# APPLYING THE SMART LITERATURE REVIEW TO SUPPLY CHAIN AND LOGISTICS RESEARCH

Tim Williams [1] *, Zena Wood [2], Roger Maull [3]
[1] Defence Data Research Centre, University of Exeter, UK, t.williams7@exeter.ac.uk
[2] Defence Data Research Centre, University of Exeter, UK, z.m.wood2@exeter.ac.uk
[3] Defence Data Research Centre, University of Exeter, UK, r.maull@exeter.ac.uk
*Corresponding Author

Disclaimer: The DDRC is funded by DSTL an agency of the UK Ministry of Defence. The views expressed in this research are solely those of the authors and do not reflect the opinions of any affiliated institutions. Any errors are the responsibility of the authors.

## Introduction

Artificial Intelligence (AI) and machine learning (ML) applications such as computer vision, natural language processing and robotics are transforming supply chain management (SCM) (McKinsey & Co, 2017). Highly cited, academic studies of supply chains offer interesting insights into technology types, algorithms and the areas where rapid advances are emerging (Kumar et al., 2023; Min, 2010; Pournader et al., 2021; Rana & Daultani, 2022; Toorajipour et al., 2021). These studies employ a range of different methodologies such as classic human-based literature reviews or network modelling such as co-citations. One significant methodological departure is by Asmussen and Møller (2019), who employ an exploratory Smart Literature Review (SLR) of '*analytics, SCM, and enterprise information systems*'. Their machine-based review optimally identifies 20 topics and illustrates a method for fast generation of SCM topics based on constituent contributing papers.

Conceptually, this may be an Ouroboros, the idea that AI is increasingly used for the analysis of AI and therefore such systems require safe, robust, effective and cyber-secure tests (MoD, 2022). A key theme of this paper is that tests, checks and human-machine teaming can be implemented in an exploratory SLR and we illustrate the challenges of the trade-offs between fast machine-based and classic, labour-intensive approaches to literature reviews in the supply chain area.

The literature review strategy in this paper uses the frequently used approach to search for the terms 'AI' OR 'ML' and combined with 'supply chain' OR 'logistics'. The method details the way high volumes of papers are collected, cleaned and screened and then summarised in a useful way. We note the process, pipeline and iterative nature of refinements that are specific to supply chain and logistics literature research.

There are an increasing number of free, open-source tools for SLR in python and R. The research described here is primarily built on litstudy (Heldens et al., 2022), which has the capacity to handle multiple literature databases through standardisation of fields. However, data cleaning which is the high fixed cost of any large or messy dataset is still helped by human observation of CSV files and best handled outside litstudy with staple python libraries like pandas.

## SLR process steps

Literature collection and screening in academic research has traditionally, been dependent and accepting that literature database providers return relevant information in a consistent order. That has been the standard but increasingly researchers will review literature samples with an interest in any AI or ML applied to select and order them. For example, an alternative to date-sorted papers is the 'most relevant' research but which is generated by algorithms that are proprietary and not transparent. Now, the SLR offers an opportunity to collect large volumes of data and then determine alternative views on what is the relevant and importantly, what is irrelevant.

Data collection for the SLR can by implemented by APIs or manually. There are advantages and disadvantages in each for efficiency, data validity and reliability. APIs are faster but less transparent, while UIs offer visual clarity but are slower. Both require significant careful attention to avoid human errors with collection that combines different literature sources: for example, literature database query syntax is often similar but not standardised between providers; basic and advanced queries *in the same service* can generate query strings that are not consistently swappable between UI and API. Applying further filters is an opportunity for further human errors when aggregating many manually exported queries. Export limits force batch operations requiring careful labelling of sets. A useful record of data provenance is a screen shot of the search conditions, filters applied and the sample size returned. A human check of the UI can identify dodgy API calls, poor search terms and actual set sizes.

Two significant issues in data collection are false negatives (omissions) and false positives (irrelevant literature). Aggregating material from multiple databases and concatenating multiple search queries can minimize false negatives. Ideally, duplicate removal is facilitated by a common unique identifier (e.g., DOI) but limitations exist. Additional data cleaning in open-source libraries (e.g., python pandas) is necessary and potentially costly.

