



Value creation in emerging technologies through text mining: the case of blockchain

Filippo Chiarello, Paola Belingheri, Andrea Bonaccorsi, Gualtiero Fantoni & Antonella Martini

To cite this article: Filippo Chiarello, Paola Belingheri, Andrea Bonaccorsi, Gualtiero Fantoni & Antonella Martini (2021): Value creation in emerging technologies through text mining: the case of blockchain, Technology Analysis & Strategic Management, DOI: [10.1080/09537325.2021.1876221](https://doi.org/10.1080/09537325.2021.1876221)

To link to this article: <https://doi.org/10.1080/09537325.2021.1876221>



View supplementary material [↗](#)



Published online: 20 Jan 2021.



Submit your article to this journal [↗](#)



Article views: 366








View related articles [↗](#)



View Crossmark data [↗](#)



Value creation in emerging technologies through text mining: the case of blockchain

Filippo Chiarello ^a, Paola Belingheri ^a, Andrea Bonaccorsi ^a, Gualtiero Fantoni ^b and Antonella Martini ^a

^aDepartment of Energy, Systems, Territory and Construction Engineering, University of Pisa, Pisa, Italy;

^bDepartment of Civil and Industrial Engineering, University of Pisa, Pisa, Italy

ABSTRACT

As technology progresses, organisations must understand where to direct their value-creating efforts to achieve or sustain competitive advantage. This is even more true in the case of emerging technologies, where innovative activities often focus on achieving a technology promise while overlooking a set of technological, operational, organisational and user-related problems that must be overcome before the technology can fulfil this promise. Through an innovative application of text-mining, this paper develops a practical methodology to identify a range of problems related to a technological field in an unsupervised manner, that may benefit firms, researchers and policymakers. We apply the methodology to the field of blockchain and compare it to traditional literature reviews.

ARTICLE HISTORY

Received 7 January 2020
Revised 16 November 2020
Accepted 6 January 2021

KEYWORDS

Sentiment analysis; value creation; emerging technology; blockchain

1. Introduction

In emerging markets, besides pursuing the promise of a technology, firms also need to identify and address the specific challenges that the technology may present, to better direct their value creation efforts. Innovative activities often focus on the promise that a technology will transform one or more sectors (Dedehayir and Steinert 2016), while overlooking a set of technological, operational, organisational and user-related problems (Wang, Han, and Beynon-Davies 2019) that must be overcome before the technology can fulfil its early promise. One of the issues that firms face is, therefore, identifying and mapping the above-mentioned problems as early as possible, to address them and focus on value creation.

Previous approaches to this issue have mainly followed two avenues. On the one hand, quantitative patent analyses (Madani, Daim, and Weng 2017) or text-mining in patents (Chiarello, Fantoni, and Bonaccorsi 2017) have been used to map competitor's research agendas, identify core and peripheral patents, and detect unsolved technical limits. On the other hand, literature reviews of scientific papers provide a broader overview of potential problems at a given moment in time, but require skilled personnel, are time-consuming and often lack in reproducibility (Snyder 2019). We combine the advantages of both approaches by developing a reproducible and efficient methodology to extract technology problems from a large and interdisciplinary body of scientific literature. Our research question is, therefore, *can an automated textual analysis approach be used to extract technology problems from scientific literature?*

Knowledge extraction from big textual data is a challenge (Chen, Chiang, and Storey 2012), due to information being embedded in technical texts such as patents and papers (Golzio 2012). On the

other side, methodologies to manage massive amounts of textual data, such as text-mining, are growing in scope and sophistication. One of the most effective techniques is sentiment analysis, the study of the polarity (positive vs. negative) of a text, and within this context, lexicon-based techniques are among the most accessible methods. Lexicons are collections of annotated words with positive or negative orientation. The overall sentiment of a text is, therefore, computed upon the polarity of the words it contains (Annett and Kondrak 2008).

In order to answer our research question, we propose an innovative methodology for the semi-automatic extraction of knowledge from academic articles, using text-mining techniques. Among many possible attributes related to technical knowledge, our methodology focuses on the collection and the analysis of the problems presented by a technology, for the purpose of informing firm's research agendas and directing their efforts towards areas where value can be created.

In addition, we propose a case study to apply the developed methodology to blockchain technology. Blockchain is radically innovative (Davidson, De Filippi, and Potts 2016), has a potential impact in many fields (Risus and Spohrer 2017) and suffers from both technical and non-technical issues (Wang, Han, and Beynon-Davies 2019).

The first contribution will be to provide a tool for the early mapping of technological problems to be solved in an emergent field. This map can be used by scholars and practitioners to identify research topics and open issues. Second, we contribute to innovation management studies since the tool can be used ex-ante to map the evolution of problems, thereby indicating where the potential for value creation lies. Finally, we offer a theoretical contribution to the natural language processing (NLP) and text-mining fields since we present an innovative method of information extraction from unstructured sources.

The application scenarios of such an output have a wide range of potential users, including firms that want to rapidly map a technological field, policymakers who wish to direct investment towards solving technology problems to boost innovation and gain technology advantages, researchers who wish to understand their positioning in their field, and research networks who may need to better define their strategy to exploit possible collaborations and synergies.

To maximise the contribution to practise of the present work it will be accompanied by code written in the statistical software R (R Core Team 2020) to ensure reproducibility of the method.¹

The remainder of the paper is structured as follows. Section 2 presents a literature review on the topics of value creation and technology problems. Section 3 presents our research setting and outlines the methodology. Section 4 illustrates the results obtained from the application of the methodology in the blockchain setting. Section 5 presents a mapping of manual literature reviews that identify blockchain problems and a comparison of the results with those stemming from our methodology. Section 6 concludes.

2. Related work

2.1. Value creation in emerging markets: promises vs. problems

Value creation occurs when a process or an activity manages to raise the novel and appropriate benefits that are received from its output or from the product or service to which it is applied (Lepak, Smith, and Taylor 2007). Firms usually concentrate on two types of value creation, either increasing processes efficiency (Amit and Zott 2001) or increasing distinctiveness or price-quality ratio of products and services, thereby better matching customer needs (Priem 2007). Actors in the marketplace need to first create value and then capture it to reap competitive benefits.

To create value, firms must first determine 'what is value/valuable, who values what, and where value resides' (Lepak, Smith, and Taylor 2007, 181). In nascent industries, value creation often coagulates around the so-called 'promise' of a technology (Hawlitschek, Notheisen, and Teubner 2018). This may be a key attribute that indicates the technology's potential suitability for a particular use case, as for example, the Blockchain's promise to create trustless and intermediary-free

systems for the sharing-economy thanks to decentralisation and security (Hawlitschek, Notheisen, and Teubner 2018). It may be the promise of changing industry architectures, such as Additive Manufacturing that was predicted to entirely decouple design from manufacturing (Goehrke 2019). It could also manifest as a performance feature of a technology, such as lithium-ion batteries in the electric vehicle sector that increase range (Young et al. 2013). Firms thus often innovate in one main ‘promising’ technological direction to achieve a planned value proposition. However, this approach may overlook several issues that can impact whether and how the product or service can be effectively adopted. This is the process described in Gartner’s hype cycle, where the technology’s potential to prove transformative and valuable in the short-term is over-estimated, ‘based on a sudden, overly positive and irrational reaction on the introduction of a new technology’ (Dedehayir and Steinert 2016, 31), until a series of issues are detected that shatters this illusion. The technology then transitions towards a trough of disillusionment, where firms must tackle and overcome the issues that have arisen.

Wang, Han, and Beynon-Davies (2019) classify potential problems into technological, operational, organisational and user related. In the above-mentioned cases, the focus of blockchain on building trust mechanisms exclusively within the software layer, ignored the behavioural component formed by the real-world interactions of the users, so far impeding the success of entirely trust-free peer-to-peer systems (Hawlitschek, Notheisen, and Teubner 2018). The focus of the automotive industry on lithium-ion batteries overlooked the importance of charging stations and determined a delay in EV adoption (Rezvani, Jansson, and Bodin 2015). Finally, in Additive Manufacturing, unforeseen problems emerged such as the need for regulation and the need for specialised design knowledge to capitalise on the benefits of 3D-printed parts (Goehrke 2019). Therefore, the solutions that focus on the advantages or promise of a new technology may find themselves at odds with the final consumers (blockchain), or may require additional technologies, products, regulations or know-how in order for their intended value proposition to materialise (additive manufacturing and electric vehicles). These problems are hard to be foreseen and are not exclusively technical. An effective search process for the most promising avenues for value creation must consider a broad range of aspects and possibly be data driven.

