Blockchain technology in supply chain management: insights from machine learning algorithms

Enna Hirata
Center for Mathematical and Data Sciences, Kobe University, Kobe, Japan

Maria Lambrou
Department of Shipping, Trade and Transport,
University of the Aegean, Chios, Greece, and

Daisuke Watanabe
Department of Logistics and Information Engineering,
Tokyo University of Marine Science and Technology, Tokyo, Japan

Abstract

Purpose – This paper aims to retrieve key components of blockchain applications in supply chain areas. It applies natural language processing methods to generate useful insights from academic literature.

Design/methodology/approach – It first applies a text mining method to retrieve information from scientific journal papers on the related topics. The text information is then analyzed through machine learning (ML) models to identify the important implications from the existing literature.

Findings – The research findings are three-fold. While challenges are of concern, the focus should be given to the design and implementation of blockchain in the supply chain field. Integration with internet of things is considered to be of higher importance. Blockchain plays a crucial role in food sustainability.

Research limitations/implications – The research findings offer insights for both policymakers and business managers on blockchain implementation in the supply chain.

Practical implications – This paper exemplifies the model as situated in the interface of human-based and machine-learned analysis, potentially offering an interesting and relevant avenue for blockchain and supply chain management researchers.

Originality/value – To the best of the knowledge, the research is the very first attempt to apply ML algorithms to analyzing the full contents of blockchain-related research, in the supply chain sector, thereby providing new insights and complementing existing literature.

Keywords LDA, Supply chain, Machine learning, Blockchain, Word2vec,
Natural language processing (NLP)

Paper type Technical paper

1. Introduction
The application of blockchain in the supply chain sector has been vastly discussed recently from both optimistic and negative perspectives. Many consider applying

blockchain in supply chain and logistics is an exciting prospect, there are also concerns that the technology lacks the maturity (Duru and Zin, 2019) today to handle global supply chain complexity. Such limitations might be discouraging, but as an innovative technology, blockchain has the potential to reform global supply chains. Practitioners should meticulously consider its existing hurdles and possible challenges as the technology matures. A number of multi-case, qualitative research works proposed the overarching theoretical model, which systematizes the technological components, the prevailing management rationales and determinant factors of digitalization including blockchain (Lambrou *et al.*, 2019a, 2019b; Wagner and Wiśnicki, 2019; Duru and Zin, 2019). However, these case studies are basically based on the deductive method and it is difficult to ensure generality.

Academic literature has become a rich source of information for researchers, practitioners and informed citizens, on various technological applications. Researchers use documents to express new ideas, theories, hypotheses, methods, approaches and experimental results with other researchers and interested parties. Text documents, such as research articles, technical reports and patents, are the preferred method of communication by researchers. Therefore, there is a lot of effort put into scientific communication, with scientific texts presenting a challenge to text mining methods, as the language used is formal and highly specialized.

From academia to industry, text mining has become a popular strategy for keeping up with the rapid growth of information. Automatic text mining methods can make the processing of extracting information from a large set of documents more efficient. However, as natural language is not easily processed by computer programs, it is necessary to develop algorithms to transform text into a structured representation, which is performed through natural language processing (NLP).

NLP is a branch of artificial intelligence (AI) that helps computers to understand, interpret and manipulate human natural languages. For example:

> NLP makes it possible for computers to read text, hear speech, interpret language, measure sentiment and determine which parts are important. Today's machines can analyze more language-based data than humans, without fatigue and in a consistent, unbiased way (SAS, 2020).

Increasingly, nowadays, experimental, emergent and mixed methods research approaches are acknowledged as constituting valid academic discourse choices, in their different nuances and sophistication (Eickhoff and Neuss, 2017; Asmussen and Charles, 2019). Both blockchain researchers and supply chain management (SCM) scholars have started to embrace topic modeling as a promising research method.

In our paper, we use a complementary, alternate perspective to investigate blockchain application in supply chains, based on text mining techniques for academic literature content analysis. The mined text is then applied with a couple of machine learning (ML) models to extract useful information.

Previous research papers have mainly analyzed bibliometric or abstracts of scientific articles. Different from the approaches existing literature took, this study presents a method of applying NLP to extract information from the full text of scientific articles. The extracted text documents are then trained by an ML algorithm, which performs automatic text classification. The rest of the paper is organized as follows: Section 2 reviews literature; Section 3 outlines the methodology; Section 4 discusses the results and findings; Section 5 concludes the research and future prospects.

2. Literature review

The multi-faceted phenomenon of blockchain technology pertaining supply chains has recently attracted considerable academic research attention and efforts.

The first stream of academic papers focuses on exemplifying both technology features and business incentives to adopt blockchain in supply chains, along with enabling and constraining factors. Sternberg and Baruffaldi (2018) reviewed supply chain blockchain initiatives and theorized the logic and challenges of blockchains in the supply chain industry. The authors concluded that while several incentives for developing and using blockchain technology exist, it was not apparent how companies can actually benefit from a materialized business advantage.

Kshetri (2018) also reviewed early supply chain industry cases and delineated the blockchain's role in SCM. The cases illustrate drivers and mechanisms for meeting cost, quality, speed, dependability, risk reduction, sustainability and flexibility objectives of supply chain organizations. Furthermore, identified determinants of blockchain adoption include the number of entities involved (viable blockchain ecosystem), participants' capabilities and the extent of industry competitive pressure.

