Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Analyzing Trends and Patterns in Supply Chain Case Studies using LDA Topic Modelling

Noor Fazilla Abd Yusof[1,*], Chee Wei Han[1], Kemas Rahmat Saleh Wiharja[2]

[1] Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[2] School of Computing, Telkom University, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat 40257, Indonesia

ARTICLE INFO

Article history:
Received 26 February 2025
Received in revised form 23 August 2025
Accepted 23 September 2025
Available online 30 October 2025

Keywords:
Topic model; LDA; supply chain

ABSTRACT

This research project applies Latent Dirichlet Allocation (LDA) topic modelling to a corpus of supply chain case study articles with the main objective of uncovering themes in the literature to improve understanding of dominant trends, challenges, and innovations in supply chain management. By using LDA, we systematically analyse various case studies collected from the Scopus database, identifying key topics and their evolution over time. Unlike previous research that typically relies on article abstracts, this project extracts full article content using advanced text extraction tools, allowing for a deeper exploration of themes. Our methodology includes comprehensive text data pre-processing methods widely used in natural language processing (NLP) today. The findings reveal critical insights into areas such as logistics optimization, risk management, sustainability practices, and technological advancement. This research demonstrates that LDA is an effective tool for analysing large volumes of textual data, providing insights into complex supply chain issues. The approach not only aids researchers in studying supply chain management but also equips practitioners with valuable information for strategic decision-making. This research emphasizes the potential of topic modelling to significantly contribute to academic discourse and practical applications in the dynamic field of supply chain management.

1. Introduction

Supply Chain Management (SCM) has become a critical component in contemporary business operations, assisting companies in the planning and execution of activities such as sourcing, production, and the delivery of goods and services to end customers [1]. SCM enhances a company's ability to provide products or services swiftly and reliably, preventing issues like overstocking or inventory shortages. The primary objective is to satisfy customers and establish company credibility. The COVID-19 pandemic significantly disrupted business operations, introducing logistic challenges due to movement restrictions and border closures [2]. As a result, companies have increasingly focused on minimizing supply chain costs and seeking solutions to maintain competitiveness.

* *Corresponding author.*
E-mail address: elle@utem.edu.my

https://doi.org/10.37934/ard.146.1.90104

To gain deeper insights into SCM, analyzing case studies offers a valuable approach. SCM case studies provide real-world examples of challenges, solutions, and best practices encountered across various industries. These case studies serve as a rich resource for practitioners, researchers, and educators, enabling them to understand the complexities of SCM in different contexts and draw applicable lessons. Despite the abundance of literature on SCM case studies, extracting actionable insights and identifying overarching themes remains challenging. Traditional methods of manual analysis are often time-consuming, subjective, and limited in scalability. Consequently, there is a need for a computational approach that can systematically categorize and analyze SCM case studies to uncover latent themes, trends, and interconnections.

Recently, there has been growing interest in utilizing computational techniques, such as text extraction and topic modeling, to analyze large collections of textual data, including articles, publications, books, case studies, and web content. Topic modeling, particularly Latent Dirichlet Allocation (LDA) [3], has emerged as a valuable method for uncovering latent themes within a corpus without predefined categories. LDA posits that each document is a mixture of topics, represented as document-topics, and each topic is a distribution over words, represented as topic-words. By iteratively estimating these distributions, LDA identifies the underlying topics within the document collection and assigns probabilities to words within each topic [4].

Previous research often relied solely on abstracts; however, analyzing the full content of case studies offers a more comprehensive understanding. Challenges in this process include selecting appropriate preprocessing methods, determining the optimal number of topics, interpreting the results meaningfully, and validating the quality of the topic model. Addressing these challenges is crucial to ensure the reliability and utility of the insights derived from the analysis.

SCM has evolved significantly since the introduction of mass production and assembly lines in the late 1920s, with the modern concept of SCM taking shape in the mid-20th century. The term "*supply chain management*" emerged in the late 1980s, preceding terms like "*logistics*" and "*operations management*" [5]. Recognizing the importance of SCM, companies have realized that it enhances business efficiency, reduces costs, and improves customer satisfaction. Through comprehensive research, companies can identify inefficiencies, uncover cost-saving opportunities, and implement strategies that enhance resilience.

To underscore the value of researching SCM, it is essential to highlight how such research can guide companies in improving efficiency, risk management, and overall performance. For instance, in 1998, BioAg, an agricultural chemicals manufacturer, experienced rapid growth from selling two products to 36, supported by 420 supply firms across 24 states. Key success factors included strong partnerships between buyers and sellers, streamlined logistics for inventory management, and effective information systems that facilitated quick decision-making [6]. Similarly, a 2005 case study emphasized the importance of inventory control in optimizing supply chain performance, highlighting various inventory policies to achieve shorter lead times and more effective strategies [7]. Moreover, SCM's benefits extend to society and the environment, as demonstrated by a study on sustainable innovation in the cosmetics industry, which emphasized the role of suppliers in mitigating negative social and environmental impacts throughout the product lifecycle [8].

However, implementing SCM effectively can be complex. Challenges such as the bullwhip effect, where small demand fluctuations at the retail level cause larger fluctuations upstream, highlight the need for better demand forecasting and inventory management strategies [9]. The COVID-19 pandemic further underscored the importance of resilient SCM strategies, as professionals had to adapt to ensure the continuity and sustainability of supply chains [10,11].

Researching SCM case studies is therefore important for providing companies with tools and insights to enhance operations, reduce costs, improve customer satisfaction, mitigate risks, and make

informed strategic decisions—all of which are crucial for sustainable growth in today's competitive environment. This research aims to develop and implement a robust methodology for applying LDA topic modeling to a corpus of SCM case studies, uncovering underlying themes and trends in SCM literature. The findings will offer valuable insights into the complexities of SCM, facilitating informed decision-making for practitioners and researchers alike.

2. Methodology

This research adopts the CRISP-DM methodology up to but does not include the deployment phase. The focus is on the topic modelling for supply chain case studies. This involves the following steps business understanding, data understanding, data preparation, modelling and model evaluation.

