A Multilayer Perceptron Architecture for Detecting Deceptive Cryptocurrencies in Coin Market Capitalization Data

Harshita Dalal
Department of Computer Science
South Asian University, New Delhi, India
harshita@ieee.org

Muhammad Abulaish
Department of Computer Science
South Asian University, New Delhi, India
abulaish@sau.ac.in

ABSTRACT

Due to increasing popularity of Bitcoin and other cryptocurrencies, proliferation of deceptive cryptocurrencies over the internet is a global concern. In this paper, we have identified a set of 24 features through analyzing Cryptocurrency Market Capitalization (CMC) data and propose a Multilayer Perceptron (MLP) architecture for detecting deceptive cryptocurrencies. The proposed MLP architecture is compared with three traditional machine learning algorithms over a real cryptocurrency dataset crawled from CMC website, and it performs significantly better.

CCS CONCEPTS

• **Security and privacy** → **Web application security**; • **Information systems** → *Data analytics*; • **Computing methodologies** → *Supervised learning*.

KEYWORDS

Data analytics, Cryptocurrency deception, Machine learning

ACM Reference Format:
Harshita Dalal and Muhammad Abulaish. 2019. A Multilayer Perceptron Architecture for Detecting Deceptive Cryptocurrencies in Coin Market Capitalization Data. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3350546.3352564

1 INTRODUCTION

Cryptocurrency (CC) is an intangible asset designed to work as a medium of exchange. It is digital, decentralized and based on the underpinning blockchain technology, which can be conceived as a revolutionary technology having cryptocurrencies as one of its promising products. To date, there are 2209 cryptocurrencies listed on Cryptocurrency Market Capitalization[1] (CMC) website with approximately 251 billion as total market capital in USD [2], and the count is still on acceleration. Figure 1 shows the increasing popularity of cryptocurrencies since 2013.

Cryptocurrencies are classified into two most common categories – *alternative cryptocurrency coins* (*altcoins*) and *tokens. Altcoins* and *coins* are often used interchangeably. Altcoins refer to coins that are an alternative to Bitcoin, i.e., variants of Bitcoin. Namecoin, Peercoin, Litecoin, Dogecoin, etc. are examples in the category. There are altcoins that are not derived from Bitcoin, instead they have designed their own blockchain. Examples in this category include Ripple, Ethereum, Omni, Nxt, Waves, etc. The commonality between all altcoins is that each coin owns an independent blockchain. Tokens, on the other hand, reside over some pre-existing blockchain (often known as platform).

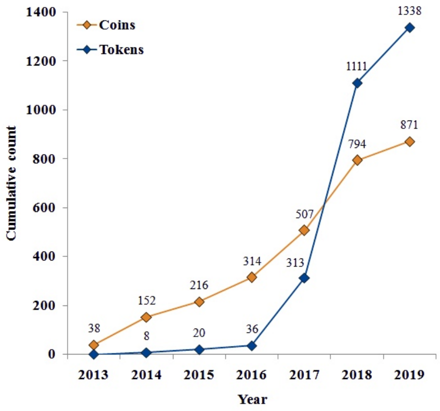

Figure 1: Growth of cryptocurrencies (*Source: CMC*)

Cryptocurrencies are purchased online only, and three common modes of purchase are – vendors' websites via Initial Coin Offering (ICO), exchanges (e.g., Poloniex, Binance, etc.), and between individuals. Vendors also utilize various social media services to promote and advertise their coins or tokens. Cryptocurrency industry is in its early stage and cryptocurrencies can be highly volatile. Cryptocurrency market alike other social media environments is vulnerable and plagued with individuals with malicious objects. It is probable that many cryptocurrencies are deceptive, they may fail or are downright scams. The cryptocurrency market would turn dark to newbies who delve into this revolution without sufficient due diligence as it is likely to come across scammy ICOs or abandoned projects that are just money-grabs. The absence of any regulations and lack of understanding of technology behind cryptocurrencies even by the tech-savvy individuals further compound this problem, creating a conducive environment for deceivers to create coins and schemes that serve to exploit the ill-informed. The realm of deceptive cryptocurrencies includes failed ICOs, scam, parody, deceased,