Saberi et al. (2019) overviewed blockchain technology and its applicability in the supply chain, systematized a comprehensive list of barriers (i.e. system-related, intra and inter-organizational and external barriers) and proposed certain directions for overcoming those salient obstacles (i.e. governance mechanisms).

The second stream of research works endeavors to analyze how the current blockchain technology applications in SCM are implemented, delineate the foundations of the technology and further articulate the value of the technology for SCM, toward identifying pertinent enablers for achieving certain business goals and broader market adoption (Blossey et al., 2019; Casey and Wong, 2017; Queiroz et al., 2019; Roeck et al., 2020; Wang et al., 2019). Gurtu and Johny (2019) also discuss the significance and applications of blockchain technology with elaborate references to both generic supply chain, transport and maritime logistics cases.

A number of papers examine specific supply chain applications, such as food or pharmaceutical supply chain blockchains, shedding light on particular aspects, such as determinants to achieve visibility and trust via blockchain (Rogerson and Parry, 2020) or the interplay of blockchain technology features (i.e. consensus protocols) and business model requirements and blockchain ecosystem governance.

Research efforts focusing on specific supply chain areas, such as ports and shipping are emerging as well. Tsiulin et al. (2020) categorize blockchain projects in shipping and SCM and discuss the interrelations between blockchain features and shipping and ports concepts, toward delivering a better understanding of suitable use scenarios.

The third stream of literature re-uses present management theory to comprehend the unfolding of supply chain transformation resultant from blockchain technology; Treiblmaier (2018) uses core economics and management theories, namely, principal agent theory, transaction cost analysis resource-based view and network theory, to address the implications stemming from applying blockchain in SCM. Roeck et al. (2020) examine in particular how blockchain technology affects transactions and governance modes in supply chains, viewed from a transactions cost economics point of view, while conducting an abductive multiple case study of five supply chain industry cases. Kummer et al. (2020) identify pertinent organizational theories used in blockchain literature in the context of SCM, in specific agency theory, information theory, institutional theory, network theory, the resource-based view and transaction cost analysis. Most importantly, the authors reframe

SCM research questions addressed from the identified organizational theories vantage point, as intertwined with blockchain technology.

The latest, more mature stream of academic research, now habitually, focuses on more fine-grained topics, such as feasibility assessment and decision-making for technology selection (Ar *et al.*, 2020). Bai and Sarkis (2020) assess the growing literature studying the application of blockchain technologies in SCM; their research findings also identify a broad range of application types and operational objectives pursued (i.e. traceability, avoiding counterfeit products or reducing carbon footprints) and associated blockchain technical characteristics (i.e. scalability, complexity, security, etc.). Furthermore, the authors propose a performance measures framework that considers how blockchain technologies can help the supply chain meet targeted key objectives, based on a hesitant fuzzy set and regret theory.

Apparently, maritime and shipping research is indeed rejuvenated in terms of both methods and research topics. Fiskin and Cerit (2020), apply bibliometric and network analysis toward identifying areas of current research interests in the entire body of shipping literature, revealing interesting publication clusters, their relationships and changes over five years. Lee and Shin (2019) apply topic modeling on port research publications. Shin *et al.* (2018) conducted a literature review study on sustainability in maritime research, with text mining.

Chang and Chen (2020) provide an elaborate review of recent blockchain studies in the SCM context, systematizing an enhanced list of topics and applications.

The identification of research problems, appropriate theories and methods to investigate the application of blockchain technology in SCM is predominantly conducted according to the social sciences, positivist, empirical research tradition. Nonetheless, alternate research approaches are also used, beyond qualitative or directed content analyzes and case study field research designs. Pournader *et al.* (2020) review the existing academic literature and industrial knowledge sources regarding the applications of blockchains in supply chains, logistics and transportation and identify the 4Ts – technology, trust, trade, traceability/transparency (research themes clusters), based on a co-citation analysis of the publications on this topic. Wang *et al.* (2019) compare and summarize 29 of blockchain studies in the logistics area, and suggest the value of blockchain in four areas, namely, extended visibility and traceability, supply chain digitalization and disintermediation.

Against this background, to the best of our knowledge, extant literature examining blockchain application to SCM has indeed reached a maturity level where a sufficient number of pertinent research questions (or themes), revealed by multidisciplinary research frameworks examining blockchain technology and its implications, have been brought into the SCM academic discourse with varying intensity, rigor and insight (Iansiti, and Lakhani, 2017).

Currently, we do have a fair understanding of how, in particular, supply chain blockchain design and features (i.e. consensus mechanisms, security configurations, immutability and decentralized control) impact:

- Organizations and industries, in specific, how blockchain enabled new business models are unfolding in the supply chain sector, and how different intertwined industries comprise the disruptive potential that blockchain technology involves.

- Platforms, different blockchain implementations and protocols (i.e. Hyperledger), as well as various types of supply chain blockchains (i.e. private and permissioned), as well as inter-platform interoperability and integration with legacy systems are unfolding.

- Intermediation, the manner alternate blockchain features and designs enact different intermediation possibilities i.e. complementing existing supply chain intermediaries rather than excluding them.
- Users and society, in particular how growing market adoption is shaping societal effects, such as sustainable development goals and eventually realize and enact a multiplicity of possibilities regarding how supply chain blockchains create value (Risius and Spohrer, 2017).

