11-2019

Automated theme search in ICO whitepapers

Chuanjie FU
Singapore Management University, chuanjie.fu.2017@mitb.smu.edu.sg

Andrew KOH
Singapore Management University, andrewkoh@smu.edu.sg

Paul GRIFFIN
Singapore Management University, paulgriffin@smu.edu.sg

Citation

FU, Chuanjie; KOH, Andrew; and GRIFFIN, Paul. Automated theme search in ICO whitepapers. (2019). *Journal of Finance and Data Science*. 1, (4), 140-158. Research Collection School Of Computing and Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/4839

Automated Theme Search in ICO Whitepapers

FU CHUANJIE, ANDREW KOH, AND PAUL GRIFFIN

FU CHUANJIE
is a master's student at Singapore Management University and a director at Savills Singapore in Singapore.
chuanjie.fu.2017@mitb.smu.edu.sg

ANDREW KOH
is the practicum manager of the Masters of IT in Business (Analytics and AI) program at Singapore Management University in Singapore.
andrewkoh@smu.edu.sg

PAUL GRIFFIN
is an associate professor of information systems (practice) and director of financial technology analytics at Singapore Management University in Singapore.
paulgriffin@smu.edu.sg

KEY FINDINGS

- Categorization of ICO whitepapers can be done via topic modeling with the latent Dirichlet allocation (LDA) model.
- Statistical and human judgment methods confirms that there is enough evidence to conclude that the LDA model appropriately categorizes ICO whitepapers.
- Statistical tests suggests that the categorization results from the LDA model provides useful information on predicting whether an ICO will be successful funded.

ABSTRACT: *The authors explore how topic modeling can be used to automate the categorization of initial coin offerings (ICOs) into different topics (e.g., finance, media, information, professional services, health and social, natural resources) based solely on the content within the whitepapers. This tool has been developed by fitting a latent Dirichlet allocation (LDA) model to the text extracted from the ICO whitepapers. After evaluating the automated categorization of whitepapers using statistical and human judgment methods, it is determined that there is enough evidence to conclude that the LDA model appropriately categorizes the ICO whitepapers. The results from a two-population proportion test show a statistically significant difference between topics in the success of an ICO being funded, indicating that the topics are usefully differentiated and suggesting that the topic model could be used to help predict whether an ICO will be successful.*

TOPICS: ***Statistical methods, simulations, big data/machine learning****

An initial coin offering (ICO) within the cryptocurrency space is similar to an initial public offering (IPO) within the regular financial markets. Just like an IPO, investors investing in an ICO hope to see the cryptocurrency token perform well so that a return can be obtained on their investment. ICOs are mainly a funding tool used by blockchain-related startups, and in 2018, a listed company also used it as a tool (Zhao 2018) for fundraising, probably because of the lower regulatory hurdles for nonsecurity crypto tokens. Unlike IPOs, the underlying rights behind ICOs generally do not represent equity ownership. Instead, these cryptocurrency tokens could represent asset ownership or provide token owners access to a product or service.

As of June 2018, there were 5,368 ICOs on the ICObench platform. Singapore ranks as the world's third largest ICO launch pad in terms of money raised (after the United

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

States and Switzerland) with $2.2 billion.[1] China was previously the main player in Asia for digital currency trading. However, because of a hostile regulatory environment in China, Singapore is now the forerunner in digital currency trading in Asia (Raza 2018).

In 2017, 884 ICOs were launched, raising a total of $6 billion. This would mean that on average there are around 73 ICO launches every month. Given the sheer volume, it would be difficult to comb through all the information provided to make an informed investment decision. This difficulty is compounded by the fact that each ICO provides a different level of information and no regulations are in place to determine the amount of information disclosure required. Based on industry norms, ICOs will typically provide the following information:

- Team—People who will take the project forward
- Advisory board—People with the relevant experience to cover for the team's shortcomings
- Whitepaper—Serious projects may also have a position paper and yellow paper, the former highlighting the technology proposal and the latter presenting in scientific detail the technology and the innovations created
- Code and/or prototype—Good code repository communicating the vision
- Roadmap—Plan for the next years with the relevant milestones
- Token sales terms—Description of the platform and utility of the token, amount being distributed and kept, duration of the sale, and hard and soft caps

With this deluge of information, it would be beneficial for potential investors to be able to segregate ICOs into different topics. This segregation would then help investors decide which ICOs they should focus on for further analysis. For example, certain investors could be interested in gaming, manufacturing, or trading ICOs. The objective of this article is to introduce a method to build an automated tool that segregates ICO whitepapers according to different topics, themes, domains, or industries.

Although this article's scope is on the application of topic modeling to ICO whitepapers, this concept could be applied to other industries in which the provision of documents to the public is prevalent (e.g., prospectuses put forward to do an IPO for a company or whitepapers published for clinical drug trials). Similar situations in which whitepapers are made available, such as an initial exchange offering or security token offering, could also capitalize on this methodology for quicker analysis.

LITERATURE REVIEW

To have a holistic view of the literature available, a literature review has been conducted at four levels. The first and highest level involves exploring what text analytic techniques have been used in the area of finance. The second level involves determining how latent Dirichlet allocation (LDA) has been applied in different areas of study and how it has been validated. The third level will then move on to look at what analytical techniques have been adopted for ICO whitepapers. The fourth and final level looks at what research work has been done for whitepapers in general, not just ICO whitepapers.

Text Analytic Techniques in Finance

In the finance domain, text mining has been applied within a few broad areas, such as foreign exchange rate prediction, stock market prediction, customer relationship management applications, and cybersecurity (Kumar and Vadlamani 2016). The algorithms vary widely among support vector machines, n-grams, self-organizing maps, LDA, and so on. The datasets mainly focus on news articles, corporate filing, and tweets.

LDA Application and Validation

Several research papers have explored the application of LDA. For the purpose of regulating the finance industry, Bastani, Namavari, and Shaffer (2019) used LDA to explore the complaints submitted by consumers to the Consumer Financial Protection Bureau. LDA allowed the regulator to extract useful insights and explore trends in these latent topics over time.

In the government sector, literature has been published on the categorization of e-petitions (Hagen 2018) using LDA models. This article relied heavily on validation of the topic models through human judges. The study also relied on validation through corresponding social events that occurred at the time the e-petitions were raised.

[1] See https://icobench.com/stats.

LDA is an unsupervised learning technique. Therefore, to validate the number of topics input into the model, many research papers have explored different means of validating the latent topics generated by the LDA model. Some articles involved the use of visualization techniques, such as those by Sievert and Shirley (2014) and Griffiths and Steyvers (2004), who suggested different visualization techniques in validating LDA models. These approaches each have their merits, which will be highlighted in the subsequent sections of this article.

Other validation methods use statistical measures and human judgment, such as automated evaluation of topic coherence through various statistical measurements (Konrad 2017) and validation through human judgments using word and topic intrusion techniques (Chang et al. 2009). This article references these techniques and adopts a bespoke validation process.

Application of Analytics in ICO Whitepapers

The use of machine learning techniques to categorize documents for ICOs is not a field that has been well researched. As of the date of the present article, no other research papers have used data science techniques to categorize ICO whitepapers. However, Bian et al. (2018) lightly touched on using LDA to categorize ICOs into 10 different topics based on their manual labeling process. Because the paper focused on detecting ICO scams by using machine learning, minimal work was done on the topic labels, and the labels were not used as a predictive feature in the authors' machine learning models.

Another paper that dealt with ICO whitepapers relates to the correlation between the readability of token whitepaper against the ICO's first-day returns (Zhang et al. 2019). The authors adopted techniques for determining reading levels to measure the readability of ICO whitepapers and regressed readability against the ICO's first-day returns. Results indicated a positive correlation between readability and returns.

Other Research on Whitepapers

Apart from the aforementioned research papers, two research papers with the word *whitepaper* were uncovered. The first research paper relates to the use of machine learning models to detect Japanese characters used in a peculiar manner or in an unorthodox combination with other words within a corpus of Japanese contemporary whitepapers (Shinnou and Sasaki 2010).

The second research paper on whitepapers relates to the measurement of trust portrayed by major cloud providers (Gantner, Demetz, and Maier 2015). This study involved manually reviewing whitepapers and websites and conducting management interviews. The findings were then translated into trust measures using social coding. A lot of manual work was involved in the creation of this study, and no machine learning algorithms were used.

DATASET

The dataset for this paper has been extracted from ICObench, which is an ICO rating platform supported by investors and financial experts. The website has an API called ICObench Data API, which provides a variety of information and data from the platform, including ICO listings, ratings, and statistics. Apart from the names of the ICOs, certain key details such as the URL of the whitepaper and whether the ICO met its funding requirements were collected. Correspondingly, the whitepapers were downloaded from the indicated URLs. These whitepapers were then processed using the SAS Cloud Analytic Server with SAS Viya.