Examining the market exploitation of emerging technologies, An and Ahn (2016) found significant distances between the technological promise and the market implementation and described the strategic actions needed to overcome the chasm in commercialisation. The literature on value creation has focused on processes of value generation (Lepak, Smith, and Taylor 2007), measuring value creation (Lieberman, Garcia-Castro, and Balasubramanian, 2017), the location of technology challenges in the firm’s environment (Adner and Kapoor 2010) and demand-side perspectives (Priem 2007); there is thus a gap in the ex-ante identification of the most promising avenues for value creation focusing on the problems or disadvantages of a technology or product.

Indeed, few studies systematically approach the extraction of problems or disadvantages associated with emerging technologies. In the effort to anticipate disadvantages and failures, Apreda et al. (2014, 2019) suggest to systematically apply FMECA and Functional Analysis, methodologies drawn from engineering design. While they significantly mitigate the cognitive distortions of decision-makers such as desirability bias and overconfidence (Bonaccorsi, Apreda, and Fantoni 2020), for the time being, they have not been subject to automatic processing in texts. Han and Shin (2014) combine QFD and trend analysis to identify product disadvantages created by subsystems whose performance hinders an entire system, while Kwon, Kim, and Park (2017) apply Latent Semantic Analysis (LSA) to anticipate shortcomings in drone technology.

We argue that firms should strive to identify potential problems in a technology sector taking a broad view of various aspects that are required to fulfil a technology’s promise. An early identification of a wide range of problems may not only provide firms with interesting avenues for innovation but may also be a good predictor of future innovation trajectories, as firms competing in the market will need to address these problems before they can deliver on the value proposition.

In this space, text mining supported by NLP techniques represents a new approach. These tools have already shown potential in the mapping of technology trajectories and patent analysis, and we argue that an extension of their use in the above-mentioned case is timely.

2.2. The text-mining approach to emerging technologies

Emerging technologies share the properties of novelty, fast growth, large impact, coherence, uncertainty and ambiguity (Rotolo, Hicks, and Martin 2015), and are the object of a large, dedicated literature. Initially, studies were based on expert panels or bibliometric methods applied to the metadata of patents and publications. Technology emergence was examined mainly through curve fitting (Shin, Coh, and Lee 2013).

Following the introduction of NLP techniques, scholars extracted indicators from the text of patents and publications, using a variety of algorithms (Xu et al. 2019; see Lee et al. 2018 and Hassan et al. 2018 for comparative analyses) and adopted several similarity measures (Kreutz, Sahitaj, and Schenkel 2020) to delineate the fields and track the emergence and growth of words in documents. As an example, technical keywords may be extracted automatically from the full-text of patents, their frequency measured using Term Frequency-Inverted Document Frequency metrics, their meaning disambiguated using general lexicons such as WordNet (Joung and Kim 2017). Words can be extracted and clustered using Topic Modelling techniques (Kyebambe et al. 2017; Kay, Kim, and Kang 2019; Son, Kim, and Kim 2020; Wang et al. 2020; Ebadi et al. 2020) while network maps can be generated, identifying technological opportunities via link prediction (Yoon and Magee 2018; Kim, Kim, and Lee 2019). Hot topics can also be identified and monitored over time (Tu and Seng 2012; Yan 2014).

Until recently, NLP techniques were able to extract single words, or n-grams, often using general lexicons such as WordNet – that have a limited ability to identify complex technological descriptions. In addition, frequency and similarity measures were defined with reference to a collection of documents, not from large scale general corpora (Hu et al. 2018a). After the introduction of pre-trained embedding languages such as Word2vec, BERT and SciBert (see below for references and technical discussion), many studies used a larger context to identify the meaning of words in the document. Recent applications include delineation of scientific fields such as AI (Dunham, Melot, and Murdick 2020), mathematics (Greiner-Petter et al. 2020) and Covid-19 related research (Kricka et al. 2020; Hope et al. 2020), scientific summarisation (Zerva et al. 2020), identification of scientific sources (Sanyal et al. 2019), and textual analysis of financial disclosures (Siano and Wysocki 2019).

Few studies deal with emerging technologies and related issues, but the field is growing rapidly. Hu et al. (2018b) and Kai et al. (2019) used Word2vec to identify domain topics in research, among which new emerging hot topics, while Choi et al. (2020) used several pre-trained models to calculate word and document similarity and to identify time-evolving product opportunities. The availability of context-aware language models allows a better integration of texts drawn from heterogeneous sources, such as product descriptions and patent databases (Lee et al. 2020).

2.3. Problems in patents and papers

The computation of technology trajectories, especially using patents, has been used to identify where companies are currently investing to create value. However, this approach presents several limitations. First, patents focus on technological solutions (Madani, Daim, and Weng 2017), providing a limited snapshot of possible avenues for value creation. Although it is possible to identify both drawbacks and advantages of technologies or products using patents (Chiarello, Fantoni, and Bonaccorsi 2017), it is highly unlikely to find problems and solutions related to policy and legal issues, consumer interactions, skills and knowledge requirements, or new business models. On the other side, sources such as social media can be used to mine polarised (positive and negative) information

about products (Chiarello, Bonaccorsi, and Fantoni 2020), but these sources lack an appropriate technical content.

Another approach considers research papers as a knowledge source. There are indeed technological fields where failures are more likely to be found in technical or grey literature, due to lack of patents or due to their recent emergence. In fields such as bioinformatics, its identification as a new IPC class promoted patent growth that enabled a detailed description of the field and focused the attention of firms on patents (Appio, Martini, and Fantoni 2017). However, in both these cases, the negative perspective is a good indicator of technology success or failure.

3. Research setting and methodology

3.1. Research setting

The blockchain is a digital, open-source, peer-to-peer transaction system, based on decentralised databases and managed by a self-governing community. It provides cryptographically secure transaction-based modifications of the machine state so that transactions are irreversibly recorded and stored. (Davidson, De Filippi, and Potts 2016). The technology is a breakthrough as it solves a well-known problem of computational systems, i.e. the impossibility to build a tamper-proof payment system with irreversible transactions, thus avoiding the illegal duplication of value.

Thanks to its advantages, blockchain has created a hype, which has fuelled a growing interest in the drawbacks of this technology (Eyal and Sirer 2018). Indeed, Gartner's Hype Cycle analysis in 2019 placed blockchain on the peak of inflated expectations, well-poised to enter the trough of disillusionment (Gartner 2019).

This sector started with an open-source mindset, and many projects still do not pursue patenting. Moreover, the impossibility to patent algorithms and other issues are still being addressed by patent offices (EPO 2018). Therefore, scientific papers constitute a more promising and less biased avenue to examine the latest developments.

For these reasons, it is particularly interesting to apply our proposed methodology to automatically extract and map problems in blockchain.

3.2. Research methodology

Considering the research objectives, the approach adopted for the scientific literature analysis (with a focus on problems identification) is different with respect to the traditional ones (i.e. bibliometric analyses and keyword approaches) that are unable to provide an automatic deeper understanding of the texts' technical content. The methodology relies on NLP, supported by a technical knowledge base (Chiarello, Fantoni, and Bonaccorsi 2017). Following a bottom up-approach, these lexicons were redesigned and updated for an optimal application to scientific papers. The workflow of the proposed methodology is shown in Figure 1.

The process starts with the collection of abstracts belonging to the same technological field (i.e. blockchain), using the Scopus API,² matching documents that contain keywords in the title, abstract or keywords. The texts are then pre-processed using NLP tools (sentence splitter, tokenizer, lemmatisation). For each sentence, the sentiment polarity is computed, to select only the sentences having a negative polarity score that are more likely to represent problems. A topic modelling algorithm is then applied to these sentences with the aim of clustering them into coherent groups. The output of the topic modelling is finally evaluated by technology domain experts that label the identified clusters of problems.