Asmussen and Charles (2020) identified the current state-of-the-art digital technologies in SCM, and enablers for competitive advantage, based on a topic modeling framework. Shahid (2020) derived Latent Dirichlet allocation (LDA) topics of blockchain research (i.e. novelty, disruption, business blockchain types, protocol development, etc.), contributing with an efficient reporting of research trends and identified potential areas for interdisciplinary blockchain research collaboration. LDA has been applied in maritime-related studies, which generated useful insights. Shin *et al.* (2018), Lee and Shin (2019) apply LDA to identify research topics and suggest future research should focus on port collaboration and environmental issues.

The existing literature has mainly analyzed bibliometric data and/or abstracts, few of them have analyzed the paper contents. By analyzing full text articles, our paper aspires to contribute in the advancement of SCM and maritime transport research with a particular application of ML techniques to blockchain literature review and rigorous, new theory building. We exemplify our model as situated in the interface of human-based and machine-learned analysis, potentially offering an interesting and relevant avenue for blockchain and SCM researchers.

3. Methodology
The proposed method is organized into three major modules, namely, pre-processing, ML and visualization. The pre-processing stage involves the techniques and processes, which conduct the task of text mining. A couple of ML models including principal component analysis (PCA), word2vec and LDA are formulated by the training modules, which conduct the learning and classification tasks. Finally, the visualization phase describes the findings of the study. The workflow of the proposed system is represented as follows (Figure 1).
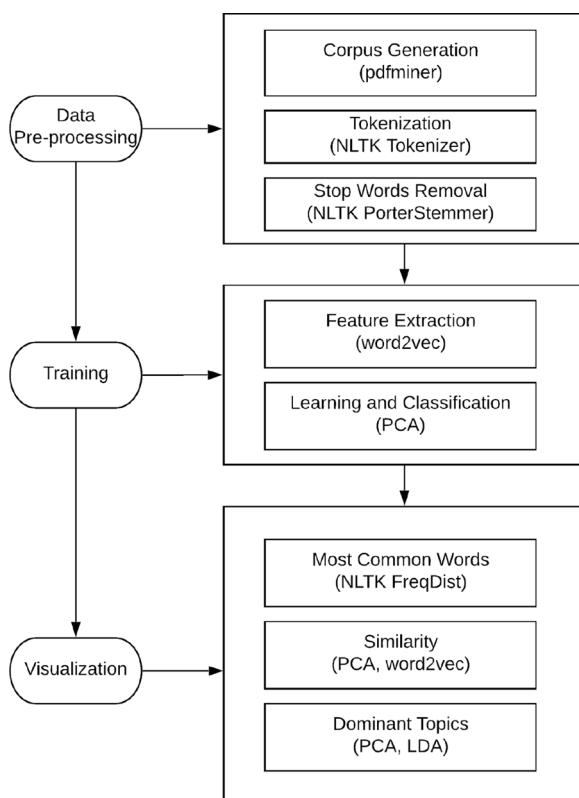
3.1 Data pre-processing
3.1.1 Text mining. Text analysis allows automatic extraction and classification of information from text (Westergaard *et al.*, 2018), such as tweets, emails, product reviews and survey responses. Popular text analysis techniques include word frequency, collocation, concordance, text classification, sentiment analysis, topic detection, language detection, clustering, keyword extraction and entity recognition, etc.

Sorting through data is a repetitive, time-consuming and expensive process if done by humans. Instead, if done by machines, high volumes of text can be analyzed with the least efforts, while providing even more accurate insights.

Text mining in this study is performed in the following steps.

3.1.2 Corpus generation. The experiment of this study is carried out on a text corpus, which is a collection of literature published in Science Direct, Emerald and Springer database with the following criteria. Time span is set to be from January 1990 to January 2020. The articles are retrieved by using keywords "blockchain" and "supply chain." In total, 422 articles hit the search. Table 1 outlines the distribution of corpora. Only literature in English language is included in this study.

Corpus Generation
(pdfminer)

Tokenization
(NLTK Tokenizer)

Stop Words Removal
(NLTK PorterStemmer)

Data
Pre-processing

Feature Extraction
(word2vec)

Learning and Classification
(PCA)

Training

Most Common Words
(NLTK FreqDist)

Similarity
(PCA, word2vec)

Dominant Topics
(PCA, LDA)

Visualization

Figure 1.
Proposed method
(tools used are
indicated in brackets)

Corpus	No. of related article	Document types
Science direct	68	Full text articles
Emerald	103	Full text articles
Springer	251	Full text articles
Total	422	Full text articles

Table 1.
Corpora

As the retrieve articles are in pdf format, we use python script pdfminer to convert the pdf data to text format for further analysis. There are other similar tools available, which should generate same result as what pdfminer does in converting pdf data to text format.

3.1.3 Tokenization. Tokenization is a critical and the most basic step to proceed with NLP. Tokenization in NLP means to split raw text into smaller units, such as words or terms, which are called tokens. These tokens are the key elements of the NLP.

Tokenization plays an important role in NLP because tokenization provides a way to easily interpret the meaning of a text by analyzing the sequence of the words in the text. To have a better understanding of tokenization, let's consider the below sentence:

Blockchain and supply chain are a match made in heaven.