[1] https://coinmarketcap.com/

WI '19, October 14–17, 2019, Thessaloniki, Greece

ACM ISBN 978-1-4503-6934-3/19/10…$15.00
https://doi.org/10.1145/3350546.3352564

etc. They are likely to come within the purview of identity forgery. The key concern here is to identify the deceptive cryptocurrencies as numerous instances were reported, where vendors deceived their customers after raising the desired fund.

In this paper, we present a set of 24 features identified through analyzing CMC data, and propose an MLP architecture for detecting deceptive cryptocurrencies. It also presents comparative analysis of MLP with three popular machine learning algorithms – *linear regression*, *softmax regression*, and *support vector machine* (SVM).

2 DECEPTION CATEGORIES

Deception is a deliberate act wherein the deceiver aims to transfer a false belief to a recipient who is unaware of the fact that the information received has been falsified [1, 3, 9]. A deceptive attack is carried with either single or any combination of the three categories of motivations such as *instrumental* (goal-driven), *relational* (relationship-driven), and *identity-driven* [1]. Though deception may occur online or offline, online deception is more devastating reason being the ease of content or identity manipulation, absence of accountability, poor moral cost, lack of many verbal or non-verbal cues, and imperfect laws dealing with the situation [7]. Online deception attacks are classified into three major categories – *content manipulation*, *communication channel manipulation*, and *identity manipulation* [7]. Amalgamation of these three can make the resultant deception more effective. Almost all online platforms are vulnerable to either of these manipulations.

Identity deception is a special category of online deception wherein a sender's true identity is deliberately concealed or altered aiming to deceive the target receiver meanwhile the receiver doesn't anticipate the sender's identity tampering. It can be classified as *identity concealment* (withholding a part of an individual's identity), *identity theft* (stealing an individual's identity), and *identity forgery* (creating a fictional identity) [9]. Identity forgery is most prevalent due to the inherent design of social media services, allowing individuals to create an unlimited number of accounts with no or limited identity verification, generating new identities in a single click [8]. Today, almost all online platforms are plagued with individuals that abuse this option and attempt to deceive potential targets against their malign intents.

3 PROPOSED APPROACH

This section presents the functional steps of our proposed approach, including data curation, feature extraction, and deceptive cryptocurrency detection.

3.1 Data Curation

The CMC website maintains a list of cryptocurrencies and it provides financial statistics for almost all cryptocurrencies since their inception.

The data utilized in this study was collected from CMC website on May 5, 2019 using a crawler implemented in Python employing `NumPy`, `pandas`, and `Beautiful Soup`. CMC website tags cryptocurrencies into two categories – *active* and *inactive*. Here, active and inactive cryptocurrencies are referred as legitimate and deceptive, respectively. Table 1 presents the statistics of the dataset curated from CMC website. The dataset is balanced and comprises 1000 cryptocurrencies data, including top 500 legitimate out of 2149 active cryptocurrencies and 500 deceptive cryptocurrencies.

Table 1: Statistics of CMC dataset

CC category	#Legitimate (active)	#Deceptive (inactive)	Total
Coins	170	435	605
Tokens	330	65	395
Total	500	500	1000

3.2 MLP Architecture

An MLP comprises of an input layer, one or more hidden layers, and an output layer [4, 5]. In this paper, we use an MLP comprising of an input layer, three hidden layers each holding 500 nodes, and an output layer. The input layer (*aka* layer 0) is defined using equation (1), where X is the training data stacked horizontally.

