International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Cryptocurrency Analysis using Social Media Sentiments

Nishchal Nandagopal , Atla Sai Abhinav , Karthik Sai BS , Kumar Ketan

Computer Science and Engineering, BMS College of Engineering,Bengaluru, Karnataka-560019

Abstract

The rapid emergence and widespread adoption of digital media, particularly social media platforms, have sparked a profound revolution in technology. This transformation has not only revolutionized the way we communicate and share information but has also opened up new avenues for understanding human emotions, sentiments, and opinions. In this dynamic landscape, the present study aims to delve deeper into the intricate fluctuations observed in Bitcoin prices within a specific timeframe and subsequently harness the power of machine learning to predict future values. To achieve this ambitious objective, the study capitalizes on the vast trove of user comments available on Reddit, one of the most influential social media platforms of our time.

The study adopts a cutting-edge deep learning framework that leverages the prowess of a Bidirectional Recurrent Neural Network (RNN) combined with the robustness of Long Short-Term Memory (LSTM) units. Furthermore, to enhance the model's capacity to comprehend and represent the complex semantic structures inherent in textual data, word2Vec embeddings are employed. This powerful combination of advanced neural network architectures and sophisticated word embeddings sets the stage for accurate and insightful predictions.

To ensure the reliability and efficacy of the predictions, the textual data obtained from Reddit undergoes a meticulous refinement process, carefully removing noise and extracting the most relevant and meaningful information. This refined data is then harnessed to train the model, enabling it to make accurate forecasts of Bitcoin prices across various time intervals. The model's predictive capabilities extend to time horizons of +1, +2, +6, +12, and +24 hours, providing valuable insights into the short-term price dynamics of Bitcoin.

The true power of the model lies in its ability to directly forecast the direction of price change, allowing it to uncover hidden patterns and underlying trends in the price movement. By deciphering the intricate relationships between user comments and Bitcoin price fluctuations, the model gains the expertise to make informed predictions that capture the essence of market dynamics.

In summary, the advent of digital media, coupled with advancements in deep learning techniques, has paved the way for groundbreaking research into the realm of cryptocurrency prediction. By harnessing the wealth of user comments on social media platforms and employing advanced neural network architectures, this study exemplifies the potential to unravel the complex dynamics of Bitcoin prices and offer valuable insights into future market trends.

1. Introduction

Social media has become an integral part of people's lives, playing a significant role in various aspects such as business, politics, communication, and socialization. However, it also has its drawbacks, including cyberbullying, addiction, and the propagation of unrealistic expectations. This research focuses on understanding the impact of social media on the cryptocurrency market, particularly in relation to Bitcoin. Cryptocurrency refers to a digital medium of exchange that utilizes blockchain technology and encryption techniques to secure transactions. The value of cryptocurrencies is subject to change based on trends, interests, and investments, with social media sentiment potentially playing a crucial role in these fluctuations.

Similar to stock markets, cryptocurrency markets are influenced by factors such as supply and demand, mining difficulty, utility, and market news. Supply and demand directly impact cryptocurrency prices, with scarcity driving prices up and abundance causing prices to drop. Mining difficulty, which relates to the computational power required for mining, also affects the value of cryptocurrencies. Perception of utility plays a role as well, as cryptocurrencies without practical use may be viewed as having little value, leading to low market prices. Market news, including sentiment derived from reviews, can heavily influence the value of cryptocurrencies, with positive sentiment driving prices up and negative sentiment causing them to decrease.

Bitcoin, the first cryptocurrency, was created by an anonymous group known as "Satoshi Nakamoto" in August 2008. It operates on a decentralized peer-to-peer transaction system, allowing for uninterrupted transactions without government interference. Bitcoin transactions are recorded in a blockchain, which is a publicly available database that maintains a history of individual Bitcoin transactions. This database ensures data redundancy, security, and transaction verification. Bitcoin trends and discussions are widely observed on social media platforms, with Reddit being one of the major platforms for Bitcoin-related discussions.

Reddit is a social media platform that facilitates news aggregation, discussions, and content rating. It provides a personalized newspaper-like experience based on users' likes and interests, allowing for interactive discussions through the posting of links and voting by other users. The focus of this research is to analyze user opinions and reviews on Bitcoin by scraping data from the Bitcoin subreddit.

Sentiment analysis plays a crucial role in this research, and it is achieved through Natural Language Processing (NLP). NLP is an interdisciplinary field that combines computer science and linguistics to process and analyze unstructured text data, which accounts for a significant portion of the world's data. NLP enables computers to learn, understand, and make predictions based on human language inputs. It encompasses various tasks such as grammar induction, lemmatization, segmentation, parts of speech tagging, parsing, stemming, and terminology extraction. Initially, NLP tasks were carried out through hand-coding rules, but with advancements in machine learning techniques, NLP has evolved significantly.

The present study examines the predictive relationship between public opinions on social media and Bitcoin returns by considering data from different social media platforms, particularly Reddit. Textual data from news headlines and messages scraped from Reddit are used as inputs for the cryptocurrency prediction model. The model classifies user sentiments based on pre-trained word embeddings using the Word2vec method. The primary objective is to predict market trend fluctuations on an hourly basis. The research addresses various research ques- tions related to this objective. // //

RQ1. Can textual data accurately find a trend in bitcoin exchange rate? To find the answer to the question this research makes use of Bi-directional RNN with LSTM units on comments extracted from reddit application. // // RQ2. Does the amount of user opinion in the form of textual data help in discovering bit coin prices over short period of time? Research is conducted on predicting bit coin prices at intervals of +1, +2, +6, +12 and +24 hours to check if the compactness of comments influences the results. // // Social media has become an integral part of people's lives, playing a significant role in various aspects such as business, politics, communication, and socialization. However, it also has its drawbacks, including cyberbullying, addiction, and the propagation of unrealistic expectations. This research focuses on understanding the impact of social media on the cryptocurrency market, particularly in relation to Bitcoin. Cryptocurrency refers to a digital medium of exchange that utilizes blockchain technology and encryption techniques to secure transactions. The value of cryptocurrencies is subject to change based on trends, interests, and investments, with social media sentiment potentially playing a crucial role in these fluctuations.

Similar to stock markets, cryptocurrency markets are influenced by factors such as supply and demand, mining difficulty, utility, and market news. Supply and demand directly impact cryptocurrency prices, with scarcity driving prices up and abundance causing prices to drop. Mining difficulty, which relates to the computational power required for mining, also affects the value of cryptocurrencies. Perception of utility plays a role as well, as cryptocurrencies without practical use may be viewed as having little value, leading to low market prices. Market news, including sentiment derived from reviews, can heavily influence the value of cryptocurrencies, with positive sentiment driving prices up and negative sentiment causing them to decrease.

Bitcoin, the first cryptocurrency, was created by an anonymous group known as "Satoshi Nakamoto" in August 2008. It operates on a decentralized peer-to-peer transaction system, allowing for uninterrupted transactions without government interference. Bitcoin transactions are recorded in a blockchain, which is a publicly available database that maintains a history of individual Bitcoin transactions. This database ensures data redundancy, security, and transaction verification. Bitcoin trends and discussions are widely observed on social media platforms, with Reddit being one of the major platforms for Bitcoin-related discussions.

Reddit is a social media platform that facilitates news aggregation, discussions, and content rating. It provides a personal-

ized newspaper-like experience based on users' likes and interests, allowing for interactive discussions through the posting of links and voting by other users. The focus of this research is to analyze user opinions and reviews on Bitcoin by scraping data from the Bitcoin subreddit.

Sentiment analysis plays a crucial role in this research, and it is achieved through Natural Language Processing (NLP). NLP is an interdisciplinary field that combines computer science and linguistics to process and analyze unstructured text data, which accounts for a significant portion of the world's data. NLP enables computers to learn, understand, and make predictions based on human language inputs. It encompasses various tasks such as grammar induction, lemmatization, segmentation, parts of speech tagging, parsing, stemming, and terminology extraction. Initially, NLP tasks were carried out through hand-coding rules, but with advancements in machine learning techniques, NLP has evolved significantly.

The present study examines the predictive relationship between public opinions on social media and Bitcoin returns by considering data from different social media platforms, particularly Reddit. Textual data from news headlines and messages scraped from Reddit are used as inputs for the cryptocurrency prediction model. The model classifies user sentiments based on pre-trained word embeddings using the Word2vec method. The primary objective is to predict market trend fluctuations on an hourly basis. The research addresses various research questions related to this objective.

