DISCUSSION PUBLISHERS
Research Center · Open Access Edition

Smart literature review: a practical topic modelling approach to exploratory literature review

Claus Boye Asmussen, Charles Mller

Discussion Research ID	DPR-2026-1185
Year	2019
Field	Social Sciences
DOI	10.1186/s40537-019-0255-7
Author ORCID	0000-0002-2998-2293
Licence	CC-BY
Original source	Journal Of Big Data
Permanent link	https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce
Publisher	Discussion Publishers discussionpublishers.com

Abstract

Abstract Manual exploratory literature reviews should be a thing of the past, as technology and development of machine learning methods have matured. The learning curve for using machine learning methods is rapidly declining, enabling new possibilities for all researchers. A framework is presented on how to use topic modelling on a large collection of papers for an exploratory literature review and how that can be used for a full literature review. The aim of the paper is to enable the use of topic modelling for researchers by presenting a step-by-step framework on a case and sharing a code template. The framework consists of three steps; pre-processing, topic modelling, and post-processing, where the topic model Latent Dirichlet Allocation is used. The framework enables huge amounts of papers to be reviewed in a transparent, reliable, faster, and reproducible way.

Republished by Discussion Publishers under the original CC-BY licence. Copyright remains with the original authors; please cite the original work.

Cite as: Claus Boye Asmussen, Charles Mller (2019). Smart literature review: a practical topic modelli. Discussion Publishers, DPR

Asmussen and Møller *J Big Data* (2019) 6:93
https://doi.org/10.1186/s40537-019-0255-7

Journal of Big Data

METHODOLOGY Open Access

Smart literature review: a practical topic modelling approach to exploratory literature review

Claus Boye Asmussen* and Charles Møller

*Correspondence:
cba@mp.aau.dk
Department of Materials and Production, Center for Industrial Production, Aalborg University, Fibigerstræde 16, 9220 Aalborg Øst, Denmark

Abstract

Manual exploratory literature reviews should be a thing of the past, as technology and development of machine learning methods have matured. The learning curve for using machine learning methods is rapidly declining, enabling new possibilities for all researchers. A framework is presented on how to use topic modelling on a large collection of papers for an exploratory literature review and how that can be used for a full literature review. The aim of the paper is to enable the use of topic modelling for researchers by presenting a step-by-step framework on a case and sharing a code template. The framework consists of three steps; pre-processing, topic modelling, and post-processing, where the topic model Latent Dirichlet Allocation is used. The framework enables huge amounts of papers to be reviewed in a transparent, reliable, faster, and reproducible way.

Keywords: Supply chain management, Latent Dirichlet Allocation, Topic modelling, Automatic literature review

Introduction

Manual exploratory literature reviews are soon to be outdated. It is a time-consuming process, with limited processing power, resulting in a low number of papers analysed. Researchers, especially junior researchers, often need to find, organise, and understand new and unchartered research areas. As a literature review in the early stages often involves a large number of papers, the options for a researcher is either to limit the amount of papers to review a priori or review the papers by other methods. So far, the handling of large collections of papers has been structured into topics or categories by the use of coding sheets [2, 12, 22], dictionary or supervised learning methods [30]. The use of coding sheets has especially been used in social science, where trained humans have created impressive data collections, such as the Policy Agendas Project and the Congressional Bills Project in American politics [30]. These methods, however, have a high upfront cost of time, requiring a prior understanding where papers are grouped by categories based on pre-existing knowledge. In an exploratory phase where a general overview of research directions is needed, many researchers may be dismayed by having to spend a lot of time before seeing any results, potentially wasting efforts that could have been better spent elsewhere. With the advancement of machine learning

Springer Open

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

methods, many of the issues can be dealt with at a low cost of time for the researcher. Some authors argue that when human processing such as coding practice is substituted by computer processing, reliability is increased and cost of time is reduced [12, 23, 30]. Supervised learning and unsupervised learning, are two methods for automatically processing papers [30]. Supervised learning relies on manually coding a training set of papers before performing an analysis, which entails a high cost of time before a result is achieved. Unsupervised learning methods, such as topic modelling, do not require the researcher to create coding sheets before an analysis, which presents a low cost of time approach for an exploratory review with a large collection of papers. Even though, topic modelling has been used to group large amounts of documents, few applications of topic modelling have been used on research papers, and a researcher is required to have programming skills and statistical knowledge to successfully conduct an exploratory literature review using topic modelling.