Tokenize the sentence, we will get:

["blockchain," "and," "supply," "chain," "are," "a," "match," "made," "in," "heaven"]

The python script of NLTK (natural language tool kit) Tokenizer is applied in this study to split the text data to tokens. The corpus applied in this study contains 4,775,532 tokens. Once sentences are tokenized, the next step is to clean the text by removing stop words to get ready for the model building part.

3.1.4 Stop words removal. Next stage of data pre-processing is stop words removal. Stop words are words, which are commonly used in any natural language. For the purpose of analyzing text data and building NLP models, those stop words might not add much value to the meaning of the document, as such they are often filtered out in the data pre-processing stage.

Stop words usually refers to the most common words in a language, however, there is no single universal list of stop words. In this study, two types of stop words are removed: common English stop words (e.g. "is," "was," "where," "the," "a," "for," "of" and "in") and some extra stop words (e.g. "ieee," "paper," "vol," "doi," "et," "al," "https," "www," etc.) that are associated with the corpus particularly. Removing these stop words help reduce the size of the corpus and identify the keywords in the corpus, as well as frequency distribution of concept words in overall context more precisely. After removing stop words, the tokenized sentence in the above example contains:

["blockchain," "supply," "chain," "match," "made," "heaven"]

In data pre-processing stage, we also conducted lemmatization using PorterStemmer of NLTK. The result shows that the corpus generated similar results in the top 30 most common words with or without stemming. We decided to adopt the result without stemming during to two reasons. First, some words may get over-stemmed (e.g. both "generous" and "general" are stemmed to "gener") or under-stemmed (e.g. "bought" remains "bought" while "buy" is stemmed to "bii," while normally these two words have same stem "buy"). Second, some words cannot be correctly stemmed (e.g. "does" is stemmed to "doe").

To improve the accuracy, we have programmed to uniform the word forms to the best possible extent. The measurements include to replace plural form with singular form (e.g. replace "systems" with "system") and to unify different expressions (e.g. replace "block chain" with "blockchain;" "SCM").

After data pre-processing, we have enough tokenized clean text for the machine to work with, and to develop algorithms to differentiate and make associations between pieces of text to make predictions.

3.2 Training with machine learning algorithm

ML is the process of applying algorithms and statistical models to find patterns in massive amounts of data (MIT, 2018). With the application of ML, computer systems can perform a specific task without having to rely on patterns or inferences. ML is seen as a subset of AI and ML algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult to develop a conventional algorithm for effectively performing the task. This study applies PCA, word2vec and LDA in the Python environment to extract useful insights such as dominant topics from the corpus.

With the recent advent of ML, topic modeling, in particular, providing an overview of themes being addressed in documents, has gained notable popularity as an innovative methodological approach, in a broad range of disciplines, including management and information systems (Hannigan *et al.*, 2019). Topic modeling aims to reduce the complexity of compiling a large literature corpus by representing text as a combination of topics. Topics are clusters of words that reappear across texts, but the interpretation of these clusters as

themes, frames, issues or other latent concepts depends on the methodological and theoretical choices made by the researchers (Jacobs and Tschötschel, 2019).

As Hannigan *et al.*, 2019 clearly illustrate, topic modeling's technical and theory-building features are distinct from those of content analysis and general NLP of text. Topic modeling is a "rendering process," for juxtaposing data and a problem domain theory, to generate new theoretical insights and artifacts such as technical and management constructs and the links between them. The topic modeling process, as a theory building tool, involves the rendering of corpora (preparing the sets of texts to be analyzed), the rendering of topics (making choices in view of how topics are identified within the text corpora) and the generation of theoretical artifacts (producing new theoretical constructs, identifying causal mechanisms and further insights out of the revealed topics).

3.2.1 Principal component analysis. PCA is originally proposed by Hotelling (1933). It is a mathematical algorithm that derives the matrix of correlation coefficients from the original data set. The purpose of PCA is to reduce the dimensionalities of the data. In this study, PCA is carried out in four steps:

(1) Standardize data.

(2) Compute the covariance matrix with the standardized data.

(3) Calculate eigenvectors and eigenvalues in the covariance matrix.

(4) Sort the eigenvalues in decreasing order.

The first principle component (PC1) will carry the most of variance, etc.

Efficiency of the PCA model is measured by cumulative contribution rate (CCR). CCR is defined by the maximal amount of variance that is explained by the principal components representing the directions of the data. A CCR of 80% or above is considered acceptable in evaluating PCA model efficiency.

3.2.2 Word2vec. Mikolov *et al.* (2013) propose word2vec model for computing continuous vector representations of words from very large data sets, and observe a large improvement in the quality of these representations measured in a word similarity. Word2vec is a two-layer neural network deep learning model that has a text corpus as input and the word vectors as output. To be more specific, it first constructs a dictionary of words from the training text data and then learns vector representation of those words.

The advantage of word2vec is that it detects similarities mathematically. It creates vectors that are distributed numerical representations of word features, such as the context of individual words. It then outputs a dictionary of words in which each word has a vector attached to it, which can be grouped to vectors of similar words or be fed into a deep learning model for further analysis. Through the algorithm, word2vec establishes a word's association with other words, which forms the basis of sentiment analysis and recommendations in various research domains.