	Rank	Market Cap. ($B)	No. of obs	Mean	Std.	Skewness	Kurtosis	Min.	Max.	L-B($l_{|y|}$)	L-B(l_{y^2})	Normality test
Bitcoin	1	67.7602	1225	0.0009	0.0362	-1.0460	11.9590	-0.2543	0.1830	439.1810 (< .0001)	211.8933 (< .0001)	4320.1650 (< .0001)
Ethereum	2	28.4994	732	0.0054	0.0742	-0.0981	7.4296	-0.3959	0.3293	244.3408 (< .0001)	122.0188 (< .0001)	599.6219 (< .0001)
Ripples	3	6.7727	1225	0.0003	0.0751	-2.0852	40.6985	-0.8844	0.6393	460.2791 (< .0001)	89.1193 (< .0001)	73426.8413 (< .0001)
NEM	4	2.3565	853	0.0046	0.0831	0.5260	7.0738	-0.3196	0.4279	227.0109 (< .0001)	135.8583 (< .0001)	629.1816 (< .0001)
Dash	5	1.5033	1219	0.0020	0.0724	0.0827	11.7466	-0.4881	0.5232	902.6657 (< .0001)	586.7064 (< .0001)	3887.1329 (< .0001)

Summary statistics of the global weighted average indices for each relevant Cryptocurrency. P values of the relevant columns are reported in parentheses

2. Literature Review

The current research aims to predict the change in the cryptocurrency price prediction by classifying the user comments and news related to cryptocurrency. Literature review is important for any research to study about the new conclusions, facts, building the model with the help of existing knowledge and it articulates the result helps to understand the problem within and view the research problem from various perspective. There are many machine learning model algorithms which are enforced in predicting the cryptocurrency value(The Law Library of Congress, 2018).

2.1. Study on Cryptocurrency

Cryptocurrencies are digital currencies that rely on cryptographic techniques and blockchain technology to regulate transactions and ensure security. The cryptocurrency market is characterized by high volatility, with significant fluctuations in value over time. Understanding the dynamics of this market is essential for comprehending how design choices impact its behavior. In 2017, the total value of cryptocurrencies experienced a substantial increase of nearly 600 billion, representing a staggering growth rate of approximately 3,300. Bitcoin, being the largest cryptocurrency, exhibited remarkable gains of around 1,364 compared to other cryptocurrencies. Its price surged from 0.01in2010toover20,000 per token. However, bitcoin has also experienced periodic declines, highlighting its volatile nature. This volatility makes cryptocurrencies, including bitcoin, a popular and debated topic of discussion on social media platforms.

In a study by Phillip, Chan, and Peiris (2018), the researchers explored the diverse nature of cryptocurrencies. They identified the advantage effect, which stems from the asymmetric relationship between returns and volatility, influenced by factors such as financial leverage and debt-to-equity ratios. To analyze the varying time volatility in Bitcoin data, the researchers employed the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model. The dataset used in their study consisted of 2,796 cryptocurrency time-series indices from the Brave New Coin (BNC) Digital Currency in- dices, with 224 cryptocurrencies being traded at least once per day. The study revealed that the five largest cryptocurrencies exhibited significant variability, particularly in terms of market capitalization. In their study, Krafft, Della Penna, and Pentland (2018) investigate the susceptibility of traders in the cryptocurrency market to peer influence on trading behavior. They examine three primary dependent variables: time, condition, and two

 DOI: 10.55041/IJSREM24175

Time	Condition	Dependent Var.	*n* Control	*n* Treat	Control Mean	Mean Effect	*t*-stat	*p*-value
15 Min.	Buy	Buy Prob.	25483	25602	0.279	0.019	4.79	2.96e-05
15 Min.	Buy	% Buy Vol.	24321	24313	0.290	0.017	4.40	1.97e-04
15 Min.	Buy	Trade Prob.	52050	51314	0.490	0.009	3.00	4.81e-02
15 Min.	Sell	Sell Prob.	25483	25987	0.721	-0.006	-1.44	1.00e+00
15 Min.	Sell	% Sell Vol.	24321	24660	0.710	-0.005	-1.31	1.00e+00
15 Min.	Sell	Trade Prob.	52050	51727	0.490	0.013	4.12	6.71e-04
30 Min.	Buy	Buy Prob.	23647	23871	0.278	0.003	0.83	1.00e+00
30 Min.	Buy	% Buy Vol.	23583	23802	0.291	0.005	1.33	1.00e+00
30 Min.	Buy	Trade Prob.	52049	51312	0.454	0.011	3.51	7.98e-03
30 Min.	Sell	Sell Prob.	23647	23809	0.722	0.001	0.15	1.00e+00
30 Min.	Sell	% Sell Vol.	23583	23735	0.709	0.002	0.58	1.00e+00
30 Min.	Sell	Trade Prob.	52049	51724	0.454	0.006	1.94	9.53e-01
60 Min.	Buy	Buy Prob.	31065	31118	0.274	0.003	0.76	1.00e+00
60 Min.	Buy	% Buy Vol.	30984	31044	0.289	0.001	0.33	1.00e+00
60 Min.	Buy	Trade Prob.	52030	51288	0.597	0.010	3.18	2.70e-02
60 Min.	Sell	Sell Prob.	31065	31351	0.726	0.000	0.14	1.00e+00
60 Min.	Sell	% Sell Vol.	30984	31275	0.711	-0.003	-1.00	1.00e+00
60 Min.	Sell	Trade Prob.	52030	51713	0.597	0.009	3.02	4.50e-02

Probability and percentage statistics

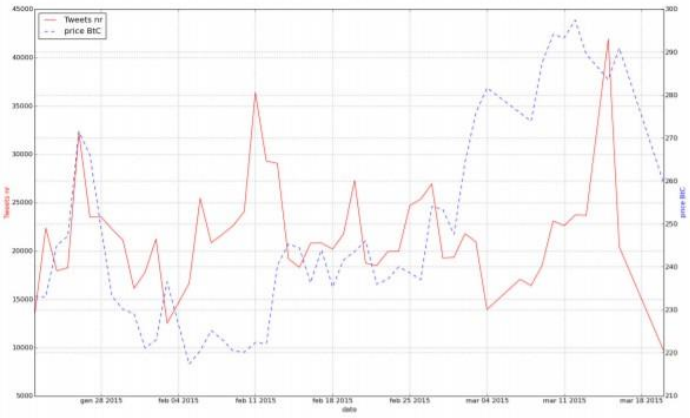
statistics to analyze the impact of peer influence on trade direction. The first statistic considers the type of transaction used in the last transaction and the associated time. This statistic aggregates whether the last transaction involved buying or selling. The second statistic calculates the fraction of the trading volume related to buying or selling. These statistics are computed in three different time intervals to assess their influence on trade direction.

2.2. Study on Machine Learning Models

In their research, Matta, Lunesu, and Marchesi (2015a) take a different approach to automatically analyze people's opinions, sentiments, and attitudes towards Bitcoin. Their main objective is to identify Bitcoin price behavior and its variations based on sentiments extracted from tweets. The study collects data from January 2015 to March 2015, comprising approximately 1,924,891 tweets. Additionally, Google Trends data is used to analyze Bitcoin's popularity. The researchers utilize the Twitter Streaming API to access real-time tweets, which are then stored in a MySQL Database. The tweets are classified as positive (score of 1), negative (score of -1), or neutral (score of 0) using the SentiStrength tool. A Java module is employed to automatically retrieve new tweets by sending requests to the Twitter Streaming API.

Through cross-correlation analysis, the researchers find that tweets related to Bitcoin price exhibit a correlation coefficient of 0.15 with a one-day lag. They also observe a significant relationship between Google Trend data and Bitcoin's price. The study concludes that these findings can be beneficial for investment professionals in the Bitcoin market. Future work is proposed to gather data over a six-month period to improve the quality of the results. The figure below illustrates the similarity between Bitcoin price and the number of tweets.

In a different approach to predicting Bitcoin price, Lamon, Nielsen, and Redondo (no date) categorized daily news and social media data based on each coin and examined their correlation with the subsequent price movements. They utilized two datasets: one from Kaggle and another obtained by scraping news from cryptocoinsnews.com. Logistic regression, Linear support vector machine, and Bernoulli Naive Bayes classifiers were employed to predict the prices of Bitcoin, Litecoin, and Ethereum. The evaluation metrics and hyperparameter tuning were also considered, revealing that Logistic regression performed best for Bitcoin and Litecoin, while Bernoulli Naive Bayes showed superior performance for Ethereum.



Similarity between Bitcoin price and Number of Tweets

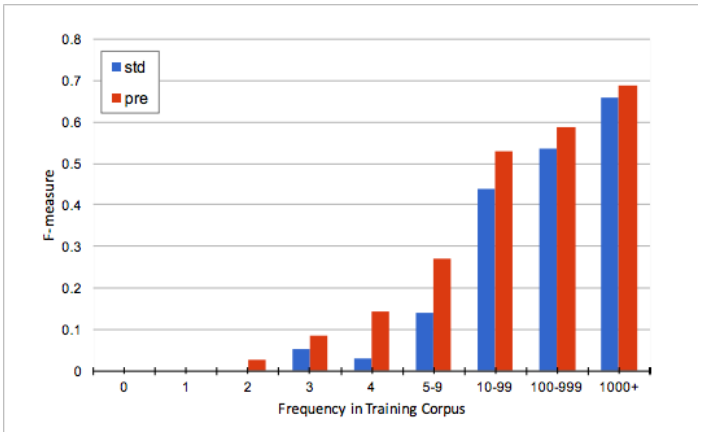
In another study, Madan, Saluja, and Zhao (2015) aimed to identify trends in the Bitcoin market that could help predict Bitcoin prices. They utilized two datasets: one from Blockchain.info containing 26 features about the Bitcoin market over a five-year period, and another obtained through scrapping data from Coinbase API and OKcoin API. Feature selection techniques were applied to identify the most significant 16 features for solving the problem. Three time series datasets with intervals of 30, 60, and 120 minutes were created to capture future market fluctuations. Binomial GLM and random forest algorithms were employed on each time series dataset to predict Bitcoin price changes, with random forest yielding higher accuracy compared to binomial GLM.