This paper presents a framework where topic modelling, a branch of the unsupervised methods, is used to conduct an exploratory literature review and how that can be used for a full literature review. The intention of the paper is to enable the use of topic modelling for researchers by providing a practical approach to topic modelling, where a framework is presented and used on a case step-by-step. The paper is organised as follows. The following section will review the literature in topic modelling and its use in exploratory literature reviews. The framework is presented in "Method" section, and the case is presented in "Framework" section. "Discussion" and "Conclusion" sections conclude the paper with a discussion and conclusion.

Topic modelling for exploratory literature review

While there are many ways of conducting an exploratory review, most methods require a high upfront cost of time and having pre-existent knowledge of the domain. Quinn et al. [30] investigated the costs of different text categorisation methods, a summary of which is presented in Table 1, where the assumptions and cost of the methods are compared.

What is striking is that all of the methods, except manually reading papers and topic modelling, require pre-existing knowledge of the categories of the papers and have a high pre-analysis cost. Manually reading a large amount of papers will have a high cost of time for the researcher, whereas topic modelling can be automated, substituting the use of the researcher's time with the use of computer time. This indicates a potentially good fit for the use of topic modelling for exploratory literature reviews.

The use of topic modelling is not new. However, there are remarkably few papers utilising the method for categorising research papers. It has been predominantly been used in the social sciences to identify concepts and subjects within a corpus of documents. An overview of applications of topic modelling is presented in Table 2, where the type of data, topic modelling method, the use case and size of data are presented.

The papers in Table 2 analyse web content, newspaper articles, books, speeches, and, in one instance, videos, but none of the papers have applied a topic modelling method on a corpus of research papers. However, [27] address the use of LDA for researchers and argue that there are four parameters a researcher needs to deal with, namely pre-processing of text, selection of model parameters and number of topics to be generated, evaluation of reliability, and evaluation of validity. The uses of topic

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Table 1 Summary of assumptions and costs of discrete text categorisation [30]

	Method				
	Reading	**Human coding**	**Dictionaries**	**Supervised learning**	**Topic model**
A. Assumptions					
Categories are known	No	Yes	Yes	Yes	No
Category nesting. If any, is known	No	Yes	Yes	Yes	No
Relevant text features are known	No	No	Yes	Yes	Yes
Mapping is known	No	No	Yes	No	No
Coding can be automated	No	No	Yes	Yes	Yes
B. Costs					
Preanalysis costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Moderate/high	High	High	High	Low
Analysis costs					
Person hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate/high	Moderate	Low	Low	Low
Postanalysis costs					
Person-hours spent interpreting	High	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

modelling are to identify themes or topics within a corpus of many documents, or to develop or test topic modelling methods. The motivation for most of the papers is that the use of topic modelling enables the possibility to do an analysis on a large amount of documents, as they would otherwise have not been able to due to the cost of time [30]. Most of the papers argue that LDA is a state-of-the-art and preferred method for topic modelling, which is why almost all of the papers have chosen the LDA method. The use of topic modelling does not provide a full meaning of the text but provides a good overview of the themes, which could not have been obtained otherwise [21]. DiMaggio et al. [12] find a key distinction in the use of topic modelling is that its use is more of utility than accuracy, where the model should simplify the data in an interpretable and valid way to be used for further analysis They note that a subject-matter expert is required to interpret the outcome and that the analysis is formed by the data.

The use of topic modelling presents an opportunity for researchers to add a tool to their tool box for an exploratory and literature review process. Topic modelling has mostly been used on online content and requires a high degree of statistical and technical skill, skills not all researchers possess. To enable more researchers to apply topic modelling for their exploratory literature reviews, a framework will be proposed to lower the requirements for technical and statistical skills of the researcher.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Table 2 Applications of topic modelling

