Scam Cryptocurrency Detections using Machine Learning Techniques

Kentaro Asaba

Keio University

It has been reported that more than 2,400 cryptocurrencies have been created by October 2018 [Investing.com, 2018], and new coins are constantly produced even today. However, a large number of so-called *scam coins* are also produced, and they are generating a huge amount of loss to the investors [Seth, 2018]. This research tries to build models that estimate whether a cryptocurrency is a scam, using the following machine learning techniques: Doc2Vec, Word2Vec, Support Vector Machine (SVM), Decision Tree, Neural Net Classifier, and Random Forest. It has been found that Doc2Vec×Random Forest has the highest performance.

1. Introduction

Since the very first cryptocurrency called Bitcoin was published in 2009, more than 2,400 cryptocurrencies have been created by October 2018 [Investing.com, 2018]. Many investors seek for investment opportunities in cryptocurrencies' Initial Coin Offering (ICO), a newly invented way to fund money using cryptocurrencies, as the prices of the newly published coins often dramatically increased after the launch of ICOs. It has previously been observed that the average return brought by ICO is 179% [Kostovetsky and Benedetti, 2018]. Nevertheless, so-called scam coins, which can be defined as cryptocurrencies that have no practical use and created solely to deceive money from investors, are also being produced. It has been estimated that almost $9 million worth of money was lost every day due to cryptocurrency frauds for the first two months of 2018 [Seth, 2018]. Therefore, there is a strong demand for methods to prevent the loss caused by cryptocurrency scams.

This paper attempts to create models which receive cryptocurrencies' whitepapers as inputs, and predict whether they are scam coins or not, by following the process illustrated by previous researches [Basavaraju and Prabhakar, 2010], [Bergsma, et al., 2012].

2. Methods

The document classification task includes the following procedures:

(1) Training data preparation
(2) Preprocessing transformation
(3) Document vectorization
(4) Application of classification models
(5) Evaluation

(1) Training Data Preparation

In total, 120 whitepapers are collected from each cryptocurrencies' website. 60 of them are labeled as "scam", which represents deceptive cryptocurrencies, and rest of 60 papers are labelled as "listed", whose coins are listed on binance.com, the world's largest cryptocurrency exchange website. The list of scam coins is extracted from deadcoins.com.

(2) Preprocessing transformation

When converting the documents into text, Reference sections are removed since they are irrelevant with the content of the papers. In addition, images, graphs, and tables are also removed from the documents before analysis.

After unnecessary parts are removed, all capital letters are converted into lowercase letters. Occasionally, in other classification tasks, words that include capital letter(s) are excluded from the texts [Bergsma, et al., 2012]. The advantage of removing them is that names, citations, and other proper nouns can be extracted which can contribute to the removal of nonessential parts. However, this research does not remove the words with capital letters in order to minimize the information loss.

(3) Document Vectorization

The next step is document vectorization. Sentences need to be transformed into vectors with this process because the classification models can only receive vectors. TF-IDF, Average vector of Word2Vec [Mikolov, et al., 2013] embedding, Doc2Vec [Le & Mikolov, 2014] are used in this research to compare their performances, and to achieve the highest accuracy as possible.

Average Word2Vec method takes the average of the word vectors of every words v_{wi} in the document and defines it as a document vector **d**.

$$\mathbf{d} = \frac{1}{n}\sum_{i=1}^{n} v_{wi},$$

where w_i is the i-th word in the document, and v_{wi} is the word embedding for the word w_i.

In this research, pre-trained 300-dimension word vectors which are obtained from training of Wikipedia documents are employed since it is found that the pre-trained word vector is useful in sentence categorization tasks [Kim, 2014].

(4) Application of classification models

This paper will compare the performance of five classification techniques: Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and Neural Net Classifier. As the

Contact: Kentaro Asaba, Keio University,
kentaroasaba@keio.jp

documents are labeled when they are collected, the training will be conducted in a supervised way.

(5) Evaluation

The models are evaluated based on four criteria: Accuracy, F1-score, Recall, and Precision. We define *True Positive* as number of scam currency correctly classified as scam, *True Negative* as number of listed currencies correctly classified as listed, *False Positive* as number of listed currencies classified as scam, and *False Negative* as number of scam currency classified as listed.

3. Results

3.1 Model Performance

Table 1 displays the model performances based on the four criteria. Among all combinations of methods, Doc2Vec×Random Forest method has the highest performance, and has 90% accuracy.

