UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Similarity analysis in the context of ICO whitepapers

Author(s): Admir Tairov
Supervisor(s): James Thewissen
Academic year 2020-2021
Dissertation for the master of Management Science (M120)
Master subject and focus Corporate & International Finance
Daytime schedule

Abstract :

The basis of the paper is to study the innovation aspect of initial coin offerings (ICOs), more specifically whitepapers, and to see how innovation has an impact on fundraising and success. We filtered these text documents into different topics using Topic Modeling to analyse the similarity of these documents. The filtering was done with Latent Dirichlet Allocation (LDA) and the categorisation of topics was done by human judgement. Using this method, we were able to analyse the similarity of the whitepapers and the results show that the similarity of the whitepapers has a positive impact on fundraising provided that the similarity is not too great. In other words, there is a curvilinear relationship between similarity and fundraising. On the other hand, a very low textual similarity implies a higher chance of encountering a potentially fraudulent ICO, thus fraudulent projects bring innovation in an industry with few competitors to attract investors looking for rapid profit.

Keywords: Initial coin offering, ICOs, whitepapers, similarity, innovation, textual analysis, regulation

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique
www.uclouvain.be/lsm

TABLE OF CONTENTS

1. INTRODUCTION

Since the explosion of Bitcoin's value in 2018, we have seen a new trend related to cryptocurrencies, the value of digital gold and its competitors has reached even higher values in 2021 and despite their high volatility, the demand is only increasing. As a result, a slew of projects has been launched on a massive scale through fundraising that can be exchanged for cryptocurrencies. More commonly known as ICOs, Initial Coin Offerings, have some characteristics as IPOs but instead of acquiring shares in a company, the investor will receive digitised tokens that will be used in the use of a product or service that the company offers or these tokens can be considered as a share that the investor owns in the company or project

As of June 2018, the ICObench platform had over 5368 ICOs. The three countries raising the most money are The United States, Switzerland and Singapore with $2.2 billion. (Chuanjie, Koh & Griffin, 2019). In 2019, there were over 5711 ICOs published with over 1785 ICOs raising funds. The US retains its place as the country with the most fundraising, Singapore comes second and the British Virgin Islands come third ahead of Switzerland. The top categories for 2019 were Cryptocurrencies $304.4M, Platform $280.3M, Investment $263.7M, Banking $241.1M. (ICOBench.com). Singapore is the new forerunner of digital currency trading in Asia as China has banned ICOs in its territory as most ICOs were fraudulent fundraisers (Raza, 2018)

This new, easier and faster way to raise money has created a shake-up in the investment world. Indeed, in 2019, ICOs raised about $31 billion in funds. This is why we are seeing more and more start-ups wanting to join this fundraising system. Indeed, when a start-up wants to raise funds, it creates a whitepaper which is a document that aims to describe the project containing key information such as the protocol or the service used, the characteristics of the token as well as the way it will be distributed and also the presentation of the team which will allow to reassure investors about their competence and their added value.

However, like any new financial innovation, ICO's have some issues. First of all, we can highlight the loss of information but also its asymmetry. Indeed, the whitepapers are not always totally explicit because the projects are new and are part of the launch or development phase of the company and can therefore evolve over time. As this phase is very sensitive to change, some information may not be disclosed. In addition, most of these companies are considered small ventures, the entrepreneurs do not have much experience or a concrete project.

A second point to highlight is that ICOs can have regulatory problems. There are many ICO projects launched every day. Still implies that the financial authorities do not have control over all these ICOs and the liquidity of the tokens is not always guaranteed as some projects may be fraudulent investments. In case of bankruptcy, unlike shareholders, token holders have no possibility to get their money back as they do not own any share in the company, which makes the investment very risky. There are many risks for investors associated with this fundraising. It can be a scam where the issuer after selling their tokens decides to resell the recovered cryptocurrencies and abandons the project. The volatility of tokens and cryptocurrencies should also be taken into consideration as this can create a sharp drop in the value of investments. Another source of risk is money laundering from criminal sources. So, the entrepreneurs' goal is mainly to make their ICO "regulatory compliant" which is a standard that meets the standards set for market authorities.

The basis of the article is to investigate the innovation aspect of these ICOs and to see how the similarity between whitepapers has an impact on the amount raised. However, there is no research on the innovation of ICOs, But we can make some links with crowdfunding, which is a fundraising system with certain similarities to ICOs. Indeed, in the context of ICOs, investment differs from traditional new venture investment. Investors are normal people who are not all professionals and have a lack of financial or industry expertise and are not looking for huge returns, they behave in a unique way by trying to believe in a project that can yield a return on investment in cryptocurrencies. The same profile can be observed as for crowdfunding investments (Chan & Parhankangas, 2017) It should be taken into consideration that their decisions are not always rational as cryptocurrencies are very volatile and have little stability over time. As consumers, investors are likely to evaluate how innovation affects their potential profits and returns (Schmidt & Keil, 2013). Using consumer insights, this study provides information on how innovation and similarity can affect the outcomes of ICO funding.

Based on data collection of 2500 ICOs, the way we will analyse the innovation is based on the similarity with previous projects. Our method will be based on machine learning which allows us to identify degrees of similarity between different whitepapers. There are many similar ICO projects. Yet implies that it is interesting to see how these projects differ from others, it is also useful to see how their whitepaper information is communicated. Therefore, all these ICOs have a different nature, but they all have certain points in common, not only in their origin but also in their currency. In this work we will particularly analyse the similarities of the categories written on the whitepapers but also the impact of innovation on fundraising.

2. CONTRIBUTION

This paper makes several theoretical contributions. It examines the relationship between innovation and ICO funding outcomes. Investors and entrepreneurs often suffer from a pro-innovation bias, viewing innovation as a universally desirable characteristic (Kimberly, 1981), but we'll see that incremental innovation has a positive impact in whitepapers. Indeed, the literature on entrepreneurial finance provides evidence that early-stage investors prefer innovative firms, in order to capture new or existing markets and generate extraordinary returns. (Kortum & Lerner, 2000; Metrick & Yasuda, 2010). Incremental innovation is a more familiar dimension because the application of existing knowledge, skills and technology offers a familiarity advantage and research shows that consumers feel more comfortable with products but also prefer to take fewer risks by buying familiar products rather than familiar unknown products (Park & Lessig, 1981; Prakash & Thukral, 1984).

In a second step, we will provide an answer to the determinants of the key success factors of ICOs that control the similarity variables, in other words, we will test different regression configurations in order to see which one can increase the positive impact of similarity on the dependent variable amount raised. Other previous research has been done to analyse the key success factors of ICOs (Aslan, Sensoy & Akdeniz, 2021) as a whole, in our approach we will study only the similarity as component"é& to see if it is relevant. The topic of similarity of ICOs is currently under-explored, so we find it interesting to look at ICOs from a different perspective by taking into consideration whitepapers as an evaluation tool. We will see that the similarity can have both a positive and a negative effect on the outcome of a whitepaper, the question what is the impact of the similarity of ICOs on its fundraising.

Finally, entrepreneurs can use this study to better understand the relationship between innovation and similarity in order to achieve better results in ICO funding and managing their interactions with investors more effectively. It will enable them to understand and be able to use the similarity score in order to evaluate different ICOs, competing or not. This similarity score will contribute to the literature to identify ICO scams and copycats.

3. LITERATURE REVIEW

The literature on ICOs is becoming more and more consistent. Many articles focus on how ICOs could be beneficial for entrepreneurs. For example (Catalini & Gans, 2018) show that ICOs can help entrepreneurs to collect information about consumers' willingness to pay which allows for higher returns compared to equity finance. Similarly (Chod and Lyandres, 2018) present a theory of agencies and elaborate on why ICOs may have an advantage over equity financing and explain why ICOs may have an advantage over traditional venture capital financing. Others present models to explain how tokens and ICOs can create value on networks by resolving coordination failures during platform construction (Li & Mann, 2018). Other studies focusing on the success of ICOs show that higher ratings from cryptocurrency experts, the addition of pre-sales as well as the implementation of bonus systems and the reduction of token selling time bring more chances of success (De Jong & al, 2018).

In the introduction we discussed some of the limitations of ICOs that make them unsafe, such as information asymmetry and regulatory issues, which is why we would like to draw a parallel with stock exchange shares which have a more advanced regulatory system as they are more mature financial products.

In this part of the work, we will introduce concepts taken from articles in the economic literature which allows us to understand two important notions that will be addressed in this master thesis. Firstly, we would like to highlight those regulatory issues are not only addressed in the context of ICO's. Indeed, we will share with you an article that explains the difficulties that IPOs have to be regulated by the authorities. ICO's are similar to IPO's in that they have tokens instead of dividends. Even if the remuneration system is different because it has a separate technology, it is still interesting to draw a parallel with IPOs.

The two types of projects with equivalent ambitions are different in the sense that listed companies are more likely to succeed when they are centralised and fully controlled by a company whereas ICOs will be very successful when the notion of freedom is applied and there is no central authority.

In a second step, we will discuss how whitepapers are analysed using artificial intelligence and tools such as Topic Modeling, it will be useful to understand the process of textual analysis as we need it for our research.

3.1 Regulation

As mentioned in the introduction, ICOs have certain similarities with IPOs. What makes ICOs different from traditional financing methods is that the investor does not receive a security or a title of ownership, but a token that gives the investor future rights of various kinds to the application developed by the provider. The token mainly provides access to future services, and then allows financial and political rights to be acquired. Once the tokens are generated, investors can buy or sell them on the secondary market if there is reasonable liquidity. The value of the token can be subject to fluctuations; the evolution of its value depends on the performance of the project (Adhami & al, 2018).