Xie, Chen, and Hu (2017) explored the impact of social media discussions on predicting Bitcoin price movements. They obtained their dataset from the CoinDesk website and social media data from bitcointalk.org. To analyze the discussions related to the Bitcoin market, a network perspective was adopted. Latent Dirichlet Allocation (LDA) method was used to identify latent topics among the messages posted on social media platforms. The study found a negative relationship between social media discussions and the prediction of Bitcoin price movements.

Another perspective on price prediction involves analyzing trends on Google and their correlation with Bitcoin trading volumes. Matta, Lunesu, and Marchesi (2015b) investigated the relationship between Google Trends data and Bitcoin trading volumes. They obtained data from Google Trends website and blockchain.info. Correlation analysis, including Pearson Correlation and Cross Correlation, was performed, and a value of 0.60 was obtained for the correlation. The authors also employed the Granger Causality method to assess whether Google Trends could predict Bitcoin trading volumes.

Abraham, Higdon, and Nelson (2018) focused on building a model to predict Bitcoin and Ethereum buying and selling decisions. Their dataset was collected from Twitter API and Google Trends. VADAR (Valence Aware Dictionary and Sentiment Reasoner) was used for data cleaning, classification, and measuring polarity and intensity. The study found a strong correlation (0.817) between Twitter data and Google Trends data

 DOI: 10.55041/IJSREM24175



Standard Embedding v/s Pre-trained Embedding

using the Pearson correlation method. Linear modeling techniques were initially applied, and then multiple linear regression was used to address non-linearity and high variance. The authors concluded that search volume index was highly correlated with cryptocurrency prices during both price increases and decreases.

Fung et al. (2019) introduced new Bitcoin trading strategies based on related news and tweets. They utilized Valence Aware Dictionary for Sentiment Reasoning (VADER) to convert qualitative data into quantitative sentiment scores. Logit regression was applied to news and tweets sentiment, yielding accuracies of 62.2 and 52.11 respectively. Time series modeling was performed on price data against tweet volumes, achieving an accuracy of 61.11.

2.3. Study on Word Embedding Models

All machine learning algorithms accepts data to be in vector format. In Natural Language Processing (NLP), word embedding methods help to convert the text data into vector format. It is distributed representation of text in n-dimensional space. The pre-trained word embedding methods perform the better accuracy when compared to the standard embeddings as shown in the below figure. According to previous research in the field of NLP, the role of word embedding methods is to obtain the information retrieval, feature extraction, Analysing the sentiments and summarizing the text(Graves, A.-r. Mohamed, 2015).

(Wang et al., 2018) used window approach to identify the tasks such as: part-of-speech tags, chunks, named entity tags, semantic roles, semantically similar words. Since the long- range relations between the words, they used sentence-based approach to capture linguistic similarities into word vectors. They trained entire network together using weight sharing. The used all tasks uses labelled data expect for the language model which learnt from unlabelled data using semi-supervised learn- ing algorithms.

(Collobert and Weston, 2008)come up with a new idea Dict2vec using a huge dataset for the word similarity task for the text classification. They used Dict2vec for classifying strong and weak pairs of words to provide both novel positive sampling objective and novel negative controlled objective. The

	50M							200M							Full						
		w2v		FT		our			w2v		FT		our			w2v		FT		our	
	oov	A	B	A	B	A	B	oov	A	B	A	B	A	B	oov	A	B	A	B	A	B
MC-30	0%	697	847	722	823	840	**859**	0%	742	830	795	814	**854**	827	0%	809	826	831	815	**860**	847
MEN-TR-3k	0%	692	753	697	**767**	733	762	0%	734	758	754	**772**	752	768	0%	733	728	752	751	**756**	755
MTurk-287	0%	657	**688**	657	685	665	682	0%	642	**671**	671	661	667	666	0%	660	656	**672**	671	661	660
MTurk-771	0%	596	677	597	692	685	**713**	0%	628	669	632	675	682	**704**	0%	623	620	631	638	**696**	694
RG-65	0%	714	865	671	842	824	**875**	0%	771	842	755	829	857	**877**	0%	787	802	817	820	**875**	867
RW	36%	375	420	442	**512**	475	489	16%	377	408	475	**507**	467	478	2%	407	427	464	468	**482**	476
SimVerb	3%	165	371	179	374	363	**432**	0%	183	306	206	329	377	**424**	0%	186	214	222	233	**384**	379
WS353-ALL	0%	660	739	657	739	738	**753**	0%	694	734	701	735	**762**	758	0%	705	721	729	723	756	**758**
WS353-REL	0%	619	**700**	623	696	679	688	0%	665	706	644	685	**710**	699	0%	664	681	687	686	702	**703**
WS353-SIM	0%	714	**797**	714	790	774	784	0%	743	**792**	758	792	784	787	0%	757	767	775	779	781	**781**
YP-130	3%	458	679	415	674	666	**696**	0%	449	592	509	639	616	**665**	0%	502	475	533	553	**646**	607
W.Average		453	562	467	582	564	**599**		471	533	503	563	569	**592**		476	488	508	512	**573**	570
AG-News		**874**	871	868	871	871	866		**886**	882	880	881	880	880		885	885	**887**	887	881	884
DBPedia		936	942	942	944	**944**	944		952	956	957	958	**960**	959		966	966	967	967	968	**969**
Yelp Pol.		808	835	821	**842**	832	834		837	855	852	859	856	**859**		865	867	872	874	**876**	875
Yelp Full		451	469	460	**473**	471	472		477	491	488	495	499	**501**		506	506	512	514	516	**518**

Spearman's rank correlation coefficient

below table shows that Dtct2vec model outperforms state-of-the-art approaches.

Predicting the value of fluctuating currencies in the real world is a challenging task. Kim et al. (2015) addressed this challenge by using user opinions from virtual gaming environments to study sentiments and emotions. The researchers collected data based on topics, content, number of views, and posting time. They employed the Stanford Parser and Stanford Log-linear Part-Of-Speech Tagger to extract parts of speech from user opinions. The pretrained Glove word embedding method was then used to represent the words in a 300-dimensional space. Eight primary emotions, including joy, trust, fear, surprise, sadness, disgust, anger, and anticipation, were compared against each word to determine the number of blanks between the target and emotive word. The primary emotion of the opinion data was defined as the emotion with the smallest value, and if the emotion value decreased with the decrease in the number of blanks between the target and emotive word, it indicated a strong association. If there were zero blanks between all emotive words, the opinion data was classified as neutral. Granger causality test and Support Vector Machine were employed for training and evaluation.

Twitter data related to cryptocurrencies can be utilized to develop crypto trading strategies. Colianni, Rosales, and Signorotti (2015) explored supervised learning algorithms to identify crypto market movements using Twitter API. Real-time data, including user ID, unique identifier, timestamp, and tweets, were collected. The data cleaning process involved removing non-alphabetic characters and duplicates, resulting in a processed dataset with 350,000 rows. Feature vectors were used to extract unique words and create a lexicon to store binary values for each word. Text-processing.com API was used to classify tweets into positive, negative, and neutral categories on a word-by-word basis. Naive Bayes classifier and Support Vector Machine algorithms were employed. Naive Bayes classifier performed better overall with an accuracy of 95 and an hour-to-hour accuracy of 76.23, while logistic regression achieved an overall accuracy of 86 and an hour-to-hour accuracy of 98.58. Logistic regression outperformed the Naive Bayes classifier.

Ahn and Kim (2019) focused on the psycholinguistic approach to understanding investor sentiment during the high volatility of the Bitcoin market in 2017 and 2018. They utilized data from the bitcointalk.org website over a one-year period. Three dictionaries were used: one consisting of predefined

 DOI: 10.55041/IJSREM24175

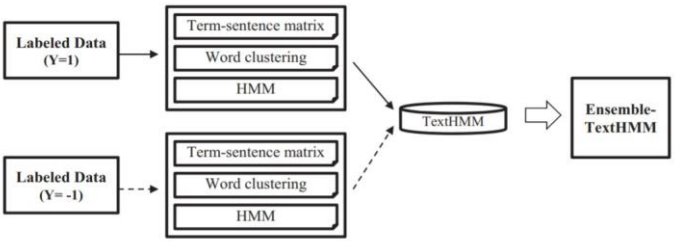
words unrelated to finance, another containing words with detailed inferences of cryptocurrency, and a third dictionary called semantic orientation from point-wise mutual information (SO-PMI) to identify specific words related to cryptocurrency. A polarity score of 0.2 was considered positive, while scores below -0.1 were considered negative. Two sentiment-related variables, attention and disagreement, were classified, achieving an accuracy of 81.6

2.4. Study on Neural Networks

In their study, Jang and Lee (2017) examined the impact of Bayesian Neural Networks on Bitcoin price analysis. They utilized a dataset from September 5, 2011, to August 2017, obtained from bitcoincharts.com, as well as Blockchain information from blockchain.info. The researchers employed ordinary least squares (OLS) to determine the long-term influence of variables such as the Dow Jones index, euro-dollar exchange rate, WTI oil price, and other factors on Bitcoin price. Bayesian Neural Networks were then used to investigate the nonlinear features of input variables, Blockchain information, and macroeconomic factors in Bitcoin price formation. They trained the Bayesian Neural Network using relevant features and evaluated its performance through non-linear methodologies, such as Support Vector Regression (SVR) and linear regression, in terms of training and testing errors. Additionally, they developed a prediction model for near-future Bitcoin prices using a rollover framework, which proved to be faster and more straightforward than sequential neural networks like LSTM and RNN. Through time series analysis, they compared the BNN model against linear regression and SVR, considering 25 explanatory variables and two response variables: the logarithmic price of Bitcoin and Bitcoin price volatility. The study concluded that the BNN model performed relatively accurately compared to linear regression and SVR, and future work would explore other extended machine learning algorithms and incorporate new input capabilities related to Bitcoin.

