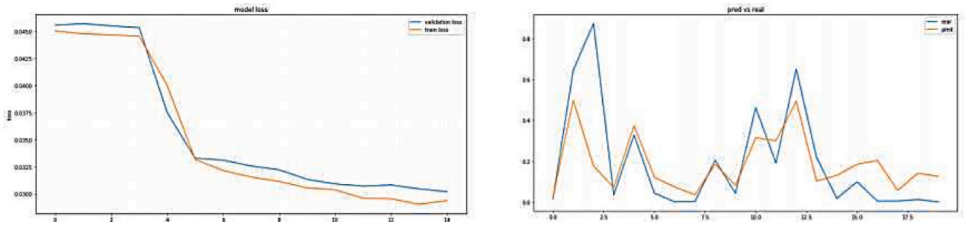

Fig. 11. Model 4.

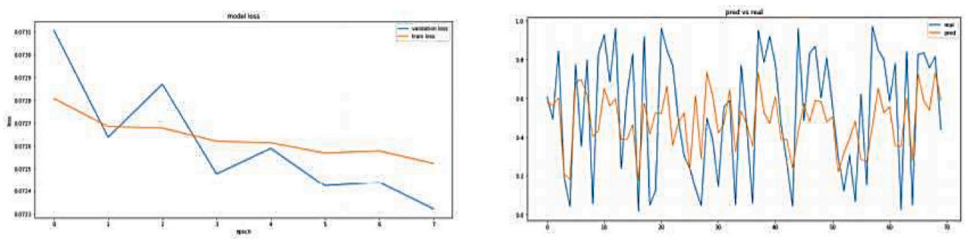

Fig. 12. Model 5.

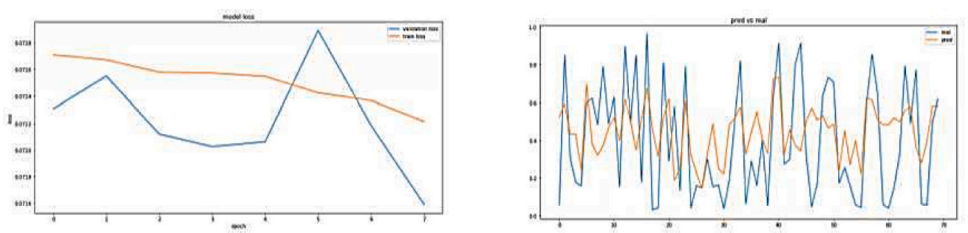

Fig. 13. Model 6.

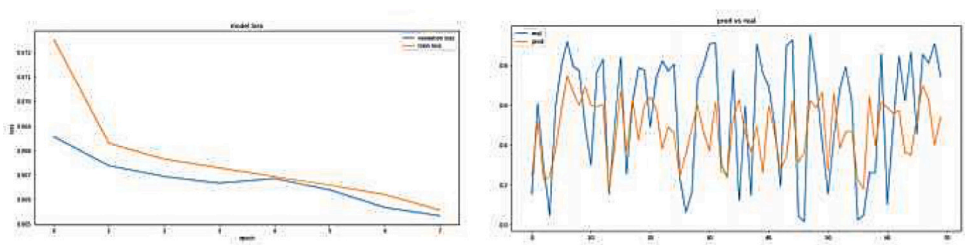

Fig. 14. Model 7.

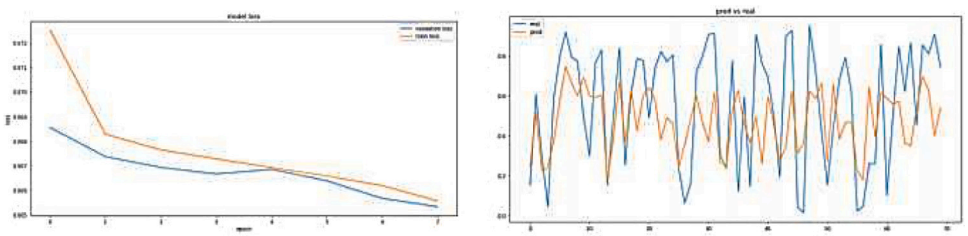

Fig. 15. Model 8.

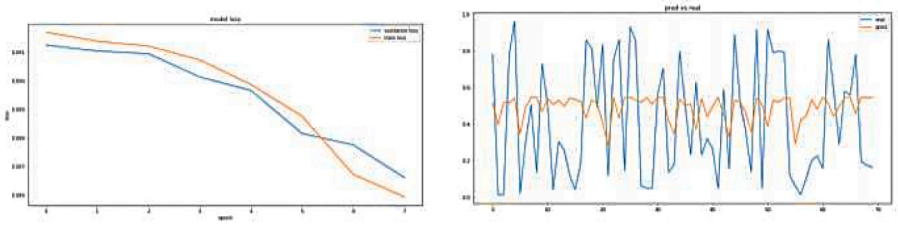

Fig. 16. Model 9.

This experiment is done to attempt distinctive blends of assistant factors if there should be an occurrence of the uniform cost marking (uniform min max).