Nevertheless, it is wrong to say that they are both similar types of fundraising. Indeed, even if the value of the token and the value of the share fluctuate over time, the market in which the shares and the tokens evolve is not the same. On the one hand, ICOs are mainly on an unregulated market and the companies are recently compared to IPOs which are on a regulated market and the companies that want to enter have already a few years behind them as well as a great stability.

A research paper shows the regulatory difficulties and information asymmetry faced by A-listed (nonfinancial) companies on the Chinese stock exchange. Their topic addresses the similarity of financial reporting and the likelihood of administrative sanctions. They analysed the annual reports, but more specifically the management discussion and analysis, which describe the company's performance. They found that similarity to the previous MD&A can lead to fraud or administrative sanctions. (Qiana & Zhu, 2019)

This study can indirectly contribute to the ICO regulatory bodies to understand disclosure and administrative sanctions in case of fraud. It can provide information for creating or refining disclosure policies as the ICO regulatory system is in a constant state of evolution.

Between 2008 and 2017, there were more than 4011 sanctions on A-listed companies with 1644 companies punished by regulators for serious misconduct such as unauthorised use of funds, illegal information disclosures or delayed disclosures. Information asymmetry can be mitigated with better disclosures to protect the legitimate rights and interests of shareholders (Hu & Tan, 2013). It is also important to mention that being an emerging market economy, China's judicial system and market supervision are in the improvement stage and are not yet perfected to have the best way to regulate these listed companies. (Jiang & Kim, 2015; Kato & Long, 2006).

The value of non-financial disclosure is as a useful resource for investors to help them better assess the value of the company and reduce analysts forecasting errors, it also improves audit quality and investment efficiency which is an element of risk (Cheng, Tan, & Liu, 2012). Their characteristics such as tone, length, readability and similarity of financial reports can affect the decision-making of these actors but also of external users, especially stock returns or trading volume. (Lawrence, 2013; Loughran & Mcdonald, 2014; Reuven et al., 2011).

By providing more information in financial reporting, it will help to reduce information asymmetry and help information users understand a company's strategic and business decisions. Information such as research and development activities, social responsibility and accounting policies provide value to investors, analysts and other stakeholders.

According to the authors of the article, there are two reasons for the textual similarity of the current annual report to the previous one. The first is that the disclosure behaviour of companies is said to be negative. Indeed, the information in the previous MD&A is copied because the company does not add any objective analysis of business operations or additional information.

The second reason is that few changes have occurred in the current or future period of the company. No recent investment projects are added, there are no planned mergers and acquisitions and no information about research and development. As a result, a strong similarity in the management report leads to a poor impression of corporate governance, reliable accounting information and business performance to regulators. As a result, the risk of the company being sanctioned increases during the current period.

Their main hypothesis is therefore whether there is a positive or negative correlation between the similarity of a listed company's management reports and the likelihood of administrative sanctions being applied.

The findings of their work show that a high degree of similarity with the current period's MD&A compared to the previous period of listed companies is more likely to be investigated and sanctioned by regulatory authorities. These regulators take into consideration the information in the MD&A as well as other information outside the MD&A. The analysis also shows that regulators have a lower regulatory requirement for state-owned companies. Administrative sanctions are less applied when it comes to publicly owned companies. They also found that the readability of the text has no impact on the relationship between similarity and administrative sanctions.

Their paper can greatly help regulators to be fully aware of the importance of textual information disclosure by listed companies. Adding information to financial reports allows information users to better understand operational situations and also improves the information asymmetry between companies and external information users. Chinese regulators should improve the information disclosure requirements of companies. The same applies to whitepaper regulators for ICOs

3.2 Similarity

The topic of similarity in ICOs is not yet discussed in detail in the literature. We found one article (Florysiak & Schandlbauer, 2019) that deals with the topic of whitepapers and their similarity. We therefore analysed their work to obtain key information to share in this paper. They found that the similarity of whitepapers is related to different criteria such as belonging to the same industry or having national restrictions that regulate the way investors participate in ICO projects. Also, ICOs with a similar number of staff and a similar product or service idea are more likely to have similar whitepapers. Their contribution also focuses on the readability of these written documents in order to see the relation with the success of ICOs but also scams.

It should be noted that their database comes from the same source as ours and we have learned that ICOBench removes ICOs that are considered fraudulent or potentially fraudulent, which reduces the possible sample in the database. However, there are projects that fall below the threshold and may be present in the database. This brings them to 444 observations with 41 ICOs that exhibit fraud. They also found that the more images in the whitepapers, the more likely it is that there is potential fraud. Having a possibility to pre-order a token decreases the chance of a potential fraud. Being in an industry with few competitors can also increase the chances of potential fraud. Fraudulent ICOs want to attract more investors by showing themselves to be unique, which is why they do not use the same words as whitepapers of ICOs in the same industries.

3.3 Textual Analysis

The field of Topic Modeling and textual analysis in the field of ICOs is recent but there are some literary contributions. Chuanjie, Koh & Griffin, 2019 highlight how topic modelling can automate the classification of ICOs into distinct topics. In addition, they added the number of documents that are present in these categories. This model provides a textual representation of a document using probabilities and the results are presented in the form of text classification and filtering. Each document presents different topics where each topic is characterised by a distribution of words, the aim being to ultimately categorise the texts into several mutually exclusive classes. (Blei, Ng & Jordan, 2003).

The aim of this literature is to prove that the LDA model can effectively categorise ICO whitepapers into different topics, domains or industries. Several research papers have explored the application of LDA in the finance sector. Nevertheless, few research articles have used data science techniques to classify ICO whitepapers. One research article aiming to detect ICO scams used LDA to categorise 10 different topics, but labelling was not used as a main predictive feature by the authors (Bian & al, 2018)

The data in the article by (Chuanjie, Koh & Griffin, 2019) is extracted from ICObench, which is a platform that provides a variety of information and data on ICOs as well as links to the whitepapers. By extracting the full text of the whitepapers, they created a cloud of the most used words in these texts. Words such as "platform", "user", "data", "service", "system", "network", "contract", "exchange", "company" are considered as empty words as they have no impact on the information in the whitepapers. The documents are filtered a second time in order to obtain a greater disparity of keywords.

Then they created a table that groups the five most important words for each topic. The topics were manually studied in order to be the most likely to encompass the five words. The composition of the words in the whitepapers is important to provide similarities between the different topics of the whitepapers. For example, there are 46 documents with the words "energy", "mining", "power", "project", "production" and are considered in the category "Natural Resources: Energy, Precious Metals and Rare Earth". Using this table, they found that in 3 of the 10 subjects there were significant differences in the proportion of successful whitepapers compared to other whitepapers. They proved the validity of the LDA model for the analysis of ICO whitepapers by relating them to the success of ICOs.

4. METHODOLOGY

To better understand similarity and innovation in the context of ICOs, it is useful to conduct a textual analysis of the whitepapers. This analysis will allow us to filter, observe and quantify the information that has a similar essence between projects. Our main hypothesis is based on the fact that ICOs with similar whitepapers have a higher chance of raising funds. Nevertheless, it can be assumed that too much similarity leads to a decrease. We therefore expect a curvilinear relationship.

'Hypothesis: the similarity between whitepapers has a positive impact on the fundraising and token tradability and therefore guarantees ICO success.'

In this section, we will look at how to measure incremental innovation using similarity as an analysis tool. To do this, we will focus our methodology on Topic Modeling with LDA (Latent Dirchlet Allocation). This tool gives us another point of view on the way we approach the analysis of textual documents, it allows us to identify the topics discussed and it also gives us more information about the relationships between the different topics discussed in all the documents we propose.

The objective is to classify a collection of documents and to gather them according to the number of topics we want. This classification allows us to organise, filter and search for the information we need. (Blei, Ng & Jordan, 2003). This method is based on machine learning that models text corpus and other discrete data collections without supervision, i.e. the algorithm is run through the software R. and allows creating categories based on the filters we inject to it, but the results can be biased because we control the information we want to filter and the algorithm depends on human judgment. Therefore, we need to have a detailed approach to each element to get more meaningful results. Thanks to the results created by the algorithm, we can link these insights to external variables.

In other words, we will create tables that will allocate words found by the algorithm to topics, these words will be measured with probabilities to know if the word has some importance for this or that topic. The second step will be to observe the percentage of presence of each topic in the documents, which will allow us to see which topics are the most present in the documents. With this information we will be able to observe the similarities between the different whitepapers because they have topics in common in their text.

To implement this algorithm, we will work with the software R and we will propose you the different tables in the appendices of this article. We have at our disposal 2000 whitepapers from 2015 to 2020 in text format that we will inject into the software. These whitepapers come from the ICObench platform which has a large amount of data on ICOs. These whitepapers are not standardised and therefore have no rules in terms of length, style or content. (Florysiak & Schandlbauer, 2019)

The first problem we have at the moment is that the documents we have uploaded are in PDF format, they have to be converted into a text format and this transition cannot be totally accurate. This is why our role is to have a methodology that is as clean as possible to reduce the risk of errors and to try to get closer to reality.