Word2vec conducts semantic comparisons (Mikolov *et al.*, 2013) ranging from country-currency (e.g. "India" is to "Rupee" as "Japan" is to "Yen") and male-female (e.g. "man" is to "king" as "woman" is to "queen").

Word2vec is applied in this study to assess word similarities, but the paper does not aim to discuss word2vec model in detail. Interested parties may refer to Rong (2014), which explains word2vec parameters in detail.

3.2.3 Latent Dirichlet allocation. LDA is an unsupervised generative probabilistic method for modeling a textual corpus. It is used as a language model to cluster co-occurring words into topics. LDA builds a topic per document model and words per topic model, modeled as Dirichlet distributions (Blei, 2012). LDA assumes that each document can be

represented as a probabilistic distribution over latent topics, and that the topic distributions in all documents share a common Dirichlet prior probability.

The basic idea of LDA is to compute the probability distribution over words. For a document in study, topics and their distributions in the text database are considered as latent variables or hidden structures. LDA model allows sets of observations to be explained by unobserved variables. When observations are words collected into documents, each document is a mixture of a small number of topics and each word's presence is attributable to one of the document's topics.

LDA is considered to be one of the most effective approaches (Blei *et al.*, 2003) to model topics. Detailed explanation on the LDA model is available in Blei *et al.* (2003), Steyvers and Griffiths (2007) and Blei (2012).

3.3 Visualization of research findings

3.3.1 Most common words. The most commonly used words in research data are reported in Figure 2. The top 10 most common words are, "blockchain," "data," "technology," "system," "information," "management," "service," "transaction," "business," "model," respectively. This indicates that the top concerns surrounding blockchain are relating to technology, information management, transaction and business model.

3.3.2 Word similarity. Word similarity is measured by cosine similarity, which is the cosine of the angle between two non-zero vectors of an inner product space. No similarity is expressed as a 90-degree angle, while a total similarity of 1 is a 0-degree angle, complete overlap. Table 2 lists top 20 words associated with "blockchain" using Word2vec, in order of proximity.

4. Results and discussions

4.1 Word classification

Text classification is the task of assigning a set of predefined categories to free-text. Text classifiers can be used to organize, structure and categorize words. For example, chat conversations can be organized by language, brand mentions can be organized by sentiment, etc.
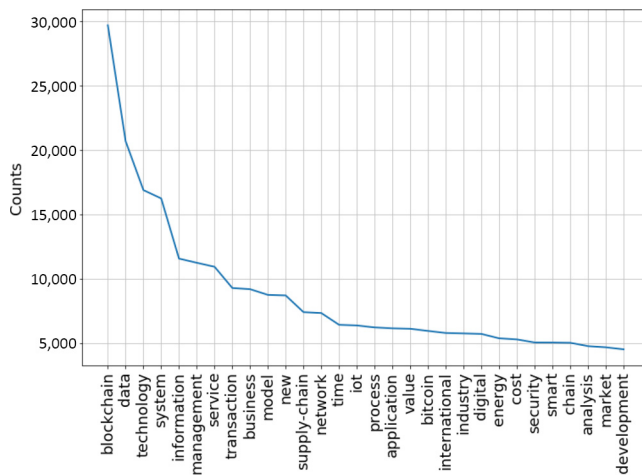

Figure 2.
Plot of most common words

By reducing the dimensions of word vectors by using PCA, most common words are classified (Figure 3). The plot illustrates three groups of keywords:

(1) "Blockchain," "data," "bitcoin," "security" and "application" are associated with each other closely.

(2) "Business," "industry," "information," "management" and "system" are associated with each other closely.

(3) "International," "cost," "transaction" and "energy" are outliers, which may be explained that these are popular topics, however, are not topics particularly related to supply chain blockchain.

4.2 Dominant topics

Applying to the LDA model, four dominant topics are identified as listed in Table 3. In LDA models, each document is composed of multiple topics. However, typically only one of the

Word	Similarity	Word	Similarity
Bitcoin	0.906903	Analysis	0.693042
Ledger	0.897149	Challenges	0.683677
Security	0.821995	Privacy	0.674539
Distributed	0.797730	Trust	0.643097
Block	0.795304	Implementation	0.640840
Application	0.756384	Iot	0.637737
Peer	0.754152	Smart	0.630747
Public	0.748685	Access	0.628037
Private	0.724296	Technology	0.617000
Design	0.715209	Adoption	0.613373

Table 2.
Word similarity
(top 20)
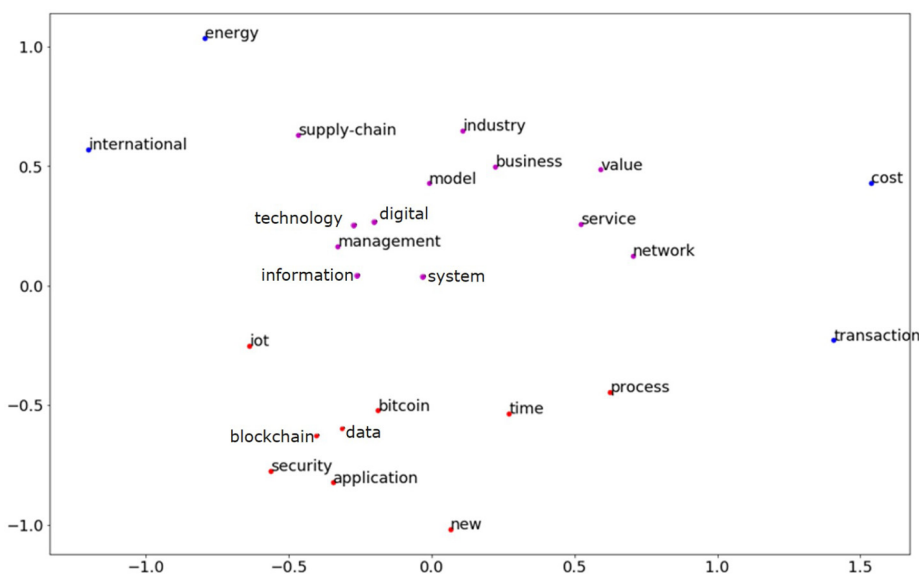
Figure 3.
Word classification