The ending dates of these ICOs ranges between May 31, 2016 and August 20, 2018. However, most of the data points were between April 2018 and July 2018. See Exhibit 1 for a histogram of the number of ICO records in ICObench by months of the year.

It was observed that most of the unsuccessful ICOs (those that did not meet their minimum funding level) within the dataset had ICO funding deadlines ending in 2018 (see Exhibit 2), which suggests that there is survivorship bias within the dataset wherein unsuccessful ICOs have been removed from the database.

For a preliminary view on the absolute frequency of words used in the whitepapers, a word cloud has been constructed based on the full-text extraction from the whitepapers. As observed in Exhibit 3, words such as "token," "market," and "platform" were, unsurprisingly, the most common.

TECHNOLOGY ADOPTED

With the purpose of ensuring the data were collected and analyzed in a structured manner, the key tools and libraries listed in Exhibit 4 were used.

EXHIBIT 1
Histogram of ICO Records by Month of the Year

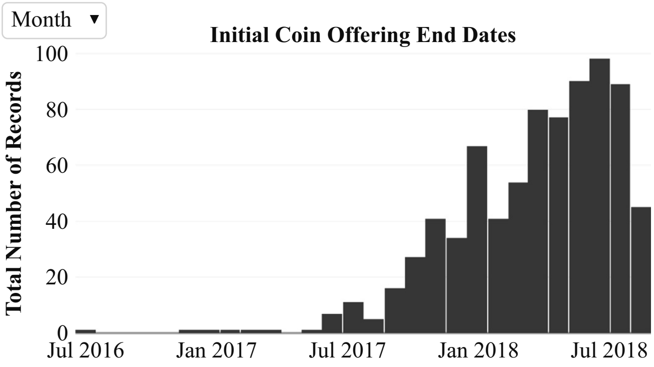


EXHIBIT 2
Proportion of Successful and Unsuccessful ICOs by Year

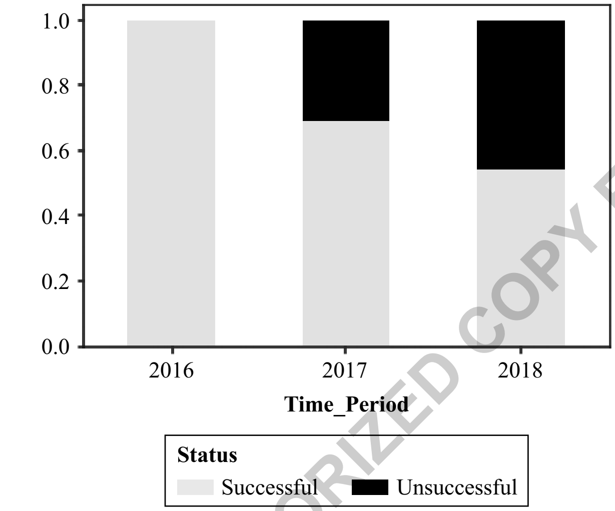


EXHIBIT 3
Word Cloud for All Whitepapers

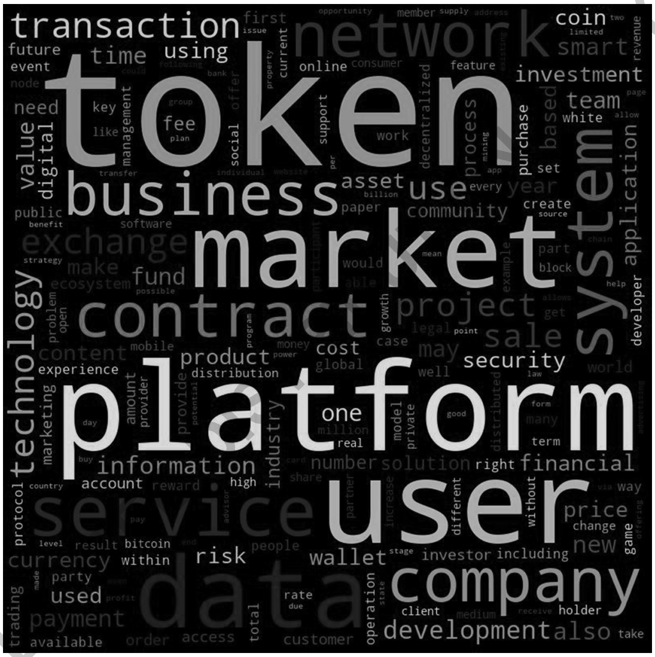


EXHIBIT 4
List of Software Used

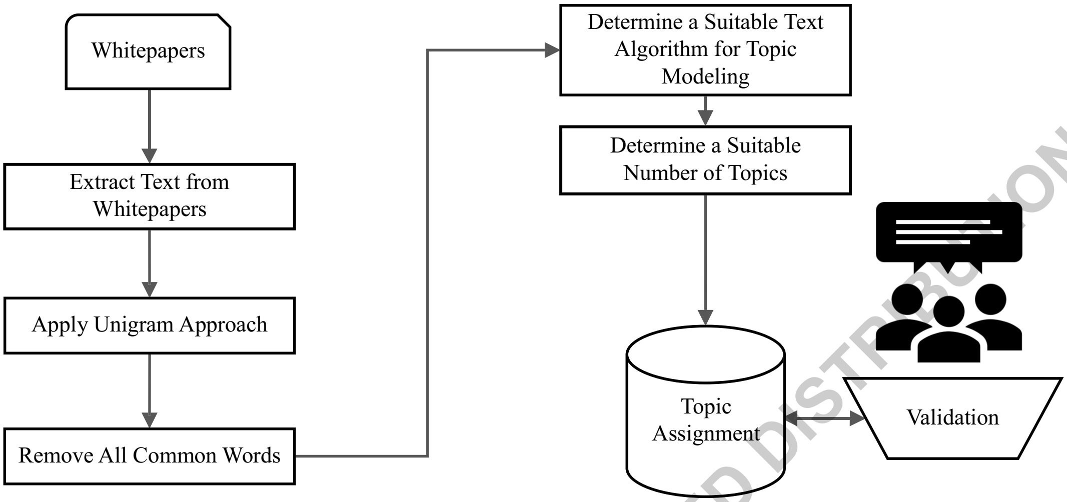
Software Tools	Purpose
Ubuntu Server 18.04	Establishing an analytical platform
Python 3.7	Performing data analysis
Jupyter Server	Presenting and performing data analysis
Docker Container	Running of MongoDB and Jupyter servers
MongoDB	Storing data from ICObench
GridFS	Storing whitepapers in PDF format
SAS Viya Release V.03.04	Extracting the text from the whitepapers formatted in PDF and plotting certain word clouds

Some difficulties were encountered with using open-source libraries to perform text extraction tasks, such as the nonrecognition of text in certain fonts and the subpar quality of optical character recognition functions. Thus, as mentioned earlier, text extraction was performed using SAS Cloud Analytic Service. This produced better quality text extractions compared with solutions based on open-source libraries (PyPDF2 and PDFMiner).

METHODOLOGY

Text analysis is a process of deriving meaning from language to serve a particular purpose. The purpose, in this case, is to isolate the textual information that differentiates one whitepaper from another depending on the nature of the ICO. Because text is a form of unstructured data and not fit for consumption by computers in its raw form, there is a need to preprocess the data into a structured format for computers to understand.

EXHIBIT 5
Analytic Flow Diagram


Whitepapers
Extract Text from Whitepapers
Apply Unigram Approach
Remove All Common Words
Determine a Suitable Text Algorithm for Topic Modeling
Determine a Suitable Number of Topics
Topic Assignment
Validation

The analytical flow diagram (Exhibit 5) illustrates the full process.

Labeling of Each ICO

The dataset available at ICObench does not provide information on whether the ICOs are successful in their project outcomes. However, information on the fundraising goals (i.e., soft caps and hard caps) and data fields on the funds raised are available in the dataset. These fields have been used to determine whether the ICO is deemed successful, which, for the purpose of this paper, is defined as having met its soft cap requirements.

Text Preprocessing: Unigram Approach

In textual data science tasks, text preprocessing is a fundamental requirement (Mayo 2017). Two main approaches were adopted in the preprocessing of text data. The first is to consider a unigram approach, in which each word is considered individually; second is the bigram approach, in which two words in sequence are considered as one unit. The topic models produced by the bigram approach were not ideal because of the approach's inability to segregate documents into different topics (i.e., more than 90% of the documents were clustered into a single topic). Therefore, the results of this approach were not included in this article. See Exhibit 6 for an example of a bigram versus unigram approach.

