Initial Coin Offerings: risk or opportunity?

Toma Anca Mirela, Paola Cerchiello

RegTech Workshop on risks in blockchain and cryptocurrencies
Fintech SandLab
University of Pavia
www.fintech-ho2020.eu
230020
ancamirela.toma01universitadipavia.it

Fraud detection - ICO case study

- Initial Coin Offerings are a new yet uncovered mean to raise funds through tokens at the interplay of **crowdfunding** and **blockchain**.
- The acronym stands for initial offering of any crypto asset (ESMA).
- Few numbers (based on Coinschedule.com)
  - around **6** bi USD raised in 2017 by **456** ICOs
  - around **21.7** bi USD raised till the end of 2018 by **1076** ICOs
  - around **1** bi USD raised in Q1 2019 by **328** ICOs
- The crypto assets bring benefits but also risks due to the presence of criminal activity.
- Financial market authorities are very prudent and some countries ban straightaway all ICOs from their jurisdiction.

ICO risks and regulation activities

In the current hype, the risks of ICOs are a dangerous cocktail

- Exaggeration of expected returns
- The knowledge and expertise required is underestimated
- Lack of transparency
- Market driven by speculation and even manipulation

The crypto assets bring benefits but also risks due to the presence of criminal activity.

Financial market authorities are very prudent and some countries ban straightaway all ICOs from their jurisdiction.

Do we need to regulate ICOs?

- Regulation activities started in 2017 with different typologies of advises from the institutions, covering the potential and the mechanisms of the finance mechanism.
- Worldwide jurisdictions have opted for one of the following three solutions for ICO regulation:
    - proactive approach
    - careful consideration
    - undefined approach
    - Dura lex sed lex

"Washington D.C., Oct. 21, 2020 The Securities and Exchange Commission announced today that a federal district court has entered a final judgment on consent against Kik Interactive Inc. to resolve the SECs charges that Kiks unregistered offering of digital Kin tokens in 2017 violated the federal securities laws." For further references see here

ICOs peculiarities: Success, Failure or Scam

- The success of ICOs relays on the decentralized nature of P2P technology and on the lack of regulation.
- By February 2018 almost half of ICOs sold in 2017 failed (Hankin, 2018).
- Recent scientific studies are few, relying mainly on financial data or on the legal side (Adhami et al, Zeztche et al.)
- Our main goal is to contribute with an ensemble of alternative data and statistical approaches to the jigsaw puzzle of alternative crowdfunding systems, detecting which characteristics of an **ICO are significantly related to success and fraudulent behaviours.**

Data

- Data collecting process involves structured and unstructured information
  - websites specialized in providing financial information and in listing the existing ICOs (icobench.com, TokenData.io, ICODrops.com, CoinDesk.com)
  - Telegram social channel
- Typology of data collected:
  - categorical, numerical and textual data.
  - characteristics of white papers: elicited through textual analysis;
  - team members: quantitative and qualitative information;
  - type of business;
  - geographical distribution;
  - the supporting community: social channels;
  - Telegram's chat text.

Methodology - Response Variable

The response variable representing the status of an ICO is made up of 3 classes, intended as follows:

- **Success**: the given ICO collects the predefined cap within the time horizon of the campaign;
- **Failure**: the given ICO does not collect the predefined cap within the time horizon of the campaign;
- **Scam**: the given ICO is discovered to be a fraudulent activity during the campaign and described as such by all the platforms we use for data gathering (namely ICObench and Telegram).