2.1 Business Understanding

The business understanding phase is a critical component of the CRISP-DM methodology, serving as the foundation for defining research objectives and goals. Traditional methods of manual analysis are frequently criticized for being time-consuming, subjective, and limited in scalability. Additionally, previous research has often applied topic modeling exclusively to abstracts, which may inadequately capture the nuanced themes present in full-text articles. Topic modeling, particularly Latent Dirichlet Allocation (LDA), was selected for this project due to its proficiency in managing extensive textual datasets and its capacity to uncover latent structures within text.

The primary objective of this project is to employ LDA topic modelling techniques to uncover latent themes within a corpus of supply chain case study articles. Supply chain management is a complex and multi-dimensional field that includes logistics, procurement, inventory management, risk management, and sustainability. By applying topic modeling, this project seeks to identify predominant topics and trends within the case studies, offering valuable insights for both academic and industry stakeholders. For researchers, the identified topics may highlight emerging areas of interest and highlight potential gaps in the existing body of knowledge. For industry practitioners, understanding these topics can enhance decision-making processes strategic planning in supply chain management. To address this problem, several key questions need to be answered:

- What are the main topics discussed in supply chain case study articles?
- How do these topics evolve over time?

By clearly defining the objectives and transforming them into a well-articulated problem statement, this phase ensures that the research is aligned with stakeholder needs and expectations, thereby facilitating the generation of relevant and actionable insights.

2.1 Data Understanding

The data understanding phase is crucial for gaining insights into the dataset that will serve as the input for the solution. This phase involves a comprehensive exploration of the data to ensure its suitability for modeling and to identify potential challenges early in the process. For this study, the focus is on compiling a corpus of supply chain management (SCM) case study articles for topic modeling. The data utilized in this project comprises a collection of SCM case study articles. Data was collected by accessing the Scopus database, a comprehensive and renowned abstract and citation database with over 90 million publication records, including approximately 3 million new records added annually. The content of Scopus spans more than 39,000 serial titles [25].

The data collection process for this project involved searching for article records in Scopus on April 6, 2024. The search criteria were set to "Article Title Only" with the keywords: "supply" AND "chain" AND "case" AND "study". This query specifically targeted titles containing the terms "supply chain case study", ensuring the retrieval of records relevant to the project. The initial search yielded 2135 documents. Subsequently, the search was refined using filters to include only articles and conference papers, in English, and with open access status. This refined search returned 597 documents, spanning from 1998 to 2024.

Out of the 597 documents, 565 were successfully downloaded to a local file. The text of each document was then extracted and stored in a dataframe. Several text extraction tools, such as pyPDF2, pdfminer.six, and PyMuPDF, were considered. PyMuPDF was selected for its superior performance, compact size, and high-quality rendering of documents. It allows programmatic interaction with PDF documents using Python, including efficient text extraction [26]. PyMuPDF demonstrated the best performance in capturing the complete content of PDF articles with faster extraction times.

For text extraction, the content needed for topic modeling was delineated by removing extraneous information, such as reference sections and abstracts, to avoid redundancy. Regular expressions (re) were utilized to locate keywords like "*References*" and "*Introduction*" with text extraction commencing from "*Introduction*" and terminating at "*References*." For documents lacking an introduction section, alternative keywords such as "Background" and "Motivation" were incorporated into the search pattern. Upon completing the text extraction for all 565 documents, the text was stored in the dataframe for subsequent analysis.

2.2 Data Preparation

The data preparation phase involves transforming raw data into a clean, structured format suitable for topic modeling. This phase encompasses data cleaning, data processing, and feature engineering, culminating in the creation of a dictionary and document-term matrix. Several steps are necessary to preprocess the text. According to Churchill *et al*., [27], the preprocessing rules include pattern-based preprocessing, which matches patterns in terms such as URLs, capitalized words, hashtags, and punctuation. These elements are removed as they do not contribute meaningful information to the topic modeling process. Another preprocessing rule is dictionary-based preprocessing, which involves removing stopwords—common words like "*the*" "*are*" and "*will*" that do not convey significant meaning. The final rule involves natural language preprocessing techniques, including stemming, lemmatization, and part-of-speech (POS) tagging. Stemming reduces tokens to their base forms by removing suffixes (e.g., "*giving*" to "*giv*"), whereas lemmatization reduces tokens to their lemma forms using linguistic knowledge (e.g., "*better*" to "*good*" and "*giving*" to "*give*"). Lemmatization is generally more accurate than stemming, which simply truncates words [28]. POS tagging identifies and categorizes words into parts of speech such as nouns, adjectives, and adverbs, allowing for selective inclusion or exclusion based on the project's objectives.

In this research, spaCy was primarily used due to its advantages over NLTK. SpaCy is object-oriented, while NLTK primarily functions as a string-processing library. SpaCy offers a more user-friendly experience, whereas NLTK requires more specialized customization for text processing tasks like tokenization and lemmatization [29]. According to Muriuki [30], NLTK takes strings as input and returns lists, while spaCy's functions return objects, facilitating more efficient data manipulation. Additionally, spaCy demonstrates superior speed and performance in tasks such as word tokenization and POS tagging. Table 1 details the preprocessing functions utilized, along with corresponding code and explanations.

Table
Applied preprocessing function

Process	Code	Description
Tokenization	doc = nlp(text)	"nlp" in nlp(text) is the pipeline preprocessing function used in spaCy. The first step is tokenization.
Punctuation Removal	not token.is_punc	Detect and remove the punctuation token.
Url Removal	not token.like_url	Remove url link.
Email Removal	not token.like_email	Remove email.
Stopwords Removal	token.text not in stopword_list	Remove the words that exist in stopword_list.
Alphabetic characters	token.is_alpha	Remain the alphabetic characters.
POS tag	token.pos_ in ["NOUN", "ADJ", "VERB", "ADV"]	Remain the words that are nouns, adjectives, verbs and adverbs.
Lemmatization	token.lemma_	Transform the words into their lemma form.

In addressing the issue of special pattern terms, it was observed that certain words in the articles were split across lines due to line breaks (e.g., the word "accelerator" might appear as "acceler-\nator"). The sequence "-\n" indicates a hyphenated word split across lines, which can cause tokenization processes to erroneously treat it as two distinct words. To resolve this issue, a regular expression (regex) was employed to identify and remove the "-\n" pattern, subsequently concatenating the separated segments into a single coherent word.