4.1 Cleaning Process

In order to have more meaningful and easier to analyse results, we need to prepare our data. Text pre-processing is a fundamental requirement for textual analysis (Mayo 2017). To do this, we will clean the documents step by step with the software. It is important to know that the software does not handle special characters and does not translate, so we have to work only with whitepapers written in English. Then, we will have to standardise the text in order to have words without punctuation, without capital letters, without numbers and unnecessary spaces. Then we will lemmatise the words of the documents, which means grouping all words into one single family called item in order to have only one word to analyse. We must also remove stop words that are words that are so common such as "the", "is", "as", "is", "at", etc. that it is unnecessary to include them in the analysis.

Once the text was cleaned, we excluded words that have fewer than 3 letters and words that appear in fewer than 5 documents. We also used the “tf-idf” (term frequency-inverse document frequency) function which increases proportionally to the number of times the word appears in the document, but is compensated by the frequency of the word in the corpus, allowing us to adjust to the fact that some words appear more frequently in general.

We were able to produce a word cloud of the most used terms in all whitepapers. As you can see in Exhibit 1, there are words that are obviously repeated in all the whitepapers. Furthermore, some of these words are part of the essence of certain projects, notably for topics concerning the cybersecurity of cryptocurrencies, or the launch of a cryptocurrency with innovative technology. That's why it's worth keeping them to get an overview of the most used terms.

EXHIBIT 1
Word cloud of all the terms used for all whitepapers

4.2 Latent Dirichlet Allocation

Thanks to the cleaning process, we can have a first view on the terms that will be studied in this paper. The objective now is to use Topic Modeling which is a method for unsupervised classification of textual documents that will allow us to cluster our whitepapers into different topics and we have observed how each word is weighted in the different topics. In Exhibit 2, there is a list of the 15 most used topics in the whitepapers and it is interesting to note that some words are used in more than one topic such as 'node', 'trade' or 'coin' as these topics have similarities. Topics related to online services will have similar keywords to topics related to blockchain, as will topics related to financial products which will be comparable to words in the trading topic. For more information, please see *appendix 1.*

In order to identify the weight of each word, we have implemented a function called 'beta' which allows us to evaluate the presence of each word in a topic. In the appendices, you will find table 1 which shows the weight of each word in the 15 most frequent topics found in the whitepapers. The next step is to associate each topic with whitepapers. To do this, we need to use another function called 'gamma' which will be used to define the proportion that the topic has in our textual database. Each document has a proportion of topics and we will in the next section which topics are the most used and which have a positive correlation on fundraising.

EXHIBIT 2
The 15 most discussed topics in whitepapers

Topic	Terms	Subject
Topic 1	*node, device, cloud, server*	Online Services
Topic 2	*credit, loan, lend, consumer*	Financial Product
Topic 3	*coin, mine, gold, miner*	Cryptocurrency Mining
Topic 4	*ico, asset, trade, crypto*	Payment Technology
Topic 5	*chain, food, app, restaurant*	Food Industry
Topic 6	*medical, patient, care, healthcare*	Health
Topic 7	*player, bet, tournament, casino*	Gaming and Gambling
Topic 8	*bitstamp, estate, poloniex, btc*	Cryptocurrency Platform
Topic 9	*buyer, marketplace, reward, provider*	Trade
Topic 10	*learn, travel, education, train*	Education and Transport
Topic 11	*confidential, copyright, license*	Legal
Topic 12	*protocol, chain, node, chain*	Blockchain
Topic 13	*video, app, creator, campaign*	Advertising
Topic 14	*plant, solar, electricity, green*	Energy and Sustainability
Topic 15	*trader, asset, liquidity, broker*	Trading Investment

4.3 Cosine Similarity

Our research study focuses mainly on the similarity of ICOs. From our results, we can say that the whitepapers are similar because they have topics in common because they are part of the same sector of activity and have similar criteria such as national restrictions or having a same product idea and therefore have a greater chance to describe the project in a similar way. Afterwards we need a tool that allows us to identify the textual similarity of the documents. We will use as a tool the cosine similarity which is considered one of the most popular similarity measures applied to textual documents. It consists of calculating the cosine of the angle between the vector representations of the documents to be compared which corresponds to the correlation between the vectors of terms of the whitepapers. This correlation is translated into a value between 0 and 1 where 0 shows that the whitepaper has no textual similarity with all the other documents. The closer the value is to 1, the more similar the whitepaper is considered to be in other documents and therefore lacks textual innovation and may present a risk of a scam.

5. DATA AND APPLICATIONS

The sample we used in our paper is taken from an ICO scoring platform called ICObench. We have at our disposal a sample of 2000 ICOs from April 2015 to June 2020. By including our data in the R software, we can retrieve information about the ICOs such as their name, their country of origin, the number of people working on the project, the total duration of the ICO and other essential information, such as the general rating by ICObench, the volatility of the token, the amount raised which is the variable we have studied.

In Exhibit 3, there is a table divided into 5 parts, the first part is dedicated exclusively to the independent variables related to the success of the ICO which will be studied with the other parts of the table, the control variables. Each variable has a number of observations, a mean, a standard deviation, its minimum, median and maximum value.

5.1 Dependent variables

AmountRaised: As its name indicates, this variable indicates the amount of money that ICOs have raised during their period. This variable allows us to measure the success of ICOs. In addition, it is highly skewed so we used it in logarithms to be consistent with previous research, the name of this variable is *logAmount.*

rating_general: This is the rating offered by the ICObench platform. It allows us to rate the ICO project from 1 to 5 and this can give us an additional indication of the success of the ICO. However, the scale is very small, so it does not give us accurate information about the real value of the ICO.

ssGarchVol: This variable is useful to study as it is the volatility of the token price. The token can be traded on the secondary market for other tokens or fiat currencies.

TokenTradedHistorical: Finally, the last dependent variable is a binary variable that describes the success of the ICO with a 1 and the failure of the ICO with a 0. This variable has been studied in previous research (Adhami & al, 2018)

5.2 Control variables

The control variables are the variables we will use to control similarity in different contexts. We have divided the control variables into three parts. The first part is reserved for the internal elements of the ICO.

number_of_team: This variable takes into account the number of team members listed by the ICOs. The variable looks at the aggregate capacity measured by the total number of members team involved and not the individual quality of the members. Several studies have found significant positive relationship between this variable and the ICO success. The teams are composed of an average of 12 people with a maximum of 73 people. We support the fact that the more people there are in a team, the stronger the project can appear.

ICOpriceUSD: We will also check whether the similarity of the dollar price of the ICO can have an impact on its success. The aim here is to analyse the investor's perception of the price of a token and to see if its price has an influence on its success. The average price is $17

Duration: This is the total duration of the days of the ICO and is calculated by subtracting the date of the end of the ICO from the date of the beginning. The average duration of an ICO is 64 days and the maximum can reach 760 days.

BonusDummy: The ICO could include bonuses to attract investors by offering discounted prices. The control variable, BONUS, will indicate if bonuses were offered during the ICO phases or not. Even if bonuses generate market interest and so helps to raise to a greater amount, studies do not find any significant relationship between bonuses and the amount raised.

TaxHaven: This is the variable that describes ICOs located in countries where there is a very low tax rate for investors. In our sample, we have on average 30% of ICOs that are located in these locations.

FiatAcceptingDummy: Fiat currency is any money established as legal by government based on the credit of the economy and not on physical commodities as traditional money. This control variable indicates if the ICO accepts direct fiat contributions or not. Accepting fiat currencies could increase numbers of investors but also could have a doubt about the capacity to complete the ICO with only the cryptocurrency investors and so have a negative impact.

ScamDummy: this last internal control variable gives us information on whether the ICO is a scam or not. With 1 for yes and 0 for no. In our sample, we have on average 9.1% that are considered scam. It will be interesting to analyse whether the similarity of the whitepapers has a direct link with scams.

Institutions: This variable measures the institutional development and institutional strength of the ICO country based on 6 World Bank Governance Indicators. The 6 indicators are *CostOfCorruption*, *GovernanceEffectiveness*, *PoliticalStability*, *RegulatiroyQuality*, *RuleOfLaw* and *VoiceAndAccountability*

The third part of the table is devoted to information about whitepapers. We have chosen as variables *NumberOfWords*, *WP_Readability* and *PositiveWords*. These variables will control the similarity of the whitepapers to see if they have a correlation with the success of an ICO.

NumberOfWords: As its name indicates, this is the total number of words that a whitepaper has. On average whitepapers have 6830 words and can go up to 86000 words for some documents that share more information about the project. It is interesting to see if the number of words has an impact on the similarity of whitepapers.

WP_Readability: This is a variable that determines the readability of a whitepaper, the higher the value the easier it is to understand the document.

PositiveWords: The positive words are counted and we will observe if they have an influence on the similarity of the whitepapers.

Then, as for the similarity cosine, we divided it into three parts:

Similarity Minimum (*min*): This is the category that groups whitepapers with a low similarity and is in a range between 0 and 0.192. The variable min translates as radical innovative whitepapers, which are not similar to the others and which diversify in the way they are written. This variable can be considered as the innovation variable as it represents a low similarity rate.

Similarity Mean (*mean*): This category includes all whitepapers with a similarity between 0.005 and 0.486. On average these documents have a similarity of 0.33. This variable is also considered to be an innovation variable as it has a lower than average similarity rate. This means that there is a low similarity between the documents but this similarity is reflected in the fact that the projects are similar and have similar characteristics such as regulations, national standards, using similar payment technology or other.