From the outcomes generated in Figs. 8 to 16 we can presume that the model works entirely well considering that the information was less and was taken just from one gathering. We can see that the most elevated R squared score was acquired if there should be an occurrence of having just 150 words in each report, be that as it may, this is not a decent forecast in light of the fact that a lot of posts were precluded for this situation. Next, we can see that changing that esteem from 750 to 1500 does not have much effect, on the grounds that there are not all that numerous reports with a length bigger than 750. Min max cost marking gives better R squared score, and less misfortune, be that as it may, the dispersion of the cost is correct skewed, which implies we have more opportunity to be close to the genuine incentive by anticipating the incentive close to one side pinnacle. Albeit uniform min max gives more misfortune and less R squared, yet measurably it is more precise to do the examination on. Utilizing uniform min max naming, with 750 we have R squared score about 0.22. In any case, when we included additionally LDA points' weights as an element the model worked better and we got R squared score 0.28. By this outcome we can state that our speculation made in Section 4 was valid; without a doubt LDA point weights enhanced the R squared score. Consequently, we can state that point circulation of the posts impact value variances. What is more, we ought to likewise see that in the event of including LDA subjects the uniform standardization gave a higher score than the other one. Additionally, the after effects of including the helper factors of extremity, subjectivity and volume additionally increment the well being of the model.

8. Conclusion and future work

Taking everything into account, from the loss plots of our model we can see that not withstanding not terrible correctness of significant worth expectations, the model predicts the snapshots of increments and downturn of the cost exceptionally well. In conclusion, we can confirm with sufficient evidence that global cryptocurrency trend prediction is possible given social media data particularly about the cryptocurrency at hand. The outlook of using social media data to predict global cryptocurrency prices seems quite positive. However, this research was primarily focused on Bitcoin with its corresponding forum. In the future, we can include all the other popular cryptocurrencies like Ethereum and Litecoin. We might also use a larger dataset in the form of tweets and reddit posts. Furthermore, for future upgrades we will endeavor to test our model on bigger dataset from various forums and social media. All this further the reach of our idea that social media trends is in fact a strong indicator of cryptocurrency trends.

CRediT authorship contribution statement

Poongodi M.: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing - review & editing. **Tu N. Nguyen:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing - review & editing. **Mounir Hamdi:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Korhan Cengiz:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Doc. scrapy. org (2019). Scrapy 1.6 documentation — Scrapy 1.6.0 documentation. (online) Available at: https://doc.scrapy.org/en/latest/.

Garcia, David, & Schweitzer, Frank (2015). Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science*, *2*(9), Article 150288.

Ghahramani, Z., Jordan, M. I., & Adams, R. P. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in neural information processing systems* (pp. 19–27).

Kim, Young Bin, Kim, Jun Gi, Kim, Wook, Im, Jae Ho, Kim, Tae Hyeong, Kang, Shin Jin, et al. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLoS One*, *11*(8), Article e0161197.

Kim, Young Bin, Lee, Jurim, Park, Nuri, Choo, Jaegul, Kim, Jong-Hyun, & Kim, Chang Hun (2017). When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PLoS One*, *12*(5), Article e0177630.

Lamon, Connor, Nielsen, Eric, & Redondo, Eric (2017). Cryptocurrency price prediction using news and social media sentiment. *SMU Data Science Review*, *1*(3), 1–22.

Lee, H. Y., Chen, Y. N., & Lee, L. S. (2012). Utterance-level latent topic transition modeling for spoken documents and its application in automatic summarization. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 5065–5068). IEEE.

Li, Tianyu Ray, Chamrajnagar, Anup, Fong, Xander, Rizik, Nicholas, & Fu, Feng (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics*, *7*, 98.

Li, Xin, & Wang, Chong Alex (2017). The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. *Decision Support Systems*, *95*, 49–60.

Linton, M., Teo, E. G. S., Bommes, E., Chen, C. Y., & Härdle, W. K. (2017). Dynamic topic modelling for cryptocurrency community forums. In *Applied quantitative finance* (pp. 355–372). Berlin, Heidelberg: Springer;

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 9931022.

Mai, Feng, Shan, Zhe, Bai, Qing, Wang, Xin, & Chiang, Roger H. L. (2018). How does social media impact Bitcoin value? A test of the silent majority hypothesis. *Journal of Management Information Systems*, *35*(1), 19–52.

Phillips, Ross C., & Gorse, Denise (2018). Mutual-excitation of cryptocurrency market returns and social media topics. In *Proceedings of the 4th international conference on frontiers of educational technologies* (pp. 80–86). ACM.

Poongodi, M., Hamdi, M., Sharma, A., Ma, M., & Singh, P. K. (2019). DDoS detection mechanism using trust-based evaluation system in VANET. *IEEE Access*, *7*, Article 183532-183544.

Poongodi, M., Vijayakumar, V., Al-Turjman, F., Hamdi, M., & Ma, M. (2019). Intrusion prevention system for DDoS attack on VANET with reCAPTCHA controller using information based metrics. *IEEE Access*, *7*, Article 158481-158491.

Somin, Shahar, Gordon, Goren, & Altshuler, Yaniv (2018). Social signals in the ethereum trading network. arXiv preprint arXiv:1805.12097.

Willett, P. (2006). The porter stemming algorithm: then and now. *Program*, *40*(3), 219–223.

Xie, Peng, Chen, Hailiang, & Hu, Yu Jeffrey (2019). Network cohesion and predictive power of social media in the Bitcoin market. In *Georgia Tech Scheller College of Business Research Paper 17-5*.

Xie, Peng, Wu, Jiming, & Wu, Chongqi (2017). Social data predictive power comparison across information channels and user groups: evidence from the Bitcoin market. *The Journal of Business Inquiry*, *17*(1), 41–54.