Table 2 presents the token counts for the entire dataset before and after preprocessing. Initially, the raw text data contained 5,297,731 tokens, encompassing alphabetic characters, symbols, and numbers. Of these, 3,508,721 tokens were alphabetic. Post-preprocessing, the token count was reduced to 1,757,209, demonstrating the efficacy of preprocessing in eliminating redundant data.

Table 1
Number of token data before and after preprocessing

Number of Tokens Before Preprocessing	Number of Alphabetic Tokens Before Preprocessing	Number of Tokens After Preprocessing
5297731 tokens	3508721 tokens	1757209 tokens

In natural language processing (NLP), raw text data must be transformed into a numerical format for machine learning applications, a process known as text representation. The Bag-of-Words (BoW) model is a prevalent technique for text representation in topic modeling. The BoW model represents textual data as an unordered collection of words, disregarding grammar and word order. Each document is converted into a fixed-length vector, with each dimension corresponding to the frequency of a specific word within the document [31]. Gensim's corpora.Dictionary class facilitates the creation of a dictionary, and the corpus is generated in BoW format, where each document is represented as a tuple containing word_id and word_frequency. The doc2bow() method in Gensim's corpora.Dictionary class calculates unique word frequencies, assigns integer IDs to words, and returns a sparse vector representation stored in the corpus.

In NLP, pruning involves the removal of exceedingly frequent and infrequent words from a corpus. Words with extremely high or low frequencies often contribute little to topic modeling [32]. Removing infrequent words reduces data dimensionality [33], improving computational efficiency [34] and mitigating overfitting risks. However, excessive pruning of infrequent terms can result in the

loss of critical information, especially when such terms denote significant concepts. Conversely, very frequent words, which appear across numerous topics, provide minimal discriminative value. Common practice suggests eliminating words appearing in less than 0.5% to 1% of documents for infrequent terms, and those present in over 99% of documents for frequent terms [32, 35, 36]. Practical implementations have varied, with some projects pruning words present in more than 50% to 80% of documents [37, 38], and others, such as IBM, removing words appearing in over 90% of documents [39]. In this study, the dictionary was pruned using dictionary.filter_extremes (no_below=3, no_above=0.90), targeting words appearing in fewer than 0.5% to 1% (approximately 3 to 5 documents) and those present in over 90% of documents. Table 3 summarizes the number of unique dictionary words pruned. While the list of the unique dictionaries is shown in Fig. 3.

Table 2
Number of unique dictionary after pruning

Before Apply Pruning	After Apply Pruning
31026 tokens	9365 tokens

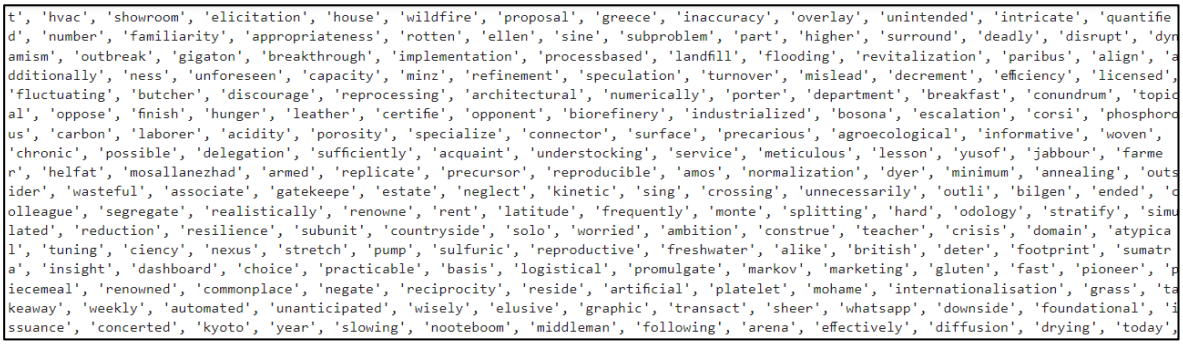

Fig 1. List of unique dictionary

2.3 Modelling

The modelling phase involves applying LDA topic modelling techniques to the prepared dataset to achieve the research objectives. In this research, Gensim library was using topic modelling to uncover latent themes within a corpus of supply chain case study articles. This phase includes training the LDA models and optimizing their parameters to ensure the best possible performance. By default setting, Gensim LDA only requires a corpus and dictionary that was created during feature engineering and requires users to input the number of topics to be determined. To obtain a better result, the hyperparameters of LDA modelling can be fine-tuned. Some common hyperparameters in LDA are illustrated in Table 4.

To identify the optimal number of topics for this dataset, a grid search was performed across a range of 1 to 15 topics. This range was chosen to cover a broad spectrum of potential topic structures, ensuring both simplicity and complexity are considered. Additionally, the alpha and beta parameters were fine-tuned utilizing a method based on Gensim's documentation [40], which recommends 'auto' settings to adapt these hyperparameters dynamically for improved model coherence and performance. However, it is important to note that these choices may introduce certain limitations or biases, such as the dependency on predefined ranges and the influence of Gensim's default optimization approach. Acknowledging these potential biases is crucial for interpreting the results and considering the robustness of the study.

Table 3
Hyperparameters description

Hyperparameters	Description
Number of Topics (K)	The hyperparameter determines the number of topics which the model should identify in the corpus.
Alpha (α)	The Dirichlet parameter controls the distribution of topics in documents (document-topic distribution). A lower value of alpha leads to documents containing fewer topics.
Beta (β)	Dirichlet parameter controlling the distribution of words in topics (topic-word distribution). A lower value of beta leads to topics containing fewer words.
Chunksize	The number of documents processed at a time during training. Larger chunksize values can lead to faster training times, but they require more memory. Smaller chunksize values reduce memory consumption but may result in slower training.
Passes	The parameter determines the number of passes through the entire corpus the algorithm will make during training. Each pass goes through the entire corpus, updating the model parameters.