EXHIBIT 6
Example of Bigram versus Unigram Approach

Raw Text	Blockchain technology and online gaming have a long-shared history
Unigram Approach	[blockchain, technology, and, online, gaming, have, a, long, shared, history]
Bigram Approach	[blockchain, technology, and, online, gaming, have, a, long, shared, history, blockchain technology, technology and, and online, online gaming, gaming have, have a, a long, long shared, shared history]

Analyzing Word Count by Unique Documents

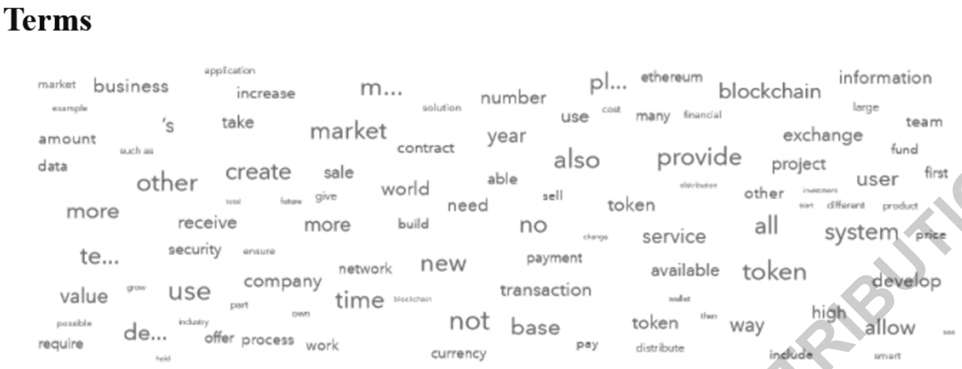
Exhibit 7 presents a word cloud drawn such that the size of the word is determined by the number of documents in which it appears. For example, if the words "market" and "blockchain" appear in all the whitepapers, the size of these words in the word cloud would be the same no matter how often these words appear.

Dealing with Common Words

From Exhibit 7, it can be observed that words such as "blockchain," "token," and "market" appear in almost all documents. Hence, these words would

Exhibit 7

Word Cloud by Document Count



generate minimal value in differentiating one category of whitepapers from another.

Based on this analysis, those words and others such as "platform," "user," "data," "service," "system," "network," "contract," "exchange," "business," and "company" have been included in the list of stop words in the preprocessing of the text documents to achieve more disparity in the topic words.

Model Algorithm Selection

Bastani, Namavari, and Shaffer (2019) discussed four main text corpora modeling techniques—namely, term frequency–inverse document frequency, latent sematic analysis, probabilistic latent semantic analysis, and LDA. Each of these techniques has its pros and cons. However, only LDA can allocate new documents based on a pretrained model, among other advantages. Hence, for this article, LDA has been adopted to create a generative model for segregating the ICO whitepapers into topics. With this generative model, a document is viewed as a distribution over topics, whereas a topic is a distribution over words. First, this model samples a document-specific multinomial distribution over topics from a Dirichlet distribution, and then it repeatedly samples the words in the document from the corresponding multinomial distribution. LDA can capture correlations between words but not correlations between topics (Cao et al. 2009).

Determining the Number of Topics

LDA is an unsupervised learning technique. To apply the model to a corpus of 972 documents, a selected number of topics (K) will need to be predetermined for the topic model to be generated. However, the model's best K is not determined only by the size of the dataset. Instead, it is also sensitive to the inherent correlations in the document collection. These inherent correlations exist in documents that feature different topics but share common words such as "blockchain" or "whitepaper." Therefore, it is important to select the optimal number of topics. Three measures have been computed to introduce objectivity into the selection of the optimal number of topics: cosine distance, topic coherence, and perplexity.

Comparing cosine distance. To achieve the optimal number of topics, the method discussed by Cao et al. (2009) has been applied. The principle for this method is that the best K of LDA is correlated with the distance between topics, T. Thus, Cao et al. adopted a measure that is based on the average cosine distance between every pair of topics to determine the stability of topic structure in its selection of K. The formula to compute the average cosine distance is

$$ave\_dis(structure) = \frac{\sum_{i=0}^{K}\sum_{j=i+1}^{K} corre(T_iT_j)}{K \times (K-1)/2} \tag{1}$$

The aim of this method is to maximize similarity within clusters and minimize similarity between clusters. However, the objective of keeping the model available for practical usage takes precedence. Therefore, the number of topics cannot be so large that the model becomes unusable for practical usage.

From Exhibit 8, as expected, the average cosine distance decreases with the increasing number of topics

Exhibit 8

Average Cosine Distance between Every Pair of Topics

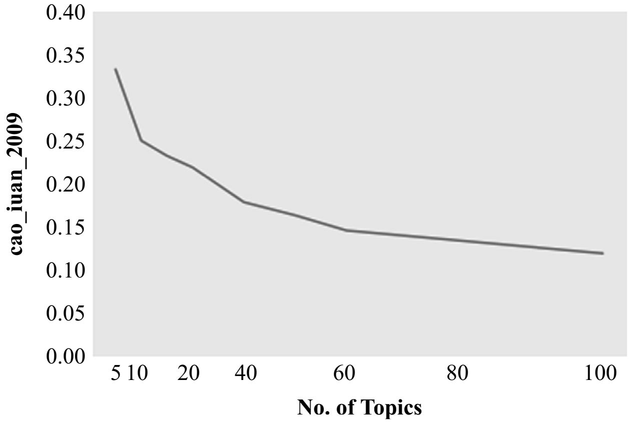


Exhibit 9

NPMI by Number of Topics

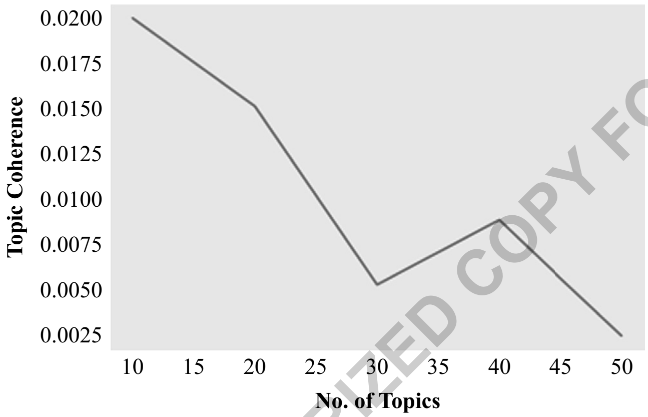
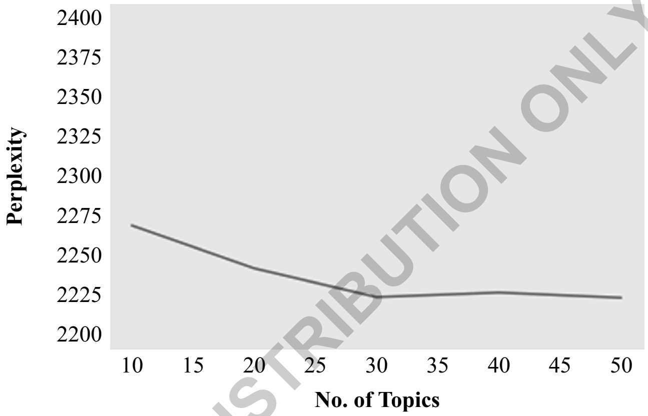


Exhibit 10

Perplexity by Number of Topics



used within the LDA model. Ten topics were chosen because the average cosine distance's gradient of descent is shallower and the spread between documents is more even.

Comparing topic coherence. Topic coherence provides a measure to compare different topic models for human interpretability. There are a few different ways of computing this measure. Normalized pointwise mutual information (NPMI) was selected because it gave the largest correlations with human ratings on topic interpretability (Röder, Both, and Hinneburg 2015). The idea behind the measure is to identify in a corpus a combination of words that show some idiosyncrasy in their linguistic distribution (Bouma 2009). The NPMI ranges between 0 and 1, with 1 signifying that the words are perfectly collocated to each other throughout the documents.

From Exhibit 9, the highest NPMI value observed, at 0.02, is found with a 10-topic model. This further substantiates the selection of the 10-topic model.

Comparing perplexities. Perplexity measures how well a probability model predicts a sample ("Perplexity" 2019). This measure is usually used to compare probability models to determine the best model. With a lower score signifying a better model (Machine Learning Plus 2018), the perplexity is computed as

$$e^{(-1*log-likelihood\ per\ word)} \tag{2}$$

As observed in Exhibit 10, the lowest perplexity exists when the model computes for 30 topics, reflecting a score of 2,230. However, the score difference between 10 topics (2,272) and 30 topics is only 1.8%. Therefore, the lower number of topics was selected to make the analysis more manageable.