$$A^{[0]} = X \tag{1}$$

MLP learns in two steps – a feed-forward pass followed by a back-propagation pass. In forward propagation, each layer (except l=0) passes on its output to the next layer using equation (2), where $W^{[l]}$ and $b^{[l]}$ are randomly initialized weights and biases, respectively that are assigned to layer l, $A^{[l-1]}$ is the output obtained from the previous layer $(l-1)$, $f^{[l]}$ is the non-linearity operating at layer l, and L is the total number of layers (in our case, L=4).

$$\begin{aligned} Z^{[l]} &= W^{[l]} \cdot A^{[l-1]} + b^{[l]} \\ A^{[l]} &= f^{[l]}(Z^{[l]}) \\ \forall l &= 1, 2, ..., L \end{aligned} \tag{2}$$

Mathematically, a non-linear activation function is defined using two key characteristics – *differentiability* and *monotonicity*. In this paper, *sigmoid* is placed at the output layer while *ReLU* is employed to all the hidden layers. *Sigmoid* or *logistic sigmoid* activation function squashes real numbers in the range [0, 1], whereas *ReLU* activation function is simply the half-wave rectifier and its range is $[0, \infty)$. It learns much faster than smoother non-linearities, such as *sigmoid* or *tanh* in multilayer architectures [6].

A cost function measures the performance of an MLP on entire training dataset, and resultantly governs the learning process. Mean Cross Entropy (MCE) is the cost function, which is used in our proposed architecture. It is defined using equation (3), where m is the total count of training samples, y is the expected class, and $\hat{y}$ is the predicted class.

$$L(y, \hat{y}) = -\frac{1}{m} \sum \left[y \cdot log\hat{y} + (1-y) \cdot log(1-\hat{y}) \right] \tag{3}$$

Backward propagation, an application of the chain rule of differentiation, is used to differentiate functions of more than one variable with the help of partial derivatives. Mathematically, it is defined using equation (4), where $dA^{[l]}$, $dZ^{[l]}$, $dW^{[l]}$, and $db^{[l]}$ are the error derivatives with respect to various intermediate quantities of layer l. It may be noted that at output layer (l = 4), $dA^{[4]}$ is determined by finding the derivative of the cost function. Finally, the Stochastic Gradient Descent (SGD) learning algorithm is used to tweak the parameters W and b to minimize the cost.

$$\begin{aligned} dZ^{[l]} &= dA^{[l]} \cdot f^{[l]'}(Z^{[l]}) \\ dW^{[l]} &= \frac{1}{m} dZ^{[l]} \cdot A^{[l-1]} \\ db^{[l]} &= \frac{1}{m} dZ^{[l]} \\ \forall l &= L, (L-1), \dots, 1 \end{aligned} \tag{4}$$

3.3 Feature Extraction

We have identified a total number of 24 features pertaining to seven different categories from CMC dataset. Since, to the best of our knowledge, no literature on deceptive cryptocurrency detection exists, all identified features in this study are new. A brief description of the identified features is presented in the following sub-sections.

3.3.1 Type. The *type* feature category aims to capture the *class*, *group*, and *platform-related* information about a cryptocurrency. Accordingly, it consists of three boolean features whose values are set to either 0 or 1. The *class* feature is devised to capture whether a CC is coin or token, and accordingly its value is set to 1 (if coin) or 0 (otherwise). The *group* feature is applicable only to coins, and it is used to represent whether a CC is mineable or non-mineable. Accordingly, its value is set to 1 (if mineable) or 0 (otherwise). Finally, the *platform* feature is applicable only to tokens, and it is used to capture whether a CC is ethereum-based or non-ethereum-based.