3.2.1. Document collection and selection

The documents were extracted from Scopus in September 2019 through the query: *TITLE-ABS-KEY ('blockchain' OR 'block chain' OR 'block-chain')* which resulted in 6893 documents. As the term

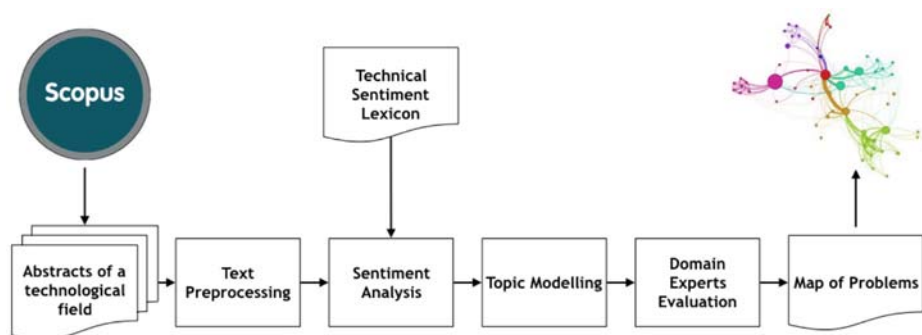


Figure 1. Workflow of the proposed methodology.

blockchain and its variations are used also in chemistry, all the papers belonging to chemistry-related All Science Journal Classification (ASJC) classes, as well as their keywords, were excluded with a result of 6601 papers.

3.2.2. Text preprocessing and sentiment analysis

A sentence splitter was applied to the abstracts to divide the text into 46,250 sentences, applying a state-of-the-art system (Straka and Straková 2017).

Then, the sentiment polarity of the sentences was measured (Rinker 2018) assigning a polarity score between -1 (strongly negative) and 1 (strongly positive) for each sentence depending on the dictionary entries found by the algorithm in the sentences, resulting in 7085 negative sentences.

Following this, state-of-the-art natural language processing was applied to the negative sentences, resulting in 218,618 tokens (single words).

To extract meaningful words (that add information about the problem that the sentence is describing) a series of filtering steps were applied, largely used in knowledge extraction from technical documents (Fantoni et al. 2013), using the lemma of the word as a meaningful unit of analysis. This resulted in 64,051 final lemmas. Considering the 7085 sentences in input, it implies a mean of 9 lemmas per sentence (from an initial mean of 25). This output constitutes a clean input for the topic modelling phase.

3.2.3. Topic modelling

Topic modelling is an unsupervised technique for documents clustering (Griffiths and Steyvers 2004). It generates statistical models able to capture word correlations in a collection of documents with a set of topics. Latent Dirichlet Algorithm (LDA) is a particularly popular method for fitting a topic model (Blei, Ng, and Jordan 2003). It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to overlap each other in terms of content, rather than being separated into discrete groups, mirroring the typical use of natural language.

The choice of LDA for topic fitting was compared to more advanced techniques for NLP. In the past years, transformer architectures and embedding language models such as BERT (Devlin et al. 2018) and SciBERT (Beltagy et al., 2019) have revolutionised NLP, evolving from standard word-embedding approaches (Pennington, Socher, and Manning 2014; Mikolov et al. 2013). Being relatively new, these approaches suffer from limitations such as lack of reproducibility, need for large amounts of computational resources and lack of transparency (i.e. black-box effect). In particular, the last point led to our decision to choose LDA, despite recent NLP literature having attempted to understand the encoding performed by the most effective language models (Radford et al. 2019). The focus is on finding out if the vectors, fitted by the model to encode each sentence, carry a deeper meaning than plain vocabulary aggregation. This is achieved via diagnostic classifiers (Conneau et al. 2018), used to inspect the vectors that the networks generate. Instead, LDA gives as

output both document-level and word-level embeddings, thus making the results easier to interpret for technical (i.e. blockchain) domain experts. Using the probability of a word to belong to a topic, β , non-NLP-experts have been able to validate our results. Furthermore, an important parameter of the Topic Model is α , the probability for a document to belong to a topic. This method is conceptually better than assigning each document to a problem since the same document can contain information about different problems.

Our approach is, therefore, complementary to the one followed by Kim, Park and Lee's (2020) Word2vec solution to detect general technology trends: we decided to favour the interpretability of the results, considering the well know interpretability-effectiveness trade off.

We started by forming a document-term matrix (DTM) using sentences as documents and lemmas as terms, counting how many times each token (lemma) occurs in each document. In our blockchain application, the DTM had 6834 rows representing the relevant sentences and 6299 unique tokens.

To fit an LDA model, we require the number of topics k that best represent the corpus under analysis. Different methods exist to compute an optimal value of k ; however, experts disagree with their results in terms of interpretability and coherence. Therefore, a novel hybrid approach is proposed using state-of-the-art k -tuning methods, starting from a graphical output computed for the five best k values to be evaluated by four different experts.

We employed the four most-used methods to evaluate the output of a topic model for different values of k (Cao et al. 2009; Arun et al. 2010; Griffiths and Steyvers 2004; Deveaud, SanJuan, and Bellot 2014).

These four measures were computed fitting an LDA model for every K between 2 and 20 (Figure 2) which constituted a sufficient range for the algorithms to converge. It is evident how the measures to be maximised intersect at a value of 6 topics; while the measures to be minimised are less stable (especially Cao et al. 2009) and intersect at 6, 8 and 9 topics, providing 4 candidates for an optimal output.

3.2.4. Domain expert evaluation

To make the final decision among the four models through consensus-building, experts need to be visually assisted so they can obtain a birds-eye view of the topics. A visual map of the topics (Appendix 1), was produced for each of the potential optimal values of K . These maps were analysed by a panel of four experts that deliberated on the optimal output. Greater details on the experts'

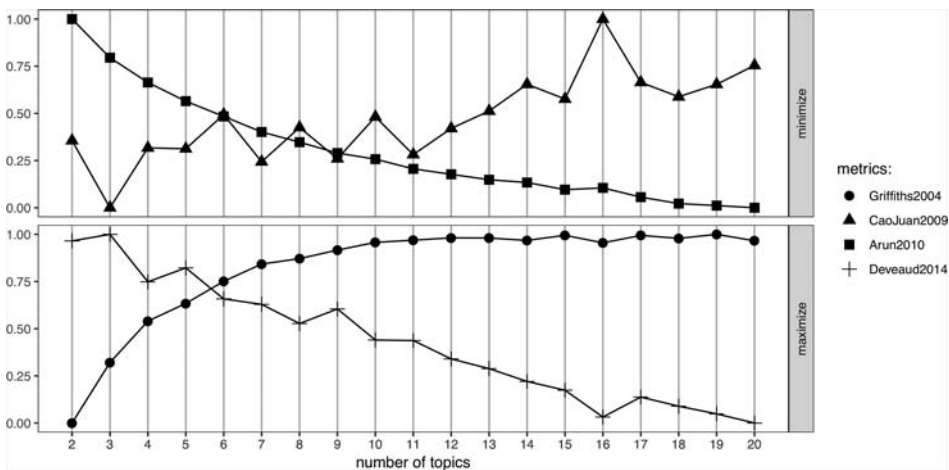


Figure 2. Quality measures of the topic modelling results for 2–20 topics, using minimising (top) and maximising (bottom) metrics.

evaluation methodology are shown in Appendix 2. The final decision to retain nine topics was taken considering that each topic should consistently represent a blockchain problem (i.e. easy to label) while avoiding overlaps between the topics. These topics were then labelled considering the word that had the maximum value of β for that topic.

3.2.5. Comparison of outcomes with previous research

The advantages obtained through our methodology include the possibility to rapidly examine a large body of documentation with minimum effort, classifying the results into meaningful clusters. However, these advantages must be measured against humans, rather than algorithms, identifying technology problems. Therefore, we performed a semi-systematic literature review (Snyder 2019), identifying articles that list problems in blockchain technology. We performed a search in Scopus for the words ‘literature’ AND ‘review’ AND ‘blockchain’ in the title and abstract resulting in a list of 179 papers and conference proceedings from 2016 to 2019. We retained 20% of the articles that received 90% of the citations, for a total of 38 articles. One of the authors, not involved in the creation and execution of the methodology, read through the abstracts and excluded those that did not comprise an overview of problems, issues or challenges related to blockchain. The same person then examined in detail the remaining 18 papers, excluding a further two, while adding two from the citations, finally highlighting mentioned problems, issues, and challenges (Figure 4). These were compared with the results illustrated in Section 4.