Similarity Maximum (*max*): Finally, we have grouped the whitepapers with a similarity between 0.150 and 1 in a category with an average of 0.65. We consider this category as the one that gives us the most information on similarity. Indeed, whitepapers in this category are more likely to have the same writing style, discuss similar topics and have common sectors of activity. The whitepapers with a high similarity close to 1 are also more likely to be judged as scams or copycat.

EXHIBIT 3
ICO summary statistics

Dependant Variables	N	Mean	St. Dev.	Min	Max
AmountRaised	1,659	15,167,796.000	116,049,000.000	189.000	4,197,956,135.000
rating_general	3,933	3.001	0.750	0.800	5.000
ssGarchVol	679	1.170	3.398	0.00000	70.223
TokenTradedHistorical	3,933	0.200	0.400	0	1
Internal Control Variables					
number_of_team	3,739	12.520	7.630	0.000	73.000
ICOpriceUSD	3,749	17.063	660.692	0.000	39,384.000
Duration	3,917	63.702	69.657	0.000	760.000
BonusDummy	3,933	0.163	0.369	0	1
TaxHaven	3,933	0.309	0.462	0	1
FiatAcceptingDummy	3,933	0.020	0.139	0	1
ScamDummy	3,933	0.091	0.287	0	1
External Control Variables					
Institutions	3,741	2.357	1.869	−4.805	4.448
CostOfCorruption	3,725	1.083	0.965	−1.627	2.216
GovernanceEffectiveness	3,725	1.196	0.783	−1.767	2.231
PoliticalStability	3,725	0.494	0.747	−2.747	1.615
RegulatoryQuality	3,725	1.212	0.865	−2.334	2.206
RuleOfLaw	3,725	1.070	0.880	−2.339	2.046
VoiceAndAccountability	3,725	0.657	0.820	−1.958	1.733
Whitepapers Control Variables					
NumberOfWords	1,856	6,829.825	4,817.340	0.000	86,551.000
WP_Readability	1,853	45.418	14.611	−353.672	121.220
PositiveWords	1,856	79.919	60.260	0.000	710.000
Similarity Control Variables					
SimilarityMin	1,367	0.025	0.039	0.000	0.192
SimilarityMean	1,282	0.331	0.073	0.005	0.486
SimilarityMax	1,367	0.658	0.137	0.150	1.000

5.3 The different features of similarity

EXHIBIT 4
The direct impact of similarity on the *logAmount*

Variables	***t-value***	***correlation***	***r-squared***	***p-value***
logAmount ***mean***	2,331	0,09142	0,008357	0,02
Variables	***t-value***	***correlation***	***r-squared***	***p-value***
logAmount ***max***	1,746	0,06569	0,004316	0,0813
Variables	***t-value***	***correlation***	***r-squared***	***p-value***
logAmount ***min***	-0.062	-0,00233	5,428e-06	0.951

In Exhibit 4 we find the main relationship between the similarity of the whitepapers and the amount raised. We have created three regressions with a confidence interval of 95%, this table gives important information about the group of whitepapers with an average similarity. It can be observed that there is a correlation of 9.14% between the mean and the *logAmount* with a t-value of 2.331 which is higher than 2. The p-value is 0.02 which means that there is a 2% chance of being wrong.

This first result shows that the similarity of the whitepapers has a positive impact on the Raised amount. The second piece of information is that the group of whitepapers with an average similarity of 65% also has a positive impact on the Amount Raised. Nonetheless, we have a higher chance of being wrong because the p-value is 0.0813. It is interesting to observe a result like this, we have a positive relationship when documents are similar, which means that there is some form of standardisation in the way whitepapers are written which leads to raise funds.

As for the last group, whitepapers with a low similarity which are considered to be whitepapers with a radical innovation in the way they are written, we can see that there is a negative relationship between the *min* and the *logAmount*. On the other hand, we do not have significant evidence for this correlation. Indeed, as the p-value is too large, we only have a 5% chance of proving that this correlation is real.

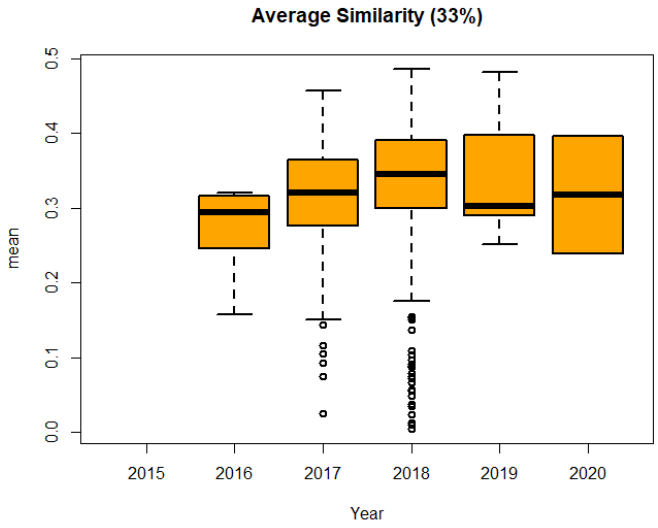
Let's have a look at the evolution of similarity over time to understand its trend a little better. We see in Exhibit 5 that the average similarity of whitepapers with mean similarity to other documents (*mean*) has increased from 2016 to 2018. From the year 2019, the average similarity of this group has decreased almost to the level of the year 2016.

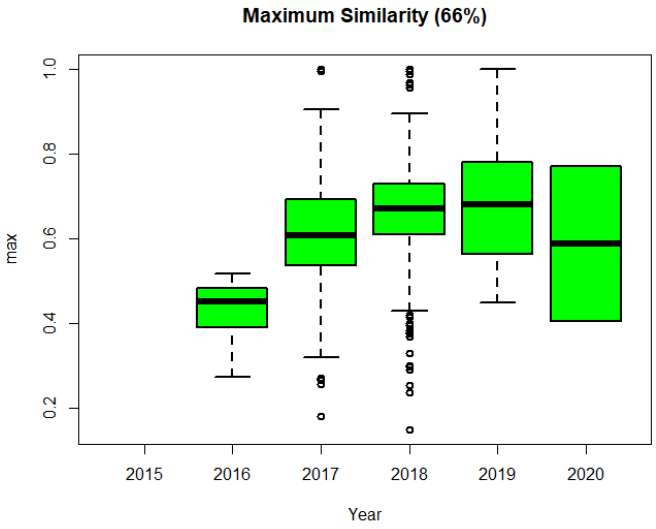
Finally, in 2020, the similarity increased by a level comparable to 2017. We can therefore observe an average trend for this group which does not have any major changes and where the similarity moves between 25% and 40%.

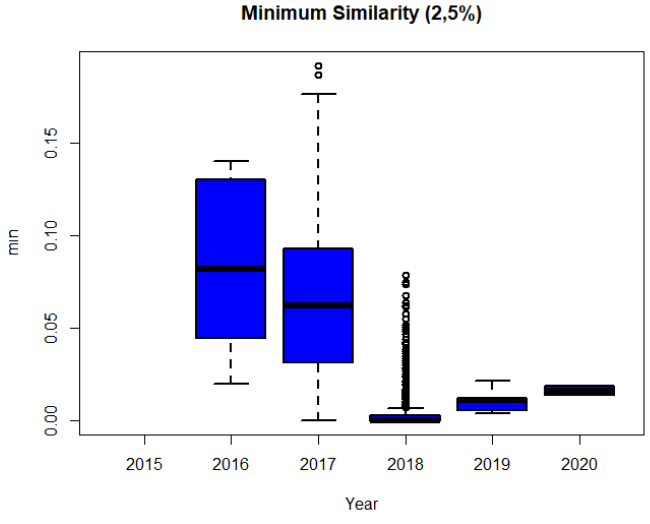
For documents with a high degree of similarity between them, the trend is almost similar to that of the average group. We can see that in 2016, the similarity of this group was around 40%, then the similarity increased drastically to an average level of 60%, the other years we see an increase in this average level to 65% until 2020 where we see a decrease. It should be noted that in 2020 the similarity gap is very large, ranging from 40% to almost 80% similarity. We can therefore note a form of standardisation that has set in from 2019 because the boxplot is larger than in previous years.

Finally, for the last group, the whitepapers which are very unique. We see the opposite effect compared to the other two similarity groups, in 2016 there was a greater disparity in similarity ranging from 5% to 14% compared to the years 2018 to 2020 where the documents have become totally unique with an average percentage of 2.5% similarity.

EXHIBIT 5
The evolution of similarity score over time







.

EXHIBIT 6
Which variables correlate with the three similarity groups?