McNally, Roche, and Caton (2018) utilized the CRISP-DM methodology to predict Bitcoin prices. They collected data from bitcoinpriceindex.com spanning from August 19, 2013, to July 19, 2016. The ensemble method Boruta, similar to a random forest classifier, was employed to evaluate which features should be included. Deep learning models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) were utilized for Bitcoin price prediction. Hyperparameter tuning techniques like stochastic gradient descent (SGD) and the RMSprop optimizer were applied to enhance the models. LSTM outperformed RNN, achieving an accuracy of 52.78

Kim et al. (2017) aimed to predict variations in Bitcoin price by extracting keywords from user comments. The data used consisted of comments and replies from December 1, 2013, to September 21, 2016. The researchers employed topic modeling techniques, such as lexicon building, to extract keywords from the corpus. They expanded the lexicon via word recommendation and employed kernel density estimation (KDE) for smoothing during document analysis. Granger causality tests were conducted between discussions of Bitcoin transaction count and



Overview of the proposed method, an ensemble of HMMs by text clustering (EnsembleTextHMM)

price, with lags ranging from 1 to 12 days. The results indi- cated a significant causal relationship between China and Bit- coin price. Deep learning architecture, specifically the Convolutional Neural Network (CNN) algorithm, was employed for price prediction. The researchers experimented with different hidden layer configurations and found that a two-layer neural network achieved the highest accuracy at 81.37

Sin and Wang (2018) investigated the relationship between Bitcoin features and price changes using Ensemble Artificial Neural Networks. They collected time series data from two sources: blockchain.info and bitcoinity.org. The ensemble ANN model consisted of five multi-layer perceptron (MLP) models with the same specifications but different numbers of nodes in each layer for training purposes. The accuracy of each MLP model ranged from 58 to 63. The input layer of each MLP had 190 nodes, followed by two hidden layers with varying numbers of nodes and a single output layer. The trained data was then tested using backtesting methods, resulting in an accuracy of 64.Kang, Ahn, and Lee (2018) proposed a unique approach to sentiment classification using an ensemble text-hidden Markov model (HMM) clustering technique.

Initially they used supervised learning method to classify the positive and negative sentences by training labelled sentence data. And then, they built hidden Markov model using latent semantic analysis by clustering words and use transition probability between them. Finally, they found 90

2.5. Conclusion

The comprehensive literature review conducted in various fields related to this research provides valuable insights and knowledge. It serves as a foundation for better understanding the concepts and requirements, thus contributing positively to the research. The examination of multiple papers on cryptocurrencies has allowed for the identification of prevailing trends and the factors influencing market dynamics. Additionally, the research papers on neural networks have provided a deep understanding of the functioning of these algorithms and the different types available, enabling the selection of the most suitable algorithm for the study. The chosen architecture for the current project emphasizes the significance of combining Bidirectional Recurrent Neural Networks (BRNN) with Gated Recurrent Unit (GRU). Through the literature review, a roadmap for the project has been established to achieve enhanced performance, encompassing essential considerations prior to implementing the model.

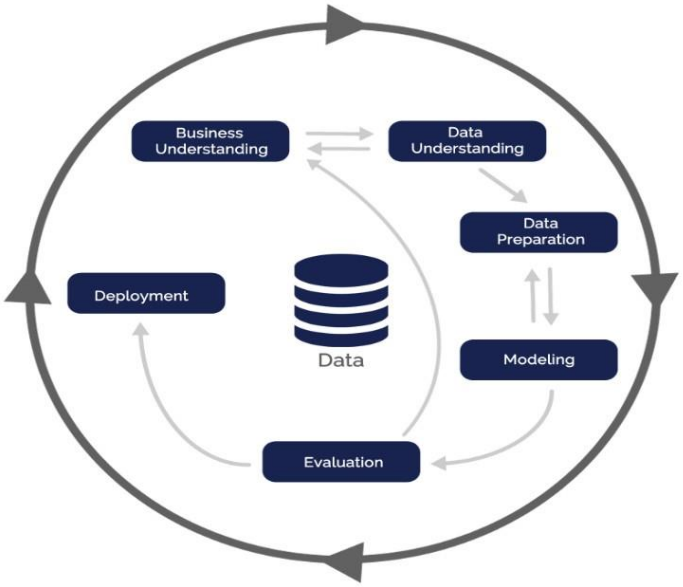


Figure 1: Flow of CRISP-DM

3. Methodology

Data analysis involves analyzing data to uncover concealed patterns and make predictions about future trends. It requires specific skills and knowledge to forecast an organization's future. Data mining requires a standardized approach to address business problems through data mining tasks and recommend suitable data transformation and mining techniques. Scholars have identified various techniques used in text mining, including Knowledge Discovery in Database (KDD) and Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM aims to make large data mining projects more cost-effective, reliable, manageable, and efficient (Wirth, 2000). Therefore, this research employs the CRISP-DM method to gain insights into text mining.

3.1. Structure of research methodology

The CRISP-DM (Cross Industry Standard Process for Data Mining) is an established data mining framework utilized by companies to guide the development of data mining projects. It consists of six stages that outline the necessary steps to undertake during the construction of a data mining project (Ncr et al., 2004). These stages include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, as illustrated in Figure 7. Each stage is accompanied by analogies and visual aids to enhance comprehension. The following subsections provide a summary of each phase.

3.1.1. Business Understanding

The initial phase of CRISP-DM assists in comprehending the business objectives, transforming information into knowledge, and utilizing that knowledge to formulate a plan for achieving those objectives. In order to proceed, it is important to acquire knowledge about cryptocurrencies: their functioning, customer base, and the factors that impact market fluctuations. Additionally, it is crucial to explore methods for predicting the factors that contribute to sudden changes in the cryptocurrency market.

3.1.2. Data Understanding

In the second phase of CRISP-DM, data collection is undertaken by gathering information from diverse sources relevant to the research. This stage also entails data exploration, where various graphs are plotted to gain insights into data distribution and attribute relationships. Additionally, data quality is assessed to determine if it meets the requirements for the research. For the current research proposal, publicly available data from Reddit is being utilized, comprising time, messages, and open market price information.

3.1.3. Data Preparation

Data preparation is a crucial and labor-intensive task in the data mining process. In this phase, the prepared data is utilized for modeling and subsequent analysis. The process begins with the selection of relevant data for the research requirements. Next, data pre-processing is conducted, which involves removing noise from the data by retaining only the attributes necessary for model building. Additionally, data integration and merging are performed to create a cohesive dataset ready for modeling. This study incorporates the utilization of various word embeddings such as word2Vec, Glove, and Concept net Number batch.

3.1.4. Modelling

The first step in this phase entails selecting suitable modeling techniques to be employed. Subsequently, a procedural mechanism is generated to evaluate the effectiveness and robustness of the model. The data is then split into training and testing sets for further analysis. Finally, various machine learning algorithms are applied. The present study focuses on the implementation of a Bidirectional Recurrent Neural Network as part of the modeling process.

3.1.5. Evaluation

The evaluation phase is one of the critical stages in the data mining process. It involves assessing the models built in the previous phase to determine if they meet the research objectives defined at the beginning of the CRISP-DM process. This phase entails comparing the results obtained from different algorithms and analyzing them using various techniques. If any issues or discrepancies are identified, they need to be addressed, and the process may need to be rerun. In the current study, K-fold validation techniques are being employed to evaluate the models.

3.1.6. Deployment

The final phase of CRISP-DM places emphasis on the customer perspective rather than the technical side. Prior to deployment, it is crucial to ensure that the model functions as specified by the customer. This phase also involves ongoing monitoring and maintenance to ensure the model's continued performance and effectiveness.