3.3.2 Trade Volume. A cryptocurrency dominates market space with its *market cap* and *volume*. *Market cap* and *volume* of a cryptocurrency cc_i, is represented by $MC(cc_i)$ and $V(cc_i)$, respectively, and they are generally measured in USD. Formally, they are defined using equations (5) and (6), respectively, where $Cir(cc_i)$ is the total units of cc_i that are available for circulation in the exchanges/markets, $Price(cc_i)$ is the per unit price of cc_i in USD, and $Trade(cc_i)$ is the total units of cc_i that are currently employed for buying/selling in the exchanges.

$$MC(cc_i) = Cir(cc_i) \times Price(cc_i) \tag{5}$$

$$V(cc_i) = Trade(cc_i) \times Price(cc_i) \tag{6}$$

The *trade volume* feature category captures three prominent features for all cryptocurrencies through which they vogue in cryptocurrency market space – *market cap dominance*, *average market cap*, and *average volume*.

Market Cap Dominance: The *market cap dominance* of a cryptocurrency cc_i is denoted by $MCD(cc_i)$ and it ascertains the percentage of $MC(cc_i)$ to the *total market cap* on CMC website, as defined in equation (7).

$$MCD(cc_i) = \frac{MC(cc_i)}{\sum_{j=1}^{n} MC(cc_j)} \times 100 \tag{7}$$

Average Market Cap: The *average market cap* of a cryptocurrency cc_i is denoted by $AMC(cc_i)$, and it is used to measure the daily average of the *market cap* of cc_i. Formally, it is defined using equation (8), where $MCV_{DoD}(cc_i)$ and $MCV_{DoB}(cc_i)$ are the *market cap values* of cc_i that are recorded on two different dates – *death* (the date it was last listed) and *birth* (the date it was first listed), respectively. $MC_{DoD}(cc_i)$ refers to the death date of the *market cap* of cc_i, i.e., the date when its *market cap* was last listed on CMC website. $MC_{DoB}(cc_i)$ refers to the birth date of the *market cap* of cc_i, i.e., the date when its *market cap* was first listed on CMC website. For legitimate cryptocurrencies, the value of $MC_{DoD}(cc_i)$ is considered as May 4, 2019.

$$AMC(cc_i) = \frac{MCV_{DoD}(cc_i) - MCV_{DoB}(cc_i)}{MC_{DoD}(cc_i) - MC_{DoB}(cc_i)} \tag{8}$$

Average Volume: The *average volume* of a cryptocurrency cc_i is represented as $AV(cc_i)$, and it is used to measure the daily average of the *volume* of cc_i. Formally, it is defined using equation (9), where $VV_{DoD}(cc_i)$ and $VV_{DoB}(cc_i)$ are the volumes of cc_i that are recorded on two different dates – death and birth, respectively. $V_{DoD}(cc_i)$ and $V_{DoB}(cc_i)$ refer to the death and birth dates, respectively, of the volume of cc_i. For legitimate cryptocurrencies, the value of $V_{DoD}(cc_i)$ is considered as May 4, 2019.

$$AV(cc_i) = \frac{VV_{DoD}(cc_i) - VV_{DoD}(cc_i)}{V_{DoD}(cc_i) - V_{DoB}(cc_i)} \tag{9}$$

3.3.3 Rank. Since a cryptocurrency can be either a coin or a token, we have defined two different features – *intra-class rank* and *inter-class rank* based on the *market capital contribution* (MCC) values under this feature category.

Intra-class Rank: The intra-class rank (*aka* local rank) defines the position of a cryptocurrency within the class based on its MCC value. Formally, the intra-class rank of a cryptocurrency cc_i is denoted as $\rho_{intra}(cc_i)$, and defined using equation (10), where $position_{local}(cc_i)$ is the position of cc_i within its class based on the MCC value.

$$\rho_{intra}(cc_i) = 1 - \frac{position_{local}(cc_i) - 1}{\max\limits_{\forall j} \{position_{local}(cc_j)\} - 1} \tag{10}$$

Inter-class Rank: Inter-class rank (*aka* global rank) of a cryptocurrency cc_i is denoted as $\rho_{inter}(cc_i)$, and calculated using equation (11), where $position_{global}(cc_i)$ is the position held by cc_i across its category, i.e., among coins and tokens both.