Topic_Num	Topic_Perc_Contrib	Keywords	Representative text
0	0.99553	Iot, system, datum, blockchain, base, network, security, device, service, application	[Network, available, network, homepage, locate, challenge, way, forward, technology, internet, t. . .
1	0.99536	Technology, supplychain, management, industry, datum, blockchain, system, business, service, model	[Purchase, supply, management, available, purchase, supply, management, homepage, never, walk, a. . .
2	0.99672	Blockchain, transaction, technology, system, information, process, base, platform, datum, business	[Express, author, intend, represent, position, opinion, wto, member, prejudice, member, obligati. . .
3	0.99678	Supplychain, management, system, sustainability, service, performance, product, model, customer,. . .	[Index, note, cross, refer, subentry, main, entry, main, entry, repeat, space, index, arrange, s. . .

Table 3.
Dominant topics

topics is dominant. Setting the selection criteria as CCR ratio greater than 98%, 4 dominant topics are obtained. The CCR of the dominant topic in the relevant document is higher than 99.5%, which indicates the ML process was effective. The keywords in dominant topics concur with the top common words generated in 3.3.1.

To crosscheck against the top common words obtained in 3.3.1, we further plotted topics and weights (Figure 4). Notably, internet of things ("IoT,") "security," "device," "performance," "customer," "sustainability" and "food" are of higher weight (importance) regardless of lower frequency of appearance than other keywords.

The text mining and ML are performed via Python version 3.7.3 in macOS Catalina 10.15.3, MacBook Pro, Processor 2.4 GHz Quad-Core Intel Core i5, Memory 16GB 2133 MHz LPDDR3. The computing time of each model is summarized in Table 4 as **a** reference.

5. Conclusion, research implications and future work

The research findings can be summarized in four perspectives. First, for studies on blockchain application in supply chain, the top topics seem to be related to "data," "technology," "system," "information" and "management" (Figure 2). Second, blockchain is considered of higher similarity to "bitcoin," "distributed ledger," "security" and "application" (Table 2). Third, "IoT," "security," "device," "performance," "customer," "sustainability" and "food" are of higher weight (importance) regardless of lower frequency of appearance than other key words (Figure 4). Fourth, "design," "trust," "implementation," "challenges" and integration with "IoT" are of higher concern than other perspectives such as "standardization", "interoperability" and "regulation" (Tables 2 and 3).

The research insights and implications, as derived from our study, are three-fold. First, while predominantly, generic technology and management challenges are of concern, focus should be given to particular design and implementation aspects of blockchain in the supply chain field. Blockchain deployments in practices are mostly in the pilot stage as of yet (Queiroz *et al.*, 2019). Future focus should be given to develop the architecture of blockchain
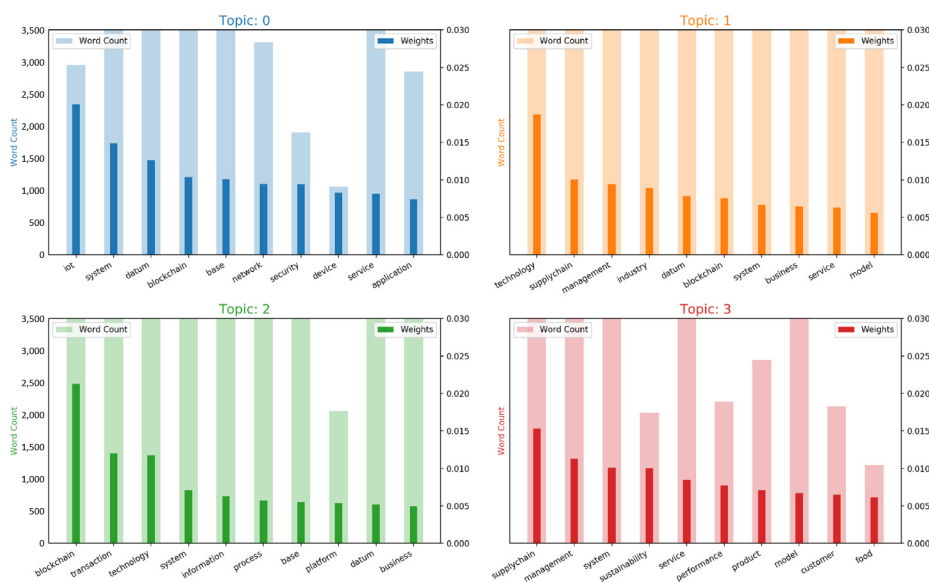
Figure 4.
Topics and weights

Models	PCA	Word2vec	LDA
Computing time (seconds)	7.39088	181.52167	105.01739

Table 4.
Computing time

solutions to provide seamless network and transparency in supply chains that benefit public safety and security.

