Contents

1. Introduction

ICO is the abbreviation of Initial Coin Offering, which is similar to IPO (Initial Public Offering), is a unique way of financing for those projects in early stage, especially in blockchain industry. The “coin” in ICO actually refers to cryptocurrency, or token. In ICOs, such kind of tokens is the return for the investors, compared to securities in traditional IPOs. Most of the ICOs are financed by Bitcoin or Ethereum, and the investors can buy Bitcoin or Ethereum with legal tenders, and then invest in ICOs for certain amount of tokens, which could be traded in digital currency exchange after the financing is over.

The nature of ICO is still a way of financing, and is featured for rapid financing even if the projects are still at starting stage and even without any information of performance, profit or market scale. In this aspect, ICO is really different to IPO, because the companies which are eligible for IPOs should be mature companies and meet certain conditions set by local government or laws. However, IPOs are regulated by sound laws and regulations, such as JOBS ACT in the U.S. and Securities Act in China, while ICOs are not regulated by any laws or government sectors at all. That is another difference between ICOs and IPOs. In addition, the entities of ICOs are usually foundations or non-profit organizations and everyone can invest in, while the entities of IPOs are strictly regulated and the investors are also limited to some extent.

Nowadays, the scale of ICOs are growing rapidly and the cryptocurrencies are getting more and more popular. Although ICOs has both positive side and negative side, it does offer a big and uncertain market for investors to speculate in.

In this report, we are going to introduce seven variables which can to some extent predict the success or failure of an ICO before it starts, using and applying the relevant text mining and manipulation knowledge we acquired from the course. The seven variables are Sentiment Analysis GI, KWIC(Key Words in Context), Investors’ Passion, Team Size, Advisor Board, Document Clustering and Readability.

2. Preliminary Data Processing

According to the definition of "Success in ICOs", in the given 337 ICOs, firstly we filtered out those successful ones, in which the tokens managed to be listed within 60 days after the start. Then we also filtered out those unsuccessful ICOs for the comparison to successful ones.

Because we are only provided with whitepapers of 220 firms but we have 337 firms to analyze, we wrote a simple crawler to download whitepapers of those firms without whitepapers provided. We downloaded whitepapers from a website named ICOSBULL <https://icosbull.com>. Firstly we formed the download URL by grabbing the target part from source code of each firm's website, and used a loop-over to download them. Then we created a tibble named "final.ICO" to store the data we got. Finally we managed to collect 328 whitepapers for the ICOs we need to research. When we study the whitepapers we got, we found that some of them mostly consist of pictures, where the textual data is very hard to read in R, hence we excluded that kind of whitepapers.

Below is the basic statistical information of the whitepaper documents:

nr of doc	Min word	Max word	Avg word	Median word	Sd word
315	566	47624	7589	6414	5470
nr of doc	Min sentence	Max sentence	Avg sentence	Median sentence	Sd sentence
315	34	2673	440	348	398
nr of doc	Min size	Max size	Avg size	Median size	Sd size
328	120	448896	60403	49972	48103

Table 1: Whitepaper Document Summary Statistics

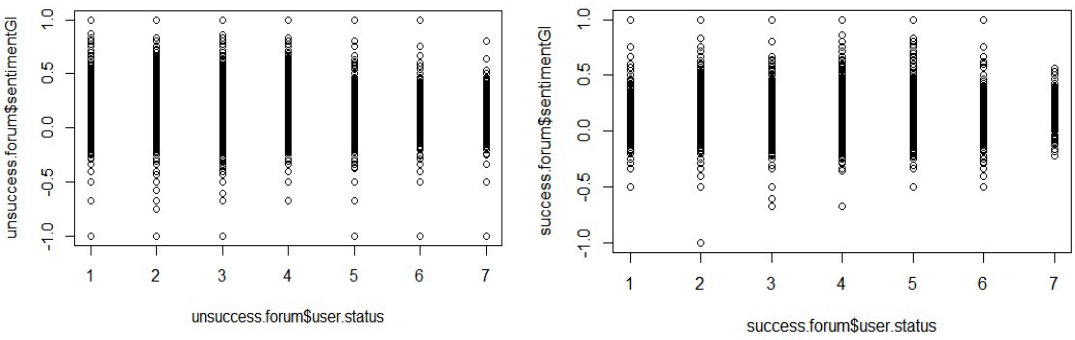
3. Variables to Predict Success

3.1 Sentiment Score (GI)

Firstly we divided the content of tibble "forum" into successful ICOs and unsuccessful ICOs according to the column "token". Secondly we remove the text posted by the token issuer in case it causes some disturbances of the sentiment score. Thirdly we filter out those texts posted before the start dates of ICOs because we only use the text information before the ICOs to predict success or failure of ICOs.