$$\rho_{inter}(cc_i) = 1 - \frac{position_{global}(cc_i) - 1}{\max\limits_{\forall j} \{position_{global}(cc_j)\} - 1} \tag{11}$$

3.3.4 Supply. A cryptocurrency is introduced in the markets with a count of units available for its supply. From *supply* feature category, we have ascertained three counts for each cryptocurrency – *maximum supply limit*, *total supply*, and *circulation*. The *maximum supply limit* of a CC refers to the upper bound on its units. The *total supply* refers to the portion of the supply limit which is held by various stakeholders. Finally, *circulation* refers to the portion of the total supply which is in circulation across various exchanges/markets.

3.3.5 *Age.*

Age is an important characteristic that ascertains the longevity of a cryptocurrency in years it performed/survived in such a highly volatile market space. We have defined three features under *age* feature category that are briefly described in the following paragraphs.

Life-span: Life-span of a cryptocurrency cc_i is denoted as $Life(cc_i)$, and it refers to the age (in years) of cc_i on CMC website. It is defined using equation (12), where $DoD(cc_i)$ and $DoB(cc_i)$ refer to the death and birth dates of cc_i, respectively. For legitimate cryptocurrencies, the value of $DoD(cc_i)$ is considered as May 4, 2019.

$$Life(cc_i) = \left\lfloor \frac{DoD(cc_i) - DoB(cc_i)}{365} \right\rfloor \tag{12}$$

Market Cap Age: Market cap age of a cryptocurrency cc_i is denoted as $MCA(cc_i)$, and it refers to the age (in years) of the *market cap* of cc_i on CMC website. Formally, it is defined using equation (13).

$$MCA(cc_i) = \left\lfloor \frac{MC_{DoD}(cc_i) - MC_{DoB}(cc_i)}{365} \right\rfloor \tag{13}$$

Volume Age: Volume age of a cryptocurrency cc_i is denoted as $VA(cc_i)$, and it refers to the age (in years) of the *volume* of cc_i on CMC website. Formally, it is calculated using equation (14).

$$VA(cc_i) = \left\lfloor \frac{V_{DoD}(cc_i) - V_{DoB}(cc_i)}{365} \right\rfloor \tag{14}$$

3.3.6 *Trade Junction.*

Under *trade junction* feature category, we have defined two features – *number of exchanges* and *number of markets*. An exchange is a place, where a cryptocurrency is listed for trading. On the other hand, a market refers to a pair of cryptocurrencies that are available on an exchange to facilitate trading between them. For instance, Ripple/Bitcoin (XRP/BTC) is a market available on Poloniex which means if one holds XRP s/he can trade in BTC or vice-versa.

3.3.7 *Social Media.*

A cryptocurrency is introduced to the world through an official website presenting credentials, such as white paper, ICOs, exchanges and potential markets, etc. Vendors utilize online social media platforms to provide regular financial updates, and future trends rendering quality services to the various stakeholders. The *social media* feature category captures all social media services that are being used by a cryptocurrency. A total number of eight boolean features viz. *website*, *source code*, *white paper*, *announcement*, *explorer*, *explorer 2*, *twitter presence*, and *reddit presence* are identified in this category and their values are set to either 1 (if active/available) or 0 (otherwise).

4 EXPERIMENTAL SETUP AND RESULTS

This section presents experimental results using MLP architecture and three traditional machine learning algorithms, Linear Regression, Softmax Regression, and SVM for detecting deceptive cryptocurrencies in CMC data. The dataset is partitioned in the ratio 8:2 for training and test phases, respectively. All experiments were conducted on a LINUX system having Intel(R) Xeon(R) 3.4 GHz CPU and 16 GB RAM. Table 2 enlists the manually-assigned hyper-parameters for the machine learning algorithms.