Variables	***mean***	*p-value*	***max***	*p-value*	***min***	*p-value*
logAmount	0,09142	0,02003	0,06569	0,08132	-0,00233	0,95076
rating_general	0,19082	0,00000	0,15886	0,00000	-0,17602	0,00000
ssGarchVol	0,03971	0,51040	-0,02936	0,60365	0,02848	0,61451
TokenTradedHistorical	0,02898	0,29984	0,02528	0,35038	0,20698	0,00000
number_of_team	0,23162	0,00000	0,10579	0,00017	-0,08028	0,00438
Duration	0,06021	0,03111	0,00812	0,76427	-0,10927	0,00005
BonusDummy	0,03898	0,16308	0,01606	0,55299	-0,18751	0,00000
TaxHaven	0,11564	0,00003	0,07749	0,00415	-0,11473	0,00002
FiatAcceptingDummy	0,03892	0,16374	0,02181	0,42043	-0,04057	0,13382
ScamDummy	0,04029	0,14937	0,02123	0,43287	0,11867	0,00001
Institutions	0,00490	0,86406	0,03003	0,27819	-0,07452	0,00705
CostOfCorruption	0,00955	0,73856	0,03542	0,20088	-0,07336	0,00800
GovernanceEffectiveness	-0,00451	0,87464	0,02414	0,38331	-0,05082	0,06635
PoliticalStability	0,04976	0,08168	0,05516	0,04626	-0,07773	0,00495
RegulatoryQuality	0,00318	0,91150	0,01888	0,49534	-0,06456	0,01963
RuleOfLaw	-0,01159	0,68542	0,02216	0,42352	-0,05952	0,03150
VoiceAndAccountability	-0,01069	0,70866	0,01206	0,66336	-0,08368	0,00247
NumberOfWords	0,28922	0,00000	0,17236	1,41E-10	-0,04790	0,07667
Positive	0,31662	0,00000	0,16553	7,41E-10	-0,07020	0,00942
WP_Readability	-0.18841	0,00000	-0.09872	0.00030	-0.03898	0.14990

Using the correlation matrix, we were able to extract some information about the relationship between these variables and the three similarity groups. In the table, we have highlighted in colour the correlations above 10%.

rating_general: Here we see that similarity has an impact on how ICOs are rated on the ICObench platform. For the first group and the second group, we observe a correlation of 15.8% to 19%, i.e. whitepapers with a similarity higher than 33% are more likely to be rated well. The third group shows the opposite, which is normal, because the way they are written is very unique and this impacts the way they are rated by the platform.

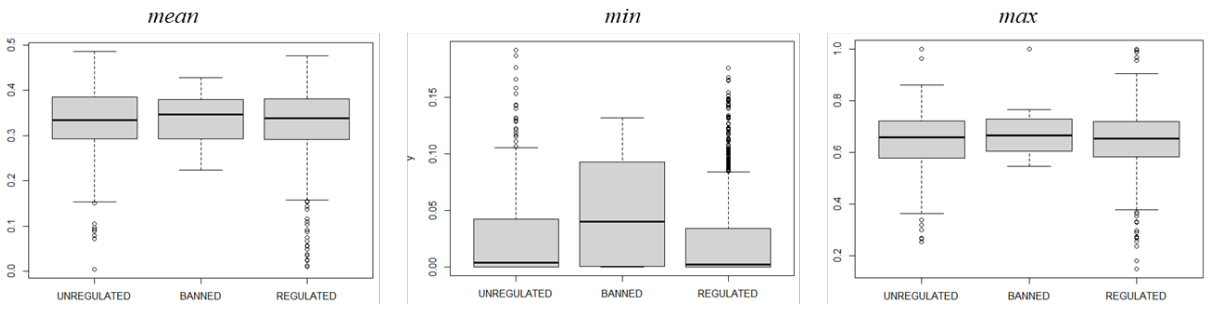
number_of_team: the number of people also has an influence on the similarity of whitepapers. From 10 to 23% for the first two groups, we see that the more similar the number of people, the more similar the documents will be written.

This is also reflected in the way the teams are presented in the whitepapers. They highlight their specialists as well as the members of the team.

TaxHaven: We also find that similar whitepapers are located in countries with lower taxation. This should certainly be highlighted in the written material. In contrast, unique and very similar whitepapers do not highlight this kind of information or are not located in countries with lower taxation.

ScamDummy: We observe here that there is a 10% probability that whitepapers with a very low similarity are considered as Scam. This is an important point to take into account, the p-value is very low which proves a positive correlation of these two variables. We can also mention the institutions variable which has a negative correlation with this category. This means that these are whitepapers located in countries or localities where institutions do not have a strong presence in the regulation of ICOs. Indeed, in Exhibit 6 you can see a large disparity of ICOs that have been banned in the category of whitepapers with low similarity. Compared to the other two categories, the boxes have a similar size, while for the minimum we have a significant unbalance.

EXHIBIT 6
The visual correlation between similarity and the ScamDummy variable.



NumberOfWords: The number of words is also correlated with similarity, as texts are similar, so is their word count. With a correlation of 28.9% in the category of documents with mean similitude. And 17.3% of documents with high similarity

Positive: The same applies to positive words that are repeated in similar whitepapers. Even though we notice that positive words are repeated twice as much in documents similar to 33% as in those similar to 66%.

5.4 The relation between similarity and fundraising

In Exhibit 4 we saw that similarity has a positive impact on the natural logarithm of fundraising. For both the average similarity and the maximum similarity whitepaper categories, the p-value indicates a medium significance with a 2% chance of being wrong with the mean and an 8% chance of being wrong with the max. Nevertheless, we had no concrete evidence for the min variable because the p-value is too large.

In this section we will go further by adding control variables to our regressions. To begin with we will see the impact of similarity on fundraising using control variables external to the ICOs, in other words, the institution variables.

In Exhibit 5 we notice that the control variables related to the regulatory institutions can have an influence on the impact of similarity on the Amount Raised. Yet this only applies to the mean group. As the max group does not have a significant p-value as well as the min group, we can conclude for this part that the external institutional variables are not the variables that explain the positive impact of similarity on fundraising.

One point to note is that the variable for similar documents at weak similarity has a positive impact when they are subject to government rules. We cannot say that this is a significant relationship as we have a 40% chance of being wrong. However, the direct relationship between the *min* variable and the *logAmount* variable has a p-value of 0.951 (Exhibit 4) whereas here its significance has increased to 0.401 (Exhibit 7.0)

When we control our similarity variable with internal control variables (Exhibit 7.1), i.e. the number of people in the team, the duration of the ICO, the bonuses they offer, the countries with lower taxation and the possibility to exchange their token with fiat currencies. It can be seen that some variables have an important weight to raise funds like the *number_of_team* variable or the *TaxHaven* variable. However, we do not notice any particular impact of the similarity on the Amount Raised. The category with a high similarity to the other whitepapers has a positive impact with a chance of being wrong of 17%. In addition, the influence is not very strong. As for the category with medium similarity, it has a positive relationship but we cannot guarantee anything because the p-value is very high. The same applies to the very low similarity group which has a negative relationship but where the p-value is also very high. Therefore, we will try another regression to see how our similarity variables can perhaps have a positive impact on the Amount Raised.

EXHIBIT 7.0
***logAmount* regression: Similarity and Worldwide Governance Indicator variables**

		Mean		*Max*		*Min*	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*logAmount*						
+	***Similarity Variable***	**2.219**	**0.0269***	**1.159**	**0.247**	**0.841**	**0.401**
+	Institutions	-0.194	0.8465	-0.206	0.837	-0.192	0.848
+	CostOfCorruption	-0.530	0.5960	-0.120	0.905	-0.011	0.991
+	GovernanceEffectiveness	1.218	0.2238	0.973	0.331	0.901	0.368
+	PoliticalStability	1.053	0.2927	0.992	0.322	1.100	0.272
+	RegulatoryQuality	-1.016	0.3099	-1.288	0.198	-1.317	0.188
+	RuleOfLaw	0.844	0.3987	0.941	0.347	0.885	0.377
	R-square	0,04056		0.03226		0.03134	
	Total p-value	0.0006327		0.002553		0.003284	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

EXHIBIT 7.1
***logAmount* regression: Similarity and internal variables**

		Mean		*Max*		*Min*	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*logAmount*						
+	***Similarity Variable***	**0.252**	**0.80112**	**1.343**	**0.179622**	**-0.533**	**0.59396**
+	number_of_team	6.328	4.83e-10 ***	7.365	3.21e-13 ***	6.688	4.83e-11 ***
+	Duration	-2.248	0.02494 *	-4.445	9.57e-06 ***	-2.021	0.04372 *
+	BonusDummy	0.014	0.98893	-0.397	0.691079	-0.461	0.64507
+	TaxHaven	3.106	0.00199 **	2.630	0.008640 **	3.297	0.00103 **
+	FiatAcceptingDummy	0.104	0.91738	-0.844	0.398776	0.111	0.91177
+	Year2017	1.943	0.05244 .	-2.027	0.042904 *	1.928	0.05426 .
+	Year2018	1.234	0.21760	-3.295	0.001012 **	1.050	0.29431
+	Year2019	1.095	0.27396	-2.211	0.027212 *	0.881	0.37858
+	Year2020	0.120	0.90485	-0.683	0.494867	0.118	0.90642
	R-square	0.105		0.1046		0.1002	
	Total p-value	1.294e-10		< 2.2e-16		4.236e-11	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

EXHIBIT 7.2
***logAmount* regression: Similarity and random variables**

		Mean		*Max*		*Min*	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*logAmount*						
+	***Similarity Variable***	**2.207**	**0.027705 ***	**1.699**	**0.089744 .**	**0.944**	**0.345330**
+	Duration	-2.589	0.009858 **	-2.521	0.011947 *	-2.385	0.017383 *
+	ICOprice2	0.048	0.961341	0.092	0.927120	0.001	0.999187
+	FiatAcceptingDummy	0.197	0.843789	0.328	0.742889	0.431	0.666321
+	Institutions	3.539	0.000432 ***	3.877	0.000116 ***	3.876	0.000117 ***
	R-square	0.04134		0.03758		0.03467	
	Total p-value	0.0001083		0.0001221		0.0002954	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

In Exhibit 7.2 we have created another regression with variables that do not match each other. We decided to add the duration of the ICO, another variable called ICOprice2, which is the price of the ICO in dollars, we added the variable concerning the possibility of exchange with fiat currencies, the last variable that was added is the institutions.