4. Results

4.1 Topic model interpretation

Each of the nine topics identified by the topic modelling phase represents an issue in the context of blockchain, that can be interpreted as follows:

1. **Power:** Blockchain has an environmental cost and indeed this topic is strongly related with topic 5.
2. **Storage:** refers to the infrastructure itself, or how data is stored, transformed, and shared in servers. This creates issues to be solved in terms of trust and scalability of the systems.
3. **Design:** The need to consider requirements coming from different disciplines (e.g. how to use sensors, how to certify transactions, how to scale the system).
4. **Communication:** Problems related to data and meta-data exchange in the network are crucial to ensure scalability, integrity, and energy efficiency. This is particularly true for cryptocurrencies such as Ethereum and Bitcoin.
5. **Cost:** The cost of the technology is still a major issue, but related problems also include policymaking.
6. **Bitcoin:** Bitcoin issues are discussed at length as it is the first and most notable blockchain application. Bitcoin’s main issues are related to consensus, trust, transparency and safe storage, especially in the cloud.
7. **Environment:** Blockchain technology could be a game-changer for the environment if properly managed, mitigating its high energy consumption.
8. **Consensus:** Consensus protocols are paramount since they ensure a common, unambiguous ordering of transactions and blocks and guarantee the integrity and consistency of the blockchain across geographically distributed nodes.
9. **Trust:** The principles of encryption and distributed ledgering behind blockchain, are hard to understand and thus hard to trust for most potential users.

4.2. Evolution of topics over time

Figure 3 shows the sum of a per year and per topic. We can see that *Power*, *Environment*, *Cost*, and *Trust* have garnered increasing attention, whereas *Storage* and *Bitcoin* have received less attention; finally *Design* and *Consensus* have a steady presence. This can provide firms with an indication of which problems are heading towards a solution or irrelevance and which are increasingly discussed. We considered these results to build the discussion made in the next section.

5. Comparison with literature reviews

Regarding the manual and bibliometric literature reviews examined, only four of them examine the whole field of blockchain (Yli-Huumo et al. 2016; Risus and Spohrer 2017; Casino, Dasaklis, and Patsakis 2018; Hughes et al. 2019). The remaining ones focus on specific perspectives such as authentication (Mohsin et al. 2018) and trust, or specific applications such as cyber-security (Taylor et al. 2019) and AI (Salah et al. 2019), or specific problems such as security (Alketbi, Nasir, and Talib 2018; Taylor et al. 2019).

The literature reviews included between 29 and 260 papers, the bibliometric study from Miao and Yang (2018) analysed 801 papers, whereas our algorithm was able to process 6601 papers. While there are many overlaps between the reviews, none presents a comprehensive view of the problems experienced by blockchain technology as we found 18 problems in total and the most exhaustive review lists 14.

Regarding the structure of the papers, most of them do not summarise the results in tables or graphs, therefore, the disadvantages are identified by carefully reading through the text. This creates a lack of synthesis, that slows down the comprehensions and exploration of problems, especially for companies.

Finally, the papers cited in the literature reviews span from 2008 to 2018. Since blockchain research is increasing exponentially (Mohsin et al. 2018; Miao and Yang 2018), all studies providing a snapshot become rapidly obsolete (Risus and Spohrer 2017), whereas our methodology can easily be replicated periodically with a limited effort.

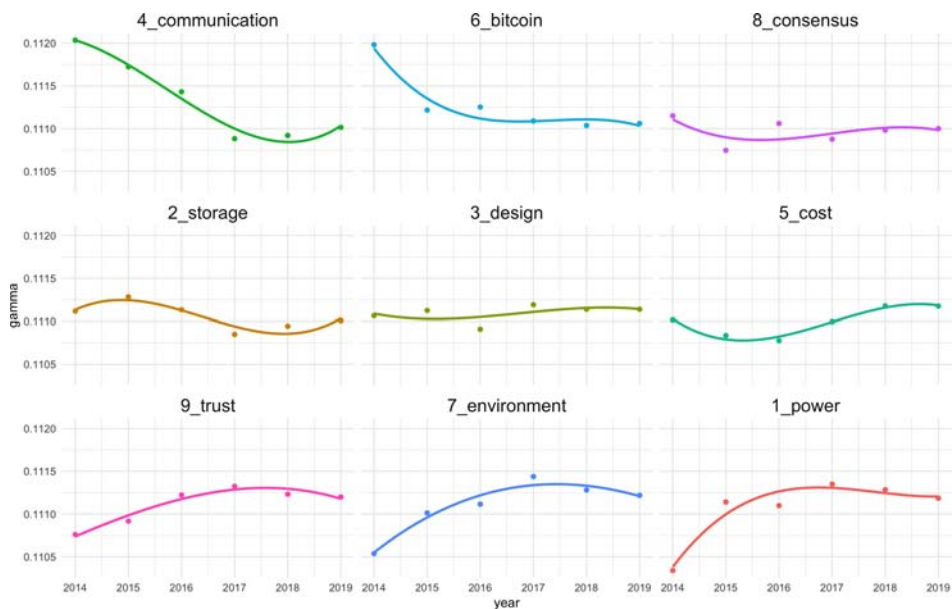


Figure 3. Mentions of topics over time.

Problems (Keywords)	Literature Review (2010-2021)											Citation (This paper)
	Ahmadova & Sheng (2010)*	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	Alm and Shalunov (2011)	
Applications (usability, need to develop more user-friendly APIs to facilitate the development of applications, understanding blockchain implementations and their potential for specific use cases)				x		x		x	x		x	1,2,3,4,6,7,8
Behavioral Layer (problems with user-software interactions, behavior changes required)			x					x				4,7,8
Cost (Energy consumption, computational cost, overhead costs created by consensus algorithms, environmental concerns, wasted resources)			x		x			x	x	x	x	1,5,7,8,9
Decentralization (consequences for maintenance, updates, hard forks, irreversibility, centralization tendencies, orphaning)	x								x			5,6,8,9
Design (which features are relevant for which industries and how to design them, computational and system design)			x					x		x		3,4,7,8,9
Governance (standards, interoperability, regulations, administration, troubleshooting, settlement of disputes, rights, power, transaction supervision, oversight)			x	x	x			x	x		x	2,5,6,7,8,9
Infrastructure (communication protocols, centralized storage, mining hardware, network administration/address management)								x		x		5,7,8,9
Integration (aligning real-world processes with blockchains, data transmission to and from blockchains, integration of complementary blockchains, interoperability)						x		x			x	3,4,6,7
Latency (large numbers of transaction per time unit, throughput, congestion)			x		x			x	x	x	x	5
Legal Rules (on blockchain usage, to prevent illegal activities through blockchain, service provision and fraud liability, policy)		x		x				x	x	x	x	5
Privacy (anonymity)	x		x	x		x		x	x	x	x	5,6,8,9
Quantum Resilience					x						x	-
Scalability (memory capacity, computational efficiency, storage, block size, bandwidth, side chains)		x	x		x		x	x	x	x	x	2,3,4,5,6,7,8,9
Security (preservation of data integrity, also on sidechains, hacking, intermediaries, availability)	x		x		x	x		x	x	x	x	2,4,6,7,8,9
Smart Contract Vulnerabilities and Deterministic Execution						x				x	x	-
Tradeoffs/Interdependencies between blockchain features, (e.g. between consensus, permissioning, scalability, decentralization, anonymity and interoperability, security and performance) and their consequences for use cases			x		x					x		-
Trust (trusting blockchain providers, smart contract developers, network validators, understanding the technology enough to enable trust, transparency)			x				x		x	x		2,4,5,6,8,9
Volatility of transaction currencies	x									x		-

Figure 4. Papers, problems mentioned and correspondence with methodology output topics. *Used also grey literature.