Methodology - Explanatory variables

Table: Employed Covariates

class0	f=failed, sc=scam su=success
class1	0=success, 1=scam
class2	0=failed, 1= success
w_site	Website (dummy)
tm	Telegram (dummy)
w_paper	White paper (dummy)
usd	presale price in USD
tw	Twitter (dummy)
fb	Facebook (dummy)
ln	Linkedin (dummy)
yt	Youtube (dummy)
gith	Github (dummy)
slack	Slack (dummy)
reddit	Reddit (dummy)
btalk	Bitcointalk (dummy)
mm	Medium (dummy)
nr_team	Number of Team members
adv	Existence of advisors (dummy)
nr_adv	Number of advisors
project	Official name of the ICO
nr_tm	Number of users in Telegram
tot_token	Number of Total Tokens
Pos_Bing	Standardized number of positive words for BL list
Neg_Bing	Standardized number of negative words for BL list
Sent_Bing	Standardized sentiment for BL list
Pos_NRC	Standardized number of positive words for NRC list
Neg_NRC	Standardized number of negative words for NRC list
Sent_NRC	Standardized sentiment for NRC list

Methodology

- Using supervised classification models we will get insights for discriminating and classifying ICOs by their probability of success.
- At the same time, text mining methods will be the tools for dealing with the large corpus of text coming from the Telegram chats and the white papers.

Analysis – I Logistic Regression for Successful ICOs

Logistic regression aims at classifying the dependent variable into two groups, characterized by a different status [1=success vs 0=scam or 1=success vs 0=failure]:

$$ln(\frac{p_i}{1-p_i}) = \alpha + \sum_j \beta_j x_{ij}, \quad (1)$$

where p_i is the probability of the event of interest, for ICO i, $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})$ the J covariates from which the probability of success (or scam) can be obtained as:

$$p_i = \frac{1}{1 + exp(\alpha + \sum_j \beta_j x_{ij})}, \quad (2)$$

Analysis – II Multilogit Regression

Since the target variable is naturally categorized according to 3 classes, success, failure and scam we extend the aforementioned binary logistic regression to a multinomial one. Such model assesses all the categories of interest at the same time as follows:

$$ln(\frac{p_k}{1-p_K}) = \alpha_k + \sum_j \beta_k x_{ij}, \qquad (3)$$

where p_k is the probability of kth class for $k = 1, ..., K$ given the constraint that $\sum_K p_k = 1$.

Textual Analysis

We have applied a Bag of Word (BoW) approach, where a text is represented as an unordered collection of words, considering only their counts in each comment of the chat.

The word and document vectorization has been carried out by creating a Term Document Matrix (TDM).

Classical text cleaning procedures have been put in place like: stop-words, punctuation, unnecessary symbols and space removal, specific topic words addition.

For descriptive purposes we have used wordclouds for each and every Telegram chat according to the general content and to specific subcategories like sentiments and expressed moods.

Textual analysis – I

Figure: Wordcloud with negative words – failed – scam – success

Textual analysis – II

Figure: Wordcloud with positive words – failed – scam – success

Sentiment Analysis

We decided to focus on a dictionary based approach, adapting appropriate lists of positive and negative words relevant to ICOs topics in English language. We employ 3 vocabularies from the R package 'tidytext':

- AFINN from Finn Arup Nielsen;
- BING from Bing Liu and collaborators;
- NRC from Saif Mohammad and Peter Turney.

By applying the above lexicons, we produce for each and every ICO a sentiment score as well as counts for positive and negative words. All these indexes are used as additional predictors within the logistic models.

Sentiment Analysis - II

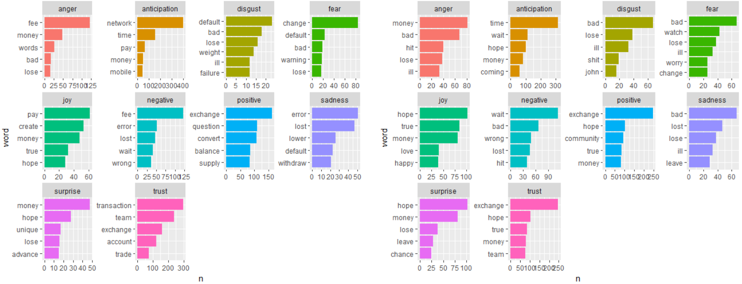

Figure: NRC emotion lexicon – success – scam

Results - I

Table: Results from Logistic regression on Success/Failure

	Dependent variable:
	class2
tw	2.481*
	(1.381)
Paper_du	1.351**
	(0.635)
nr_adv	0.461***
	(0.135)
nr_team	0.233***
	(0.088)
Sent_NRC_sc	2.187***
	(0.595)
Constant	−3.601**
	(1.458)
Observations	196
Akaike Inf. Crit.	89.41
McFadden pseudo R^2	0.63
McFadden Adj. pseudo R^2	0.57
Cox & Snell pseudo R^2	0.49