MODEL RESULTS

Three methods were adopted to visualize the model's results. No methods is superior to the others; each has its advantages and disadvantages. The first method is a tabular method, presenting the total count of documents within each topic with the top five words for each topic. The second method relies on multidimensional scaling to present the topics on an intertopic distance

EXHIBIT 11

Words for All 10 Topics within the LDA Model

Topic	Words	Category	No. of Documents
Topic 1	Game, player, team, bet, ticket	Media: Gaming, Gambling	81
Topic 2	Energy, mining, power, project, production	Natural Resources: Energy, Precious Metals and Rare Earth	46
Topic 3	Information, may, white, paper, risk	Information: Data and Analytics	93
Topic 4	Transaction, node, smart, protocol, block	Information: Distributed Computing	134
Topic 5	Community, marketing, project, new, reward	Media: Social Networks	155
Topic 6	May, technology, sale, financial, risk	Finance: Exchange	108
Topic 7	Information, health, medical, patient, technology	Health and Social: Health and Medical	47
Topic 8	Content, video, advertising, publisher, digital	Media: Smart Contract	62
Topic 9	Payment, transaction, wallet, coin, sale	Finance: Payment Processing	105
Topic 10	Currency, fund, asset, project, investor	Finance: Banking and Other Financial Services	141

map. The last method uses a heatmap with the probability of words within each topic presented accordingly.

Tabulation of Results

In Exhibit 11, the corresponding top five words for each topic have been tabulated. These words present the themes behind those whitepapers. Because LDA will compute the topic distribution among the 10 topics for each whitepaper, each document will have a probability result for each of the topics. The category column is based on Smith and Crown's industry segregation (Smith + Crown 2018). For the purpose of segregating whitepapers into topics, each document has been allocated to the topic to which it is most likely (i.e., has the highest probability) to belong.

As shown in Exhibit 11, there are more than 40 documents allocated within each of the 10 topics. This suggests that the topic model does function as a useful categorization tool for ICO whitepapers.

Intertopic Distance Map

Chuang et al. (2012) suggested the use of multidimensional scaling techniques to visualize the topics obtained from the LDA model. This was subsequently implemented in Python by Sievert and Shirley (2014). The topics are plotted in the two-dimensional plane as circles whose centers are determined by computing the distance between topics. Multidimensional scaling is then used to project the intertopic distances onto two dimensions. The topic's overall prevalence, within the corpus, is denoted using the areas of the circles.

For the purpose of the present article, the visualization of topics with this technique has been used to compare the bag-of-words approach and term frequency–inverse document frequency approach. The visualization shown in Exhibit 12, based on the bag-of-words approach, portrays a good segregation of topics owing to the following properties: minimal overlap between circles and circles with similar sizes. The former observation denotes that the words in the topics have minimal overlap, and the latter denotes that the documents are evenly split among topics. The same visualization method applied on a term frequency–inverse document approach produced a classification wherein more than 90% of the documents were classified into a single topic (not shown here). The remaining nine topics also were shown to overlap. This overlapping was due to the repeat of keywords between these topics. Thus, the present article has adopted a bag-of-words approach.

Word Diagnostic Heatmap

As suggested by Griffiths and Steyvers (2004), the makeup of the words within the ICO whitepapers is key to providing the differentiators and similarities between different whitepaper themes. Finding significant word distribution differences between topics would suggest that the topics have differences that can be expressed in terms of the statistical structure uncovered by the LDA model.

EXHIBIT 12
Unigram Intertopic Distance Map

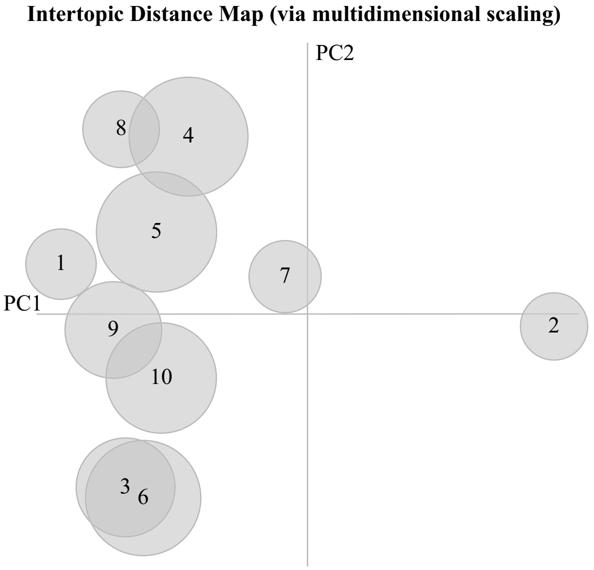


Exhibit 12 shows how the model has segregated the ICO whitepapers into various topics. An LDA model segregates the whitepapers into various topics by fitting a probability distribution of words for each topic. Thus, for every topic, a unique probability distribution of words can be identified. These differences in probability distribution can then be observed by reviewing Exhibit 13. As shown, there are observable differences in the word distribution among topics. Only a few words span across more than 1 topic (e.g., "technology," "coin"). This chart further proves that the model has served its purpose because minimal overlap occurs in words across topics.

MODEL VALIDATION

LDA model evaluation is difficult given that it is an unsupervised learning technique in which unlabeled data are used. In this article we have adopted two methods for validating the models. The first method is to assess the model practically using the human-in-the-loop approach (Chuang et al. 2012), which is based on a method found in the literature review. The second method focuses on observing the proportion of successful whitepapers within each topic compared with the rest of the whitepapers not classified under the same topic.

Human-in-the-Loop

Some research papers have concluded that higher predictive likelihoods do not directly translate into improved model interpretability. Chang et al. (2009) made such an argument and proposed that human subjects be made available as a resource to validate model interpretability. Word intrusion and topic intrusion have been suggested as methods to give the participants a framework to comment on the output of the model. The core concept of this framework is to have the test administrators introduce foreign words into the topic words suggested by the LDA model for each topic. The human subjects are then asked to identify these word intrusions. Successful identification indicates that the LDA model has performed up to expectations.

Without extensive resources to conduct a similar survey, we adopted the following methodology for this article:

1. Identify the top three documents with the highest probability, bottom three documents with the lowest probability, and 10 randomly selected documents from the selected topic.
2. Read all the selected documents to understand the content of the ICOs.
3. Assess whether the contents of the whitepaper are aligned to the respective themes within each topic.
4. Perform steps 1 to 3 for all the topics identified by the LDA.
5. Ascertain that there is overall consistency in the documents identified within each topic.

Based on the results tabulated in Exhibit 14, it can be observed that the top three documents for each topic have a consistent theme behind them. This finding helps to justify that the LDA model has appropriately categorized the ICO whitepapers by topics.

To further validate ICO whitepapers as categorized by the topics, three documents with the lowest probability in the selected topic will be reviewed to identify the themes behind those papers and to look for consistency with the topic model's results.

Two whitepapers, Cypher ICO (no. 406) and Crypto-ISBN ICO (no. 3010), were removed from the dataset because they were allocated a 10% probability to each of the 10 topics. This was due to the lack of keywords within these documents to allow the model to