	Doc2Vec				TF-IDF				Avg. Word2Vec			
	Acc.	F1	Recall	Prec.	Acc.	F1	Recall	Prec.	Acc.	F1	Recall	Prec.
SVM	0.892	0.901	0.933	0.878	0.850	0.860	0.883	0.865	0.733	0.712	0.667	0.780
LR	0.883	0.894	0.933	0.867	0.842	0.857	0.900	0.836	0.758	0.747	0.733	0.775
DT	0.825	0.818	0.833	0.823	0.733	0.777	0.900	0.702	0.650	0.591	0.617	0.591
RF	**0.900**	**0.915**	**0.900**	**0.931**	0.892	0.915	0.833	0.993	0.800	0.750	0.733	0.794
NN	0.825	0.855	0.933	0.794	0.833	0.855	0.950	0.786	0.767	0.748	0.733	0.795

Table 1: *Model performances*

The confusion matrix for the best-performing method (Doc2Vec×Random Forest) is illustrated in Table 2 This method has almost the same numbers of FP and FN, thus has potential to be used in the real practices.

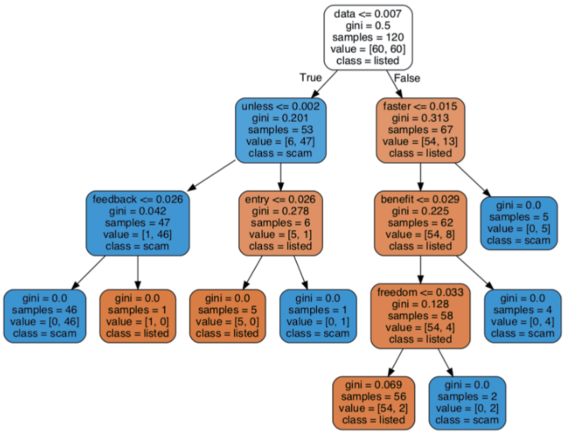
		Predicted Class	
		Scam	Listed
Actual Class	Scam	**54**	6
	Listed	4	**56**

Table 2: *Confusion matrix of Doc2Vec×Random Forest*

3.2 Visualization

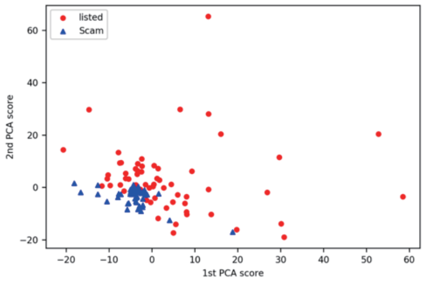
The Decision Tree thresholds when TF-IDF is used is visualized in Figure 1. The top words in the top of the boxes represent the feature and the threshold value, and the gini values show the impurities. From this figure, it seems clear that scam cryptocurrency documents are likely to have words *faster*, *benefit*, and *freedom*, and that of listed cryptocurrencies are likely to have words *data* and *unless*.

Figure 1: *Decision Tree Thresholds*



In order to visualize the distribution of vectorized documents, PCA is applied after Doc2Vec to reduce the dimensions to 2 (Table 2). The red dots, which represent the listed coins, are spread in upper right area, and the difference in distribution between scam and listed coins can still be seen even after the dimension reduction. The difference between listed and scam coins can still be detected even after the dimension reduction.

Figure 2: *Mapping of two dimension document vector by Doc2Vec and PCA*



4. Conclusion

This study has developed multiple models for scam cryptocurrency detection and compared their performance. Among all models, Doc2Vec×Random outperformed other methods and had 90% accuracy rate. However, it was not possible to include various kinds of domains other than whitepaper's texts as coin's representations due to the availabilities. Further research is required to build more effective models which includes other domains such as number and contents of images, graphs, mathematical formulae in the whitepapers, as well as domains from the coins themselves such as the types of coins (e.g. Ethereum based, non-Ethereum based), coin-names, and author's traits which can be utilized when classifying coins.

References

[Investing.com, 2018] Investing.com: All Cryptocurrencies , October 17, 2018, Retrieved from https://www.investing.com/crypto/currencies

[Seth, 2018] Seth, S: 80% of ICOs Are Scams: Report, Investopedia, 2018.

[Basavaraju and Prabhakar, 2010] Basavaraju, M., and Prabhakar, D. R.: A novel method of spam mail detection using text based clustering approach, International Journal of Computer Applications, 5(4), 15-25, 2010.

[Bergsma, et al., 2012] Bergsma, S., Post, M., and Yarowsky, D.: Stylometric analysis of scientific articles, 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 327-337, 2012

[Kostovetsky and Benedetti, 2018] Kostovetsky, L., and Benedetti, H.: Digital Tulips? Returns to Investors in Initial Coin Offerings, 2018

[Mikolov, et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, Workshop at the International Conference on Learning Representations, 2013

[Le & Mikolov, 2014] Le Q. and Mikolov T.: Distributed representations of sentences and documents, The 31st International Conference on Machine Learning, 1188–1196, 2014

[Kim, 2014] Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014