Second, integration with IoT is considered to be of high importance. IoT has been rapidly applied in various areas of SCM in the past two years, security of information is of paramount issue. Blockchain technology has been explored as one option to effectively address those security concerns, allowing the advantage of decentralized data management. Kshetri (2017) suggests that the integration of IoT data to a blockchain platform could potentially further improve the overall efficiency. Blockchain can play a key role in tracking the sources of vulnerability in supply chains and in handling crisis situations such as product recalls that occur after safety and security vulnerabilities are found. The IoT ecosystem is evolving quickly, developing several applications in different sectors. As such, a future research agenda may be set to explore secure technical solutions to integrate IoT in supply chains in the context of increasing malicious IoT treats. Regulation of IoT security and data protection need to be developed and strengthened.

Third, blockchain plays a crucial role in food sustainability. The technology could help consumers and businesses understand whether their products were produced sustainably and avoid environmentally damaging, illegal or unethical products. The blockchain-based supply chain traceability and transparency technology helps drive increased responsible production and consumption. New technologies such as IoT and blockchain can accelerate the progress of supply chain sustainability. How quick these technologies are adopted and implemented is becoming a key to protect the environment and relax pressure on food

shortage. Modern supply chains are complex and require digital connectivity and agility across participants, business leaders need to understand what needs to change in their organization to leverage blockchain implementation effectively.

Unlike previous research studies that have mainly applied analysis to bibliometric data and/or abstract, this paper analyzes full text contents of paper with ML models to generate insights.

Because of access constrains, only scientific papers published in Science Direct, Springer and Emerald are included in this study. This may potentially bias the research findings. Future studies may consider applying a larger size of corpus. In addition, it could also be of value to train with a different set of ML models.

To the best of our knowledge, our research is the very first attempt to apply ML to blockchain-related research in the supply chain sector, thereby providing new insights and complementing existing literature.

References

Ar, I.M., Erol, I., Peker, I., Ozdemir, A.I., Medeni, T.D. and Medeni, I.T. (2020), "Evaluating the feasibility of blockchain in logistics operations: a decision framework", *Expert Systems with Applications*, Vol. 158 No. 113543.

Asmussen, C.B. and Charles, M. (2019), "Smart literature review: a practical topic modelling approach to do exploratory literature review", *Journal of Big Data*, Vol. 6 Article No. 1.

Asmussen, C.B. and Charles, M. (2020), "Enabling supply chain analytics for enterprise information systems: a topic modelling literature review and future research agenda", *Enterprise Information Systems*, Vol. 14 No. 5, pp. 563-610.

Bai, C. and Sarkis, J. (2020), "A supply chain transparency and sustainability technology appraisal model for blockchain technology", *International Journal of Production Research*, Vol. 58 No. 7, pp. 2142-2162.

Blei, D.M. (2012), "Probabilistic topic models", *Communications of the ACM*, Vol. 55 No. 4, pp. 77-84.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Blossey, G., Eisenhardt, J. and Hahn, G. (2019), "Blockchain technology in supply chain management: an application perspective", *In Proceedings of the 52nd HI International Conference on System Sciences, 2019.*

Casey, M.J. and Wong, P. (2017), "Global supply chains are about to get better, thanks to blockchain", *Harvard Business Review*, Vol. 13, pp. 1-6.

Chang, S.E. and Chen, Y. (2020), "When blockchain meets supply chain: a systematic literature review on current development and potential applications", *IEEE Access*, Vol. 8, pp. 62478-62494.

Duru, O. and Zin, M.I.M. (2019), "Blockchain roaming in the Maritime industry", available at: https://splash247.com/blockchain-roaming-in-the-maritime-industry/ (accessed 27 June 2020).

Eickhoff, M. and Neuss, N. (2017), "Topic modelling methodology: its use in information systems and other managerial disciplines", *In Proceedings of the 25th European Conference on Information Systems (ECIS)*, Guimarães, Portugal, June 5-10, 2017, pp. 1327-1347.

Fiskin, C.S. and Cerit, A.G. (2020), "Comparative bibliometric and network analysis of Maritime transport/shipping literature using the web of science literature", *Scientific Journals of the Maritime University of Szczecin*, Vol. 61 No. 133, pp. 160-170.

Griffiths, T.L., Steyvers, M. and Tenenbaum, J.B. (2007), "Topics in semantic representation", *Psychological Review*, Vol. 114 No. 2, pp. 211-244.

Gurtu, A. and Johny, J. (2019), "Potential of blockchain technology in supply chain management: a literature review", *International Journal of Physical Distribution and Logistics Management*, Vol. 49 No. 9, pp. 881-900.

Hannigan, T.R., Haans, R.F., Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S., Kaplan, S. and Jennings, P.D. (2019), "Topic modeling in management research: rendering new theory from textual data", *Academy of Management Annals*, Vol. 13 No. 2, pp. 586-632.

Hotelling, H. (1933), "Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychology*, Vol. 24 No. 6, pp. 417-441.

Iansiti, M. and Lakhani, K.R. (2017), "The truth about blockchain", *Harvard Business Review*, Vol. 95 No. 1, pp. 118-127.