Figure 4 shows the 18 problems identified, the relevant keywords extracted from each paper, and their correspondence to the 9 clusters identified in our methodology. Figure 5 shows how many literature reviews mentioned each problem per year.

Out of a total of 18 topics mentioned in literature reviews, our methodology was able to identify 14, equalled only by Risus and Spohrer (2017). Because of its reliance on term frequency to form

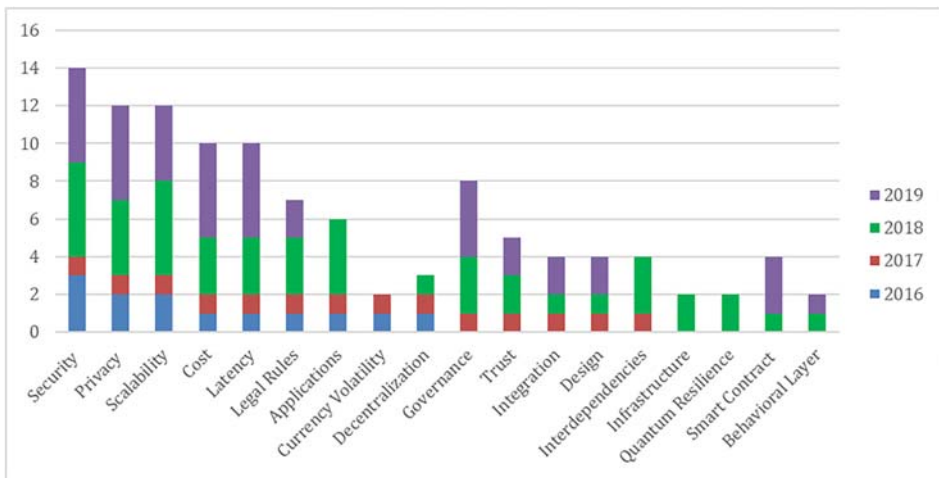


Figure 5. Number of papers mentioning each problem per publication year.

clusters, our methodology was proficient in extracting issues that were mentioned consistently for several years, or by a larger number of reviews (e.g. privacy, security, scalability), while it did not identify topics that were just emerging (e.g. smart contracts, quantum resiliency) or that were mentioned by fewer publications (e.g. volatility of transaction currencies).

6. Conclusions, limitations and further research

Novel technologies often attract attention based on an initial hype, built around advantages that promise to provide benefits in the short term. As technology development continues, new problems are identified that may impact the technology's ability to fulfil this promise. Overcoming these problems can enable firms to create value in an emerging sector and to position themselves as the technology gains market acceptance.

In this perspective, our first contribution is demonstrating that scientific literature contains valuable information regarding technology problems which can be automatically mined using NLP techniques. Through an innovative application of text-mining, 14 problems currently faced by blockchain-based solutions, out of 18 mentioned in the most-cited literature reviews, were identified and classified in an unsupervised manner. This tool, therefore, provides the opportunity to map the problems related to a technological field in a less time-consuming and more reproducible manner than current standard practices, thereby contributing to the NLP and text-mining fields.

As discussed by literature, technology problems can indicate where the potential lies for value-creating efforts. Therefore, we also contribute to the field of innovation research by proposing a novel technique to identify where the potential may lie for value creation. This could benefit a wide range of potential users, including firms who wish to map technological fields, policymakers who wish to direct investment to boost innovation and gain technology advantages, and researchers who wish to understand their positioning.

Our methodology presents a series of limitations. It may overlook problems that have been mentioned by few scientific publications or that have recently emerged. This could be mitigated by performing a network analysis of topics, linking them through mentions in the same paper (Madani, Boussaid, and Zegour 2015), or considering the importance of papers in terms of citations. Moreover, these emerging problems could be identified by repeating the methodology regularly and highlighting emerging weak signals. Furthermore, the use of certain words is not necessarily associated with content that is useful for business, such as words that are introduced because of a hype (e.g. artificial intelligence). Such expressions can be listed and under-weighted from the analysis.

Regarding generalisability, specific fields may have a particular language or content of scientific publications, that could make the described methodology more or less successful and we encourage its application beyond blockchain. The code used for the analysis is publicly available (see Section 1) and we offer our support to all the authors that want to reproduce this research over time and in other domains.

Notes

1. The code is found on git-hub at <https://github.com/FilippoChiarello/scientific-paper-analysis>
2. <https://dev.elsevier.com/>.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The data that support the findings of this study are openly available in figshare at https://figshare.com/articles/value_creation_blockchain_data/11398305, doi:10.6084/m9.figshare.11398305

Notes on contributors

Filippo Chiarello is an Assistant professor at the School of Engineering, University of Pisa. His research focuses on the use of Natural Language Processing techniques for studying technological and HR-related phenomena.

Paola Belingheri is an Assistant professor at the School of Engineering, University of Pisa. Her research focuses on markets for technology, and innovation ecosystems in the context of smart cities.

Andrea Bonaccorsi is a Full Professor of Economics and Management at the School of Engineering of the University of Pisa. He has authored in the most important journals in Economics of Science and Technology, Innovation Policy, Research Metrics and Evaluation and is ranked among the top 2% world scientists according to *PLoS ONE*.

Gualtiero Fantoni is an Associate professor at the School of Engineering, University of Pisa. His research focuses on data science and its applications for management practice. He coordinates several European projects on skills mapping and evaluation. He is co-founder of several university spin-offs in the fields of technology foresight and IoT.

Antonella Martini, PhD is a Full Professor of Business Economics and Strategic Analysis at the School of Engineering, University of Pisa. She is also President of CIMEA for academic mobility and equivalence. Her research interests include organisational ambidexterity, competitive intelligence, and strategic foresight. She has published in, among others: *California Management Review*, *Technological Forecasting and Social Change*, *Long Range Planning*, *Journal of Business Research* and *International Journal of Management Reviews*.

ORCID

Filippo Chiarello  <http://orcid.org/0000-0001-9857-0287>

Paola Belingheri  <http://orcid.org/0000-0001-9291-9302>

Andrea Bonaccorsi  <http://orcid.org/0000-0001-7425-9988>

Gualtiero Fantoni  <http://orcid.org/0000-0003-0772-600X>

Antonella Martini  <http://orcid.org/0000-0002-2006-4293>

References

- Abramova, S., and R. Böhme. 2016. "Perceived Benefit and Risk as Multidimensional Determinants of Bitcoin Use: A Quantitative Exploratory Study." In 2016 International Conference on Information Systems, December 11–14. Dublin.
- Adner, R., and R. Kapoor. 2010. "Value Creation in Innovation Ecosystems: How the Structure of Technological Interdependence Affects Firm Performance in new Technology Generations." *Strategic Management Journal* 31 (3): 306–333.
- Alketbi, A., Q. Nasir, and M. A. Talib. 2018. "Blockchain for Government Services – Use Cases, Security Benefits and Challenges." In 2018 15th Learning and Technology Conference (L&T), 112–119, February. IEEE.
- Amit, R., and C. Zott. 2001. "Value Creation in e-Business." *Strategic Management Journal* 22 (6–7): 493–520.
- An, H. J., and S. J. Ahn. 2016. "Emerging Technologies – Beyond the Chasm. Assessing Technological Forecasting and Its Implication for Innovation Management in Korea." *Technological Forecasting and Social Change* 102: 132–142.
- Andoni, M., V. Robu, D. Flynn, S. Abram, D. Geach, D. Jenkins, P. McCallum, and A. Peacock. 2019. "Blockchain Technology in the Energy Sector: A Systematic Review of Challenges and Opportunities." *Renewable and Sustainable Energy Reviews* 100: 143–174.
- Annett, M., and G. Kondrak. 2008. "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs." In *Conference of the Canadian Society for Computational Studies of Intelligence*, 25–35. Berlin, Heidelberg: Springer.
- Appio, F. P., A. Martini, and G. Fantoni. 2017. "The Light and Shade of Knowledge Recombination: Insights from a General-Purpose Technology." *Technological Forecasting and Social Change* 125: 154–165.
- Apreda, R., A. Bonaccorsi, F. Dell'Orletta, and G. Fantoni. 2019. "Expert Forecast and Realized Outcomes in Technology Foresight." *Technological Forecasting and Social Change* 141: 277–288.
- Apreda, R., A. Bonaccorsi, G. Fantoni, and D. Gabelloni. 2014. "Functions and Failures, how to Manage Technological Promises for Societal Challenges." *Technology Analysis and Strategic Management* 26 (4): 369–384.
- Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy. 2010. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: SOME Observations." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 391–402, June. Berlin, Heidelberg: Springer.