Results – II

Table: Results from multilogit regression: failure and scam compared to success

	Dependent variable:	
	f	sc
	(1)	(2)
Oweb_dum	-1.962^{**}	0.093
	(0.977)	(0.773)
adv_dum	-0.899	-1.707^{***}
	(0.809)	(0.571)
Paper_du	-0.728	-2.158^{***}
	(0.915)	(0.657)
Sent_NRC_sc	-1.390^{*}	-2.606^{***}
	(0.731)	(0.703)
Constant	-0.628	-0.572
	(0.997)	(0.925)
Akaike Inf. Crit.	166.339	166.339
Pseudo R square	McFadden 0.43 - McFadden Adj. 0.36- Cox & Snell 0.44	
Note:		*p<0.1; **p<0.05; ***p<0.01

Variance inflation factor

Table: VIF index for logistic regression model

tw	Paper_du	nr_adv	nr_team	Sent_NRC_sc
1.229	1.033	1.067	1.053	1.228

Table: VIF index for multilogit regression model

Oweb_dum	adv_dum	Paper_du	Sent_NRC_sc
6.395	2.207	3.822	7.034

Preliminary conclusions and ongoing research

From the logistic regression the relevant variables are: the presence of a white paper, of a Twitter account, number of elements of the team, number of advisors, and scaled sentiment score.

From text analysis: the net sentiment based on NRC lexicon has a positive impact in discriminating success ICOs from failure and scam ones.

From the multilogit regression we report results for fraudulent and scam ICOs compared to successful ones.

Preliminary conclusions and ongoing research

This paper represents a preliminary work and we are running a more detailed and complete NLP analysis by:

- increasing the size of the sample bu using the API access to the IcoBench Platform, and therefore analyzing the 5000 projects published there.
- refining the sentiment analysis and the dictionary based method.
- through topic modelling we aim at producing a quality index for white-paper to be included in the classification models, as a possible driver of success and/or scam activity.

Thank you for your attention.

Appendix

Table 6 Logistic regression full model

	Dependent variable:
Variables	Coeff.& s.e.
	class2
Oweb_dum	0.096
	(1.155)
Code_dum	−0.416
	(0.923)

Appendix

Continuation of Table 8	
Variables	Coeff.& s.e.
Telegram_du	−0.492
	(1.073)
tw	2.858
	(1.773)
fb	0.107
	(0.817)
ln	−1.236
	(0.927)
yt	1.495
	(1.002)
gith	−0.432
	(0.865)
slack	−0.915
	(1.105)
reddit	−0.067
	(0.902)

Appendix

Table 7. VIF index for the full model

Oweb_dum	Code_dum	Telegram_du	tw	fb
2.034	1.394	2.499	1.572	1.419
ln	yt	gith	slack	reddit
2.058	1.649	1.805	1.348	2.009
btalk	ew	mm	Neg_Bing_sc	Pos_Bing_sc
1.585	1.000	2.003	73.564	9.982
Neg_NRC_sc	Pos_NRC_sc	Paper_du	nr_adv	nr_team
70.371	12.830	2.255	1.703	1.809

Appendix

Table 9. VIF index multionomial full model

Oweb_dum	Code_dum	Telegram_du	tw	fb
15.583	3.165	9.422	126.167	5.164
ln	yt	gith	slack	reddit
3.883	2.276	4.194	2.408	4.640
btalk	ew	mm	Neg_Bing_sc	Pos_Bing_sc
3.443	1.007	6.435	575.279	1, 293.780
Neg_NRC_sc	Pos_NRC_sc	Paper_du	nr_adv	nr_team
241.850	770.032	7.955	3.908	8.418