Exhibit 13

Heatmap of Word Probabilities by Topics

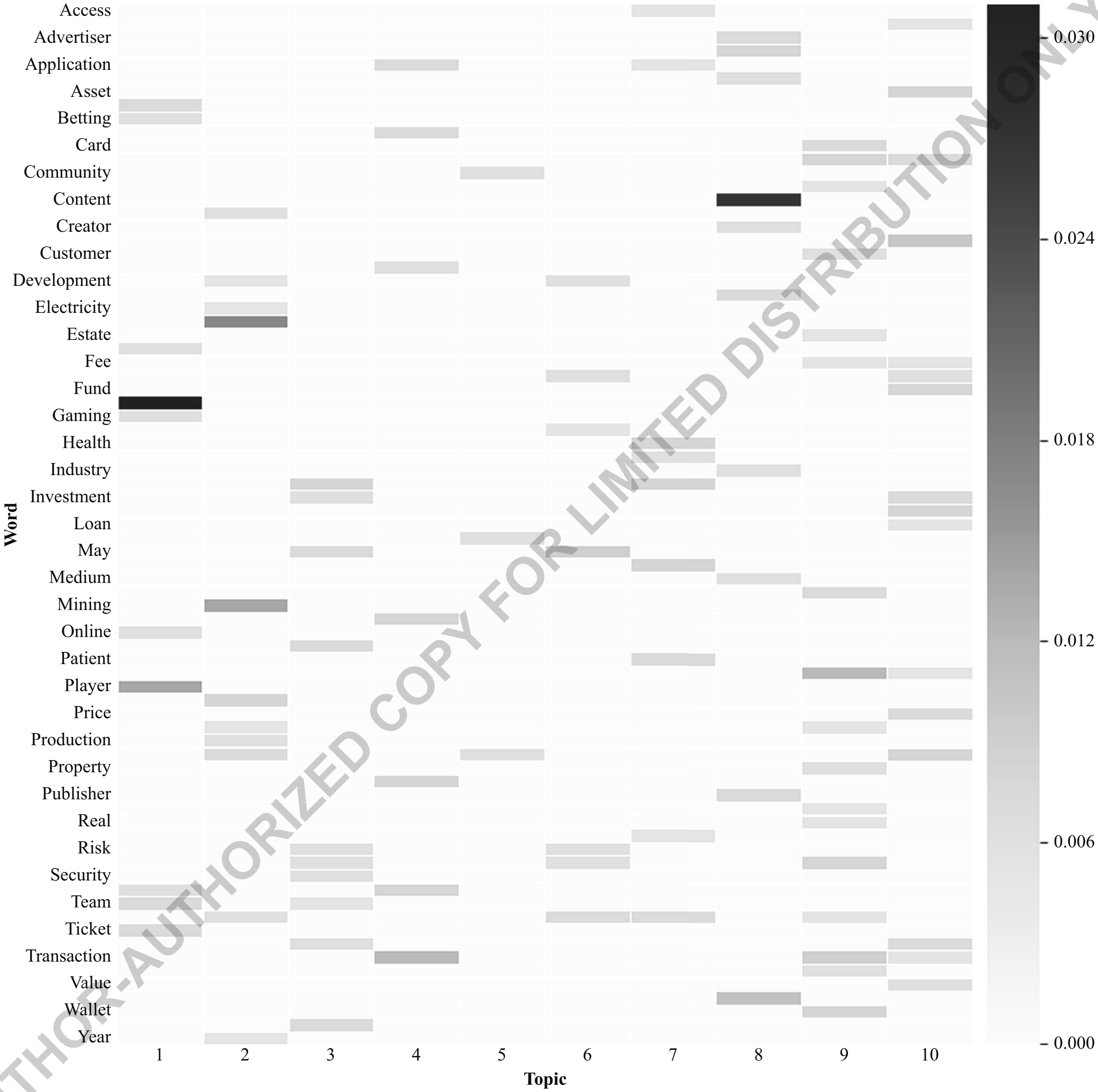


EXHIBIT 14

Human Validation of LDA Unigram Approach with the Top Three Documents

Topic	Topic Words/Description	Consistent
Topic 1	*[game, player, team, bet, ticket]* *ICOs relating to gaming for items with monetary value and/or betting* • 2966, DPLAY: Platform for eSports and its participants and organizers to cash out [Yes] • 2752, Cryptonia Poker: Platform for online poker games [Yes] • 422, Eloplay: Platform for players to organize and participate in tournaments [Yes]	Yes
Topic 2	*[energy, mining, power, project, production]* *ICOs relating to the production of foods, materials, energy, etc.* • 3060, 3cCoin: Investment vehicle in the production of composite cellular concrete [Yes] • 2903, BananaPapa: Investment in the production of banana chips in Colombia [Yes] • 4577, APIS: Investment vehicle in the production of honey from bees [Yes]	Yes
Topic 3	*[information, may, white, paper, risk]* *ICOs relating to the trading/analysis of real-time data* • 3373, AMO Blockchain: Generating value from car data [Yes] • 50, Indorse: Selling of professional social networking data of yourself [Yes] • 270, Intelligent Trading Technologies: Providing analysis of real-time data [Yes]	Yes
Topic 4	*[transaction, node, smart, protocol, block]* *ICOs relating to improvements in the blockchain technology* • 2532, Quant: Addressing the problem of blockchain interoperability [Yes] • 1178, Blockstack: Decentralization of the storage of data to minimize hacking risk [Yes] • 4, Tezos: Blockchain technology with the ability to evolve its own protocols [Yes]	Yes
Topic 5	*[community, marketing, project, new, reward]* *ICOs relating to the provision of a social media platform or community for its users* • 2813, Life Change: Provision of a Christian community for people to interact with one another [Yes] • 1980, Smoke: Provision of a community for cannabis-related activities [Yes] • 1761, Sphere: Social media platform with the ability to give privacy to users [Yes]	Yes
Topic 6	*[may, technology, sale, financial, risk]* *ICOs relating to the provision of platforms for investment* • 623, Fairgrounds: Provision of a platform for betting games that can be proven to be fair [Yes] • 3692, GBX: Provision of a digital asset marketplace for people to trade [Yes] • 624, Protos: Cryptocurrency investment fund that invests in other cryptocurrencies [Yes]	Yes
Topic 7	*[information, health, medical, patient, technology]* *ICOs relating to the provision of a shared platform to store data collectively* • 1744, Nam Coin: Creation of an open medical data system in Japan for doctors to track patients' progress [Yes] • 91, Wolk SWARMDB: Creation of an open database system for companies to use and take advantage [Yes] • 1438, Healthureum: Creation of an open database for healthcare records to be kept [Yes]	Yes
Topic 8	*[content, video, advertising, publisher, digital]* *ICOs relating to linking content with payment without the intermediary* • 1901, Tipper: Platform for video distribution where users will decide how to allocate advertising dollars [Yes] • 577, StreamSpace: Linking independent film producers with audiences [Yes] • 2978, LikeCoin: Providing an avenue for creative content providers to be paid for their content [Yes]	Yes
Topic 9	*[payment, transaction, wallet, coin, sale]* *ICOs relating to facilitation of payments for different form of transactions* • 2198, ANN: Peer-to-peer platform for real estate transactions [Yes] • 1963, IAMHERO: Platform for employers and job seekers [Yes] • 2625, FOP Coin: Payment platform for retail transactions [Yes]	Yes
Topic 10	*[currency, fund, asset, project, investor]* *ICOs relating to the provision of financial services through blockchain* • 956, CEDEX: Global exchange for diamond trading [Yes] • 1297, Datarius: Platform for financial services using cryptocurrency assets [Yes] • 393, coinloan: Platform for obtaining loans on cryptocurrency assets [Yes]	Yes

EXHIBIT 15
Human Validation of LDA Unigram Approach with the Bottom Three Documents

Topic	Topic Words/Description	Consistent
Topic 1	*[game, player, team, bet, ticket]* ICOs relating to gaming for items with monetary value and/or betting • 2018, Florafic: Platform for eSport gaming allowing players to earn money and create a career [Yes] • 1010, Play2Live: Platform for monetizing a game streamer's content and allowing for interaction with users [Yes] • 765, Refereum: Provision of referral marketing in the video gaming industry [Yes]	3 out of 3
Topic 2	*[energy, mining, power, project, production]* *ICOs relating to the production of foods, materials, energy, etc.* • 665, BitRent: Platform allowing for purchase/monitoring/sale of real estate construction projects at all stages of development worldwide [Yes] • 2271, Rouge Project: Platform for distributing of vouchers for sale purposes [Yes] • 745, AKM Global: Production of pizza for sale [Yes]	3 out of 3
Topic 3	*[information, may, white, paper, risk]* *ICOs relating to the trading/analysis of real-time data* • 3644, BCG Prediction Markets: Forecasting with crowd intelligence through the trading of information [Yes] • 3376, Ophircoin: Establishment of coin mining operations and sharing of profits with coin holders [Yes] • 3614, Tatatu: Social media platform for the creation and provision of content with monetization [No]	2 out of 3
Topic 4	*[transaction, node, smart, protocol, block]* *ICOs relating to improvements in the blockchain technology.* • 29, Primalbase: Establishment of coworking spaces using cryptocurrency to establish rights to the space [Yes] • 2462, ChainRepublik: Massive multiplayer online game running on the browser with blockchain technology without a central server [Yes] • 3965, ICST: Blockchain technology used to store, share,and protect digital creative content for artists while providing an ecosystem for revenue sharing [Yes]	3 out of 3
Topic 5	*[community, marketing, project, new, reward]* *ICOs relating to the provision of a social media platform or community for its users* • 1608, ALTTEX: Establishment of a community of users to allow for the exchange of messages and monies [Yes] • 2512, Global Spy: Splitting research costs with the community of investors [Yes] • 3329, Online.io: Creating a community of users who can browse an alternative world wide web without advertising [Yes]	3 out of 3
Topic 6	*[may, technology, sale, financial, risk]* *ICOs relating to the provision of platforms for investment* • 333, Notary: Establishment of a smart contract process [Yes] • 1546, KPR Medical Solutions: Tokens allow for trade of marijuana in Australia [No] • 3231, Foodcoin Ecosystem: Trading platform for stakeholders to deal with food and agricultural products [Yes]	2 out of 3
Topic 7	*[information, health, medical, patient, technology]* *ICOs relating to the provision of a shared platform to store data collectively* • 1410, LevelNet: Data platform that allows the sharing of cybersecurity threats among antivirus programs [Yes] • 2991, Rotharium: Provides a shared platform to allow users to create smart contracts [Yes] • 2713, Aitheon: Adopts machine learning to resolve problem statements within the blockchain and introduces human intervention when necessary [No]	2 out of 3
Topic 8	*[content, video, advertising, publisher, digital]* *ICOs relating to linking content with payment without the intermediary* • 2584, Arqute: Links the producer and the authors of children animation content [Yes] • 2212, Smart Little Machine: Ecosystem for users to obtain for one-stop services for creating their own websites [Yes] • 2279, Prover: Verification of authenticity of user-generated video files [No]	2 out of 3
Topic 9	*[payment, transaction, wallet, coin, sale]* *ICOs relating to facilitation of payments for different form of transactions* • 308, databroker dao: Enabling the trading of Internet of Things sensor data between parties [Yes] • 3393, Biggdata Crypto Currency: Facilitates payments between merchants for various cryptocurrencies [Yes] • 436, Humaniq: Provision of payment facilities in places without payment infrastructures [Yes]	3 out of 3
Topic 10	*[currency, fund, asset, project, investor]* *ICOs relating to the provision of financial services through blockchain* • 3747, Freedom Coin: Cryptocurrency with the ability to send coins anonymously without others knowing that such a transaction occurred [No] • 2274, Carboneum: Social trading platform for users [Yes] • 2514, AIREXE: Cryptocurrency to replace fiat currency [Yes]	2 out of 3