- Batubara, F. R., J. Ubacht, and M. Janssen. 2018. "Challenges of Blockchain Technology Adoption for E-government: A Systematic Literature Review." In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, May, 76. ACM.
- Beltagy, I., Lo, K., & Cohan, A., 2019. "SciBERT: A Pretrained Language Model for Scientific Text." arXiv preprint arXiv:1903.10676.
- Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.
- Bonaccorsi, A., R. Apreda, and G. Fantoni. 2020. "Expert Biases in Technology Foresight. Why They Are a Problem and How to Mitigate Them." *Technological Forecasting and Social Change* 151: 119855.
- 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.
- Casino, F., T. K. Dasaklis, and C. Patsakis. 2018. "A Systematic Literature Review of Blockchain-Based Applications: Current Status, Classification and Open Issues." *Telematics and Informatics* 36: 55–81.
- Chen, H., R. H. Chiang, and V. C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36 (4): 1165–1188.
- Chiarello, F., A. Bonaccorsi, and G. Fantoni. 2020. "Technical Sentiment Analysis. Measuring Advantages and Drawbacks of New Products Using Social Media." *Computers in Industry* 123: 103299.
- Chiarello, F., G. Fantoni, and A. Bonaccorsi. 2017. "Product Description in Terms of Advantages and Drawbacks: Exploiting Patent Information in Novel Ways." In *DS 87-6 Proceedings of the 21st International Conference on Engineering Design (ICED 17)*. Vol. 6: Design Information and Knowledge, 101–110, Vancouver, August 21–25.
- Choi, J., S. Oh, J. Yoon, J. M. Lee, and B. Y. Coh. 2020. "Identification of Time-Evolving Product Opportunities via Social Media Mining." *Technological Forecasting and Social Change* 156: 120045.
- Conneau, A., G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. 2018. "What You Can Cram into a Single Vector: Probing Sentence Embeddings for Linguistic Properties." arXiv preprint arXiv:1805.01070.
- Conoscenti, M., A. Vetro, and J. C. De Martin. 2016. "Blockchain for the Internet of Things: A Systematic Literature Review." In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 1–6, November. IEEE.
- Davidson, S., P. De Filippi, and J. Potts. 2016. "Economics of Blockchain." doi:10.2139/ssrn.2744751.
- Dedehayir, O., and M. Steinert. 2016. "The Hype Cycle Model: A Review and Future Directions." *Technological Forecasting and Social Change* 108: 28–41.
- Devaud, R., E. SanJuan, and P. Bellot. 2014. "Accurate and Effective Latent Concept Modeling for ad hoc Information Retrieval." *Document Numérique* 17 (1): 61–84.
- Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.
- Dunham, J., J. Melot, and D. Murdick. 2020. "Identifying the Development and Application of Artificial Intelligence in Scientific Text." arXiv:2002.07143v1.
- Ebadi, A., S. Tremblay, C. Gouffe, and A. Schiffauerova. 2020. "Application of Machine Learning Techniques to Assess the Trends and Alignment of the Funded Research Output." *Journal of Informetrics* 14: 101018.
- European Patent Office. 2018. "Talking About a New Revolution: Blockchain." In *Report of the Patenting Blockchain Conference*, The Hague, December 4. Accessed 22 November 2019. <https://tinyurl.com/rhqan87>.
- Eyal, I., and E. G. Sirer. 2018. "Majority Is Not Enough: Bitcoin Mining Is Vulnerable." *Communications of the ACM* 61 (7): 95–102.
- Fantoni, G., R. Apreda, F. Dell'Orletta, and M. Monge. 2013. "Automatic Extraction of Function–Behaviour–State Information from Patents." *Advanced Engineering Informatics* 27 (3): 317–334.
- Gartner. 2019. "Gartner 2019 Hype Cycle Shows Most Blockchain Technologies Are Still Five to 10 Years Away from Transformational Impact." Gartner. Accessed 2 November 2019. <https://tinyurl.com/y4gd3okl>.
- Goehrke, S. 2019. "Will 2019 Fulfill the Promises of 3-D Printing?" *Forbes*. Accessed 14 November 2019. <https://tinyurl.com/vgckr75>.
- Golzio, D. 2012. "WWWHOW (Why, When, Who, Where, What, How) Read a Patent!" European Patent Office. Accessed 14 November 2019. <https://tinyurl.com/uz6ee7t>.
- Greiner-Petter, A., A. Youssef, T. Ruas, B. R. Miller, M. Schubotz, A. Aizawa, and B. Gipp. 2020. "Math-word Embedding in Math Search and Semantic Extraction." *Scientometrics* 125 (3): 3017–3046. doi:10.1007/s11192-020-03502-9.
- Griffiths, T. L., and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (Suppl 1): 5228–5235.
- Han, K., and J. Shin. 2014. "A Systematic Way of Identifying and Forecasting Technological Reverse Salient Using QFD, Bibliometrics, and Trend Impact Analysis. A Carbon Nanotube Biosensor Case." *Technovation* 34: 559–570.
- Hassan, S. U., M. Imram, S. Iqbal, N. R. Aljohari, and R. Nawaz. 2018. "Deep Context of Citations Using Machine-Learning Models in Scholarly Full-Text Articles." *Scientometrics* 117: 1645–1662.
- Hawliczek, F., B. Notheisen, and T. Teubner. 2018. "The Limits of Trust-Free Systems: A Literature Review on Blockchain Technology and Trust in the Sharing Economy." *Electronic Commerce Research and Applications* 29: 50–63.