Reference	Data	Method	Intended use	Size
DiMaggio et al. [12]	Newspapers	LDA	Identify central concepts in news coverage	8000
Grimmer [17]	Press releases	Own developed method	Development of a Bayesian topic model	24,000
Quinn et al. [30]	Speeches (Text)	Own developed method	Development of a statistical learning model	118,000
Jockers and Mimno [21]	Books	LDA	Identify literature themes	3346
Baum, [5]	Speeches (Video)	LDA	Identify topics of German politicians	2581
Ghosh and Guha [16]	Tweets	LDA	Identify tweets related to obesity	2,581,283
Evans [15]	Newspapers	LDA	Identify subjects of discussion	14,952
Guo et al. [19]	Tweets	Dictionary-based analysis and LDA	Compare dictionary-based analysis vs. LDA	77,000,000
Jacobi et al. [20]	Newspapers	LDA	Show the usefulness of LDA	Newspaper articles from 1945 to 2013
Maier et al. [27]	Web pages	LDA	Investigate the validity and reliability of LDA	344,456
Bonilla and Grimmer [9]	Newspapers	LDA	Investigate impact of terror alerts in US media	50,000
Elgesem et al. [13]	Blogs	LDA	Identify topics in blogs regarding the arrest of Edward Snowden	15,000
Elgesem et al. [14]	Blogs	LDA	Investigate how climate change is discussed in blogs	1,300,000
Koltsova and Koltcov [24]	Web forum posts	LDA	Investigate the political agenda of Russians in LiveJournal	> 100,000
Welbers et al. [4]	Newspapers	LDA	Validate the use of LDA	99,572
Parra et al. [29]	Tweets	LDA	Investigate how Twitter has been used in academic conferences	109,076

Topic modelling for exploratory literature review

Topic modelling has proven itself as a tool for exploratory analysis of a large number of papers [14, 24]. However, it has rarely been applied in the context of an exploratory literature review. The selected topic modelling method, for the framework, is Latent Dirichlet Allocation (LDA), as it is the most used [6, 12, 17, 20, 32], state-of-the-art method [25] and simplest method [8]. While other topic modelling methods could be considered, the aim of this paper is to enable the use of topic modelling for researchers. For enabling topic modelling for researchers, ease of use and applicability are highly rated, where LDA is easily implemented and understood. Other topic

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

modelling methods could potentially be used in the framework, where reviews of other topic models is presented in [1, 26].

The topic modelling method LDA is an unsupervised, probabilistic modelling method which extracts topics from a collection of papers. A topic is defined as a distribution over a fixed vocabulary. LDA analyses the words in each paper and calculates the joint probability distribution between the observed (words in the paper) and the unobserved (the hidden structure of topics). The method uses a 'Bag of Words' approach where the semantics and meaning of sentences are not evaluated. Rather, the method evaluates the frequency of words. It is therefore assumed that the most frequent words within a topic will present an aboutness of the topic. As an example, if one of the topics in a paper is LEAN, then it can be assumed that the words LEAN, JIT and Kanban are more frequent, compared to other non-LEAN papers. The result is a number of topics with the most prevalent topics grouped together. A probability for each paper is calculated for each topic, creating a matrix with the size of number of topics multiplied with the number of papers. A detailed description of LDA is found in [6].

The framework is designed as a step-by-step procedure, where its use is presented in a form of a case where the code used for the analysis is shared, enabling other researchers to easily replicate the framework for their own literature review. The code is based on the open source statistical language R, but any language with the LDA method is suitable for use. The framework can be made fully automated, presenting a low cost of time approach for exploratory literature reviews. An inspiration for the automation of the framework can be found in [10], who created an online-service, towards processing Business Process Management documents where text-mining approaches such as topic modelling are automated. They find that topic modelling can be automated and argue that the use of a good tool for topic modelling can easily present good results, but the method relies on the ability of people to find the right data, guide the analytical journey and interpret the results.

Method

The aim of the paper is to create a generic framework which can be applied in any context of an exploratory literature review and potentially be used for a full literature review. The method provided in this paper is a framework which is based upon well-known procedures for how to clean and process data, in such a way that the contribution from the framework is not in presenting new ways to process data but in how known methods are combined and used. The framework will be validated by the use of a case in the form of a literature review. The outcome of the method is a list of topics where papers are grouped. If the grouping of papers makes sense and is logical, which can be evaluated by an expert within the research field, then the framework is deemed valid. Compared to other methods, such as supervised learning, the method of measuring validity does not produce an exact degree of validity. However, invalid results will likely be easily identifiable by an expert within the field. As stated by [12], the use of topic modelling is more for utility than for accuracy.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Framework

The developed framework is illustrated in Fig. 1, and the R-code and case output files are located at https://github.com/clausba/Smart-Literature-Review. The smart literature review process consists of the three steps: pre-processing, topic modelling, and post-processing.