### Special challenges with "logistics" false positives

Handling false positives is dependent on the area and exact nature of the literature search. In many fields, false positives may be insignificant but in supply chain research the term 'logistics' is a challenge. Literature databases include stemming even in exact match text queries[1]. Queries for "logistics" return both singular and plural forms, which is problematic due to the prevalence of "logistic regression" in academia.

To minimize false negatives in a logistics SLR, broad search queries are used, but this increases false positives, necessitating extensive filtering of irrelevant logistic regression studies. This issue is compounded in AI and ML research, where logistic regression (LR) is a crucial classification method that may also be relevant to supply chain logistics problems.

Traditional literature reviews include practical decisions for paper inclusion. For example, Rana and Daultani (2022) exclude supply chain articles with "*a sole focus on technical / engineering perspective*'. Potentially, this could be an exclusion decision by applied expert judgement, one enhanced by subject categories suggested by literature databases or, a combination of the two. In an SLR with a large set of supply chain logistics AI ML papers, determining inclusion/exclusion boundaries remains imperfect, often requiring practical human-based decisions.

[1] Stemming within exact match search was confirmed by correspondence with Elsevier.

## Options for screening false positives

There are several options for screening false positives. Firstly, literature databases offer filters such as for subject area classification. Quick but opaque, likely to be search limiting these filters will result in false negatives unless a search is technically narrow.

Secondly, search can be limited post-collection by using customised text matching. Transparent and potentially, low effort text pattern matches will generate some false positives and false negatives. For example, persistent false positive papers might include authors referring to 'logistical workload' or a dataset as an 'inventory' when they are not supply chain papers.

Thirdly, large language model (LLM) classifications can be applied to papers, titles or abstracts to score semantic meaning. Some expertise and resource are required. One approach for identifying false positives is to use the zero-shot classifier (ZSC) to screen abstracts (Moreno-Garcia et al., 2023; Wang et al., 2023). Potentially, a local, free open-source model can be used to score abstracts against any number of user-selected categories by a pre-trained model. Category scores can then be grouped and used to filter likely candidates for exclusion, inclusion or further review. In theory, the zero-shot approach is cheap because it reduces the expensive and difficult task of labelling data.

Fourthly, humans can label literature datasets to train models that classify specific subject areas, potentially offering a more cost-effective alternative to a full literature review. However, determining a consistent boundary for false positives is challenging due to emerging new areas and the subjective nature of reviewing papers. For AI and ML applied to supply chains, expertise is needed across diverse areas like hospital supply chains, energy supply chains, and credit supply chains which presents a complex task.

## Topic Modelling, Synthesis and Summarisation

In summary, there are several challenging parts to the pipeline for the SLR. If we can get a better, large set of valid and reliable literature for a SLR we can then look at different ways to synthesise the research into understandable concepts, topics or case studies. It may be that the noise of false positives is too small to affect output models but it is important to understand if it is occurring and what biases result.

In this paper, we only describe a Non-negative Matrix Factorisation (NMF) approach topic modelling. NMF is not probabilistic, not easy to mathematically optimise the number of topics but is fast. Models for several thousand papers can be generated in only a few seconds on a standard computer. The speed of NMF permits the comparison of multiple topic models. As larger ones are created it is possible to see the way finer granularity introduces new topics, adjusts topic size and topic order. For reviews, applications and alternative approaches for topic modelling there is a growing literature (Cao et al., 2023; Egger & Yu, 2022; Guo et al., 2024; O'Callaghan et al., 2015).

Asmussen and Møller (2019) conclude that topics produced by SLR have limitations. However, recent advances in LLMs suggest better synthesis of topics is possible. LLM summarisation and emerging bibliometric libraries could offer new outputs from a SLR. A key issue will be maintaining the development of human expertise in research with new machine tools.

## Method

Four searches of the Scopus database were conducted to find journal articles published between 2013-2023 that intersect the terms 'AI' or 'ML' with 'supply chain' or 'logistics.' Search results (n=3333) were combined after duplicates and papers with missing abstracts were removed. Further cleaning removed copyright text from abstracts for topic modelling.

Unfortunately, the sample contained a significant volume of false positive papers about logistic regression not logistics. We used variations of three of the methods described above to screen false positives: text matching, human labels and ZSC classifiers.