- Hope, T., J. Portenoy, K. Vasank, J. Borchardt, E. Horvitz, D. S. Weld, M. A. Hearst, and J. West. 2020. "SciSight: Combining Faceted Navigation and Research Group Detection for Covid-19 Exploratory Scientific Search." *bioRxiv*. doi:[10.1101/2020.05.23.112284](https://doi.org/10.1101/2020.05.23.112284).
- Hu, K., K. Qi, S. Yang, S. Shen, X. Cheng, H. Wu, J. Zheng, S. McClure, and T. Yu. 2018a. "Identifying the 'Ghost City' of Domain Topics in a Keyword Semantic Space Combining Citations." *Scientometrics* 114: 1141–1157.
- Hu, K., H. Wu, K. Qi, J. Yu, S. Yang, T. Yu, J. Zheng, and B. Liu. 2018b. "A Domain Keyword Analysis Approach Extending Term Frequency-Keyword Active Index with Google Word2vec Model." *Scientometrics* 114: 1031–1068.
- Hughes, L., Y. K. Dwivedi, S. K. Misra, N. P. Rana, V. Raghavan, and V. Akella. 2019. "Blockchain Research, Practice and Policy: Applications, Benefits, Limitations, Emerging Research Themes and Research Agenda." *International Journal of Information Management* 49: 114–129.
- Joung, J., and K. Kim. 2017. "Monitoring Emerging Technologies for Technology Planning Using Technical Keyword Base Analysis from Patent Data." *Technological Forecasting and Social Change* 114: 281–292.
- Kai, H., L. Qing, Q. Kunlun, Y. Siluo, M. Jin, F. Xiaokang, Z. Jie, W. Huai, G. Ya, and Z. Qibing. 2019. "Understanding the Topic Evolution of Scientific Literatures Like an Evolving City: Using Google Word2vec Model and Spatial Autocorrelation Analysis." *Information Processing and Management* 56: 1185–1203.
- Kay, H. J., C. Kim, and K. Kang. 2019. "Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA)." *Processes* 7: 379–399.
- Kim, J., S. Kim, and C. Lee. 2019. "Anticipating Technological Convergence. Link Prediction Using Wikipedia Hyperlinks." *Technovation* 79: 25–34.
- Kim, S., H. Park, and J. Lee. 2020. "Word2vec-based Latent Semantic Analysis (W2V-LSA) for Topic Modeling. A Study of Blockchain Technology Trend Analysis." *Expert Systems with Applications* 152: 113401.
- Konstantinidis, I., G. Siaminos, C. Timplalexis, P. Zervas, V. Peristeras, and S. Decker. 2018. "Blockchain for Business Applications: A Systematic Literature Review." In *International Conference on Business Information Systems*, 384–399, July. Cham: Springer.
- Kreutz, C. K., P. Sahitaj, and R. Schenkel. 2020. "Evaluating Semantometrics from Computer Science Publications." *Scientometrics* 125: 2915–2954. doi:[10.1007/s11192-020-03409-5](https://doi.org/10.1007/s11192-020-03409-5).
- Kricka, L. J., S. Polevikov, Y. Park, P. Fortina, S. Bernardini, D. Satchkov, V. Kolesov, and M. Grishkov. 2020. "Artificial Intelligence-Powered Search Tools and Resources in the Fight Against COVID-19." *eJIFCC* 31 (2): 106–116.
- Kwon, H., J. Kim, and Y. Park. 2017. "Applying LSA Text Mining Technique in Envisioning Social Impacts of Emerging Technologies. The Case of Drone Technology." *Technovation* 60–61: 15–28.
- Kyebambe, M. N., G. Cheng, Y. Huang, C. He, and Z. Zhang. 2017. "Forecasting Emerging Technologies. A Supervise Learning Approach Through Patent Analysis." *Technological Forecasting and Social Change* 125: 236–244.
- Lee, C., D. Jeon, J. M. Ahn, and O. Kwon. 2020. "Navigating a Product Landscape for Technology Opportunity Analysis. A Word2vec Approach Using an Integrated Patent-Product Database." *Technovation* 96–97: 102140. doi:[10.1016/j.technovation.2020.102140](https://doi.org/10.1016/j.technovation.2020.102140).
- Lee, C., O. Kwon, M. Kim, and O. Kwon. 2018. "Early Identification of Emerging Technologies. A Machine Learning Approach Using Multiple Patent Indicators." *Technological Forecasting and Social Change* 127: 291–303.
- Lepak, D. P., K. G. Smith, and M. S. Taylor. 2007. "Value Creation and Value Capture: a Multilevel Perspective." *Academy of Management Review* 32 (1): 180–194.
- Lieberman, M. B., R. Garcia-Castro, and N. Balasubramanian. 2017. "Measuring Value Creation and Appropriation in Firms: The VCA Model." *Strategic Management Journal* 38 (6): 1193–1211.
- Madani, A., O. Boussaid, and D. E. Zegour. 2015. "Real-time Trending Topics Detection and Description from Twitter Content." *Social Network Analysis and Mining* 5: 59.
- Madani, F., T. Daim, and C. Weng. 2017. "'Smart Building' Technology Network Analysis: Applying Core-Periphery Structure Analysis." *International Journal of Management Science and Engineering Management* 12 (1): 1–11.
- Miau, S., and J. M. Yang. 2018. "Bibliometrics-based Evaluation of the Blockchain Research Trend: 2008–March 2017." *Technology Analysis & Strategic Management* 30 (9): 1029–1045.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems*, 3111–3119.
- Mohsin, A. H., A. A. Zaidan, B. B. Zaidan, O. S. Albahri, A. S. Albahri, M. A. Alsalem, and K. I. Mohammed. 2018. "Blockchain Authentication of Network Applications: Taxonomy, Classification, Capabilities, Open Challenges, Motivations, Recommendations and Future Directions." *Computer Standards & Interfaces* 64: 41–60.
- Pennington, J., R. Socher, and C. D. Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, October.
- Priem, R. L. 2007. "A Consumer Perspective on Value Creation." *Academy of Management Review* 32 (1): 219–235.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 9.
- R Core Team. 2020. "R: A Language and Environment for Statistical Computing." In *R Foundation for Statistical Computing*. Vienna. Accessed 14 November 2019. <http://www.R-project.org/>.
- Rezvani, Z., J. Jansson, and J. Bodin. 2015. "Advances in Consumer Electric Vehicle Adoption Research: A Review and Research Agenda." *Transportation Research Part D: Transport and Environment* 34: 122–136.

- Rinker, T. 2018. "Sentimentr: Calculate Text Polarity Sentiment – An R Package." Accessed 24 September 2019. <http://github.com/trinker/sentiment>.
- Risus, M., and K. Spohrer. 2017. "A Blockchain Research Framework: What we (Don't) Know, Where we go from Here, and how we Will get There." *Business & Information Systems Engineering* 59 (6): 385–409.
- Rotolo, D., D. Hicks, and B. Martin. 2015. "What Is an Emerging Technology?" *Research Policy* 44 (10): 1827–1843.
- Salah, K., M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha. 2019. "Blockchain for AI: Review and Open Research Challenges." *IEEE Access* 7: 10127–10149.
- Sanyal, D. K., P. K. Bhowmick, P. Pratid Das, S. Chattopadhyay, and T. Y. S. S. Santos. 2019. "Enhancing Access to Scholarly Publications with Surrogate Resources." *Scientometrics* 121: 1129–1164.
- Shin, J., B. Y. Coh, and C. Lee. 2013. "Robust Future-Oriented Technology Portfolios: Black–Litterman Approach." *R&D Management* 43 (5): 409–419.
- Siano, F., and P. Wysocki. 2019. "Transfer Learning and Textual Analysis of Accounting Disclosures. Applying Big Data Methods to Small(er) Data Sets." Paper presented at the AH Conference, December.
- Snyder, H. 2019. "Literature Review as a Research Methodology: An Overview and Guidelines." *Journal of Business Research* 104: 333–339.
- Son, C., J. Kim, and Y. Kim. 2020. "Developing Scenario-Based Technology Roadmap in the Big Data era. An Utilization of Fuzzy Cognitive map and Text Mining Techniques." *Technology Analysis and Strategic Management* 32 (3): 272–291.
- Straka, M., and J. Straková. 2017. "Tokenizing, POS Tagging, Lemmatizing and Parsing ud 2.0 with UDPipe." In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99, August.
- Taylor, P. J., T. Dargahi, A. Dehghantanha, R. M. Parizi, and K. K. R. Choo. 2019. "A Systematic Literature Review of Blockchain Cyber Security." *Digital Communications and Networks* 6: 147–156.
- Tu, Y. N., and J. L. Seng. 2012. "Indices of Novelty for Emerging Topic Detection." *Information Processing and Management* 48: 303–325.
- Wang, Y., J. H. Han, and P. Beynon-Davies. 2019. "Understanding Blockchain Technology for Future Supply Chains: A Systematic Literature Review and Research Agenda." *Supply Chain Management: An International Journal* 24 (1): 62–84.
- Wang, X., X. Yang, X. Wang, M. Xia, and J. Wang. 2020. "Evaluating the Competitiveness of Enterprise's Technology Based on LDA Topic Model." *Technology Analysis and Strategic Management* 32 (2): 208–222.
- Xu, S., L. Hao, X. An, G. Yang, and F. Wang. 2019. "Emerging Research Topics Detection with Multipla Machine Learning Models." *Journal of Informetrics* 13: 100983.
- Yan, E. 2014. "Research Dynamics: Measuring the Continuity and Popularity of Research Topics." *Journal of Informetrics* 8 (1): 98–110.
- Yeow, K., A. Gani, R. W. Ahmad, J. J. Rodrigues, and K. Ko. 2017. "Decentralized Consensus for Edge-Centric Internet of Things: A Review, Taxonomy, and Research Issues." *IEEE Access* 6: 1513–1524.
- Yli-Huuma, J., D. Ko, S. Choi, S. Park, and K. Smolander. 2016. "Where Is Current Research on Blockchain Technology? – a Systematic Review." *PloS One* 11 (10): e0163477.
- Yoon, B., and C. L. Magee. 2018. "Exploring Technological Opportunities by Visualizing Patent Information Based on Generative Topographic Mapping and Link Prediction." *Technological Forecasting and Social Change* 132: 105–117.
- Young, K., C. Wang, L. Y. Wang, and K. Strunz. 2013. "Electric Vehicle Battery Technologies." In R. Garcia-Valle & J. Peças Lopes (Eds.), *Electric Vehicle Integration Into Modern Power Networks*, 15–56. New York: Springer.
- Zerva, C., M. Q. Nghiem, N. T. H. Nguyen, and S. Ananiadou. 2020. "Cited Text Span Identification for Scientific Summarization Using Pre-trained Encoders." *Scientometrics*. 125: 3109–3137. doi:10.1007/s11192-020-03455-z.