draw a relation to any of these topics. Hence, the LDA model was not able to establish differences regarding how these documents should be segregated among the 10 topics. The whitepaper Crypto-ISBN ICO (no. 3010) relates to the establishment of a global standard for International Standard Book Number (ISBN) to be used by publishers to sell media or publications globally. Cypher ICO (no. 406) is a mobile application that uses geolocation and augmented reality technology to role-play the mining of the cryptocurrency tokens created by the company.

From the results in Exhibit 15, it can be seen that the LDA model did not perform as well in terms of the topic probabilities for the last three documents classified in a particular topic. This is especially so for documents within topics 3, 6, 7, 8 and 10. Some of these whitepapers were not classified appropriately because their themes could not be clearly distinguished from the text. For example, the whitepaper for Aitheon ICO (no. 2713) described the use of machine learning, but there was no clear industry that the ICO was targeting or service or product it was selling. Therefore, the categorization was not as distinct as the rest of the ICO whitepapers. For these particular ICOs, it should be acceptable that if humans are not able to segregate them, machines would likewise not be able to do a proper job.

Based on the results in Exhibit 16, the topic model performed well (i.e., more than 7 of 10 papers have a consistent theme) with most of the topics, except for topics 5 and 6. This likely was due to the evenly balanced mixture of topics within those whitepapers. For example, the Krosscoin (ICO no. 2034), the first random selection of topic 5, had a probability distribution of 31% in topic 5 and 25% in topic 3. This made it difficult for the whitepaper to be classified into a specific topic.

The results from this validation process were positive. The reading of the top three documents successfully determined the topics to be interpretable and successfully categorized the whitepapers into different themes. Even though the bottom three documents were not as clearly categorized, these documents were not easily segregated even from a human's perspective because of their lack of a clear business model. Hence, such exceptions should be considered acceptable. To further validate the documents processed, 10 documents from each topic were randomly selected and verified in the same manner, also with good results.

Proportion of Successful and Unsuccessful ICOs

Research on validation of topic models is limited to mathematical formulas that advocate likelihood derivations. The present article proposes another way of looking at the validity of the topic model developed.

As noted in the bar chart of the proportion of successful and unsuccessful papers within each topic (Exhibit 17), there are significant differences in the proportion of the successful and unsuccessful documents within 3 out of 10 topics. This observation suggests that the ICOs have been successfully split into their consistent themes, which are nonrandom, where interest levels from investors differ.

Topic 1 (69.1%), topic 4 (67.2%), and topic 10 (66.0%) had a higher proportion of successful ICOs. These ICOs relate to gaming, blockchain technology, and financial services, respectively. This is approximately 10% higher than the overall population, where the proportion of successful whitepapers was 58.5%.

To determine whether the model performed significantly better than a random model, a two-population proportion test was conducted with the following hypothesis:

Null hypothesis, H_0: $p_1 - p_2 = 0$

Alternative hypothesis, H_1: $p_1 - p_2 \neq 0$

where p_1 represents the proportion of whitepapers that are successful among the whitepapers classified under the selected topic, and p_2 represents the proportion of whitepapers that are successful among the remaining whitepapers that have not been classified under the selected topic. The null hypothesis is that the proportion of successful whitepapers is the same for the ICOs selected by this model and the remaining ICOs. The alternative hypothesis is that the proportion of successful whitepapers is not the same for the ICOs selected by this model compared with the remaining ICOs.

With n_1 denoting the number of whitepapers in the topic selected and n_2 denoting the number of whitepapers not within the topic selected, $n_1\ p_1(1 - p_1)$ and $n_2\ p_2(1 - p_2)$ have been determined to be greater than 10. Therefore, the distribution can be determined to be approximately normal, allowing the z-test to be conducted (Sullivan 2017).

EXHIBIT 16
Human Validation of LDA Unigram Approach with Randomly Sampling Approach

Topic	Topic Description	Consistent
Topic 1	*[game, player, team, bet, ticket]* *ICOs relating to gaming for items with monetary value, ticketing, and/or betting* • 581, Guaranteed Entrance Token: Platform for the direct sale and resale of event tickets [Yes] • 2018, Florafic: Creation of a training program for gamers to improve their skills and to create a currency to represent gaming items and rewards for trading [Yes] • 2243, Virtual Reality Games Company: Funding for the expansion of virtual reality gaming arcades; tokens can be used to buy gaming time at the arcades or be kept for profit-sharing purposes [Yes] • 1364, TokenStars: Platform for management of celebrity talent, advertisers, sponsors, and fans [No] • 2515, dueltoken.io: Platform for challenging people to a duel with a wager [Yes] • 2952, Crazy Shapes: Platform for trading items for the game Crazy Shapes [Yes] • 932, EtherSport: Platform for sports or gaming wagering [Yes] • 185, Rasputin: Funding for adult entertainment platform with profits distributed to token holders [Yes] • 2929, Play Bunk: Platform to connect gamers and developers without a mediator [No] • 1385, ETHORSE: Platform for betting on cryptocurrency price changes and other events [Yes]	8 out of 10
Topic 2	*[energy, mining, power, project, production]* *ICOs relating to the production of foods, materials, energy, goods, etc.* • 1739, Smart Gold: Production of gold in Tanzania [Yes] • 745, AKM Global: Production of pizza for sale [Yes] • 2001, Valena-SV Coin: Production of motor and industrial oils [Yes] • 1530, Restart Energy: Platform for trading/production of electrical energy [Yes] • 2697, Hive Power: Platform for electrical energy sharing [Yes] • 2271, Rouge Network: Platform for voucher trading [No] • 860, econan: Production of metals through recycling of the dust from the electric arc furnaces that are produced from steel manufacturing [Yes] • 2903, Banana Papa: Production of banana chips in Colombia [Yes] • 2810, I am Aero: Production of manned and unmanned aerial vehicles [Yes] • 1099, Energi Mine: Incentivizing renewable energy production [Yes]	9 out of 10
Topic 3	*[information, may, white, paper, risk]* *ICOs relating to the trading/analysis of real-time data* • 3136, Mandala: Trading platform for cryptocurrencies; trading data are used to provide users with trading advice and risk management strategies [Yes] • 1742, Bitcoinus: Online payment processing platform [No] • 159, easyMine: Facilitating the cryptocurrency mining industry [No] • 3706, ParkinGo: Facilitating car rental and sharing for airports [Yes] • 2691, Coinstocks: Trading platform for cryptocurrency [No] • 988, Token Pay: Platform for provision of banking services and closed-end private exchange for money transfers [Yes] • 2152, SRCOIN: Platform for the collection of health data from individuals through massage chairs; information is then traded with the government [Yes] • 549, AUTONIO: Automated trading platform that leverages artificial intelligence to trade cryptocurrencies [Yes] • 633, United Traders: Platform for trading of cryptocurrencies [Yes] • 4603, Bino Exchange: Platform for crypto-to-crypto and crypto-to-fiat exchange [Yes]	7 out of 10
Topic 4	*[transaction, node, smart, protocol, block]* *ICOs relating to improvements in the blockchain technology* • 1405, Hicky Network: Dating network that relies on peers for verification of dating experiences [No] • 1261, TravelChain: Blockchain designed to allow tagging of specific information for sale to others [Yes] • 694, BLOCKv: Blockchain interface allowing developers to issue experiential smart tokens to users [Yes] • 2532, Overledger: Blockchain technology that allows for ledgers across different silos to come together [Yes] • 1968, Thought: Blockchain technology allowing the processing of data without the need for servers [Yes] • 2339, Blockchain Terminal: Blockchain technology using a private ledger and global ledger [Yes] • 1805, Dock: Blockchain technology used to trade professional data with others [No] • 957, Lamden: Framework for the rapid deployment of blockchain technologies [Yes] • 1417, Fabric: Framework for easy adoption of blockchain technology and smart contracts [Yes] • 1178, Blockstack: Technology for the creation of the Internet without remote servers [Yes]	8 out of 10