Jacobs, T. and TschöTschel, R. (2019), "Topic models meet discourse analysis: a quantitative tool for a qualitative approach", *International Journal of Social Research Methodology*, Vol. 22 No. 5, pp. 469-485.

Kshetri, N. (2017), "Can blockchain strengthen the internet of things?", *IT Professional*, Vol. 19 No. 4, pp. 68-72.

Kshetri, N. (2018), "Blockchain's roles in meeting key supply chain management objectives", *International Journal of Information Management*, Vol. 39, pp. 80-89.

Kummer, S., Herold, D.M., Dobrovnik, M., Mikl, J. and Schäfer, N. (2020), "A systematic review of blockchain literature in logistics and supply chain management: Identifying research questions and future directions", *Future Internet*, Vol. 12 No. 3, pp. 1-15.

Lambrou, M., Watanabe, D. and Iida, J. (2019a), "Maritime blockchains: decoding diverse strategies for value extraction", *27th Annual Conference of the International Association of Maritime Economists (IAME), Paper ID 114*, pp. 1-22.

Lambrou, M., Watanabe, D. and Iida, J. (2019b), "Shipping digitalization management: conceptualization, typology and antecedents", *Journal of Shipping and Trade*, Vol. 4 No. 1.

Lee, S.W. and Shin, S.H. (2019), "A review of port research using computational text analysis: a comparison of Korean and international journal", *The Asian Journal of Shipping and Logistics*, Vol. 35 No. 3, pp. 138-146.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781.

MIT (2018), available at: www.technologyreview.com/2018/11/10/139137/is-this-ai-we-drew-you-a-flowchart-to-work-it-out/ (accessed 02 October 2020).

Pournader, M., Shi, Y., Seuring, S. and Koh, S.L. (2020), "Blockchain applications in supply chains, transport and logistics: a systematic review of the literature", *International Journal of Production Research*, Vol. 58 No. 7, pp. 2063-2081.

Queiroz, M.M., Telles, R. and Bonilla, S.H. (2019), "Blockchain and supply chain management integration: a systematic review of the literature", supply chain management", *Supply Chain Management: An International Journal*, Vol. 25 No. 2, pp. 241-254.

Risius, M. and Spohrer, K. (2017), "A blockchain research framework", *Business and Information Systems Engineering*, Vol. 59 No. 6, pp. 385-409.

Roeck, D., Sternberg, H. and Hofmann, E. (2020), "Distributed ledger technology in supply chains: a transaction cost perspective", *International Journal of Production Research*, Vol. 58 No. 7, pp. 2124-2141.

Rogerson, M. and Parry, G.C. (2020), "Blockchain: case studies in food supply chain visibility", *Supply Chain Management: An International Journal*, Vol. 25 No. 5, pp. 601-614.

Rong, X. (2014), "word2vec parameter learning explained", arXiv preprint arXiv:1411.2738.

Saberi, S., Kouhizadeh, M., Sarkis, J. and Shen, L. (2019), "Blockchain technology and its relationships to sustainable supply chain management", *International Journal of Production Research*, Vol. 57 No. 7, pp. 2117-2135.

SAS (2020), available at: www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html (accessed 27 June 2020).

Shahid, M.N. (2020), "A cross-disciplinary review of blockchain research trends and methodologies: topic modeling approach", *In Proceedings of the 53rd HI International Conference on System Sciences*.

Shin, S.H., Kwon, O.K., Ruan, X., Chhetri, P., Lee, P.T.W. and Shahparvari, S. (2018), "Analyzing sustainability literature in Maritime studies with text mining", *Sustainability*, Vol. 10 No. 10, pp. 1-19.

Sternberg, H. and Baruffaldi, G. (2018), "Chains in chains: logic and challenges of blockchains in supply chains", *In 51st HI International Conference on System Sciences*.

Steyvers, M. and Griffiths, T. (2007), "Probabilistic topic models", *Handbook of Latent Semantic Analysis*, NJ: Lawrence Erlbaum Associates.

Treiblmaier, H. (2018), "The impact of the blockchain on the supply chain: a theory-based research framework and a call for action", *Supply Chain Management: An International Journal*, Vol. 23 No. 6, pp. 545-559.

Tsiulin, S., Reinau, K.H., Hilmola, O.P., Goryaev, N. and Karam, A. (2020), "Blockchain-based applications in shipping and port management: a literature review towards defining key conceptual frameworks", *Review of International Business and Strategy*, Vol. 30 No. 2, pp. 201-224.

Wagner, N. and Wiśnicki, B. (2019), "Application of blockchain technology in Maritime logistics", *In DIEM: Dubrovnik International Economic Meeting*, Vol. 4 No. 1, pp. 155-164.

Wang, Y., Han, J.H. and Beynon-Davies, P. (2019), "Understanding blockchain technology for future supply chains: a systematic literature review and research agenda", supply chain management", *Supply Chain Management: An International Journal*, Vol. 24 No. 1, pp. 62-84.

Westergaard, D., Stærfeldt, H.H., Tønsberg, C., Jensen, L.J. and Brunak, S. (2018), "A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts", *PLoS Computational Biology*, Vol. 14 No. 2, p. e1005962.

Corresponding author
Enna Hirata can be contacted at: enna.hirata@platinum.kobe-u.ac.jp