Again, we have more or less the same results as in the Exhibit 7.0 regressions with the worldwide governance indicator. We see that the duration of the ICO does not really impact the similarity. The ICO price in dollars does not really offer any weight in the relationship with the *logAmount*. On the other hand, we can observe that institutions retain considerable influence in fundraising.

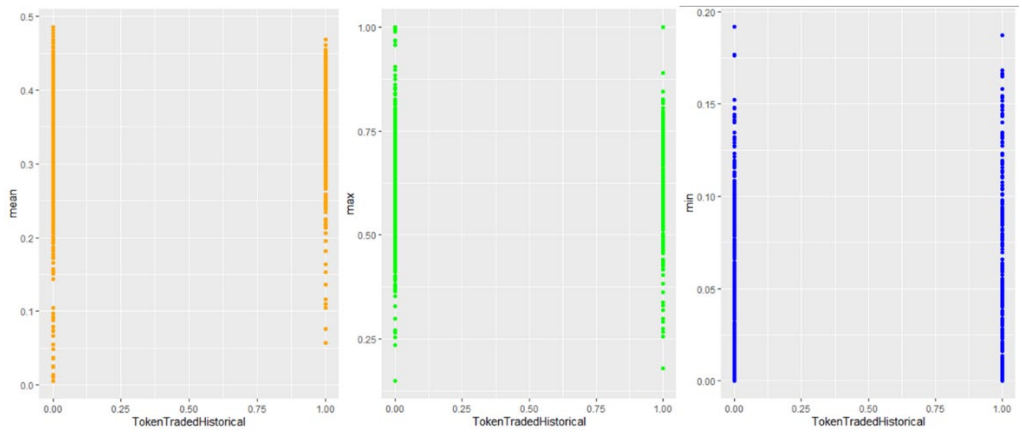
In this regression with these random variables, we have a positive influence from the group with a strong similarity with the other whitepapers. And this influence is significant. The same is true for the medium similarity group which also influences fundraising when controlled for the variables mentioned. One last point, the last group still has no impact on fundraising and the p-value is still not significant.

5.5 The relation between similarity and token tradability

The second independent variable we wish to study is the *TokenTradedHistorical* variable which is our dichotomous variable that denotes the success of the ICO. In a previous study, the researchers used trading as a dependent variable which is similar to the one we are studying (Amsden & Schweizer, 2018).

We saw in Exhibit 5 that there was a strong correlation of 20% between our *min* variable and *TokenTradedHistorical*. It is therefore important to go further to see if there is a positive or negative relationship of similarity on our success variable. This is explained by Exhibit 8 which is a table showing the differences between the three similarity groups with the success or failure of the ICO. We see that in the mean category in orange, we have a more prominent line on the failure side (0) and a less prominent line on the right (1). As for the maximum similarity group in green, we can observe that there is almost as much success as failure. But failure remains more important than success. For the last group we are interested in now, we can see that they have a large part of failure but also a very large part of success.

EXHIBIT 8
Boxplot of the three similarity groups on the variable TokenTradedHistorical



First, we can see that the min variable has a significant impact on the success of being controlled by other variables such as the *TaxHaven*, the regulation of the institutions, by the price of the ICO as well as by the number of words that the whitepapers have a low similarity.

Going back to the Exhibit 9.0 regression table, the first observation we can make is that low similarity has a strong impact on token trading. We see here that the control variables related to the worldwide governance indicator do not really have an influence on the tradability of the token but that the variable of the low similarity of the whitepaper in itself has a great power of influence. As for the average similarity variable, we note that it has a very weak influence on the tradability of the token. However the p-value is not to be neglected. We observe the same phenomenon for the variable of maximum similarity which has a weak impact with also high p-value. Then we decided as for the variable *logAmount* to add to it the variables of internal controls of the ICOs, so the number of people in the team, the duration of the ICO, if it offers bonuses, if it is in a country where the tax is favourable and if the value of the token can be exchanged with fiat currencies. Adding all these variables, only the maximum similarity has a positive impact on token's tradability but again we lack data and evidence to guarantee this influence.

Finally, in table 9.2 we have highlighted that the minimum similarity variable, to recall this variable has a very low similarity of 2.5% on average. This category of whitepapers when controlled with variables of duration, the price of the ICO in dollars, regulations of the institutions, remain nevertheless influential for the tradability of the token. We also find that the influence impact on the average and maximum similarity is much larger and the p-values are smaller.

EXHIBIT 9.0
***TokentradedHistorical* regression: Similarity and Worldwide Governance Indicator variables**

		Mean		***Max***		***Min***	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*TokentradedHistorical*						
+	***Similarity Variable***	**0.885**	**0.37617**	**0.764**	**0.4449**	**8.190**	**6.18e-16*** **
+	Institutions	0.073	0.94148	-0.195	0.8456	0.341	0.733
+	CostOfCorruption	0.236	0.81310	-0.008	0.9936	0.116	0.908
+	GovernanceEffectiveness	0.293	0.76972	-0.175	0.8609	-0.554	0.579
+	PoliticalStability	1.516	0.12972	1.576	0.1152	-1.250	0.111
+	RegulatoryQuality	-1.128	0.25971	-1.259	0.2081	-1.317	0.211
+	RuleOfLaw	0.334	0.73836	1.079	0.2810	0.490	0.625
	R-square	0.01586		0.01278		0.06087	
	Total p-value	0.006769		0.01922		6.269e-15	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

EXHIBIT 9.1

***TokentradedHistorical* regression: Similarity and internal variables**

		Mean		*Max*		*Min*	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*TokentradedHistorical*						
+	***Similarity Variable***	**-0.156**	**0.87570**	**1.343**	**0.179622**	**-0.533**	**0.59396**
+	number_of_team	7.053	2.99e-12 ***	7.365	3.21e-13 ***	6.688	4.83e-11 ***
+	Duration	-4.381	1.29e-05 ***	-4.445	9.57e-06 ***	-2.021	0.04372 *
+	BonusDummy	-0.152	0.87901	-0.397	0.691079	-0.461	0.64507
+	TaxHaven	2.944	0.00330 **	2.630	0.008640 **	3.297	0.00103 **
+	FiatAcceptingDummy	-0.565	0.57226	-0.844	0.398776	0.111	0.91177
+	Year2017	-2.078	0.03795 *	-2.027	0.042904 *	1.928	0.05426 .
+	Year2018	-3.221	0.00131 **	-3.295	0.001012 **	1.050	0.29431
+	Year2019	-1.937	0.05303 .	-2.211	0.027212 *	0.881	0.37858
+	Year2020	-0.656	0.51194	-0.683	0.494867	0.118	0.90642
	R-square	0.105		0.1046		0.1002	
	Total p-value	1.294e-10		< 2.2e-16		4.236e-11	

EXHIBIT 9.2

***TokentradedHistorical* regression: Similarity and random variables**

		Mean		*Max*		*Min*	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*TokentradedHistorical*						
+	***Similarity Variable***	**1.328**	**0.184591**	**1.058**	**0.29041**	**7.573**	**6.98e-14 *** **
+	Duration	-5.220	2.11e-07 ***	-5.403	7.79e-08 ***	-4.677	3.23e-06 ***
+	ICOprice2	1.783	0.074797 .	1.737	0.08267 .	1.140	0.254518
+	FiatAcceptingDummy	-0.506	0.613034	-0.740	0.742889	-0.444	0.657286
+	Institutions	3.117	0.001870 **	2.730	0.00642 **	3.378	0.000751 ***
	R-square	0.03571		0.03317		0.07388	
	Total p-value	2.723e-08		3.425e-08		< 2.2e-16	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

5.6 Textual similarity can detect potentially fraudulent ICOs

Another feature of similarity is that it can show us whether it influences the probability of having a scam in ICO campaigns. We also ran a regression adding several possible control variables that could positively or negatively influence the probability of having a scam. So, we added the variable TaxHaven, another variable named EthereumTokenDummy which is a variable that describes whether the token can be exchanged in the form of Etherum, we also assume that the duration of an ICO can define whether it is a scam or not and we also assume that the number of people matters too.

In Exhibit 10, we find that there is indeed a significant probability of scams as the similarity of documents decreases. Indeed, documents with high similarity do not have a significant impact on the probability of being a scam. However, the documents with a medium similarity, in other words, the whitepapers being similar to more or less 33% as well as those which have almost no similarity have a high risk of being scams. This is probably due to the fact that scams are trying to attract investors to new and totally revolutionary projects that aim to be eye-catching and therefore different from other projects. Of course, this is an idea based on the psychology of entrepreneurship.

EXHIBIT 10
The influence of similarity on scams

		Mean		***Max***		***Min***	
	Variables	***t-value***	***p-value***	***t-value***	***p-value***	***t-value***	***p-value***
=	*ScamDummy*						
+	***Similarity Variable***	**1.773**	**0.0765 .**	**1.322**	**0.18654**	**2.957**	**0.00316 ****
+	TaxHaven	-0.002	0.9988	-0.313	0.75436	0.065	0.94805
+	EthereumTokenDummy	-5.129	3.4e-07 ***	-4.949	8.49e-07 ***	-3.986	7.12e-05 ***
+	Duration	-0.834	0.4042	-1.263	0.20666	-0.970	0.33229
+	number_of_team	2.512	0.0121 *	2.728	0.00646 **	3.047	0.00236 **
	R-square	0.03571		0.02833		0.03372	
	Total p-value	2.723e-08		9.393e-07		3.692e-08	

LIMITATIONS

In this research we studied similarity and its impact on the probability of having fundraising and having a token that is tradable. We also demonstrated how Topic Modeling can be created in groups of documents such as whitepapers. We have not examined the similarity of each type of subject in detail, as this is another type of work that should be studied in its own right. We have added in the appendix's different tables and information about the topics and their relation to the similarity of the three groups (*mean, max, min*) in order to give you some additional information that may help people who want to have a starting point to learn more.