We also add a column stands for time difference of the text post date and ICO start date to see if the date of text post early or late have something to do with the sentiment score, we use quantiles, but we find that it is of nonsense.

Then we try to analyze the sentiment score of different status's users. Following is the plots of GI sentiment score of forum text about successful ICOs and unsuccessful ones, and the x-axis is user's status in the forum:



Graph 1: Sentiment Score GI for users in all status

It is clear that the standard deviation of sentiment scores of text vary significantly as the user status differ. Higher the user status is, less the standard deviation is. In addition, generally the users with higher status in the forum tend to be more experienced, so studying the text posted by higher standard should be more meaningful.

Hence, we add another constraint when filtering the forum text, we only filter out the texts posted by users with a certain high status, and the detailed data is in the following table:

	Min.	Max.	Median	Mean	s.d.
Text from users of all status					
success	-1.0000	1.0000	0.1333	0.1449	0.1595
unsuccess	-1.0000	1.0000	0.1429	0.1591	0.1350
Text from users of status >= 4					
success	-0.6667	1.0000	0.1364	0.1473	0.1311
unsuccess	-1.0000	1.0000	0.1364	0.1460	0.1388
Text from users of status >= 5					
success	-0.5000	1.0000	0.1364	0.1453	0.1359
unsuccess	-1.0000	1.0000	0.1296	0.1369	0.1350
Text from users of status >= 6					
success	-0.5000	1.0000	0.1333	0.1415	0.1273
unsuccess	-1.0000	1.0000	0.1250	0.1330	0.1306

Table 2: Summary Statistics of Sentiment Score GI (Filtered Status)

It is obviously that we cannot include all the texts posted by users of any status because in that way the sentiment score does not make any sense, as shown in the above table, the median of sentiment score of success ICOs is even lower than that of unsuccessful ones. Hence, we regard the constraint of filter users of a forum status equal or higher than 5 as reasonable, that is, in order to predict the success or failure of an ICO, we can gather the relevant text posted by the users with a status of no lower than 5 on the forum, of course we should also exclude the text posted by the issuer, and then do a sentiment analysis using GI dictionary.

For the data filtered out, we also conducted a Welch two sample t-test, and the result is: t = 3.0962, p-value = 0.001969. Hence, the means of GI sentiment score between successful and unsuccessful ICOs are significantly different at 95% confidence level. For this reason, statistically, GI sentiment score of forum texts should be a powerful indicator to predict the success or failure of an ICO. More specifically, if the sentiment score of one specific ICO is above 0.13, the ICO is very possible to be a successful one, and otherwise, it has great possibility to fail.

3.2 KWIC (Key Words in Context)

Given the features of “high risk” and “lack of regulations” of ICOs, we suppose that the frequencies of some keywords in the texts on forum, such as fraud and MLM, may

have a close relation to the success or failure of a certain ICO (Michiel 2018, para. 7). To verify our hypothesis, we firstly set a list of keywords, including the following words: scam, con, MLM, fraud, trick, deceive, lie, swindle, cheat. The keywords written in code are as followed:

^scam.*?|^con.*?|^mlm.*?|^fraud.*?|^trick.*?|^deceiv.*?|^lie.*?|^swindl.*?|^cheat.*?

According to the previous section of sentiment score analysis, we should only consider the forum texts posted by users of no lower than status five, so in this section, we also filtered the texts using this criteria. We firstly make two tibbles for all the words appear in the forum texts regarding both successful and unsuccessful ICOs, and did some cleaning steps removing whitespaces and stop words. Then we used regex to count the frequencies of keywords. Finally we found that in the text sample of successful ICOs, the frequency of keywords is 10 times per 1000 words, while in the text sample of unsuccessful ICOs, the frequency is 11 times per 1000 words. Overall the keywords regarding MLM or fraud appear more frequently in unsuccessful ICOs, but actually the difference in frequencies is very insignificant, so in practice, this KWIC indicator is not so strong but should be taken into consideration, for example, when the investors are evaluating an ICO project, they should do some research on the relevant information disclosed and rate the risk of the project, especially in the MLM or fraud aspect.