Table 2: Employed hyper-parameters for the machine learning algorithms

Hyper-parameters	Linear regression/ Softmax regression/ SVM	MLP
Cell size	NA	[500, 500, 500]
Batch size	200	200
Learning rate	1e-2	1e-2
Epochs	500	500
Dropout	NA	0.1
Data points	150000	150000

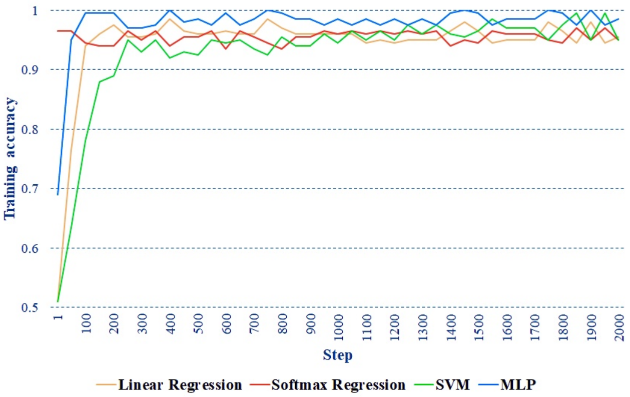

Figure 2: Training accuracy of the ML algorithms at different steps

Table 3: Evaluation results of the ML algorithms over test dataset

Eval. metrics	Linear reg.	Softmax reg.	SVM	MLP
Accuracy	93.00%	94.00%	95.00%	**98.00%**
Precision	91.35%	93.14%	93.27%	**98.98%**
TPR	95.00%	95.00%	97.00%	**97.00%**
FNR	05.00%	05.00%	03.00%	**03.00%**
TNR	91.00%	93.00%	93.00%	**99.00%**
FPR	09.00%	07.00%	07.00%	**01.00%**

Figure 2 shows the visualization of training accuracy of the machine learning algorithms recorded at different step values over the CMC dataset. Table 3 summarizes the experimental evaluation results of all machine learning algorithms over the CMC dataset. It can be observed that MLP outperforms others with highest test accuracy.

5 CONCLUSION AND FUTURE WORKS

In this paper, we have identified a set of 24 features through analyzing legitimate and deceptive cryptocurrency data curated from CMC website. We also propose an MLP architecture for detecting deceptive cryptocurrencies with better accuracy in comparison to the traditional machine learning algorithms. This study remains a challenging problem in detecting identity deceptions in cryptocurrency data to safeguard the cryptocurrency community and paramount interest of various stakeholders against all potential damages. Devising more discriminating features and their evaluation on larger datasets seems a promising direction of research for detecting identity deceptions in cryptocurrency data.

REFERENCES

[1] David B. Buller and Judee K. Burgoon. 1996. Interpersonal Deception Theory. *Communication Theory* 6, 3 (1996), 203–242.

[2] Cryptocurrency Market Capitalization (CMC) 2019. *Global Charts*. Retrieved May 26, 2019 from https://coinmarketcap.com/charts/

[3] Paul Ekman. 1997. Deception, Lying, and Demeanor. *States of Mind: American and Post-Soviet Perspectives on Contemporary Issues in Psychology* (1997), 93–105.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Vol. 1. MIT Press Cambridge.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444.

[6] Vinod Nair and Geoffrey E. Hinton. Jun., 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*. Omnipress, Haifa, Israel, 807–814.

[7] Michail Tsikerdekis and Sherali Zeadally. 2014. Online Deception in Social Media. *Commun. ACM* 57, 9 (2014), 72–80.

[8] Michail Tsikerdekis and Sherali Zeadally. 2014. Multiple Account Identity Deception Detection in Social Media using Nonverbal Behavior. *IEEE Transactions on Information Forensics and Security* 9, 8 (2014), 1311–1321.

[9] G. Alan Wang, Hsinchun Chen, Jennifer J. Xu, and Homa Atabakhsh. 2006. Automatically Detecting Criminal Identity Deception: An Adaptive Detection Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 36, 5 (2006), 988–999.