The pre-processing steps are getting the data and model ready to run, where the topic-modelling step is executing the LDA method. The post-processing steps are translating the outcome of the LDA model to an exploratory review and using that to identify papers to be used for a literature review. It is assumed that the papers for review are downloaded and available, as a library with the pdf files.

Pre-processing

The pre-processing steps consist of loading and preparing the papers for processing, an essential step for a good analytical result. The first step is to load the papers into the R environment. The next step is to clean the papers by removing or altering non-value-adding words. All words are converted to lower case, and punctuation and whitespaces are removed. Special characters, URLs, and emails are removed, as they often do not contribute to identification of topics. Stop words, misread words and other non-semantic contributing words are removed. Examples of stop words are "can", "use", and "make". These words add no value to the aboutness of a topic. The loading of papers into R can in some instances cause words to be misread, which must either be rectified or removed. Further, some websites add a first page with general information, and these contain words that must be removed. This prevents unwanted correlation between papers downloaded from the same source. Words are stemmed to their root form for easier comparison. Lastly, many words only occur in a single paper, and these should be removed to make computations easier, as less frequent words will likely provide little benefit in grouping papers into topics.

The cleansing process is often an iterative process, as it can be difficult to identify all misread and non-value adding-words a priori. Different papers' corpora contain different words, which means that an identical cleaning process cannot be guaranteed if a new exploratory review is conducted. As an example, different non-value-adding words exist for the medical field compared to sociology or supply chain management (SCM). The cleaning process is finished once the loaded papers mainly contain value-adding words.

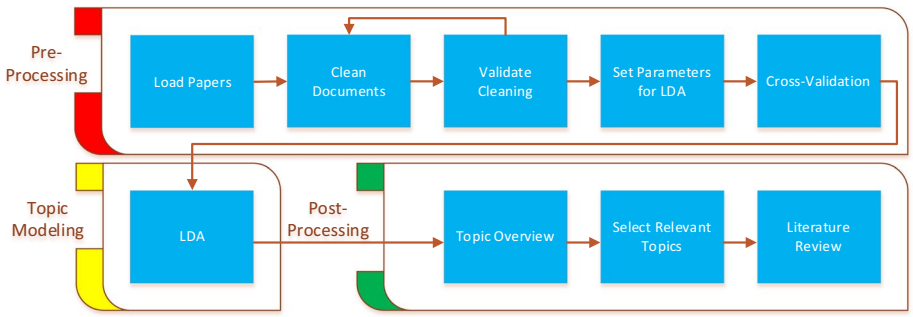


Fig. 1 Process overview of the smart literature review framework

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

There is no known way to scientifically evaluate when the cleaning process is finished, which in some instances makes the cleaning process more of an art than science. However, if a researcher is technically inclined methods, provided in the preText R-package can aid in making a better cleaning process [11].

LDA is an unsupervised method, which means we do not, prior to the model being executed, know the relationship between the papers. A key aspect of LDA is to group papers into a fixed number of topics, which must be given as a parameter when executing LDA. A key process is therefore to estimate the optimal number of topics. To estimate the number of topics, a cross-validation method is used to calculate the perplexity, as used in information theory, and it is a metric used to evaluate language models, where a low score indicates a better generalisation model, as done by [7, 31, 32]. Lowering the perplexity score is identical to maximising the overall probability of papers being in a topic. Next, test and training datasets are created: the LDA algorithm is run on the training set, and the test set is used to validate the results. The criteria for selecting the right number of topics is to find the balance between a useable number of topics and, at the same time, to keep the perplexity as low as possible. The right number of topics can differ greatly, depending on the aim of the analysis. As a rule of thumb, a low number of topics is used for a general overview and a higher number of topics is used for a more detailed view.

The cross-validation step is used to make sure that a result from an analysis is reliable, by running the LDA method several times under different conditions. Most of the parameters set for the cross-validation should have the same value, as in the final topic modelling run. However, due to computational reasons, some parameters can be altered to lower the amount of computation to save time. As with the number of topics, there is no right way to set the parameters, indicating a trial-and-error process. Most of the LDA implementations have default values set, but in this paper's case the following parameters were changed: burn-in time, number of iterations, seed values, number of folds, and distribution between training and test sets.