Text replacements were implemented (e.g. replace 'logistics regressions' with 'logistic regression') and string matches were used to filter the sample using the original terms (‘supply chain’, ‘logistics’) in the title and abstract. That is a stricter exact match than literature databases who use stemming. We did not include the index keywords field because that appeared to be where stemming may be applied. 2338 are included, 995 papers are excluded using text replacements.

To examine the validity of inclusion or exclusion of those two sets, we labelled a 10% random sample of each as relevant or irrelevant. Labels were allocated using a single person review of title/ abstract/ keyword. The review was blind of author names, country and institutional affiliations and the journal publication source to try to minimise any bias. Table 1 shows the sample sizes against human labels and text matching.

We also applied a ZSC to score the titles and abstracts in the two sets to try to identify valid papers. We used the default, most widely implemented facebook/bart-large-mnli model (Lewis et al., 2019) and we trialled many combinations of categories in batches to try to identify patterns for excluding papers (e.g. LR models not about supply chain logistics).

The ZSC distributes a score over user-supplied categories. We found that generally, using more detailed categories (> 10) was an improvement but some particular categories can be unhelpful as they seem to attract high scores and prevented a focus on supply chain issues (e.g. machine learning). By combining the scores of 20 categories for relevance or irrelevance, to a degree it is possible to apply a free LLM filter for context for supply chain research. The Include Group comprised: ["inventory management or control", "supply chain supplier selection", "demand forecasting", "production planning", "distribution management in supply chains or logistics", "warehouse management", "procurement", "supply chain logistics", "supplier networks", "health care supply chains", "military logistics"]. The Exclude Group contained: ["blockchain", "logistic regression", "biological studies", 'network security', 'health research', 'data security', 'legal and regulatory', 'cyber threats'].

The ZSC filtered documents into four groups based on classification scores using a universal threshold: Included, if any include categories were met; Excluded, if any exclude categories were met; Both, if both include and exclude conditions were met; Neither, if no categories met the threshold. Some human trial and error adjustments to the categories in the inclusion and exclusion groups helped reduce the latter two indeterminate sets. Further resource time allocated to this task is expected to improve the performance.

## Results

Applied to the 2338 papers the ZSC scoring took 7.4 hours and took the following view: 58% include; 17% exclude; 12% both; 13% neither. Applied to the 995 papers the ZSC scoring took 4.8 hours and took the following view: 14% include, 66% exclude, 13% both, 7% neither.

We can only take a real view on the performance of the ZSC using the samples with human labelled data. For the include group, the ZSC was consistent with the human labels at least 75% of the time (180/240). Greater performance is certainly possible as 80% of the inconclusive grouping had been human scored as true positives. For the reject set, 72/107 were classified correctly by ZCS (67%).

The labelled data can be used for future work on ML models in this area but is not part of this paper. Instead, we focus on the subsequent stage which is topic modelling of titles/ abstracts.

| **Collected data from Scopus n=3333** | **INCLUDE** | **EXCLUDE** |
|---|---|---|
| Text-based include / exclude e.g. "supply chain", "logistics" in Abstract or Title | 2338 | 995 |
| Human labelled sample | 240 (10%) | 107 (11%) |
| True positives/negatives by human labels | 204/240 (85%) | 88/107 (82%) |
| ZSC consistency with human labels | 180/240 (75%) | 72/107 (67%) |

TABLE 1: Sample adjustments for false positives: text-based, human labels and ZSC scores

## Topic modelling

We use the litstudy package to run topic models on a corpus formed from the titles and cleaned abstracts of the 2338 papers. Effectively, a genism NMF model is used to identify the most likely topics for a chosen number of topics. We illustrate a 30 topic model here and Table 2 shows the topics identified using 5 word summary lists. All the topics could potentially be valid and interesting but further examination is required. A further, useful observation is that a topic model can be used as a very fast way to identify a pattern of false positives.

To try to make topic models more useful and usable we have trialled a way to share them between researchers that allows them to quickly make checks and observe different ways in which topics emerge for different size models and compare them. We present the model outputs within a dashboard containing three things: firstly, documentation of the searches performed[2], secondly, non-executable code describing the collection, cleaning and topic modelling and thirdly, a tab for every topic. Topic tabs are helpful because they present the top words for that topic, cloud models, the top contributing abstracts and titles with direct URLs

[2] A data standard for smart literature reviews is a possibility. It would be useful to standardise search conditions (e.g. terms, dates, publications) for reproducibility.

to access the full paper. A helpful feature are sets of auto-text highlighting of the top topic words throughout the top abstracts which makes it clear how the abstract relates to the topic. The dashboard is portable and lightweight with no external dependencies. An important sharing limitation is that because abstracts are copyright the dashboard falls under any license agreement of the database provider.