Appendices

Appendix 1

Figure A1 represents an example of the figures provided to the group of four experts (see Appendix 2) to explore the topics representing the problems of blockchain technology and evaluate the optimal number of problems to retrieve as output of our methodology. The experts received 4 figures representing 6, 7, 8 and 9 problems respectively. Here the example is related to 9 blockchain problems, the number that they ultimately selected as most appropriate.

Each circle is a word, and words belonging to the same topic are grouped along the x-axis, where the topic label is shown. The position on the y-axis depends on the word. By reading the table along the x-axis it is possible to see which words belong to only one topic and which ones belong to multiple topics. By reading the table along the y-axis it is possible to see the top 10 words for each topic (see Table A1 in this Appendix). The label and circle sizes are proportional to β . The colour coding is used to distinguish the different topics.

Table A1. Top 20 keywords per topic identified by our methodology applied to blockchain.

Topic label	Word list
1_power	power, energy, market, trust, cloud, design, vehicle, environment, infrastructure, cryptocurrency, scalability, transparency, computing, consumption, simulation, authentication, adoption, server, interest, security, fairness, trading, participation, governance, test, payment, evaluation, growth, consumer, complexity, accountability, measure, certificate, intelligence, measurement, transportation, strategy, vulnerability, company, capability
2_storage	storage, scalability, trust, infrastructure, cryptocurrency, server, requirement, healthcare, distribution, computing, byzantine, cost, growth, strategy, certificate, consumer, authority, identity, society, disadvantage, measure, capacity, bit, environment, anonymity, consensus, communication, food, tampering, vulnerability, market, address, integrity, integration, cryptography, series, standard, frameworks, governance, payment
3_design	design, storage, cost, consensus, power, trust, sensor, certificate, identity, trading, byzantine, bitcoin, interest, capacity, latency, database, prediction, vulnerability, communication, evaluation, authentication, patient, payment, availability, server, developer, adoption, means, property, guarantee, infrastructure, recovery, interaction, hardware, requirement, variation, frequency, resolution, investigation, object
4_communication	communication, cost, bitcoin, requirement, scalability, integrity, energy, Ethereum, database, infrastructure, capability, vulnerability, cryptocurrency, money, healthcare, complexity, adoption, innovation, availability, sensor, identity, education, food, market, stage, presence, evaluation, document, trust, patient, consumer, validation, integration, interoperability, verification, anonymity, size, rule, fault, fraud
5_cost	cost, trust, traffic, design, transparency, server, integration, evaluation, governance, policy, growth, environment, infrastructure, interest, identity, customer, company, bitcoin, game, modelling, introduction, recognition, maintenance, dispute, vision, interaction, measure, patient, monitoring, regulation, prediction, congestion, rights, dataset, authority, economy, verification, reliance, traceability, strategy
6_bitcoin	bitcoin, cloud, consensus, trust, storage, payment, requirement, market, property, transparency, authentication, scalability, design, policy, patient, adoption, interoperability, sensor, power, evaluation, innovation, integration, security, rule, verification, computing, identity, cryptocurrency, decentralisation, agreement, regulation, vehicle, healthcare, diagnosis, interest, simulation, period, class, company, communication
7_environment	environment, energy, consensus, cryptocurrency, distribution, cloud, power, design, integrity, infrastructure, Ethereum, food, consumption, bitcoin, database, monitoring, trading, requirement, market, governance, verification, production, voting, healthcare, traceability, latency, company, communication, delay, authentication, dynamics, customer, strategy, storage, transportation, vehicle, certificate, capability, scalability, centralisation
8_consensus	consensus, energy, fault, property, byzantine, environment, communication, trust, transparency, bitcoin, storage, integrity, practical, voting, cloud, market, signature, identity, infrastructure, distribution, authority, prediction, cryptocurrency, database, leakage, evaluation, electricity, growth, innovation, authentication, traceability, adoption, payment, decentralisation, healthcare, analytic, Ethereum, trading, computing, other
9_trust	trust, environment, scalability, energy, storage, design, Ethereum, server, authentication, identity, consensus, vulnerability, communication, property, cryptocurrency, transparency, infrastructure, validation, fork, interest, transformation, integrity, bandwidth, signature, authority, era, anonymity, complexity, society, analytic, identification, strategy, decentralisation, comparison, verification, cloud, rights, response, ownership, reputation

Appendix 2

To identify the optimal number of topics, K , in which to cluster the blockchain problems emerging from our methodology, a panel of four experts was consulted after having used digital optimisation methods to narrow down the possible values of K between six and nine topics.

The panel of experts was composed by two Associate professors at the University of Pisa, a full professor at the University of Aalborg and an Assistant professor at the University of Bordeaux, all from Schools of Engineering. The experts were selected because of experience in the blockchain sector and considering the need of having both an academic and an industrial point of view. Two of them, in fact, are co-founders of an Internet of Things (IoT) company that uses Blockchain technology.

The consensus was reached by analysing independently the results of the four possible values of k . During this phase, the experts were visually assisted with topic maps (see in [Figure A1](#) in Appendix 1), produced for each of the potential optimal values of K . Following the personal analysis, an online meeting for consensus-building was conducted and moderated by one of the authors, where the experts discussed among themselves and came to a shared final decision.

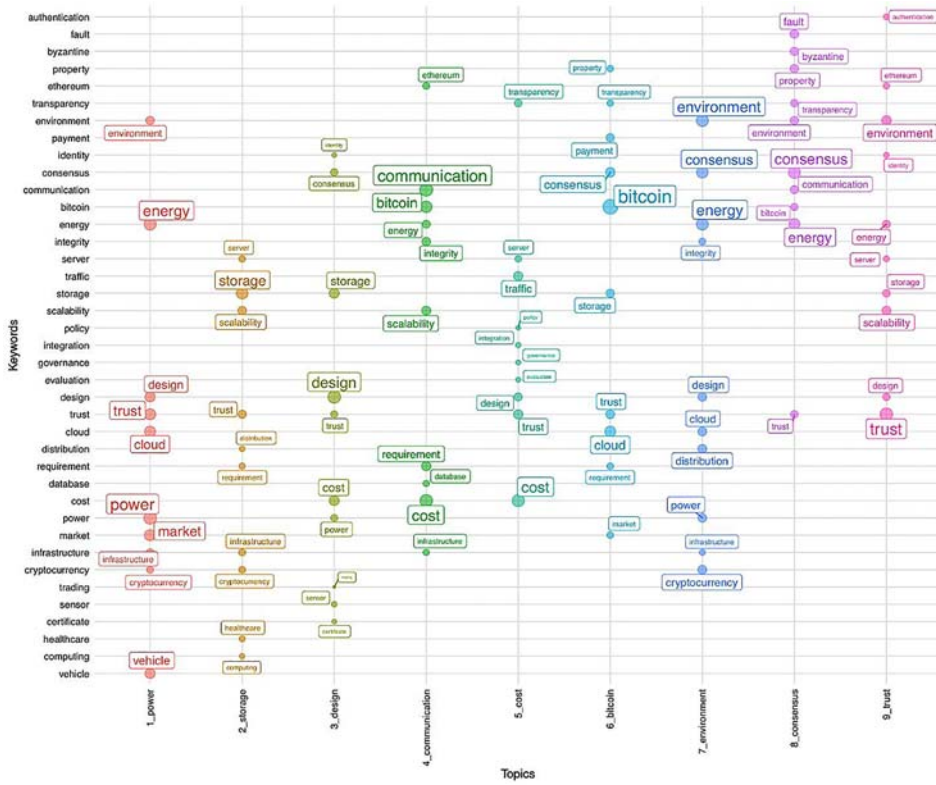


Figure A1. Content of the nine topics of blockchain problems.