2.4 Model Evaluation

The evaluation phase is critical for assessing the effectiveness and quality of the models developed during the modeling phase. This phase ensures that the models align with the research objectives and that the identified topics are both meaningful and actionable. Evaluation involves both quantitative metrics and qualitative assessments to validate the results. Quantitatively, the coherence score is commonly used, with the C_v score being a prominent metric. The C_v coherence score, which ranges from 0 (indicating complete incoherence) to 1 (indicating complete coherence), is based on a sliding window approach, cosine similarity, and normalized pointwise mutual information (NPMI). Scores above 0.5 are generally considered indicative of good coherence [41]. Qualitatively, topic interpretability is assessed by manually examining the top words associated with each topic to determine if they form a coherent and interpretable theme. To aid in this process, visualization tools such as PyLDAvis are employed. PyLDAvis is a Python library that facilitates the interactive visualization of topics generated by Latent Dirichlet Allocation (LDA) models. It provides a two-dimensional plane where each topic is represented as a bubble. The spatial proximity of these bubbles reflects the degree of similarity or dissimilarity between topics. Topics that are positioned close together share more terms and are more similar, whereas those further apart are more distinct.

3. Results

3.1 Fine-tuned LDA Topic Model

After conducting multiple iterations and fine-tuning the parameters, an optimal coherence score was attained. The optimal number of topics for result interpretation was determined to be six, as this configuration initially achieved the highest coherence score of 0.5051, as illustrated in Fig. 4. The alpha and beta parameters, set to their default 'auto' values, yielded the best results in Latent Dirichlet Allocation (LDA) topic modelling, achieving a coherence score of 0.5121, as detailed in Table 5. When set to 'auto', the model automatically optimizes the alpha parameter to achieve an optimal balance in the distribution of topics across documents, thereby adjusting the sparsity of topics per document to better align with the data. Similarly, the beta parameter was adjusted to find the most appropriate distribution of words within each topic. Additionally, the chunk size and passes parameters were set to 50 and 10, respectively, for LDA topic modelling in this research. While these parameters did not significantly affect the coherence score of 0.5121, the chunk size parameter dictated that 50 documents were processed simultaneously, thereby managing memory usage and

potentially accelerating the training process. The model was iterated over the entire corpus 10 times to ensure comprehensive training.

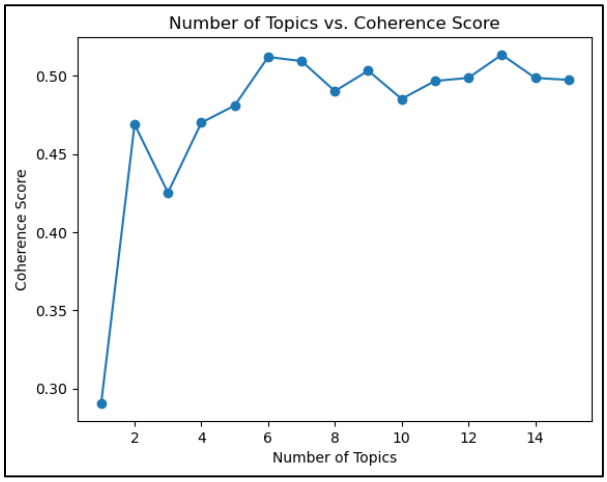

Fig. 4. Number of topics vs coherence score

Table 4
Alpha and Beta parameters with coherence score

Alpha	Beta	Coherence Score
'auto'	'auto'	0.5121

3.1.1 Topics result

The trained LDA model was then applied to the pre-processed corpus of supply chain case study articles to extract topics. Word clouds shown in Fig. 5 were used to represent topics extracted from supply chain case studies. Each word cloud corresponds to a specific topic and the size of each word indicates its importance or frequency within that topic.

In Topic 0, the predominant terms identified include "*company*," "*supplier*," "*sustainability*," "*practice*," "*industry*," "*relationship*," "*performance*," "*business*," and "*development*." This topic is characterized by a focus on supplier relationships and sustainability practices within companies and industries, emphasizing the significance of sustainable practices and performance metrics in these relationships. For example, a case study of tobacco manufacturing companies [42] illustrate how integrating sustainability practices with supplier relationships has led to improved environmental performance and stronger industry partnerships via supplier relationship management (SRM).

In Topic 1, the key terms were "*model*," "*scenario*," "*location*," "*center*," "*capacity*," "*transportation*," "*parameter*," "*demand*," and "*consider*." This topic is centered around modeling and logistics, with a particular emphasis on transportation and capacity planning, suggesting a focus on various scenarios and parameters that influence demand and location-based decision-making. An example is a case study [43] of logistics company using advanced modeling techniques to optimize transportation routes and capacity, thereby reducing costs and improving service levels through scenario-based planning.

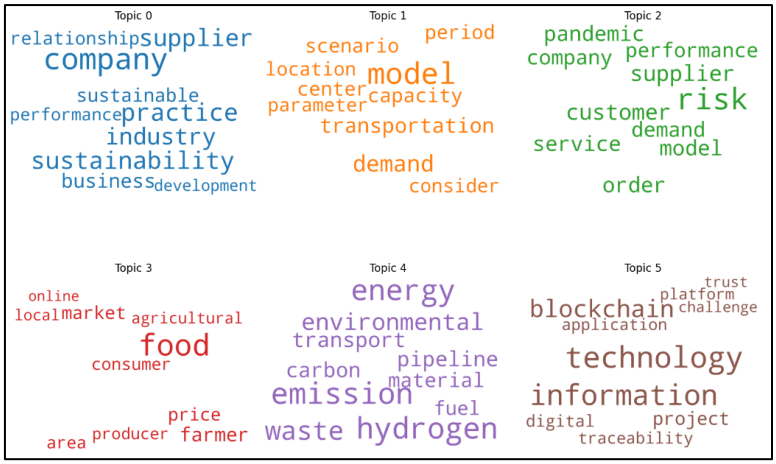

Fig. 2. Word Cloud: Dominant words in each topic

In Topic 2, the dominant terms were "*risk*," "*pandemic*," "*company*," "*performance*," "*supplier*," "*demand*," "*customer*," "*service*," and "*order*." This topic addresses risk management within supply chains, with specific reference to the COVID-19 pandemic. It encompasses aspects related to company performance, supplier relationships, and customer service during periods of risk and demand fluctuations. A specific case study [44] describes how a global retailer managed supply chain disruptions during the pandemic by diversifying supply chain network (i.e. suppliers, manufacturers, retailers) and enhancing customer communication to maintain service levels.