Topic 1: ['risk', 'credit', 'financial', 'finance', 'enterprises']
Topic 2: ['agricultural', 'production', 'farming', 'farmers', 'smart']
Topic 3: ['vaccine', 'oil', 'gas', 'transportation', 'companies']
Topic 4: ['yield', 'crop', 'prediction', 'satellite', 'vegetation']
Topic 5: ['food', 'safety', 'agri', 'fraud', 'packaging']
Topic 6: ['blockchain', 'traceability', 'security', 'technology', 'transactions']
Topic 7: ['media', 'social_media', 'social', 'sentiment', 'online']
Topic 8: ['supplier', 'selection', 'criteria', 'fuzzy', 'evaluation']
Topic 9: ['biomass', 'energy', 'uav', 'bioenergy', 'maintenance']
Topic 10: ['iot', 'smart', 'internet', 'things', 'internet_things']
Topic 11: ['fashion', 'topics', 'research', 'education', 'green']
Topic 12: ['scheduling', 'manufacturing', 'production', 'problem', 'shop']
Topic 13: ['scm', 'management', 'application', 'risk', 'systematic']
Topic 14: ['port', 'container', 'shipping', 'terminal', 'maritime']
Topic 15: ['commerce', 'cross_border', 'border', 'cross', 'logistics']
Topic 16: ['digital', 'twin', 'technology', 'manufacturing', 'innovation']
Topic 17: ['rfid', 'tags', 'identification', 'radio_frequency', 'radio']
Topic 18: ['sustainable', 'waste', 'circular', 'environmental', 'price']
Topic 19: ['urban', 'cities', 'transportation', 'smart', 'freight']
Topic 20: ['inventory', 'retail', 'demand', 'cost', 'blood']
Topic 21: ['review', 'literature', 'research', 'systematic', 'future']
Topic 22: ['forecasting', 'demand', 'series', 'sales', 'prediction']
Topic 23: ['routing', 'vehicle', 'problem', 'solve', 'delivery']
Topic 24: ['image', 'classification', 'features', 'recognition', 'svm']
Topic 25: ['decision', 'decision_making', 'support', 'making', 'fuzzy']
Topic 26: ['drug', 'pharmaceutical', 'hospital', 'patients', 'medicines']
Topic 27: ['covid', 'pandemic', 'healthcare', 'resilience', 'disruptions']
Topic 28: ['attacks', 'detection', 'security', 'hardware', 'threats']
Topic 29: ['robotics', 'agent', 'autonomous', 'systems', 'automation']
Topic 30: ['ant', 'colony', 'ant_colony', 'aco', 'algorithm']

TABLE 2: A 30 topic model of AI ML Supply Chain Logistics

## Discussion and Conclusions

This paper contributes to the SLR method and generates useful outputs for understanding the impacts of AI and ML on supply chains and logistics. The topics generated are eclectic in nature but potentially interesting. For example, some are directed at emerging technologies (6, 10, 16), others at sector specific applications (2, 4, 26) whilst other appear to focus on themes such as sustainability or resilience (18, 27).

Space precludes a detailed discussion of each topic but the first one appears to concern risk, credit and finance. On further examination using our dashboard and examining the DOI links

we see the topic has some divergent themes around risk. The top ten papers include interesting papers such as financial risks in logistics companies (Yang, 2020) and risks in digital supply chain finance (Li et al., 2022). In the top ten, there are two risk related papers that may at first sight, be anomalies. One is on general work-related risks of AI but illustrated using the example of the transport and logistics sector (Hassel & Özkiziltan, 2023). The other examines supply chain risks for prefabricated buildings (Zhu & Liu, 2023). Such risk related insights, which a human-based literature review might overlook, demonstrate the model's ability to potentially, uncover diverse yet related research areas and enhancing literature reviews. The ability to generate such insights with machine models quickly is a valuable addition to any literature research activity.