(continued)

EXHIBIT 16 *(continued)*
Human Validation of LDA Unigram Approach with Randomly Sampling Approach

Topic	Topic Description	Consistent
Topic 5	*[community, marketing, project, new, reward]* *ICOs relating to the provision of a social media platform or community for its users* • 2034, Krosscoin: Framework for the monetization of apps within the Ethereum decentralized blockchain [No] • 1250, TokenGO: Platform for the formation of a local community of actual and potential investors and cryptocurrency experts [Yes] • 204, Shine Coin: Membership coin for discounted webcam services [No] • 2172, ArmPack: Platform for maintaining a community of buyers and manufacturers to fight counterfeit goods [Yes] • 2319, Biotron: Platform for keeping control of sale of data with users [No] • 3329, Online.io: Platform for the creation of a trusted Internet through a community of users [Yes] • 293, Sense: Provision of a social media platform for the engagement of messaging across difference applications [Yes] • 391, Wings: Community of users who can identify and promote high-value proposals that have higher chances of positive financial returns [Yes] • 199, Latium: Community allowing others to hire people for various tasks for a payment [Yes] • 593, Wysker: Platform for e-commerce allowing users to protect their data while giving them the opportunity to trade data directly with advertisers [No]	6 out of 10
Topic 6	*[may, technology, sale, financial, risk]* *ICOs relating to the provision of platforms for investment* • 427, Polybius: Provision of a digital single market for investments through a digital pass [Yes] • 2031, Sentinel Chain: Provision of a business-to-business marketplace for financial services accepting the use of livestock as collateral [Yes] • 134, impak Finance: Investment platform wherein only investments for social good are considered [Yes] • 3453, Crypto Ads: Platform for the trading of ads without the need for centralized ad exchanges [No] • 3729, Horyou: Platform for promoting social entrepreneurship by bridging socially responsible companies and users for funding [Yes] • 1300, Bread: Platform for the usage of a decentralized financial institution [Yes] • 2414, Alt.Estate: Platform for investment in fractions of real estate investments [Yes] • 1025, NaviAddress: Platform to create and develop a worldwide universal digital address solution [No] • 1240, Budbo: Platform for the buying of cannabis [No]	7 out of 10
Topic 7	*[information, health, medical, patient, technology]* *ICOs relating to the provision of a shared platform to store data collectively* • 4599, Aston: Platform for secure electronic document authentication [Yes] • 2518, HighSeek: Production of neurointerface headset for mobile devices [No] • 2636, n'cloud.swiss: Cloud computing platform provider globally [No] • 1253, Lympo: Platform for fitness/health data sharing [Yes] • 2846, Oris.Space: Platform for carrying out predictions based on crowd wisdom; ability to leverage own and platform's data for predictions [Yes] • 1019, Solve Care: Platform for sharing information among consumers, insurers, and service providers to improve healthcare access [Yes] • 1759, HPLUS: Decentralized healthcare record management platform [Yes] • 2609, OGSoft: Platform for the sharing of information among hospitals, medical clinics, and patients [Yes] • 1446, CareX: Platform for payment for healthcare services and secure storage of healthcare information from medical records to insurance details [Yes] • 1410, LevelNet: Platform for sharing information on cyberthreats globally [Yes]	8 out of 10
Topic 8	*[content, video, advertising, publisher, digital]* *ICOs relating to linking content with payment without the intermediary* • 2582, Nollycoin:Platform for peer-to-peer movie streaming [Yes] • 1474, Shivers Media: Platform for the direct distribution of horror films [Yes] • 265, VuePay: Platform for viewing targeted videos while users are eligible to a share of the advertising spend [Yes] • 1311, VOXXO: Platform for music listening and music access [Yes] • 2194, MoviesChain: Platform for independent filmmakers to distribute content on attractive financial terms by being connected directly with the audience [Yes] • 1310, Oko: Platform for distribution of adult entertainment material. [Yes] • 1393, TV-Two: Platform for connecting viewers with TV show content creators directly [Yes] • 2212, Smart Little Machine: Platform for the creation of websites [No]	8 out of 10

(continued)

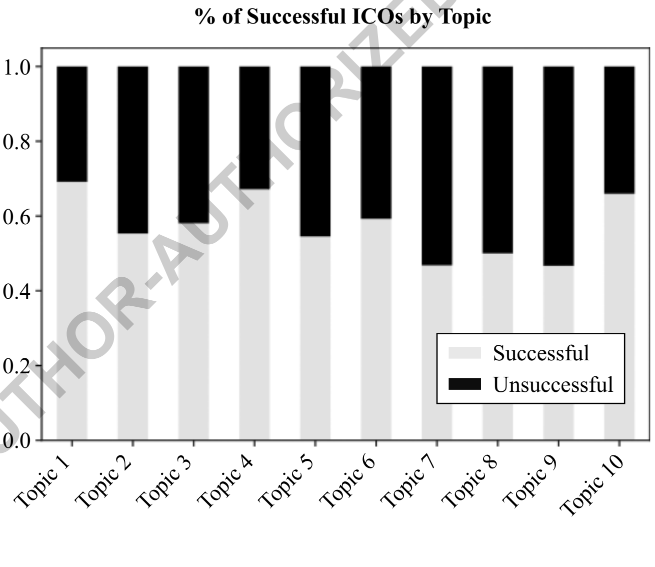
EXHIBIT 16 *(continued)*

Human Validation of LDA Unigram Approach with Randomly Sampling Approach

Topic	Topic Description	Consistent
Topic 9	*[payment, transaction, wallet, coin, sale]* *ICOs relating to facilitation of payments for different form of transactions* • 1049, Gift: Platform for giving donations to charities with confidence. [Yes] • 2443, Vendex: Platform for the listing of cryptocurrencies for trading in Africa [No] • 1797, Paygine: Platform to help consumers and businesses use cryptocurrency as simply and conveniently as any other currency [Yes] • 2838, XRED: Platform for users to invest in real estate at any moment [Yes] • 2427, SecureDonation: Platform for giving donations to charities with confidence [Yes] • 617, Global Jobcoin: Platform for paying employees and recruiting new people [Yes] • 1951, Akaiito: Platform for making payment with cryptocurrencies easy [Yes] • 2190, Thai Club: Platform for making cryptocurrency payment more accessible and available [Yes] • 388, MCO: Platform for creation of credit card payments through the use of cryptocurrencies [Yes] • 2662, Berith: Platform for the facilitation of payment [Yes]	9 out of 10
Topic 10	*[currency, fund, asset, project, investor]* *ICOs relating to the provision of financial services through blockchain* • 2506, CoinJanitor: Platform for conducting swaps for dead coins [No] • 3670, Payiza: Exchange allowing users to buy and sell cryptocurrencies directly with fiat money [Yes] • 1977, GlobCoin: An asset class pegged to a basket of currencies for users to own. [Yes] • 213, Hero: Financial services platform for the poor in Southeast Asia [Yes] • 1456, Digitex: Commission-free futures trading market [Yes] • 746, Social Financial Network: Peer-to-peer lending online service that brings creditors and borrowers together [Yes] • 1028, RAISON: Platform designed to handle investments and personal finance to maintain an optimal investment portfolio [Yes] • 2274, Carboneum: Platform for engaging in social trading, otherwise known as copy trading [Yes] • 3835, DRCG: Cryptocurrency with proof of asset ownership against gold to be used as a payment medium [Yes] • 1864, LocalCoinSwap: Platform allowing buyers and sellers to trade directly in cryptocurrency with any mode of payment [Yes]	9 out of 10

EXHIBIT 17

Proportion of Successful and Unsuccessful Whitepapers by Topic



The following formulas were used to derive the z-test statistic:

$$z - test\ \ statistic = \frac{p_1 - p_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3}$$

where

$$p = \frac{x_1 + x_2}{n_1 + n_2} \tag{4}$$

where x_1 denotes the number of successful whitepapers within the topic selected, and x_2 denotes the number of successful papers within the rest of the topics not selected.