In Topic 3, the leading terms included "*food*," "*online*," "*local*," "*market*," "*agricultural*," "*consumer*," "*price*," "*area*," "*producer*," and "*farmer*." This topic pertains to the agricultural and food supply chain, emphasizing market dynamics, the role of local producers and farmers, pricing strategies, and the impact of online platforms on the food supply chain. For instance, a case study [45] on how a local farmers' market transitioning to an online platform during the pandemic demonstrate how local producers adapted to changing consumer behaviors and market conditions.

In Topic 4, the dominant terms were "*energy,*" "*emission*," "*waste*," "*hydrogen*," "*environmental*," "*transport*," "*pipeline*," "*carbon*," "*fuel*," and "*material*." This topic focuses on environmental concerns within supply chains, particularly energy usage and emissions. It discusses issues such as hydrogen, waste management, carbon emissions, and transport infrastructure, highlighting the importance of sustainability and environmental impact. An example of a case study Erdoğan *et al*., [46] highlights how Hydrogen fuel technology has been implemented to create a sustainable transportation sector.

In Topic 5, the significant terms included "*technology*," "*information*," "*blockchain*," "*platform*," "*application*," "*digital*," "*project*," "*traceability*," "*trust*," and "*challenge*." This topic revolves around technology and digital transformation in the supply chain, underscoring the role of blockchain for traceability and trust, the use of digital platforms, and the various technological applications that pose both opportunities and challenges. An example of case study Bosona *et al*., [47] in food supply chain illustrates how blockchain technology was implemented to enhance traceability, ensuring product quality and building consumer trust through transparent supply chain practices.

These word clouds provide a concise, visual summary of the diverse themes present in the supply chain case study articles, offering insights into key areas such as sustainability, risk management, food supply, environmental impact, and technological advancements. Table 6 presents an overview of each topic along with their dominant terms.

Table 5
Summary of 6 topics and dominant words

Topic No.	Dominant Words
0	company, practice, sustainability, supplier, industry, business, sustainable, relationship, performance, development
1	model, demand, transportation, capacity, center, scenario, period, location, consider, parameter
2	risk, supplier, customer, order, model, service, pandemic, demand, company, performance
3	food, farmer, market, price, area, consumer, producer, agricultural, local, online
4	emission, hydrogen, energy, waste, environmental, pipeline, transport, carbon, material, fuel
5	technology, information, blockchain, project, digital, traceability, application, platform, trust, challenge

Besides, pyLDAvis was used for interactive topic modelling visualization to interpret the result topics as shown in Fig.6. Topic 1 and Topic 6 appeared closely related. It showed that they shared a significant number of terms or concepts. While Topic 2, located near the centre, seems to be a central topic. It potentially bridged concepts from other topics. Topic 4 and Topic 5 were more distant which indicated they covered more distinct areas within the supply chain case studies.

The bar chart on the right listed the top 30 most salient terms across all topics. The length of the bars represented the overall term frequency in the entire corpus, while the red shading indicated the estimated term frequency within the selected topic. Key terms such as "*food*", "*technology*", "*model*", and "*risk*" are dominant which showed these were critical areas of focus within the supply chain literature. Terms like "*hydrogen*", "*emission*", and "*energy*" indicated a significant interest in sustainability and environmental impact. The presence of "blockchain" suggested attention to technological innovations in supply chain management. The word "food" represented the highest overall term frequency. It reflected significant attention to supply chain issues in the food industry. While "blockchain" highlighted the importance of technological advancements and their applications in the supply chain. "*Model*", and "*Risk*" emphasized the focus on managing risks and developing models to understand and optimize supply chains. "*Environmental*" and "*Emission*" indicated the concern for the environmental aspects of supply chain management.

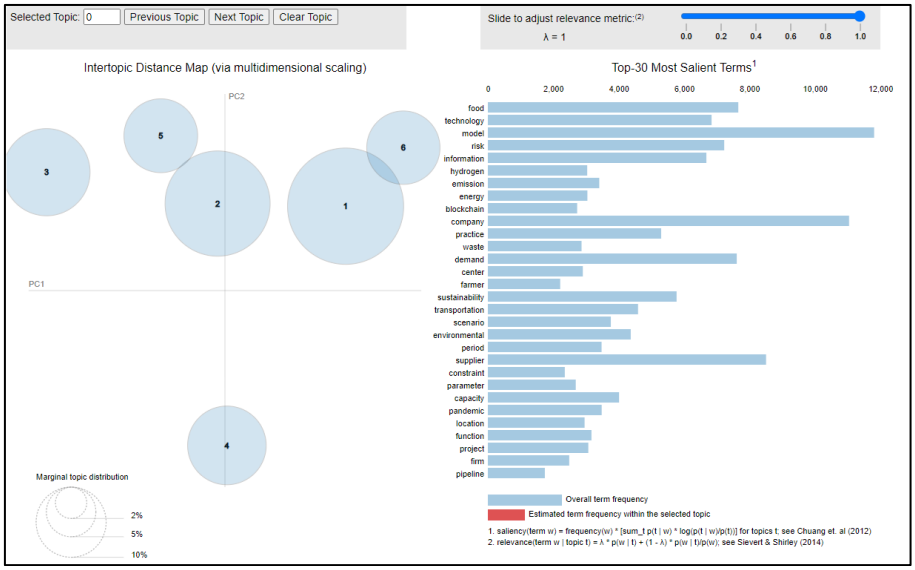

Fig. 3. PyLDAvis visualization

3.1.1 Topics evolution

To identify dominant topics and analyze their evolution over time, a stacked area chart was employed to visualize the temporal progression of categorized topics. The stacked area chart, depicted in Fig. 7, provides a comprehensive overview of the shifts in focus across various supply chain topics over time. The trends illustrate periods of heightened interest and research activity, with certain topics gaining prominence in specific years.

Overall, there has been a notable increase in the number of case studies over time, with a significant surge commencing around 2015 and peaking between 2021 and 2022. The number of articles shows a slight decline post-2022, with a noticeable drop in 2023 and 2024. As the dataset collection occurred during the second quarter of 2024, the number of case studies published in 2024 is relatively low.