Topic modelling

Once the papers have been cleaned and a decision has been made on the number of topics, the LDA method can be run. The same parameters as used in the cross-validation should be used as a guidance but for more precise results, parameters can be changed such as a higher number of iterations. The number of folds should be removed, as we do not need a test set, as all papers will be used to run the model. The outcome of the model is a list of papers, a list of probabilities for each paper for each topic, and a list of the most frequent words for each topic.

If an update to the analysis is needed, new papers simply have to be loaded and the post-processing and topic modelling steps can be re-run without any alterations to the parameters. Thus, the framework enables an easy path for updating an exploratory review.

Post-processing

The aim of the post-processing steps is to identify and label research topics and topics relevant for use in a literature review. An outcome of the LDA model is a list of topic

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

probabilities for each paper. The list is used to assign a paper to a topic by sorting the list by highest probability for each paper for each topic. By assigning the papers to the topics with the highest probability, all of the topics contain papers that are similar to each other. When all of the papers have been distributed into their selected topics, the topics need to be labelled. The labelling of the topics is found by identifying the main topic of each topic group, as done in [17]. Naturally, this is a subjective matter, which can provide different labelling of topics depending on the researcher. To lower the risk of wrongly identified topics, a combination of reviewing the most frequent words for each topic and a title review is used. After the topics have been labelled, the exploratory search is finished.

When the exploratory search has finished, the results must be validated. There are three ways to validate the results of an LDA model, namely statistical, semantic, or predictive [12]. Statistical validation uses statistical methods to test the assumptions of the model. An example is [28], where a Bayesian approach is used to estimate the fit of papers to topics. Semantic validation is used to compare the results of the LDA method with expert reasoning, where the results must make semantic sense. In other words, does the grouping of papers into a topic make sense, which ideally should be evaluated by an expert. An example is [18], who utilises hand coding of papers and compare the coding of papers to the outcome of an LDA model. Predictive validation is used if an external incident can be correlated with an event not found in the papers. An example is in politics where external events, such as presidential elections which should have an impact on e.g. press releases or newspaper coverage, can be used to create a predictive model [12, 17].

The chosen method for validation in this framework is semantic validation. The reason is that a researcher will often be or have access to an expert who can quickly validate if the grouping of papers into topics makes sense or not. Statistical validation is a good way to validate the results. However, it would require high statistical skills from the researchers, which cannot be assumed. Predictive validation is used in cases where external events can be used to predict the outcome of the model, which is seldom the case in an exploratory literature review.

It should be noted that, in contrast to many other machine learning methods, it is not possible to calculate a specific measure such as the F-measure or RMSE. To be able to calculate such measures, there must exist a correct grouping of papers, which in this instance would often mean comparing the results to manually created coding sheets [11, 19, 20, 30]. However, it is very rare that coding sheets are available, leaving the semantic validation approach as the preferred validation method. The validation process in the proposed framework is two-fold. Firstly, the title of the individual paper must be reviewed to validate that each paper does indeed belong in its respective topic. As LDA is an unsupervised method, it can be assumed that not all papers will have a perfect fit within each topic, but if the majority of papers are within the theme of the topic, it is evaluated to be a valid result. If the objective of the research is only an exploratory literature review, the validation ends here. However, if a full literature review is conducted, the literature review can be viewed as an extended semantic validation method. By reviewing the papers in detail within the selected topics of research, it can be validated if the vast majority of papers belong together.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Using the results from the exploratory literature review for a full literature review is simple, as all topics from the exploratory literature review will be labelled. To conduct the full literature review, select the relevant topics and conduct the literature review on the selected papers.

Result

To validate the framework, a case will be presented, where the framework is used to conduct a literature review. The literature review is conducted in the intersection of the research fields analytics, SCM, and enterprise information systems [3]. As the research areas have a rapidly growing interest, it was assumed that the number of papers would be large, and that an exploratory review was needed to identify the research directions within the research fields. The case used broadly defined keywords for searching for papers, ensuring to include as many potentially relevant papers as possible. Six hundred and fifty papers were found, which were heavily reduced by the use of the smart literature review framework to 76 papers, resulting in a successful literature review. The amount of papers is evaluated to be too time-consuming for a manual exploratory review, which provides a good case to test the smart literature review framework. The steps and thoughts behind the use of the framework are presented in this case section.