 DOI: 10.55041/IJSREM24175

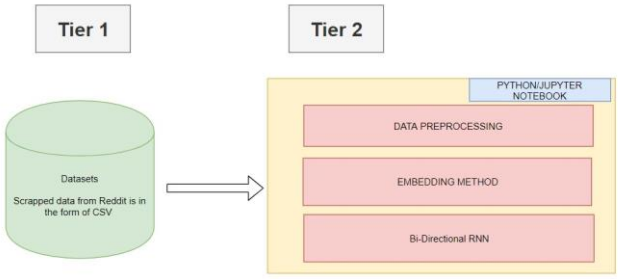


Figure 2: Two tier architecture for the proposed study

	date	subreddit	score	num_comments	title	1hr_change	2hr_change	6hr_change	12hr_change	24hr_change
0	2017-01-01 00:04:08	btc	15	2	Flight by 'ChinaPop' into bitcoin in 2017 emer...	-0.261723	-0.242146	-0.105102	0.772805	2.804769
1	2017-01-01 00:24:52	bitcoin	33	20	We all (should) by now know that SMS 2FA is ba...	-0.261723	-0.242146	-0.105102	0.772805	2.804769
2	2017-01-01 01:52:51	bitcoin	37	2	"Being able to transfer wealth with no interme...	0.019629	0.015497	0.123973	1.243866	3.517744
3	2017-01-01 02:04:56	blockchain	15	0	Fundamental concepts of Blockchain, explained ...	-0.004132	-0.126015	0.231371	1.223997	4.285537
4	2017-01-01 02:20:34	ethereum	41	15	How is Casper going to be proven working?	-0.004132	-0.126015	0.231371	1.223997	4.285537

Figure 3: Snapshot of the dataset

4. Design Specification

The current proposed study adopts a two-tier architecture. The first tier comprises a dataset obtained by scraping data from the Reddit website, which is stored in a CSV file. The data in this tier is in its raw format and is subsequently processed in the second tier of the architecture. The second tier consists of the Python IDE, Jupyter Notebook. The analysis is conducted in the client layer using different statistical measures to assess the model's accuracy and performance. The study incorporates various libraries, such as Pandas for data manipulation, Keras (which utilizes TensorFlow) for deep learning tasks, and Word2vec for natural language processing tasks.

5. Dataset

The datasets utilized in the current proposed study are obtained by scraping data from the Reddit website. To extract the data, the research employs the Python built-in package called "Beautifulsoup" and stores the data in the required CSV for- mat for further use in the model. The dataset comprises infor- mation such as the time at which users expressed their opin- ions about the cryptocurrency market, the title of the post, the number of comments, and the number of likes. However, the scraped data contains noise and missing values, making it un- suitable for direct model building. Therefore, the data must un- dergo pre-processing to remove redundant attributes. The table provided below illustrates the data used in the current research, including the date, subreddit, score, number of comments, title, and the hourly change in cryptocurrency price. The data also includes the 1st hour, 2nd hour, 6th hour, 12th hour, and 24th hour changes in cryptocurrency value.

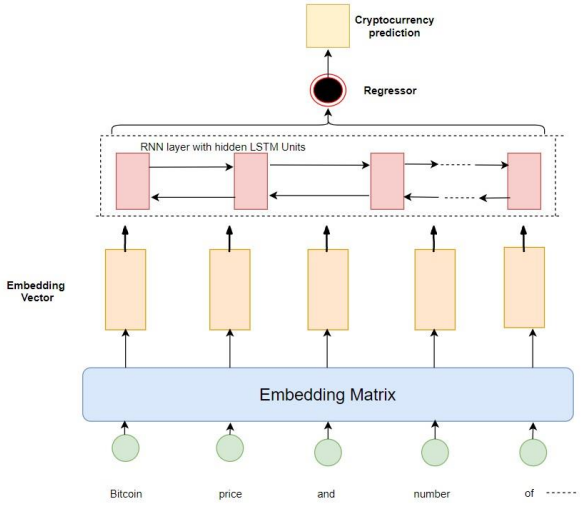


Figure 4: Snapshot of the dataset

6. Implementation

The research endeavors to construct a classification model to categorize messages as positive or negative. Additionally, the study aims to develop a real-time Bitcoin prediction model. Given that the study involves data in CSV format, a suitable approach is to utilize data training and testing. The initial step in the data pre-processing phase involves extensive work focused on removing noise from the data. By employing regular expressions, HTML links and tags have been successfully eliminated. Since the study incorporates sequential data as input, a Bidirectional Recurrent Neural Network is employed. For data processing and estimation, a machine learning model is utilized, which requires nominal data in vector format. The study encompasses both nominal and textual data, and to convert the textual data into a vector format, pre-trained Glove embeddings with a 300-dimensional vector space are employed.

The figure illustrates the conversion of textual data into a vector format for predicting cryptocurrency values. The text data sequence is inputted into an embedding matrix that contains pre-trained embedding methods, which effectively convert words into vector representations. The Word2vec method is utilized for this process, mapping input words into a 300-dimensional vector space. Since the data consists of long sequences, it becomes challenging to keep all the data in memory for an extended period. To address this, LSTM units are employed in conjunction with RNN units to avoid gradient-related issues. The Bidirectional RNN, combined with LSTM units, effectively captures context and semantics. The approach is organized into five components: 1. Data pre-processing, 2. Word Embedding, 3. Classification model, 4. Estimation, and 5. Prediction.

6.1. Data preprocessing

The implementation process begins with transforming the raw data into a format that is understandable by the model. Raw data cannot be directly sent to the model as it consists

 DOI: 10.55041/IJSREM24175

of real-world data written in human language, which may contain noise and missing values. The machine cannot process this data directly due to its inability to comprehend it, leading to errors. The study incorporates several steps to prepare the data for processing. These steps include tokenization, which involves converting sentences into words; removing punctuation tags; eliminating stop words; dropping unnecessary characters; and finally, grouping together words with similar meanings through lemmatization. By performing these steps, the data is refined and made suitable for processing by the model. The study encompasses data pre-processing, which involves preparing the data for analysis. The data is divided into training and testing sets, following a 70:30 ratio. During pre-processing, a tokenizer function is employed to convert words into tokens. This step facilitates the transformation of the data into a machine-readable format and provides additional information for subsequent development and analysis.

6.2. Word Embedding methods

Word embedding is a technique that represents text by assigning vectors of real numbers to words, capturing their semantic similarities despite differences in form. In the current research analysis, the GloVe 300d method, which is an unsupervised learning algorithm, is employed. This method plays a crucial role as it obtains vector representations for each word based on aggregate word-word co-occurrence statistics. GloVe utilizes matrix factorization to create word embeddings and is used after the data has been pre-processed and formatted appropriately. The resulting embeddings are then converted into numerical format to serve as inputs for machine learning models. However, it is important to note that GloVe 300d requires a significant amount of data to produce accurate results. One limitation of GloVe is that it is primarily a count-based model and may not be well-suited for prediction tasks.

On the other hand, Word2vec is another word embedding method that is more suitable for the current research, as it is a predictive model. It employs a two-layer neural network to generate word embeddings based on a given corpus. The objective function of Word2vec ensures that words with similar meanings have similar embeddings. This method takes a large corpus of words as input and produces vector representations for the words. Words with similar meanings within the corpus are located close to each other in the vector space. Word2vec offers two models: Continuous Bag of Words (CBOW) and Skip-gram.

The figure above shows the Continuous Bag of words (CBOW). CBOW predicts the target words based on the surrounding context words. Statistically, it has the effect that CBOW smooths over a lot of the distributional information and is suitable for smaller datasets.

The figure above shows the architecture of Skip-gram. This works as an inverse to CBOW. It predicts the surrounding context words from the target words. Skip-gram treats each context-target pair as a new observation, and this tends to do better when we have larger data. Word2vec is a simple neu- ral network with a single hidden layer and it takes large input vector and compress it down to a smaller dense vector. The