We would also like to point out that the results we have may vary from one year to another, as we have data in our possession but the ICOs are evolving from day to day and this analysis should be repeated next year and the year after as more data may give more precision to the study. We also think that a form of standardisation will be created in a few years which will push whitepapers to have their own framework and consequently be much more similar.

We also want to provide a global idea of similarity. Yet it can be used in many forms and it requires a lot of rigour and analysis. Our goal is to propose a study on a subject that is not much discussed because similarity in text analysis is still being improved, machine learning is progressing at a great speed and we will be able to have much more interesting conclusions in a few years.

CONCLUSION

The ICO market is growing more and more over the years and attracting new investors. Projects created in 2015 and others are starting to have more experience and stability. Regulation remains a sensitive issue as the essence of these projects is to be free and decentralised from any form of control. However, some form of standardisation needs to be created for whitepapers to help and protect investors from potential fraud. The question that needs to be discussed is what information should be controlled in order to prevent the risk of potential fraud.

Thanks to Topic Modeling and machine learning, we were able to produce data on similarity. We coupled them with the different variables that can explain the success of an ICO and the results are interesting. We saw that the evolution of similarity between whitepapers underwent a form of stabilisation from 2016 to 2020, both for the group with medium similarity and the group with high similarity we saw that they had a positive impact on fundraising, When similarity is controlled by institutional factors, When similarity is controlled by institutional factors, it is only the average similarity that stands out and influences fundraising, but when it is subject to different factors such as the duration of the ICO, its price, whether it can be exchanged with fiat currencies, we see that the maximum similarity also has a slight influence on the amount raised.

On the other hand, when we want to study the token tradability variable, we see the opposite effect with the group with minimum similarity having a strong impact on this dependent variable, which can be explained by the fact that ICOs with low similarity are looking to make their token available as quickly as possible without thinking about the long term and strategies to raise solid funds.

We also found a significant relationship between the low similarity group and scams. As we have seen, the purpose of potentially fraudulent ICOs is to make themselves unique to investors in order to show significant value and thus influence them to invest quickly because it is a very innovative project and is in an industry where there are very few competitors. Our results show that the lower the similarity, the higher the chance that the ICO is a scam. These scams may also explain why ICOs with low similarity to other documents seek immediate profitability by offering their token on exchange platforms.

We would like to see this study continued in the future to dig deeper into similarity as this is a topic that is very little discussed in the ICO literature.

REFERENCES

Adhami, S., Giudici, G., & Martinazzi, S. (2018). Why do businesses go crypto? An empirical analysis of initial coin offerings. *Journal of Economics and Business, 100,* 64–75.

Amsden, R., & Schweizer, D. (2018) Are Blockchain Crowdsales the New 'Gold Rush'? Success Determinants of Initial Coin Offerings. *Working Paper. Concordia University.*

Aslan, A., Şensoy, A., & Akdeniz, L. (2021). Determinants of ICO Success and Post-ICO Performance. *Central Bank of the Republic of Turkey, 21.*

Bian, S., Deng, Z., Li, F., Monroe, W., Shi, P., Sun, Z., Wu, W., Wang, S., Wang, W. Y., Yuan, A., & al. (2018). Icorating: A deep-learning system for scam ico identication. *arXiv preprint arXiv:1803.03670.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research 3,* 993–1022.

Catalini, C., & Gans, J. S. (2018). Initial Coin Offerings and the Value of Crypto Tokens. *Working Paper 24418. National Bureau of Economic Research.*

Chan, C. S. R., & Parhankangas, A. (2017). Crowdfunding Innovative Ideas: How Incremental and Radical Innovativeness Influence Funding Outcomes. *Entrepreneurship Theory and Practice*, *41(2),* 237–263.

Cheng, X.S., Tan, Y.C., & Liu, J.M. (2012). The non-financial information, the external financing, and the investment efficiency: A study based on the constraint of external systems. *Management World, 7,* 137–150.

Chod, J., & Lyandres, E. (2018) A Theory of ICOs: Diversification, Agency, and Information Asymmetry. *Working Paper. Boston University.*

Chuanjie, F., Koh, A., & Griffin, P. (2019). Automated theme search in ico whitepapers. *The Journal of Financial Data Science 1 (4),* 140–158.

De Jong, A., Roosenboom, P., & Van der Kolk, T. (2018). What Determines Success in Initial Coin Offerings? *Unpublished working paper.*

Florysiak, D., & Schandlbauer, A. (2019). The Information Content of ICO White Papers. doi: 10.2139/ssrn.3265007.

Jiang, F., & Kim, K.A. (2015). Corporate governance in China: A modem perspective. *Journal of Corporate Finance, 32(3),* 190–217.

Kimberly, J.R. (1981). Managerial innovation. In P.C. Nystrom & W.H. Starbuck (Eds.), *Handbook of organizational design,* 84–104. Oxford: Oxford University Press.

Kortum, S. & Lerner, J. (2000). Assessing the impact of venture capital on innovation. *Rand Journal of Economics, 31(4),* 674–692.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting & Economics, 56(1),* 130–147.

Li, J., & Mann, W. (2018). Initial Coin Offering and Platform Building. *Working Paper.*

Loughran, T., & Mcdonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance, 69(4),* 1643–1671.

Mayo, M. (2017). "A General Approach to Preprocessing Text Data." KDnuggets, Retrieved from https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html.

Metrick, A., & Yasuda, A. (2010). *Venture capital and the finance of innovation*. New York: Wiley.

Park, C.W., & Lessig, V.P. (1981). Familiarity and its impact on consumer decision biases and heuristics. *Journal of Consumer Research, 8,* 223–231.

Prakash, V., & Thukral, V.K. (1984). Product familiarity and consumer satisfaction. *Proceedings of the 1984 Academy of Marketing Science (AMS) Annual Conference,* Niagara Falls, NY, 77–81.

Qian, A., & Zhu, D. (2019). Financial report similarity and the likelihood of administrative punishment:based on the empirical evidence of textual analysis, *China Journal of Accounting Studies, 7:2,* 147–169.

Raza, A. (2018). Analyzing China's Ultimate Ban on All Crypto and ICO Websites. *CryptoSlate*. Retrieved from https://cryptoslate.com/analyzing-chinas-ultimate-ban-crypto-ico-websites

Reuven, L., Li, F., & Kenneth, M. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review, 86(3),* 1087–1115.

Schmidt, J., & Keil, T. (2013). What makes a resource valuable? Identifying the drivers of firmidiosyncratic resource value. *Academy of Management Review, 38,* 206–228.

APPENDICES

FIGURE 1

In this figure, you will find the information concerning the 15 subjects found through the LDA. On the X-axis, you have the 'beta' information which gives the presence of each word in each topic and on the Y-axis you have each word which describes the category.