Pre-processing

The first step was to load the 650 papers into the R environment. Next, all words were converted to lowercase and punctuation, whitespaces, email addresses, and URLs were removed. Problematic words were identified, such as words incorrectly read from the papers. Words included in a publisher's information page were removed, as they add no semantic value to the topic of a paper. English stop words were removed, and all words were stemmed. As a part of an iterative process, several papers were investigated to evaluate the progress of cleaning the papers. The investigations were done by displaying words in a console window and manually evaluating if more cleaning had to be done.

After the cleaning steps, 256,747 unique words remained in the paper corpus. This is a large number of unique words, which for computational reasons is beneficial to reduce. Therefore, all words that did not have a sparsity or likelihood of 99% to be in any paper were removed. The operation lowered the amount of unique words to 14,145, greatly reducing the computational needs. The LDA method will be applied on the basis of the 14,145 unique words for the 650 papers. Several papers were manually reviewed, and it was evaluated that removal of the unique words did not significantly worsen the ability to identify main topics of the paper corpus.

The last step of pre-processing is to identify the optimal number of topics. To approximate the optimal number of topics, two things were considered. The perplexity was calculated for different amounts of topics, and secondly the need for specificity was considered.

At the extremes, choosing one topic would indicate one topic covering all papers, which will provide a very coarse view of the papers. On the other hand, if the number of topics is equal to the number of papers, then a very precise topic description will be achieved, although the topics will lose practical use as the overview of topics will be too complex. Therefore, a low number of topics was preferred as a general overview was

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

required. Identifying what is a low number of topics will differ depending on the corpus of papers, but visualising the perplexity can often provide the necessary aid for the decision.

The perplexity was calculated over five folds, where each fold would identify 75% of the papers for training the model and leave out the remaining 25% for testing purposes. Using multiple folds reduces the variability of the model, ensuring higher reliability and reducing the risk of overfitting. For replicability purposes, specific seed values were set. Lastly, the number of topics to evaluate is selected. In this case, the following amounts of topics were selected: 2, 3, 4, 5, 10, 20, 30, 40, 50, 75, 100, and 200. The perplexity method in the 'topicmodels' R library is used, where the specific parameters can be found in the provided code.

The calculations were done over two runs. However, there is no practical reason for not running the calculations in one run. The first run included all values of number of topics below 100, and the second run calculated the perplexity for 100 and 200 number of topics. The runtimes for the calculations were respectively 9 and 10 h on a standard issue laptop. The combined results are presented in Fig. 2, and the converged results can be found in the shared repository.

The goal in this case is to find the lowest number of topics, which at the same time have a low perplexity. In this case, the slope of the fitted line starts to gradually decline at twenty topics, which is why the selected number of topics is twenty.

Case: topic modelling

As the number of topics is chosen, the next step is to run the LDA method on the entire set of papers. The full run of 650 papers for 20 topics took 3.5 h to compute on a standard issue laptop. An outcome of the method is a 650 by 20 matrix of topic probabilities. In this case, the papers with the highest probability for each topic were used to allocate the papers. The allocation of papers to topics was done in Microsoft Excel. An example of how a distribution of probabilities is distributed across