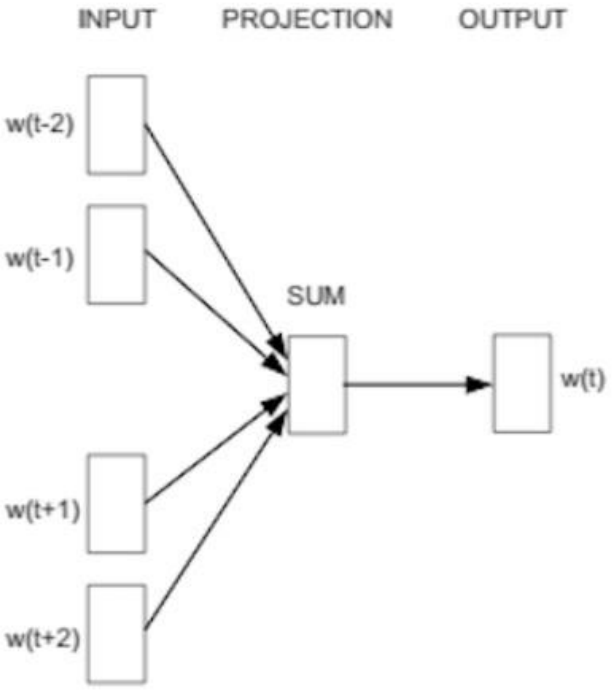


Figure 5: Continuous Bag of Words

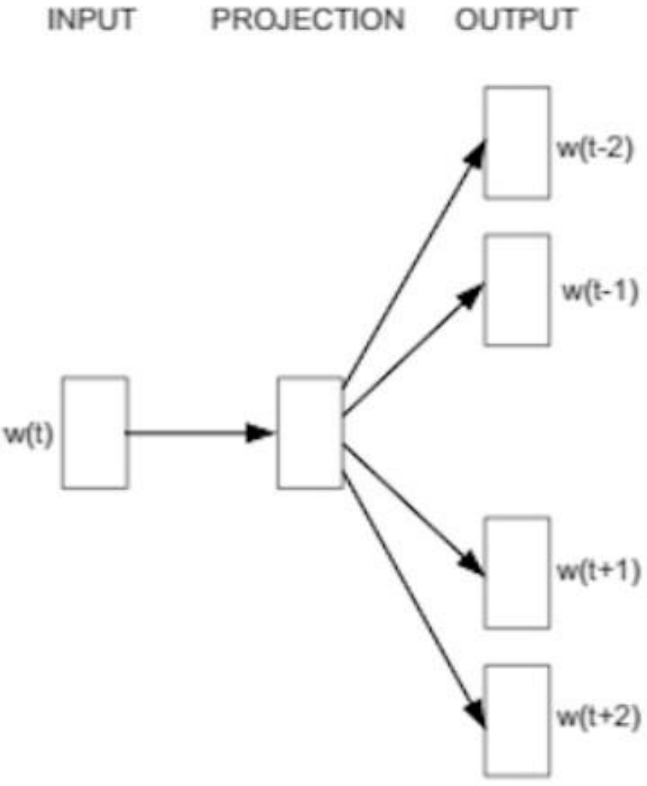


Figure 6: Skip-gram

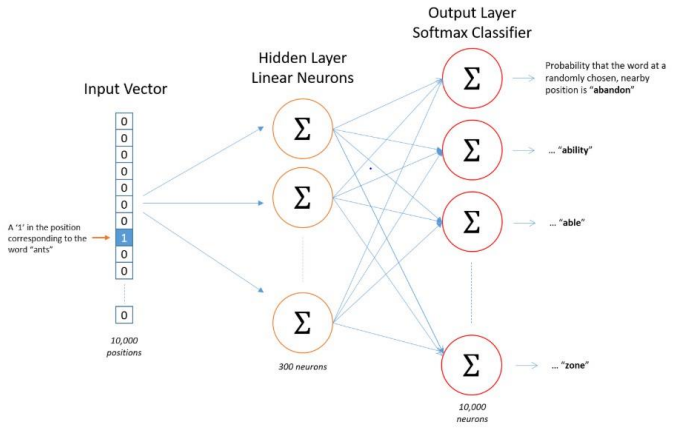


Figure 7: Architecture of Word2vec

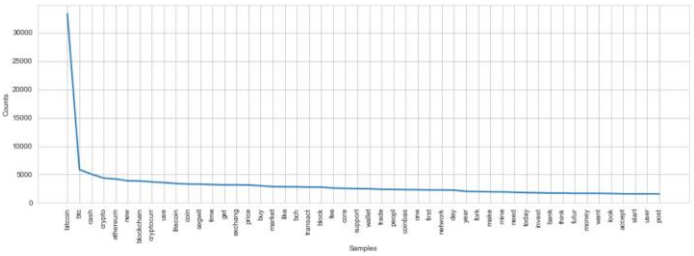


Figure 8: Frequency of words in the Data

Word2vec model consists of neural network so it cannot take string as an input instead, we pass the words to one-hot vectors, which has same length as the vocabulary filled with the zeros expect at the index. The word which we want to represent is assigned as "1". Hidden layer is fully connected dense layer which contains weights of word embeddings. The output layers contain the probabilities of the target words from the vocabulary. The word2vec embedding method learns by constructing the co-occurrence matrix that indicates the frequency of words appearing in the context.

6.3. Estimation Models

Classification involves predicting the class or category of given input data points, which are also referred to as targets, labels, or categories. Classification predictive modeling entails mapping input variables to output classes using mapping functions. In the current study, Recurrent Neural Networks (RNNs) are employed for data classification. RNNs are a type of artificial neural network designed to identify patterns in sequential data such as text, handwriting, spoken words, or time series data. The fundamental idea behind RNNs is that they learn and remember information from prior inputs to generate outputs. RNNs serve as a general framework for sequential modeling and processing of data. They excel in handling sequences of inputs rather than just single inputs and are well-suited for problems where a sequence of inputs is propagated through a model to obtain a single result.The depicted figure illustrates the architecture of an RNN, which includes a loop that enables information to persist. RNNs utilize a recurrence relation to process sequential data effectively. In the figure, the RNN takes an input vector xt, generates a prediction output yt, and updates an internal state represented by ht. This characteristic of RNNs gives them an advantage over feed-forward networks and Convolutional Neural Networks (CNNs) when dealing with sequences of input data. RNNs are specifically designed to handle sequential data and capture temporal dependencies, making them a suitable choice for modeling and processing such data.

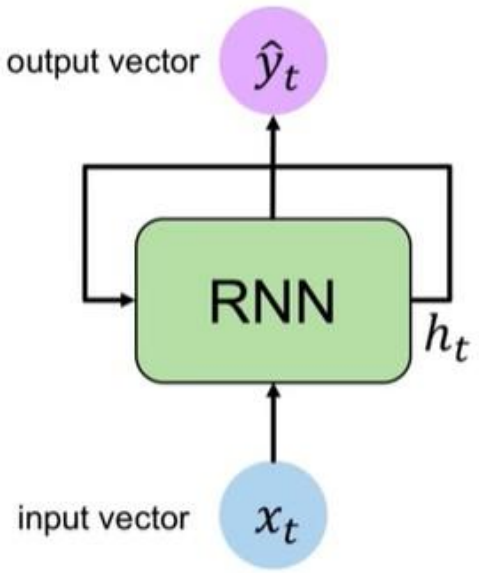


Figure 9: General architecture of RNN

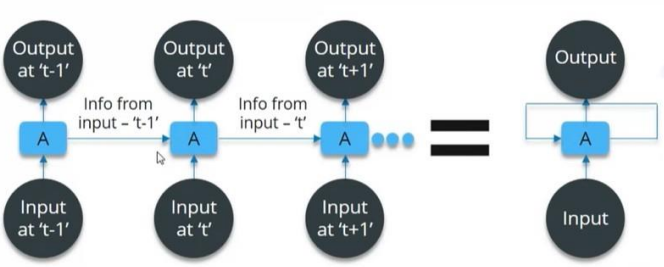


Figure 10: Many to many architectures of RNN

The input at 't-1' is fed to a network to get the output at 't-1' and the next time stamp that is time at 't' that will be given to a network along with the information from the previous time step that is 't-1' and that will help us to get the output at T similarly as output for 't+1'. We have two inputs one is a new input that we will help us to get the output at t similarly as output for 't+1' we have two inputs one is a new input that we give another is the information coming from the previous timestamps that is 't' in order to get the output at time 't+1' similarly it can go on so over.

The formula for the current state as, At = f (Outputt-1, Inputt)

 DOI: 10.55041/IJSREM24175

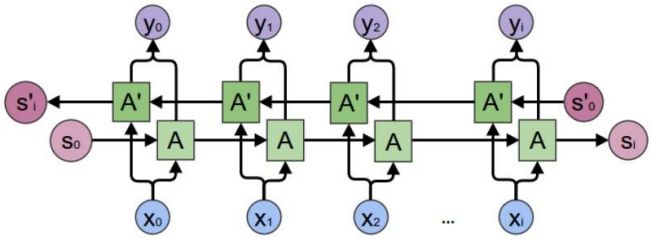


Figure 11: General structure of BRNN

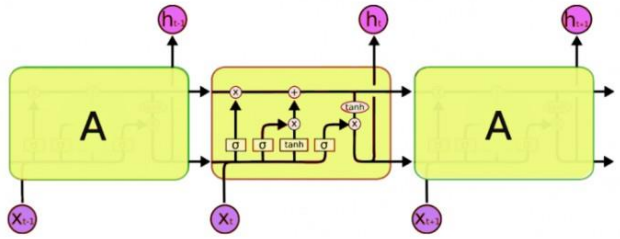


Figure 12: Architecture of Long Short-Term Memory

Along with the weighted matrixes as, At= tanh (Whh. ht-1 + Wxh. xt)
Where, tanh is considered as activation function, Whh is weighted Metrix and Wxh weight matrix of the current input. Finally, output is calculated as, ht=Why. At
In Recurrent neural network, the model learns from the prior results which may not be accurate in some circumstances. RNN suffers from vanishing gradients and exploding gradients. To overcome with, the study uses Bidirectional Recurrent Neural Network (BRNN).

BRNN is just putting two or more independent RNNs together (Zhu, 1997). The input sequence is passed in normal time order for one network and in reverse time order for another. At each time stamp both the output from two network is concatenated. This allows the network to have both forward and backword information at every time stamp.

Consider the following example below, with two sentences S1 and S2: S1: "He said, Teddy bears are on sale." S2: "He said, Teddy Roosevelt is a great president"

In the above sentence, if the model wants to predict the next word of "Teddy" just by using the input as "He said", create the ambiguity between the bears and Roosevelt. As a result, the model tends to produce the wrong results. So, it's important to learn from the previous and future representation to eliminate the ambiguity. The amalgamation of forward and backword direction resolves the problem. Along with BRNN, will be using Long Short-Term Memory (LSTM) to perform better for accuracy.