The ZSC part of the project was experimental and only a partial success for classification. The process has limitations but we conclude that further experiments here will result in higher rates of success. We also explored LLM summarisation of topics using free-open source summarisers (e.g. hugging face) but concluded that a paid LLM service may be a faster more valid summariser.

Considerable effort is required to establish pipelines for a SLR but they are tools with highly re-usable elements. Effectively, a an SLR can become a fast and useful tool for examining any area of research. That will change the balance of resources for literature research and time can be spent usefully using SLR tools alongside more traditional routes for screening papers and identifying directions for investigation.

That said, there are some special challenges in particular research areas (e.g. logistics, defence) where ambiguous words confuse contextual and semantic meaning. Iterations of data cleaning are challenging in an SLR but it is possible to get to an interesting dashboard model view of several thousand abstracts even with imperfectly screened inputs. Topic models can even be quick and useful ways to spot patterns for subsequent data cleaning activities.

It is worth noting that the boundaries of a literature review are subjective. It is frequently difficult to decide if a paper should be included in a set for topic modelling. Consequently, it would be interesting to examine the increasing number of commercial AI and ML systems used by governments and other organisations for evaluating technology trends, supply chain risks and security. How do they exclude false positives from their models when volumes of data are large?

## References


- *Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data, 6(1).*
- *Cao, Q., Cheng, X., & Liao, S. (2023). A comparison study of topic modeling based literature analysis by using full texts and abstracts of scientific articles: a case of COVID-19 research. Library Hi Tech, 41(2), 543-569.*
- *Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts [Methods]. Frontiers in Sociology, 7.*
- *Guo, Y.-T., Li, Q.-Q., & Liang, C.-S. (2024). The rise of nonnegative matrix factorization: Algorithms and applications. Information Systems, 123, 102379.*
- *Hassel, A., & Özkiziltan, D. (2023). Governing the work-related risks of AI: implications for the German government and trade unions. Transfer: European Review of Labour and Research, 29(1), 71-86.*

- *Heldens, S., Sclocco, A., Dreuning, H., van Werkhoven, B., Hijma, P., Maassen, J., & van Nieuwpoort, R. V. (2022). litstudy: A Python package for literature reviews [Article]. SoftwareX, 20, Article 101207.*
- *Kumar, S., Lim, W. M., Sivarajah, U., & Kaur, J. (2023). Artificial Intelligence and Blockchain Integration in Business: Trends from a Bibliometric-Content Analysis. Information Systems Frontiers, 25(2), 871-896.*
- *Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv pre-print server.*
- *Li, Y., Stasinakis, C., & Yeo, W. M. (2022). A Hybrid XGBoost-MLP Model for Credit Risk Assessment on Digital Supply Chain Finance. Forecasting, 4(1), 184-207.*
- *McKinsey & Co. (2017). A road map for digitizing source-to-pay.*
- *Min, H. (2010). Artificial intelligence in supply chain management: theory and applications. International Journal of Logistics Research and Applications, 13(1), 13-39.*
- 
- *Moreno-Garcia, C. F., Jayne, C., Elyan, E., & Aceves-Martins, M. (2023). A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews. Decision Analytics Journal, 6, 100162.*
- *O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications, 42(13), 5645-5657.*
- *Pournader, M., Ghaderi, H., Hassanzadegan, A., & Fahimnia, B. (2021). Artificial intelligence applications in supply chain management [Review]. International Journal of Production Economics, 241, Article 108250.*
- *Rana, J., & Daultani, Y. (2022). Mapping the Role and Impact of Artificial Intelligence and Machine Learning Applications in Supply Chain Digital Transformation: A Bibliometric Analysis [Article]. Operations Management Research.*
- *Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. Journal of Business Research, 122, 502-517.*
- *Wang, Z., Pang, Y., & Lin, Y. (2023). Large Language Models Are Zero-Shot Text Classifiers. Computation and Language.*
- *Yang, B. (2020). Construction of logistics financial security risk ontology model based on risk association and machine learning. Safety Science, 123, 104437.*
- *Zhu, T., & Liu, G. (2023). A Novel Hybrid Methodology to Study the Risk Management of Prefabricated Building Supply Chains: An Outlook for Sustainability. Sustainability (Switzerland), 15(1).*