As observed in Exhibit 18, at a significance level of 5%, all three topics (i.e., 1, 4, and 10, as highlighted within the exhibit) have statistically different proportions of successful ICOs compared with the rest of the ICOs in the sample. Hence, it can be concluded that the topic model

EXHIBIT 18
***P*-Value of ICO Segregation**

Topic	Z-Test Statistic	*P*-Value
1	2.0791	0.0188
2	–0.5970	0.7249
3	–0.1000	0.5400
4	2.3013	0.0107
5	–1.0710	0.8578
6	0.1666	0.4339
7	–1.6887	0.9544
8	–0.1430	0.9236
9	–2.6769	0.9963
10	2.0441	0.0205

has created three useful variables to statistically demonstrate that the topics are distinct. If new ICOs in the topics continue to be successfully funded, the topics could be used to determine an ICO's probability of success.

CONCLUSION

This article proves and validates that the LDA model is a useful tool for analyzing ICO whitepapers by successfully segregating the ICOs into different themes, allowing potential investors to segregate ICOs by industries. There is more than one way to validate the topic model, such as using humans, comparing complexities, measuring coherence, and observing the proportion of difference in successful and unsuccessful ICOs among the 10 topics. It was also found that, within 3 of the 10 topics, there were significant differences in the proportion of successful whitepapers compared with the remaining whitepapers.

Because of the statistical differences in the proportion of successful whitepapers across three of the topics, it is suggested that the topic model helps to create variables that might aid in the building of a prediction model for determining whether an ICO would be successful.

There were some whitepapers for which humans were unable to segregate the ICOs into categories because of the complexity of the whitepaper or the lack of a clear business model description in the whitepaper. In such instances, it is acceptable that machines would likewise not be able to segregate these whitepapers.

REFERENCES

Bastani, K., H. Namavari, and J. Shaffer. 2019. "Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints." *Expert Systems with Applications* 127 (August): 256–271.

Bian, S., Z. Deng, F. Li, W. Monroe, P. Shi, Z. Sun, W. Wu, S. Wang, W. Y. Wang, A. Yuan, T. Zhang, and J. Li. 2018. "IcoRating: A Deep-Learning System for Scam ICO Identification." ArXiv:1803.03670 [cs.CL], http://arxiv.org/abs/1803.03670.

Bouma, G. 2009. "Normalized (Pointwise) Mutual Information in Collocation Extraction." https://pdfs.semanticscholar.org/1521/8d9c029cbb903ae7c729b2c644c24994c201.pdf.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang. 2009. "A Density-Based Method for Adaptive LDA Model Selection." *Neurocomputing* 72 (7–9): 1775–1781.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

Chuang, J., D. Ramage, C. Manning, and J. Heer. "Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis." In *CHI'12: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, pp. 443–452. Austin, TX: ACM Press, 2012.

Gantner, J., L. Demetz, and R. Maier. "All You Need Is Trust—An Analysis of Trust Measures Communicated by Cloud Providers." In *On the Move to Meaningful Internet Systems: OTM 2015 Conferences,* edited by C. Debruyne, H. Panetto, R. Meersman, T. Dillon, G. Weichhart, Y. An, and C. Agostino Ardagna, Lecture Notes in Computer Science series vol. 9415, pp. 557–574. Cham, Switzerland: Springer International Publishing, 2015.

Griffiths, T. L., and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (Supplement 1): 5228–5235.

Hagen, L. 2018. "Content Analysis of E-Petitions with Topic Modeling: How to Train and Evaluate LDA Models?" *Information Processing & Management* 54 (6): 1292–1307.

Konrad, M. 2017. "Topic Model Evaluation in Python with tmtoolkit." WZB Data Science Blog, November 9, 2017, https://datascience.blog.wzb.eu/2017/11/09/topic-modeling-evaluation-in-python-with-tmtoolkit/.

Kumar, B. S., and R. Vadlamani. 2016. "A Survey of the Applications of Text Mining in Financial Domain." *Knowledge-Based Systems* 114 (December): 128–147.

Machine Learning Plus. 2018. "LDA: How to Grid Search Best Topic Models?" April 4, 2018, https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples.

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

"Perplexity." Wikipedia, 2019, https://en.wikipedia.org/w/index.php?title=Perplexity&oldid=884167633.

Raza, A. 2018. "Analyzing China's Ultimate Ban on All Crypto and ICO Websites." CryptoSlate, February 7, 2018, https://cryptoslate.com/analyzing-chinas-ultimate-ban-crypto-ico-websites.

Röder, M., A. Both, and A. Hinneburg. "Exploring the Space of Topic Coherence Measures." In *WSDM' 15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* PP. 399–408. Shanghai, China: ACM Press, 2015.

Shinnou, H., and M. Sasaki. 2010. "Detection of Peculiar Examples Using LOF and One Class SVM." https://pdfs.semanticscholar.org/72c6/b9a38c8242e49430cad80b40992b858c56dd.pdf.

Sievert, C., and K. Shirley. "LDAvis: A Method for Visualizing and Interpreting Topics." In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70. Baltimore, MD: Association for Computational Linguistics, 2014.

Smith + Crown. 2018. "2017 Token Sales/ICOs in Review: Part II." S + C intelligence, February 12, 2018, https://www.smithandcrown.com/2017-token-sales-icos-review-part-ii.

Sullivan, M. *Statistics: Informed Decisions Using Data,* 5th ed. Boston: Pearson, 2017.

Zhang, S., W. Aerts, L. Lu, and H. Pan. 2019. "Readability of Token Whitepaper and ICO First-Day Return." *Economics Letters* 180 (July): 58–61.

Zhao, W. 2018. "Public Firm Becomes First to Launch an ICO in Singapore." CoinDesk, August 10, 2018, https://www.coindesk.com/public-firm-becomes-first-to-launch-an-ico-in-singapore.

To order reprints of this article, please contact David Rowe at d.rowe@pageantmedia.com or 646-891-2157.

ADDITIONAL READING

ICO Success Drivers: *A Textual and Statistical Analysis*

Paola Cerchiello, Paolo Tasca, and Anca Mirela Toma
The Journal of Alternative Investments
https://jai.pm-research.com/content/21/4/13

ABSTRACT: *Recently, tech-savvy start-ups and SMEs have begun to finance their business by raising funds through the issuance and sale of "branded" newly minted digital currencies. Since this alternative capital market is largely unregulated, investors need to separate the drivers of business success from those that indicate failure, or worse, fraudulent money-raising. Through standard logistic regression and extreme value logistic regression, the authors are able to shed light on the riskiest business projects. In particular, they find that the existence of a white paper, the number of advisors, the number of people on the team, and the presence of a Telegram chat are significant factors useful for discerning a successful capital raising from an unsuccessful one. Conversely, the non-existence of a company website and an inactive Twitter account are indicators of a possible scam.*

New Blockchain Intermediaries: *Do ICO Rating Websites Do Their Job Well?*

Dmitri Boreiko and Gioia Vidusso
The Journal of Alternative Investments
https://jai.pm-research.com/content/21/4/67

ABSTRACT: *The fintech revolution, crowdfunding, and blockchain-based funding have dramatically reduced borrowing and lending transaction costs. Many have argued that ultimately this would lead to the complete disintermediation of financing for start-ups and SMEs. However, persistent asymmetric information and moral hazard problems have*

led to the creation of a new class of intermediaries that play a vital role in these new innovative financing methods. The authors review the new ecosystem built around initial coin offerings (ICOs), and in particular study the role of the ICO aggregators, and listing and rating portals. Using their hand-constructed database of all ICOs from inception in 2013 to September 2017, the authors find robust statistical confirmation that extensive coverage of a particular fundraising campaign in the ICO aggregators' lists is associated with more successful token sales. However, ratings data seem and appear to vary considerably across different ratings websites and appears to be of mediocre quality. Investors should therefore treat such ratings with caution.

Initial Coin Offering to Finance Venture Capital: *A Behavioral Perspective*

Jinghan Cai and Ahmed Gomaa
The Journal of Private Equity
https://jpe.pm-research.com/content/22/3/93

ABSTRACT: *An initial coin offering (ICO) is the procedure whereby ventures raise capital by selling tokens to investors. Compared with traditional financing methods, ICOs are new and less understood by both the industry and academia. For example, the literature is not clear about which factors determine the success rate or the amount raised in an ICO. Existing literature (Fisch 2019) shows that the signals of the private information of an ICO's high quality determine the amount of funding in the ICO. This article is the first in the literature that provides evidence that investor sentiment and investor awareness are determinants of the amounts raised in an ICO. Moreover, it provides evidence confirming the signaling hypothesis. Specifically, ICOs with a pre-ICO and with a higher rating are more likely to raise more funds. This article is the only one in the literature that uses the entire market of 4,367 ICOs to confirm its findings. The results extend the understanding of the dynamics of an ICO to the behavioral field and enable investors and practitioners to understand the crucial determinants of both ICO success rates and the amounts raised.*