LSTM (Long Short-Term Memory) is a type of recurrent neural network designed to address the vanishing and exploding gradient problem. Proposed by Hochreiter and Schmidhuber

```
# Get the actual embeddings
sequence_input = Input(shape=(maxlen,), dtype='int32')
embeddings = embedding_layer(sequence_input)

# Construct the model
X = Bidirectional(LSTM(128, return_sequences=False))(embeddings)
X = Dense(5)(X)

# Select y labels
y_train = np.concatenate((y_train_1hr, y_train_2hr, y_train_6hr, y_train_12hr, y_train_24hr), axis=1)
y_test = np.concatenate((y_test_1hr, y_test_2hr, y_test_6hr, y_test_12hr, y_test_24hr), axis=1)

# Define the model
model = Model(inputs=sequence_input, outputs=X)

# Compile the model
model.summary()
model.compile(loss='mean_squared_error',
              optimizer='adam',
              metrics=['accuracy'])
#(NOTE: The standard error of the estimate is a measure of the accuracy of predictions.)
```

```
Layer (type)                 Output Shape              Param #
=================================================================
input_3 (InputLayer)         (None, 78)                0
_________________________________________________________________
embedding_2 (Embedding)      (None, 78, 200)           7844000
_________________________________________________________________
bidirectional_2 (Bidirection (None, 256)               336896
_________________________________________________________________
dense_2 (Dense)              (None, 5)                 1285
=================================================================
Total params: 8,182,181
Trainable params: 338,181
Non-trainable params: 7,844,000
_________________________________________________________________
```

Figure 13: Setup of Bidirectional LSTM Model

in 1997, LSTM introduces additional components to the basic RNN architecture. In addition to the hidden state vector, LSTM cells maintain a cell state vector. At each time step, an LSTM cell can choose to read, write, or reset the cell state using explicit gating mechanisms. The LSTM cell consists of two main components: the cell state and the hidden state. The cell state is responsible for transferring information to the next cell, while the hidden state represents the output of the LSTM cell.

Forget Gate: This gate controls the removal of information from the cell state. It determines which information is important and should be retained and which information is no longer needed. The forget gate achieves this by multiplying a filter with the cell state, effectively deciding whether to set certain memory cells to zero.

Input Gate: This gate helps to add the information to the cell state. The addition of information takes place in three steps. Using Sigmoid function, the input gate regulates the values that need to be added to the cell state. It creates a vector which contains all the possible values that needs to be added to the cell state and lastly multiplying the value of the regulatory function to the vector created and then adding the useful information to the cell state using the addition operation.

Output Gate: The output gate controls whether the information of the cell state is made visible. The output gate functions in the following way, it creates a vector after applying the tanh function. Using the Sigmoid function, it makes use of a filter to regulate the values that need to be output from the vector created in the input gate. Finally, it multiplies the value of the regulatory function to the vector created in the initial step to send as the output with the hidden state of the next cell.

The depicted figure illustrates the configuration of the Bidirectional LSTM model. The model begins with a sequence layer, followed by an embedding layer. The Bidirectional LSTM layer is then applied, consisting of two stages of recurrent neural networks with 128 layers each. A dense layer is added after the Bidirectional LSTM layer. The loss function used in this model is 'mean squared error', and the 'adam' op-

	+1hr prediction	+2hr prediction	+6hr prediction	+12hr prediction	+24hr prediction
Peaks	51.11	54.76	55.74	46.32	62.5
Troughs	47.97	47.01	50	65.29	36.76
Total	48.81	50.45	50.6	54.48	50.15

Figure 14: Prediction accuracy at different time intervals

timizer is employed.

7. Results and Discussions

The research objective was to predict the hourly increase and decrease of Bitcoin prices based on user comments obtained from the Reddit application. A Bi-directional LSTM model was developed using 70 of the data for training and tested on the remaining data. The model was used to predict the Bitcoin prices for different time intervals: +1 hour, +2 hours, +6 hours, +12 hours, and +24 hours, based on the user comments at those specific time intervals.

Table 1 presents the results obtained from the research. For the 1-hour prediction, the model achieved a correct prediction rate of 48.81. The accuracy of predicting price increase was 51.11, while the accuracy of predicting price decrease was 47.97. Similar results were observed for the 2-hour and 6-hour predictions, with overall accuracy rates of 50.45 and 50.6 respectively. The model did not show significant improvement in accuracy for these time intervals.

However, a significant change in performance was observed for the +12-hour and +24-hour predictions. The model showed better estimation for price decrease in the +12-hour interval compared to price increase. The overall accuracy for the +12-hour prediction was 54.48, with accuracy rates of 46.32 for price increase and 65.29 percent for price decrease. For the
+24-hour prediction, the model achieved 62.5 accuracy in predicting price increase, but had lower accuracy in predicting price decrease at 36.76.

Overall, the +12-hour prediction showed the highest accuracy, correctly predicting 54.48percent of the values. Figure 20 provides an overview of the model's performance across all time intervals.

The graphs provide insights into the accuracy of the model's predictions based on user comments. It is observed that the graphs for the +1-hour, +2-hour, and +6-hour intervals display disproportionality in the results. This can be attributed to the unstable nature of Bitcoin prices and the limited availability of user comments within those time periods.

On the other hand, the graphs for the +12-hour and +24-hour intervals show higher variation in prediction accuracy for forecasting the rise and fall of Bitcoin prices. This indicates a potential trend in the rise and fall of Bitcoin prices based on the textual analysis of comments from social media. The visualization of the graphs reveals that no discernible pattern is formed for the +1-hour, +2-hour, and +6-hour intervals, while the +12-hour and +24-hour intervals exhibit the presence of patterns.

These findings suggest the limitations in predicting Bitcoin prices based on user comments for short intervals (+1 hour, +2 hours, and +6 hours) due to the scarcity of data within those