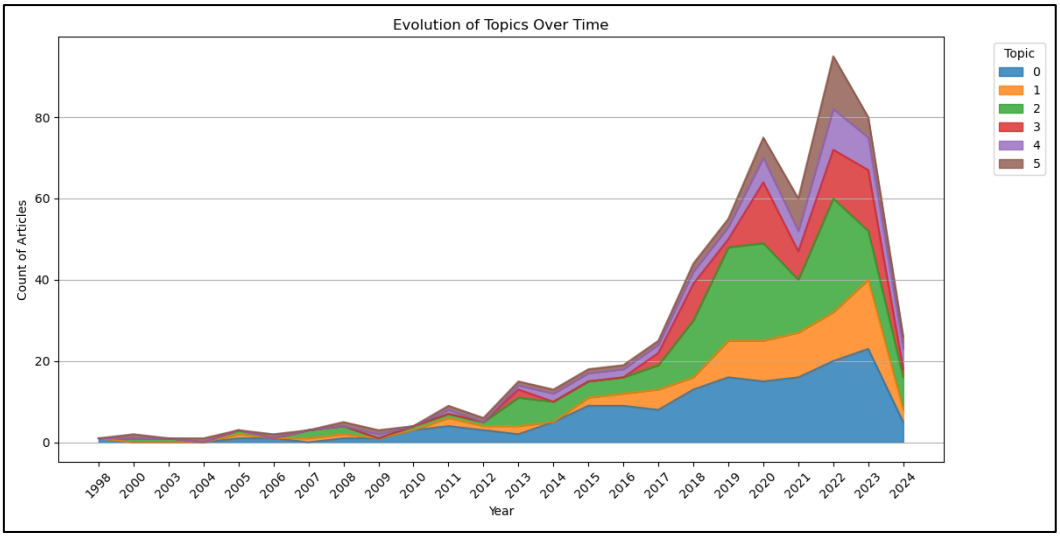

Fig. 4. Evolution of topics over time

Topic 1 (Orange) centers on logistics and transportation modelling, encompassing capacity planning and demand scenarios. The data shows a gradual increase in prominence around 2019, maintaining a steady presence through 2023. This trend indicates a growing interest in optimizing logistics and transportation networks.

Topic 2 (Green) addresses risk management, with a particular focus on the pandemic's impact. A sharp rise in case studies around 2020 correlates with the onset of the COVID-19 pandemic, reflecting heightened awareness and the need for effective risk management in supply chains.

Topic 3 (Red) pertains to the agricultural and food supply chain, including market dynamics and the role of local producers. An increase in this topic around 2018 likely corresponds to the growing focus on food supply chains, local markets, and the influence of e-commerce platforms.

Topic 4 (Purple) concentrates on environmental issues such as energy use, emissions, and waste management. The data indicates an upward trend beginning around 2012, with peaks in 2018-2019, highlighting increasing awareness and research on environmental sustainability within supply chains.

Topic 5 (Brown) revolves around technological advancements and digital transformation, including blockchain, digital platforms, and traceability. The rise in case studies from 2018 onwards underscores the growing significance of technology in driving the digitalization of supply chain operations.

During the early period from 1998 to 2010, the topics were relatively evenly distributed with low prominence, indicating a broad yet shallow research interest across various supply chain aspects.

From 2010 to 2018, there was significant growth in research interest, particularly in sustainability practices (Topic 0), logistics (Topic 1), and environmental concerns (Topic 4). Between 2018 and 2024, a sharp increase in articles related to risk management (Topic 2) and technology (Topic 5) reflects an immediate response to global challenges, such as the COVID-19 pandemic and the accelerated adoption of digital technologies.

Based on the analysis, it is recommended that the company prioritize supply chain management initiatives focusing on digital transformation, sustainability, risk management, and logistics optimization. The integration of advanced technologies such as artificial intelligence (AI), blockchain, and the Internet of Things (IoT) can significantly enhance the traceability, transparency, and efficiency of supply chain operations. Additionally, adopting green practices and measures to reduce carbon footprints will address critical environmental concerns. Developing robust risk mitigation strategies will also strengthen resilience against uncertainties, particularly those related to pandemic events like COVID-19. Furthermore, optimizing logistics networks through improved capacity planning and route optimization can reduce costs and improve delivery times.

For businesses, strategic recommendations include investing in emerging technologies to align with sustainability objectives. Implementing comprehensive risk management frameworks will enable rapid adaptation to unforeseen disruptions. For researchers, it is valuable to pursue studies that intersect supply chain management with technology, environmental science, and risk management. Investigating these areas will deepen the understanding of the long-term impacts of digital transformation and sustainability initiatives. Publishing detailed case studies on successful implementations will provide valuable insights for both academic and industrial audiences. These strategic actions collectively aim to drive innovation, efficiency, and resilience in supply chain practices.

3.1.2 Topics coherence

LDA was chosen for this study due to its superior ability to produce interpretable and coherent topics, as evidenced by a high overall coherence score of 0.513 (as shown in Table 7). This probabilistic framework effectively captures the underlying structure of complex text data, making it particularly well-suited for analyzing the diverse and nuanced content of our dataset. While alternative topic modeling methods such as Non-Negative Matrix Factorization (NMF) or Latent Semantic Analysis (LSA) were considered, LDA's capacity to model topic distributions with clear interpretability and flexibility in handling large vocabularies offers distinct advantages that align closely with the goals of this research. The exploration of these alternative methods is acknowledged but falls outside the scope of this paper.

Table 7
Summary of 6 topics and dominant words

Topic	Coherence score
Topic 0	0.501
Topic 1	0.538
Topic 2	0.348
Topic 3	0.562
Topic 4	0.612
Topic 5	0.515

4. Conclusions

In this research, we employed Latent Dirichlet Allocation (LDA) topic modeling to uncover dominant topics and trends within supply chain case studies. By customizing and optimizing the LDA model, we established a methodological framework for extracting meaningful insights from the comprehensive content of these case studies. This approach has significantly reduced the research time required for analyzing large volumes of case studies. The analysis identified key topics and trends, providing a deeper understanding of the evolution of supply chain issues and strategies. Dominant topics identified include sustainability practices, logistics optimization, risk management, environmental concerns, and technological advancements. The trends highlight the increasing adoption of sustainable practices in supplier relationships, optimization of transportation and logistics networks, enhanced risk management particularly in response to disruptions like the COVID-19 pandemic, and the leveraging of digital technologies such as blockchain and AI to improve transparency and efficiency.