3.3 Investors' Passion

Investors' passion is also a useful indicator to predict the success or failure of an ICO project. If an ICO is frequently talked about in a forum or other kinds of social media, it is more likely to succeed because more discussion and mention means more attention and more chance to attract potential investors.

In order to evaluate investors' passion, we used the text data in tibble "forum". Firstly we counted the frequencies of discussion of all the tokens and ranked it by decreasing order. Then we used quantile to see the overall distribution of discussion frequencies of successful ICOs. As shown in the below table, which indicates that more frequently a token is discussed, more possible it will succeed.

	Freq Group	Token Nr
1	[1,18]	3
2	(18,50.8]	6
3	(50.8,164]	13
4	(164,550]	10
5	(550,1030]	12

Table 3: Successful Token Nr in Freq Group of Discussion (quantile)

Next, we conducted a Welch two sample t-test, and the result is: t = 1.9456 and p-value = 0.02837, and means of successful and unsuccessful tokens discussion frequencies are 370 and 250 respectively, which indicate that the difference between the mean frequencies of successful tokens discussion and unsuccessful tokens discussion on the forum is significant at 95% confidence level, thus proving the prediction power of investors' passion. More specifically speaking, if an ICO token is talked about for over 370 times on the forum before it starts, it should have great possibility to gain success.

3.4 Team Size

Also, we took team size into consideration. If the an ICO team only contains very few members, even though those people could be very talented and they hold brilliant ideas, they might still are unable to create a successful project, because they need other members to make a joint effort (Andrey 2018, para. 5). For this reason, we evaluated the team size of both successful and unsuccessful ICOs to see if we can find a significant difference and can be a strong prediction indicator.

Firstly, we used regex to count the team members numbers of both successful and unsuccessful ICOs. Following is the basic description of the data:

	Min.	Max.	Median	Mean	s.d.
success	3	34	10	12	8
unsuccess	1	12	8	9	5

Table 4: Summary Statistics of Team Size (Both Success and Failure)

It is obviously that, according to the sample data, the successful ICOs tend to include more team members than the unsuccessful ones do. The maximum team size in unsuccessful ICOs is 12, equals the mean team size of successful ICOs. In addition, we did a Welch two sample t-test, and the result is: t = 2.2571, p-value = 0.0285, which

indicates that at 95% confidence level, the mean team size of successful ICOs is significantly different to that of unsuccessful ones. So we regard team size of an ICO as an effective predict criteria of success or failure. More specifically, we think that if the group size of one ICO project is less than 10, the investors should be alert that there could be great possibility to fail. However, there actually are certain number of cases where ICOs succeeded with small team size, we suppose that could be something to do with project budget. In the sample data, mean budget of successful ICOs with no less than 10 members is 2792000 thousand and that of those with less than 10 members is only 931500 thousand. So ICOs of small team size might be accepted, for some small budget ICO projects.

3.5 Advisor Board

Evaluating the advisor members of an ICO project can also predict the success or failure. Investors are more likely to invest in an ICO project equipped with famous names in the advisor board since they are more likely to be convinced that such project has higher possibility to achieve success under the instruction of famous and experienced advisor, which in turn makes the project get more money and more easily to success (Michiel 2018, para. 6).

Firstly, we grabbed names of famous advisors from a website named ICOHOLDER, <https://icoholder.com/en/ico-advisors>. Here we assumed those advisors appear in the first four pages of this website and have involved in at least six ICO projects as famous advisors. Then we filtered out 120 famous advisors. Next, we counted the numbers of famous advisors in each ICO project and made a frequency table:

	0	1	2	3	4	5	6
success	36	7	3	3	0	0	1
unsuccess	225	21	5	9	4	1	0

Table 5: Frequency Distribution of Famous Name

As shown in the above table, 18% of successful ICO projects are equipped with famous advisors while for unsuccessful ICOs, only 16% include famous advisors. In order to

test the significance of the difference, we conducted a Chi-square test, the result is: X-squared = 4.9323 and p-value = 0.0263. Hence the difference of famous advisor frequencies between successful and unsuccessful ICO projects is significant at 95% confidence level. For this reason, we confirm that whether a ICO's advisor board includes famous name should be a effective indicator to predict the success or failure.