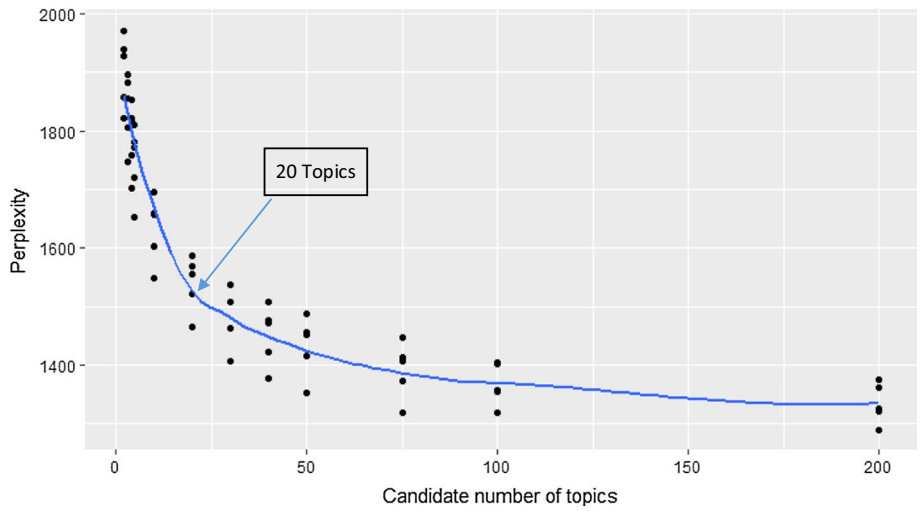


Fig. 2 5-Fold cross-validation of topic modelling. Results of cross-validation

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

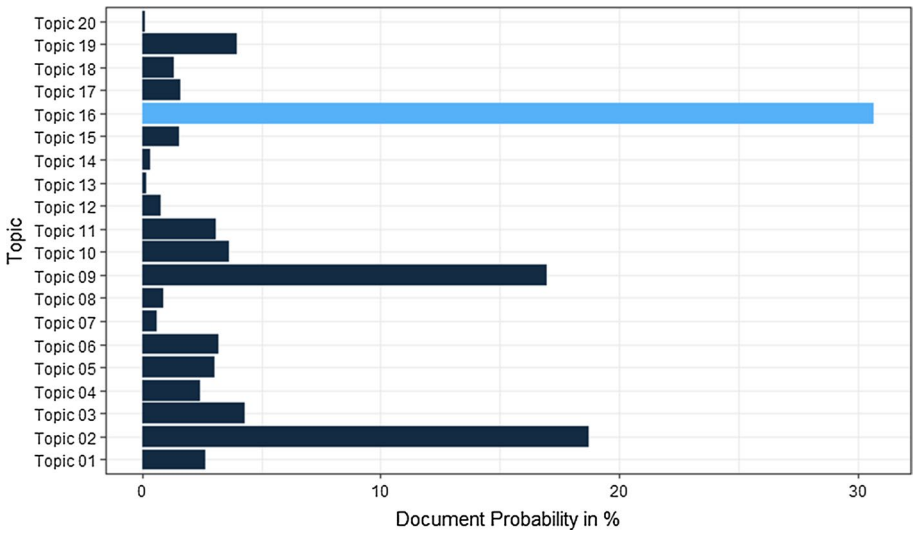


Fig. 3 Example of probability distribution for one document (Topic 16 selected)

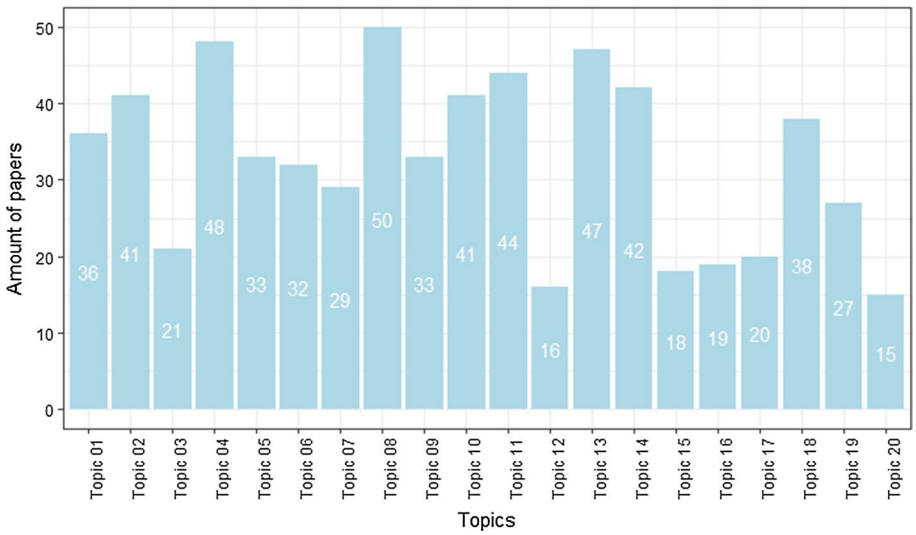


Fig. 4 Distribution of papers per topic

topics for a specific paper is depicted in Fig. 3. Some papers have topic probability values close to each other, which could indicate a paper belonging to an intersection between two or more topics. These cases were not considered, and the topic with the highest probability was selected.