For businesses, strategic investments in technology and sustainability, comprehensive risk management frameworks, and logistics optimization are crucial. Researchers should focus on interdisciplinary studies that integrate supply chain management with technology, environmental science, and risk management to generate valuable insights and drive innovation in the field.

Despite its significant contributions, this research project has several limitations. The quality of the topic modelling results is highly dependent on the quality and consistency of the input data, which can be challenging to ensure given the diverse nature of case study articles. Additionally, while LDA is a powerful tool, it has limitations related to the interpretability of topics and the necessity for manual fine-tuning of parameters.

Future research could involve integrating more advanced topic modelling techniques, such as deep learning, BERTopic, Non-Negative Matrix Factorization (NMF) or Latent Semantic Analysis (LSA) to capture more nuanced and temporally dynamic themes. Considering the unique and challenging nature of text extraction from PDFs, further utilization of advanced extraction tools is recommended to accurately capture and filter text. Collaboration with domain experts in supply chain management could provide deeper insights and validate the identified topics, enhancing the robustness and relevance of the findings. Expanding the dataset to include a broader range of sources, including industry reports and unpublished case studies, could also offer a more comprehensive view of the supply chain landscape.

Acknowledgement

The authors would like to thank Centre for Research and Innovation Management of Universiti Teknikal Malaysia Melaka (UTeM) for sponsoring this work under the Grant Tabung Penerbitan Fakulti dan Tabung Penerbitan CRIM UTeM.

References

[1] Lu LX, Swaminathan JM. Supply Chain Management. In: International Encyclopedia of the Social & Behavioral Sciences: Second Edition. Elsevier Inc.; 2015. p. 709–13. https://doi.org/10.1016/B978-0-08-097086-8.73032-7

[2] Aday, Serpil, and Mehmet Seckin Aday. "Impact of COVID-19 on the food supply chain." *Food quality and safety* 4, no. 4 (2020): 167-180. https://doi.org/10.1093/fqsafe/fyaa024

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

[4] Daniel Maier A. Waldherr PMGWANAKBPGHURTHHSP, Adam S. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. Commun Methods Meas [Internet]. 2018;12(2–3):93–118. Available from: https://doi.org/10.1080/19312458.2018.1430754

[5] Farahani, Reza, Shabnam Rezapour, and Laleh Kardar, eds. *Logistics operations and management: concepts and models*. Elsevier, 2011.
[6] Dooley FJ, Akridge JT, Chain S. International Food and Agribusiness Management Review, l(3): 435-441 Management: A Case Study of Issues For BioAg. 1998. https://doi.org/10.1016/S1096-7508(99)80010-X
[7] Musalem, Eric Porras, and Rommert Dekker. "Controlling inventories in a supply chain: A case study." *International Journal of Production Economics* 93 (2005): 179-188. https://doi.org/10.1016/j.ijpe.2004.06.016
[8] Pereira de Carvalho, André, and José Carlos Barbieri. "Innovation and sustainability in the supply chain of a cosmetics company: a case study." *Journal of technology management & innovation* 7, no. 2 (2012): 144-156. https://doi.org/10.4067/S0718-27242012000200012
[9] Anderson Jr, Edward G., Charles H. Fine, and Geoffrey G. Parker. "Upstream volatility in the supply chain: The machine tool industry as a case study." *Production and Operations Management* 9, no. 3 (2000): 239-261. https://doi.org/10.1111/j.1937-5956.2000.tb00136.x
[10] Díaz Pacheco, Raúl Antonio, and Ernest Benedito. "Supply chain response during the COVID-19 Pandemic: A multiple-case study." *Processes* 11, no. 4 (2023): 1218. https://doi.org/10.3390/pr11041218
[11] Shokrani, Alborz, Evripides G. Loukaides, Edward Elias, and Alexander JG Lunt. "Exploration of alternative supply chains and distributed manufacturing in response to COVID-19; a case study of medical face shields." *Materials & design* 192 (2020): 108749. https://doi.org/10.1016/j.matdes.2020.108749
[12] Vayansky I, Kumar SAP. A review of topic modeling methods. Inf Syst. 2020 Dec 1;94. https://doi.org/10.1016/j.is.2020.101582
[13] Alfajri, Alfajri, Donny Richasdy, and Muhammad Arif Bijaksana. "Topic modelling using non-negative matrix factorization (NMF) for telkom university entry selection from instagram comments." *Journal of Computer System and Informatics (JoSYC)* 3, no. 4 (2022): 485-492. https://doi.org/10.47065/josyc.v3i4.2212
[14] Supiadin, Muhamad Gatot, and Arif Dwi Laksito. "Evaluating LDA and LSA for Topic Modeling in the Indonesian Natural Disaster." *The Indonesian Journal of Computer Science* 12, no. 6 (2023). https://doi.org/10.33022/ijcs.v12i6.3478
[15] Kherwa P, Bansal P. Topic Modeling: A Comprehensive Review EAI Endorsed Transactions on Scalable Information Systems. 2019. https://doi.org/10.4108/eai.13-7-2018.159623
[16] Fatima-Zahrae, Sifi, and Sabbar Wafae. "Application of Latent Dirichlet Allocation (LDA) for clustering financial tweets." In *E3S Web of Conferences*, vol. 297, p. 01071. EDP Sciences, 2021. https://doi.org/10.1051/e3sconf/202129701071
[17] Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." *Journal of artificial intelligence research* 37 (2010): 141-188. https://doi.org/10.1613/jair.2934
[18] Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21, no. 3 (2013): 267-297. https://doi.org/10.1093/pan/mps028
[19] Amara, Amina, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. "Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis." *Applied Intelligence* 51, no. 5 (2021): 3052-3073. https://doi.org/10.1007/s10489-020-02033-3
[20] Prihatini, Putu Manik, I. Ketut Suryawan, and I. Nyoman Mandia. "Feature extraction for document text using Latent Dirichlet Allocation." In *Journal of Physics: Conference Series*, vol. 953, no. 1, p. 012047. IOP Publishing, 2018. https://doi.org/10.1088/1742-6596/953/1/012047
[21] Tyler D. LDA Topic Modeling: An Explanation | by Tyler Doll | Towards Data Science [Internet]. 2018 [cited 2024 Jun 14]. Available from: https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd
[22] Haaya N. Towards Data Science. 2020 [cited 2024 Apr 25]. Topic Modeling with Latent Dirichlet Allocation | by Haaya Naushan | Towards Data Science. Available from: https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8
[23] Ali, Imran, and Devika Kannan. "Mapping research on healthcare operations and supply chain management: a topic modelling-based literature review." *Annals of Operations Research* 315, no. 1 (2022): 29-55. https://doi.org/10.1007/s10479-022-04596-5
[24] Madzík, Peter, Lukáš Falát, and Dominik Zimon. "Supply chain research overview from the early eighties to Covid era–Big data approach based on Latent Dirichlet Allocation." *Computers & Industrial Engineering* 183 (2023): 109520. https://doi.org/10.1016/j.cie.2023.109520
[25] Baas, Jeroen, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi. "Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies." *Quantitative science studies* 1, no. 1 (2020): 377-386. https://doi.org/10.1162/qss_a_00019
[26] Trinh N. Extract Text From PDF Resumes Using PyMuPDF And Python [Internet]. 2022 [cited 2024 May 20]. Available from: https://www.neurond.com/blog/extract-text-from-pdf-pymupdf-and-python