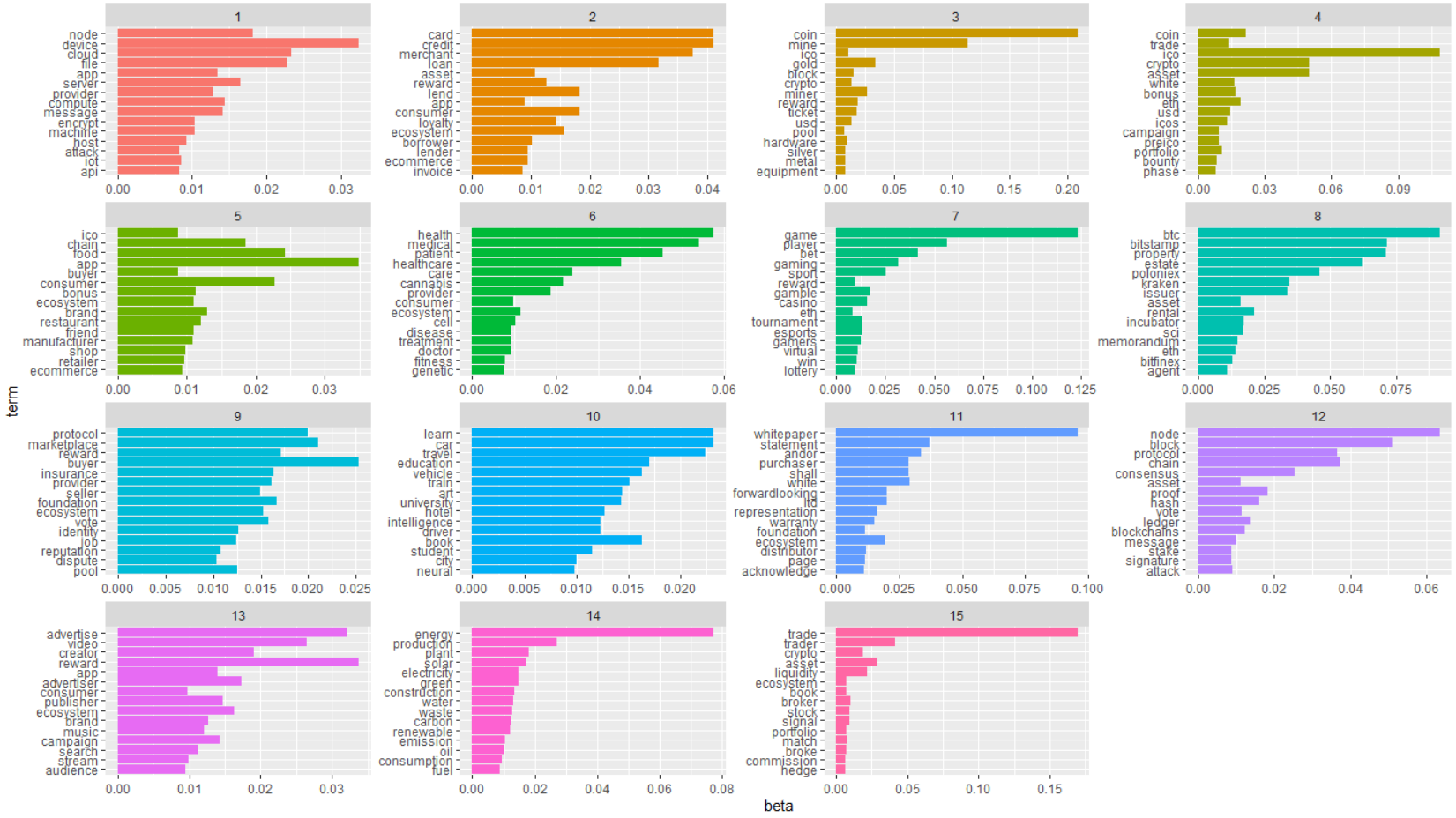


FIGURE 2

This figure describes the different correlations between the subjects (gamma.1, gamma.2, gamma.3, etc.) and the different variables such as Amount Raised, minimum, maximum and average similarity

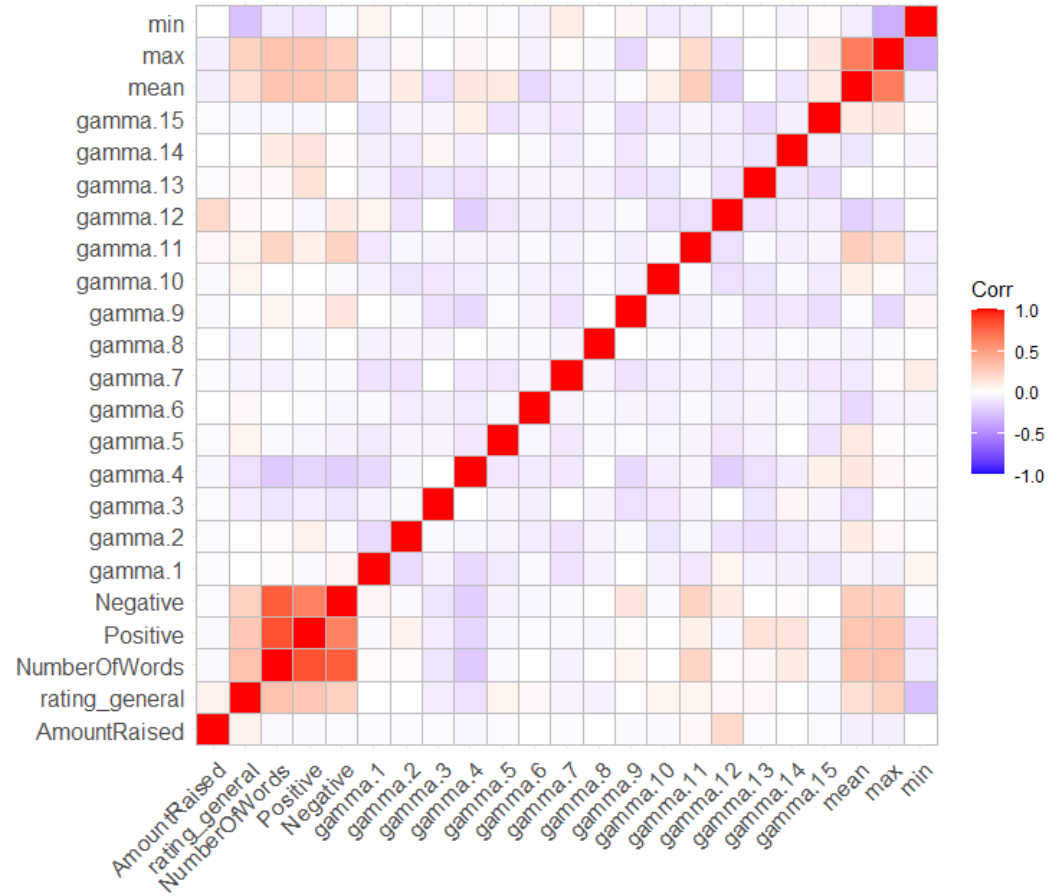


TABLE 1

Variables		***LogAmount***	*p-value*
Topic 1	*Online Services*	0,05630	0,09359
Topic 2	*Financial Product*	0,10293	0,00213
Topic 3	*Cryptocurrency Mining*	-0,04920	0,14290
Topic 4	*Payment Technology*	-0,15362	0,00000
Topic 5	*Food Industry*	-0,03325	0,32228
Topic 6	*Health*	-0,02071	0,53758
Topic 7	*Gaming and Gambling*	-0,04551	0,17544
Topic 8	*Cryptocurrency Platform*	-0,01907	0,57028
Topic 9	*Trade*	0,01713	0,61022
Topic 10	*Education and Transport*	-0,04390	0,19122
Topic 11	*Limited Company*	0,06359	0,05819
Topic 12	*Blockchain*	0,13780	0,00004
Topic 13	*Advertising*	-0,01121	0,73867
Topic 14	*Energy and Sustainability*	-0,04064	0,22633
Topic 15	*Trading Investment*	-0,00656	0,84528
mean	*Similarity Mean*	0,09142	0,02003
max	*Similarity Maximum*	0,06569	0,08132
min	*Similarity Minimum*	-0,00233	0,95076

With the help of the p correlation matrix, we were able to extract information on the influence of each topic on the fundraising. In addition, we added the three similarity variables, mean, max, min, which also have an influence on the amount. We can observe that Topic 2 has the highest correlation on fund-raising, this can be explained by the fact that the main topic is related to financial products and that investors with more capital are willing to invest more money in financial areas. Moreover, the p-value shows us that this is a significant correlation with a 0.2% chance of being wrong. In the table there are also negative correlations. However, the p-value shows us a large percentage of being wrong. The p-values with values of more than 10% chance of being wrong are not really significant and we will not give any findings on this information. On the other hand, we have some interesting results to share regarding similarity. The group of whitepapers with an average similarity between them have an influence of 9.14% on fundraising. Moreover, the p-value of 0.02 shows a chance of being wrong of 2% which can be considered as significant. Furthermore, the documents with maximum similarity, in other words, the whitepapers with a similarity of plus or minus 65%, have a correlation with the Amount Raise

TABLE 2

Variables	***mean***	*p-value*	***max***	*p-value*	***min***	*p-value*
Intercept	1		1		1	
Online Services	-0,02916	0,29715	-0,05098	0,05968	0,04538	0,09377
Financial Product	0,09077	0,00115	0,01192	0,65990	0,00795	0,76911
Cryptocurrency Mining	-0,15868	0,00000	-0,05092	0,05999	-0,04015	0,13821
Payment Technology	0,13063	0,00000	0,05528	0,04115	0,01869	0,49031
Food Industry	0,06156	0,02765	-0,04666	0,08483	-0,03154	0,24415
Health	-0,28253	0,00000	-0,02128	0,43209	-0,07682	0,00451
Gaming and Gambling	-0,05520	0,04832	0,08764	0,00119	0,04763	0,07856
Cryptocurrency Platform	-0,02934	0,29422	0,02136	0,43037	-0,00419	0,87699
Trade	0,05205	0,06263	-0,07624	0,00483	0,06713	0,01312
Education and Transport	0,06076	0,02973	-0,02178	0,42128	-0,04701	0,08251
Legal	0,25931	0,00000	0,13016	0,00000	-0,05997	0,02671
Blockchain	-0,13438	0,00000	-0,06856	0,01129	0,02707	0,31767
Advertising	0,01378	0,62226	-0,03083	0,25501	0,03626	0,18056
Energy and Sustainability	-0,12566	0,00001	-0,02893	0,28554	-0,06319	0,01956
Trading Investment	0,09704	0,00051	0,09481	0,00045	0,01371	0,61270

We analysed the correlation of each topic on similarity in order to observe whether certain topics are more likely to be used more frequently and thus induce textual similarity in the documents. We have divided the table into three parts, the first of which focuses on whitepapers with an average similarity of 33% with other whitepapers. The second part of the table focuses on whitepapers with an average similarity of 66% while the third column is devoted to whitepapers that are almost not similar with an average of 2.5% similarity. We can notice that the whitepapers with a high similarity between them (max) are written in the same way for the topics related to Legal, Gaming and Gambling, as well as Trading Investment. Probably the projects are similar and specific keywords and phrases are often repeated. Especially in the legal area, where each document has to have a section dealing with regulation and standards. We also note that there are negative correlations in the table, especially for topics related to Energy and Sustainability, Cryptocurrency Mining and Health, which means that these topics are not likely to be copied. They try to be as innovative as possible in the way they write their whitepapers and may have specific projects that are not comparable.