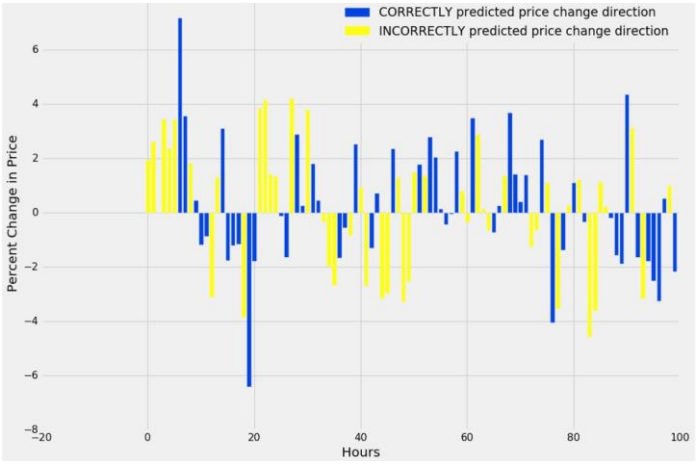


Figure 15: Prediction of price for 1 Hour

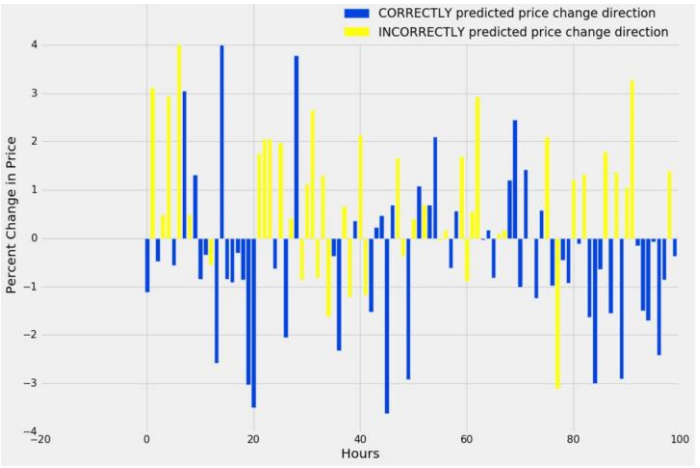


Figure 16: Prediction of price for 2 Hours

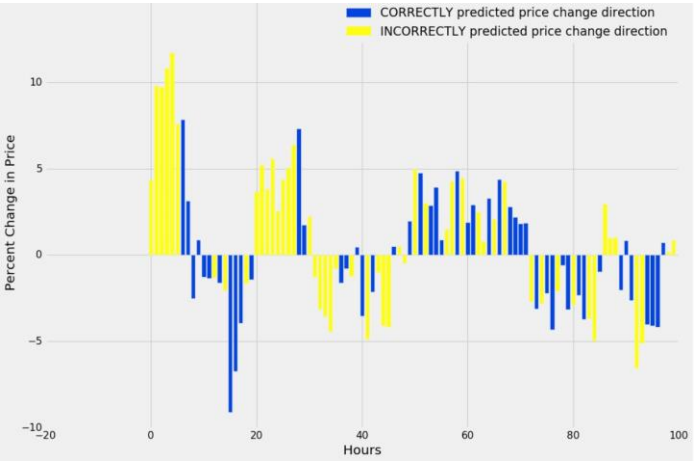


Figure 17: Prediction of price for 6 Hours

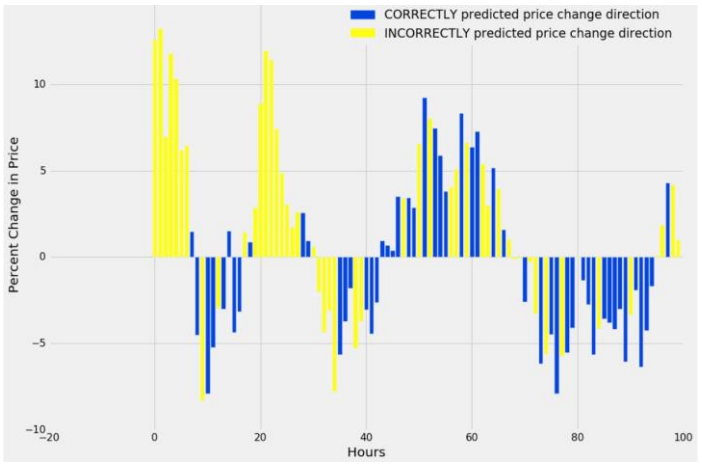


Figure 18: Prediction of price for 12 Hours

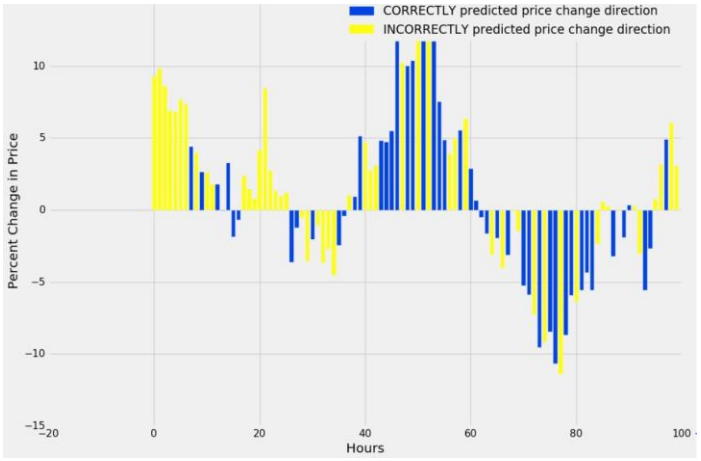


Figure 19: Prediction of price for 24 Hours

In summary, this study proposes the use of a deep learning model, specifically B idirectional R NN w ith L STM u nits, for predicting Bitcoin prices on an hourly basis using user views from the social media platform Reddit. The research aimed to identify patterns or trends in the rise and fall of Bitcoin prices and investigate the influence of the number of comments on the accuracy of the predictions.

8. Conclusion and Future Works

timeframes. It is evident that the availability of more data could potentially improve the performance of the model and increase the accuracy of price predictions. This argument is supported by the fact that as the time interval increases, there is a reduced occurrence of incorrectly predicted price directions.

Overall, the analysis underscores the importance of sufficient data and longer time intervals for more accurate predictions of Bitcoin prices based on user comments.

The model proposed in the research may not be accurate enough in predicting the bitcoin prices through textual data, but it can assure that it will be able to predict if the prices will take a dip or rise in the immediate future.

The results of the experiment showed that the Bi-directional RNN with LSTM units performed best when using comments from a 12-hour period to predict future values. Notably, the research demonstrated a greater ability to identify trends in the changes of Bitcoin prices for 12-hour and 24-hour intervals compared to shorter intervals of 1 hour, 2 hours, and 6 hours. This observation could be attributed to the varying number of comments made by users at different time intervals, which were considered in this study.

To summarize, the results of the Bi-directional RNN with LSTM units can be utilized with textual data from Bitcoin users collected over a 12-hour period or longer. This can assist customers in understanding current market trends and provide insights on when to invest in Bitcoin prices to potentially improve their profit margins.

In future work, it is recommended to conduct extended research using a larger dataset of user comments collected from different social media platforms to enhance the accuracy of the model. Additionally, exploring alternative word embedding approaches such as ConceptNet Numberbatch and XLNet within the proposed framework could be beneficial to evaluate if they yield improved results.

9. References

90 Percent of the Big Data We Generate Is an Unstructured Mess - PCMag UK (no date). Available at: https://uk.pcmag.com/news-analysis/118459/90-percent-of-the-big-data-we-generate-is-an-unstructured-mess (Accessed: 25 August 2019).

Abraham, J., Higdon, D. and Nelson, J. (2018) 'Cryptocurrency price prediction using tweet volumes and sentiment analysis', SMU Data Science Review, 1(3). Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/1.

Ahn, Y. and Kim, D. (2019) 'Sentiment disagreement and bitcoin price fluctuations: a psycholinguistic approach', Applied Economics Letters. Routledge. doi: 10.1080/13504851.2019.1619013.

Alessandretti, L. et al. (2018) 'Anticipating Cryptocurrency Prices Using Machine Learning', Complexity, 2018, pp. 0–2. doi: 10.1155/2018/8983590.

Collobert, R. and Weston, J. (2008) 'General Deep Architecture for NLP ICML 2009.pdf'. doi: 10.1145/1390156.1390177.

Graves, A.-r. Mohamed, and G. H. (2015) 'Speech recognition with deep recurrent neural networks. In ICASSP'2013', Ieee, 2013, pp. 6645–6649. doi: 10.1109/ICORR.2015.7281186.

Jang, H. and Lee, J. (2017) 'An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks Based on Blockchain Information', IEEE Access, 6(c), pp. 5427–5437. doi: 10.1109/ACCESS.2017.2779181.

Kang, M., Ahn, J. and Lee, K. (2018) 'Opinion mining

using ensemble text hidden Markov models for text classification', Expert Systems with Applications. Elsevier Ltd, 94, pp. 218–227. doi: 10.1016/j.eswa.2017.07.019.

Kim, Y. Bin et al. (2015) 'Virtual world currency value fluctuation prediction system based on user sentiment analysis', PLoS ONE. Public Library of Science, 10(8). doi: 10.1371/journal.pone.0132944.

Kim, Y. Bin et al. (2017) 'When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation', PLoS ONE, 12(5), pp. 1–14. doi: 10.1371/journal.pone.0177630.

Krafft, P. M., Della Penna, N. and Pentland, A. S. (2018) 'An experimental study of cryptocurrency market dynamics', in Conference on Human Factors in Computing Systems - Proceedings. Association for Computing Machinery. doi: 10.1145/3173574.3174179.

Lamon, C., Nielsen, E. and Redondo, E. (no date) Cryptocurrency Price Prediction Using News and Social Media Sentiment.

Madan, I., Saluja, S. and Zhao, A. (2015) 'Automated Bitcoin Trading via Machine Learning Algorithms', URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan, 20, pp. 1–5. Available at: http://cs229.stanford.edu/proj2014/Isaac Madan, Shaurya Saluja, Aojia Zhao,Automated Bitcoin Trading via Machine Learning Algorithms.pdf.

Mai, F., Bai, Q., Shan, Z., Wang, X. (Shane), et al. (2015) 'From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance', SSRN Electronic Journal, pp. 1–46. doi: 10.2139/ssrn.2545957.

Mai, F., Bai, Q., Shan, Z., Wang, X., et al. (2015) 'The impacts of social media on bitcoin performance', 2015 International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015, pp. 1–16.

Matta, M., Lunesu, I. and Marchesi, M. (2015a) Bitcoin Spread Prediction Using Social and Web Search Media, UMAP Workshops. Available at:

(Accessed: 25 August 2019).

Matta, M., Lunesu, I. and Marchesi, M. (2015b) 'The predictor impact ofweb search media on bitcoin trading volumes', IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 1, pp. 620–626.

McNally, S., Roche, J. and Caton, S. (2018) 'Predicting the Price of Bitcoin Using Machine Learning', in Proceedings - 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018. Institute of Electrical and Electronics Engineers Inc., pp. 339–343. doi: 10.1109/PDP2018.2018.00060.

Nadkarni, P. M., Ohno-Machado, L. and Chapman, W. W. (2011) 'Natural language processing: An introduction', Journal of the American Medical Informatics Association, 18(5), pp. 544–551. doi: 10.1136/amiajnl-2011-000464. Ncr, P. C. et al. (2004) 'Crisp-Dm 1.0', pp. 1–76. Available at: ftp://ftp.software.ibm.com/software/analytics/spss/support

Phillip, A., Chan, J. and Peiris, S. (2018) 'A new look at Cryptocurrencies', Economics Letters. Elsevier B.V., 163, pp. 6–9. doi: 10.1016/j.econlet.2017.11.020.

Sin, E. and Wang, L. (2018) 'Bitcoin price prediction using ensembles of neural networks', ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. IEEE, pp. 666–671. doi: 10.1109/FSKD.2017.8393351.

Singer, P. et al. (2016) 'Evidence of online performance deterioration in user sessions on Reddit', PLoS ONE. Public Library of Science, 11(8). doi: 10.1371/journal.pone.0161636.

The Impact of Social Media Marketing Today - Social Media Impact - Social Media Impact (no date). Available at: http://www.socialmediaimpact.com/impact-social-media-marketing-today/ (Accessed: 25 August 2019).

The Law Library of Congress (2018) 'Regulation of Cryptocurrency Around the World', The Law Library of Congress, 5080(June). Available at: http://www.law.gov.

Tsujii, J. (2008) 'Computational Linguistics and Intelligent Text Processing', 4919(February 2011). doi: 10.1007/978-3-540-78135-6.

Wang, Y. et al. (2018) 'A comparison of word embeddings for the biomedical natural language processing.', Journal of biomedical informatics, 87, pp. 12–20. doi: 10.1016/j.jbi.2018.09.008.

What Affects The Price Of Cryptocurrency? - Mycryptopedia (no date). Available at: https://www.mycryptopedia.com/what-

https://markets.blockchain.info0Ahttps://
pdfs.semanticscholar.org/1

Xie, P., Chen, H. and Hu, Y. J. (2017) 'Network Structure and Predictive Power of Social Media for the Bitcoin Market', SSRN Electronic Journal. Elsevier BV. doi: 10.2139/ssrn.2894089.

Zhu, S. (1997) 'Bidirectional Recurrent Neural Networks as Generative models', 45(May), pp. 1–10.