[27] Churchill, Rob, and Lisa Singh. "textprep: A text preprocessing toolkit for topic modeling on social media data [textprep: A text preprocessing toolkit for topic modeling on social media data]." In *Proceedings of the 10th International Conference on Data Science, Technology and Applications*. 2021. https://doi.org/10.5220/0010559000002993
[28] Khyani, Divya, B. S. Siddhartha, N. M. Niveditha, and B. M. Divya. "An interpretation of lemmatization and stemming in natural language processing." *Journal of University of Shanghai for Science and Technology* 22, no. 10 (2021): 350-357.
[29] Kumar Rishi. Natural Language Processing | Text Preprocessing | Spacy vs NLTK | by Rishi Kumar | Nerd For Tech | Medium [Internet]. 2021 [cited 2024 May 4]. Available from: https://medium.com/nerd-for-tech/natural-language-processing-text-preprocessing-spacy-vs-nltk-b70b734f5560
[30] Muriuki P. COMPARISON OF NLTK AND SPACY LANGUAGE PROCESSING LIBRARIES. 2022; Available from: https://www.researchgate.net/publication/375632138
[31] Qader, Wisam A., Musa M. Ameen, and Bilal I. Ahmed. "An overview of bag of words; importance, implementation, applications, and challenges." In *2019 international engineering conference (IEC)*, pp. 200-204. IEEE, 2019. https://doi.org/10.1109/IEC47844.2019.8950616
[32] Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." *Political analysis* 26, no. 2 (2018): 168-189. https://doi.org/10.1017/pan.2017.44
[33] Bystrov, Victor, Viktoriia Naboka-Krell, Anna Staszewska-Bystrova, and Peter Winker. "Analysing the impact of removing infrequent words on topic quality in LDA models." *arXiv preprint arXiv:2311.14505* (2023).
[34] Daniel Maier A, Niekler A, Wiedemann G, Stoltenberg D. How document sampling and vocabulary pruning affect the results of topic models. 2019. https://doi.org/10.31219/osf.io/2rh6g
[35] Grimmer, Justin. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political analysis* 18, no. 1 (2010): 1-35. https://doi.org/10.1093/pan/mpp034
[36] Hopkins, Daniel J., and Gary King. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54, no. 1 (2010): 229-247. https://doi.org/10.1111/j.1540-5907.2009.00428.x
[37] Samuel cortinhas. NLP6 - Topic Modelling with LDA [Internet]. 2023 [cited 2024 May 7]. Available from: https://www.kaggle.com/code/samuelcortinhas/nlp6-topic-modelling-with-lda
[38] Shashank K. Topic Modeling in Python: Latent Dirichlet Allocation (LDA) | by Shashank Kapadia | Towards Data Science [Internet]. 2019 [cited 2024 May 7]. Available from: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
[39] Jacob M. Train an LDA topic model for text analysis in Python - IBM Developer [Internet]. 2024 [cited 2024 May 7]. Available from: https://developer.ibm.com/tutorials/awb-lda-topic-modeling-text-analysis-python/
[40] Radim Ř. models.ldamodel – Latent Dirichlet Allocation — gensim [Internet]. 2022 [cited 2024 May 5]. Available from: https://radimrehurek.com/gensim/models/ldamodel.html
[41] McLevey J. Doing Computational Social Science. London: Sage; 2022.
[42] Adesanya, Ayotunde, Biao Yang, Farok Wanes Bin Iqdara, and Ying Yang. "Improving sustainability performance through supplier relationship management in the tobacco industry." *Supply Chain Management: An International Journal* 25, no. 4 (2020): 413-426. https://doi.org/10.1108/SCM-01-2018-0034
[43] Chung, Sai-Ho. "Applications of smart technologies in logistics and transport: A review." *Transportation Research Part E: Logistics and Transportation Review* 153 (2021): 102455. https://doi.org/10.1016/j.tre.2021.102455
[44] Shahed, Kazi Safowan, Abdullahil Azeem, Syed Mithun Ali, and Md Abdul Moktadir. "A supply chain disruption risk mitigation model to manage COVID-19 pandemic risk." *Environmental Science and Pollution Research* (2021): 1-16. https://doi.org/10.1007/s11356-020-12289-4
[45] Frank, Markus, Brigitte Kaufmann, Mercedes Ejarque, María Guadalupe Lamaison, María Virginia Nessi, and Mariano Martin Amoroso. "Changing conditions for local food actors to operate towards agroecology during the COVID-19 pandemic." *Frontiers in sustainable food systems* 6 (2022): 866004. https://doi.org/10.3389/fsufs.2022.866004
[46] Erdoğan, Ahmet, and Mehmet Güray Güler. "Optimization and analysis of a hydrogen supply chain in terms of cost, CO2 emissions, and risk: the case of Turkey." *International Journal of Hydrogen Energy* 48, no. 60 (2023): 22752-22765. https://doi.org/10.1016/j.ijhydene.2023.04.300
[47] Bosona, Techane, and Girma Gebresenbet. "The role of blockchain technology in promoting traceability systems in agri-food production and supply chains." *Sensors* 23, no. 11 (2023): 5342. https://doi.org/10.3390/s23115342