3.6 Document Clustering

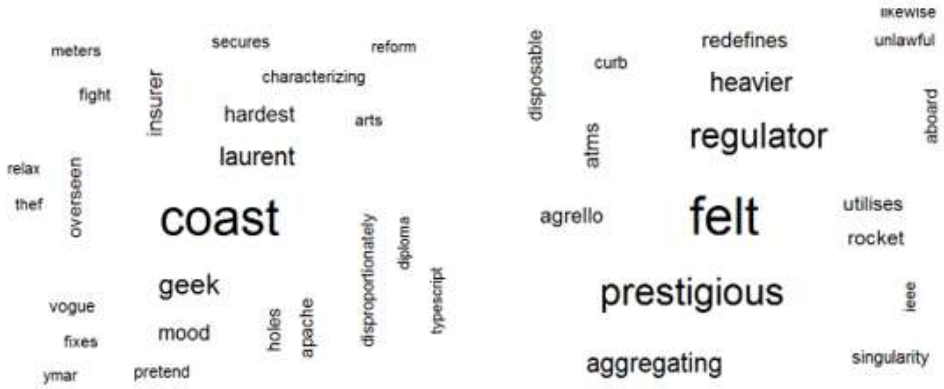
In this section, we used both LDA and K-Means to classify the topic of each whitepaper, by which classify ICOs based on industry they mainly focused on. Then we focused on the successful ICOs and unsuccessful ICOs separately, in order to find out which industry successful firms are most likely assigned to.

3.6.1 LDA

Firstly we removed those firms whose number of words are less than 300, because content of these firms are mostly pictures. Then we made contents of all remaining firms into a document term matrix. Here we set that terms appear less often in documents than three times and more often than 30 times are discarded. Then we used LDA function to construct a LDA model. We separated the topics into 25 categories (https://icowatchlist.com/statistics/categories, 2018, para. 1).

However, after we checked most frequent words of each topic, we found it's very hard to define which industry this topic describes. For instance, for topic one the most frequent words are coast, geek, lauren, insurer, hardest, mood. For topic two are felt, prestigious, regulator, heavier, aggregating, atms, redefines, agrello, coincidence, rocket, as shown in graph 2. We supposed several factors that could lead to such problem. First, a whitepaper contains not only introduction of this project, but also other information regarding roadmap, team member profiles, funds allocation and so on. Hence, it makes sense that a large proportion of words are unrelated to the industry of a certain ICO project but are common in other whitepapers. Second, the features of the service provided by different ICO projects may lead to the classification of industries

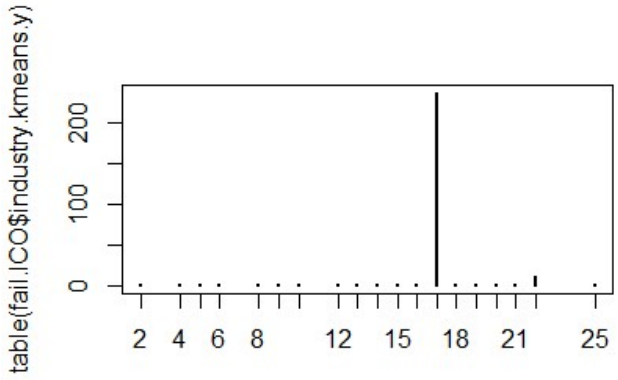
overlap, thus causing inaccuracy (https://icowatchlist.com/statistics/categories, 2018, para. 1). In addition, it makes us difficult to write a crawler and grab only the introduction part from whitepapers due to the lack of regulated format of whitepaper.



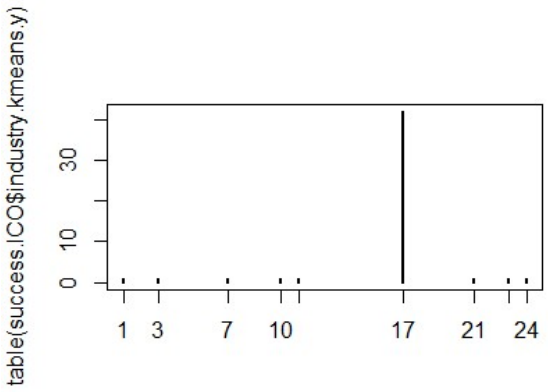
Graph 2: Word Clouds for Topic 1 and Topic 2

3.6.2 K-Means

Then we do a K-means to further proof our analysis in LDA part. As shown in graph 3 and graph 4, most projects are assigned to cluster 17 no matter a project is successful or fail. What's more, WCSS of our K-means analysis is 954683, which is very high.



Graph 3: Frequency of Each Unsuccessful ICO's assignment



Graph 4: Frequency of Each Successful ICO's assignment

To conclude this section, we regard using whitepapers to cluster documents as an ineffective way to predict the success or failure of an ICO project because of the unregulated format of whitepapers, and ambiguous classification of assigned industries of ICO projects.

3.7 Readability

Term	Frequency	syllable
biomedical	9	5
singularity	9	5
democratically	7	6
managerial	5	5
cooperative	4	5
evolutionary	4	6
appreciation	3	5
biological	3	5
inflationary	3	5
sophistication	3	5

Table 6 Complex Words in Whitepapers

To assess readability of whitepaper, we chose to use number of verbs, Fichtner's C, Yule's K and document length. We didn't choose variables related to complex words since whitepaper has a high percentage of complex words concerning computer science, finance, high technology and so on, which are not complex for investors and

professionals (shown in Table 6).

Then we used two for-loops to loop over successful firms and fail firms separately, adding calculate results to new column we create. Summary statistics of these four variables are as follows:

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Verbs	182	1601	2050	2559	3386	7537
Fichtner's C	6.82	75.75	103.97	100.7	129.32	246.32
Yule's K	38.23	79.01	87.53	100.02	98.15	297.66
Doc Length	7.78	8.61	8.87	8.95	9.34	10.31

Table 7: readability statistics for success firms

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Verbs	10	1119	1791	2024	2618	13285
Fichtner's C	0.15	80.82	105.07	108.14	134.63	424.04
Yule's K	37.92	77.52	89.11	99.28	103.89	1167.95
Doc Length	6.34	8.26	8.74	8.67	9.15	10.771

Table 8: readability statistics for fail firms

As shown in Table 7 and Table 8, average number of verb and length of success firms are greater than fail firms. We think this is because all ICO projects are funding for start-ups, which normally have more complex and uncertain business models. For investors, besides whitepapers, there are only few official information concerning this ICO project, so they need to find as much information as they can from whitepaper. As a result, if the length and the number of verbs of a whitepaper are higher, it will contain more information about this project and investors are more likely to think this project is reliable.

We conducted Welch two sample t-tests for the four variables, and the results showed that the p-values of variable "document length" and variable "number of verbs" are 0.004125 and 0.02467 respectively, which indicates that the means of "document length" and "number of verbs" between successful and unsuccessful ICOs' whitepapers are significantly different at 95% confidence level. Hence, statistically, the variable

“document length” and variable “number of verbs” are eligible to predict the success or failure of an ICO project.

4. Conclusion

To conclude, statistically, out of our seven variables, only the variables “Key Words in Context”, “Investors’ Passion”, “Team Size”, “Advisor Board” and “Readability” have prediction power of success or failure of an ICO project. However, in real word cases, those variables with statistical significance are not necessarily practical and effective to predict success or failure of ICO. Investing in this unregulated and newly-developed market should call great alert and make a overall evaluation of all the information available in many aspects, including those variables we introduced in this report.

5. Reference List

Andrey, S. 2018, “9 Keys for ICO Team Evaluation”, viewed 3 November 2018, < https://hackernoon.com/9-keys-for-ico-team-evaluation-fcfd537b64fc>.

ICO Watch List. 2018, “ICO Statistics – By Industry”, viewed 4 November, < https://icowatchlist.com/statistics/categories>.

Michiel, M. 2018, “10 Keys for Evaluating Initial Coin Offering (ICO) Investments”, viewed 3 November, < https://cryptopotato.com/10-keys-evaluating-initial-coin-offering-ico-investments/>.

Shuqing, B., Zhenpeng, D. & Fei, L. 2018, “ICO-rating, A Deep Learning System for Scam ICO Identification”, *arXiv:1803.03670v1[cs.CL]*, 8 March, 2018.