The allocation of papers to topics resulted in the distribution depicted in Fig. 4. As can be seen, the number of papers varies for each topic, indicating that some research areas have more publications than others do.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Table 3 Paper titles and ten most frequent words for topic 17

Document titles	Top 10 words by frequency
A perspective on applications of in-memory analytics in supply chain management	Data
A tool to evaluate the business intelligence of enterprise systems	Big
A service oriented approach to Business Intelligence in Telecoms industry	Analyt
An analytic infrastructure for harvesting big data to enhance supply chain performance	Applic
Big data applications in operations/supply-chain management: a literature review	Analysi
Big data driven customer insights for SMEs in redistributed manufacturing	Decis
Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives	Busi
Click here for a data scientist: big data, predictive analytics, and theory development in the era of a maker movement supply chain	Predict
Coping with demand volatility in retail pharmacies with the aid of big data exploration	Comput
CRM in social media: predicting increases in Facebook usage frequency	Manag
Holistic approach to machine tool data analytics	
Impact of business analytics and enterprise systems on managerial accounting	
Industrial materials informatics: analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain	
Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research	
Intelligent business processes in CRM	
Machine-learning techniques for customer retention a comparative study	
Managing a Big Data project The case of Ramco Cements Limited	
Past, present and future of contact centers a literature review	
Utilizing enterprise systems for managing enterprise risks	
Visual analytics for supply network management: system design and evaluation	

Table 4 Names of topics 1–10

Topic 1: Demand and inventory decision models and systems	Topic 2: Fuzzy decision models	Topic 3: SCM and manufacturing planning systems	Topic 4: Scheduling and optimisation models	Topic 5: Data and enterprise system configuration
Topic 6: Price, supplier and contract policies and plans	Topic 7: Optimisation and configuration of switched reluctance drive	Topic 8: Route and job scheduling	Topic 9: Performance measurement	Topic 10: Implementation and system selection

Post-processing

Next step is to process the findings and find an adequate description of the topics. A combination of reviewing the most frequent words and a title review was used to identify the topic names. Practically, all of the paper titles and the most frequent words for each topic, were transferred to a separate Excel spreadsheet, providing an easy overview of paper titles. An example for topic 17 can be seen in Table 3. The most frequent words for the papers in topic 17 are "data", "big" and "analyt". Many of the paper titles also indicate usage of big data and analytics for application in a business setting. The topic is named "Big Data Analytics".

The process was repeated for all other topics. The names of the topics are presented in Tables 4 and 5.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Table 5 Names of topics 11–20

Topic 11: Marketing and CRM	Topic 12: Misc	Topic 13: Implementation and integration of ERP systems	Topic 14: MRP planning methods	Topic 15: Transportation methods and models
Topic 16: Supplier selection	Topic 17: Big data analytics	Topic 18: Production and manufacturing system and models	Topic 19: Knowledge and process management of IT-systems	Topic 20: Misc

Table 6 Overview of sub-topic of papers

Name	Count
ERP Implementation and Post-Implementation	37
EIS and Analytics	16
Data and System integration	7
Literature Review	7
RFID	6
Evaluating and Selection of IT Systems	7
Analytical Methods	6
Networked Manufactoring and ERP systems	6
Performance Measurement	4
Data and Analytics	3

Based on the names of the topics, three topics were selected based on relevancy for the literature review. Topics 5, 13, and 17 were selected, with a total of 99 papers. In this specific case, it was deemed that there might be papers with a sub-topic that is not relevant for the literature review. Therefore, an abstract review was conducted for the 99 papers, creating 10 sub-topics, which are presented in Table 6.

The sub-topics RFID, Analytical Methods, Performance Management, and Evaluation and Selection of IT Systems were evaluated to not be relevant for the literature review. Seventy-six papers remained, grouped by sub-topics.

The outcome of the case was an overview of the research areas within the paper corpus, represented by the twenty topics and the ten sub-topics. The selected sub-topics were used to conduct a literature review. The validation of the framework consisted of two parts. The first part addressed the question of whether the grouping of papers, evaluated by the title and keywords, makes sense and the second part addressed whether the literature review revealed any misplaced papers. The framework did successfully place the selected papers into groups of papers that resemble each other. There was only one case where a paper was misplaced, namely that a paper about material informatics was placed among the papers in the sub-topic EIS and Analytics. The grouping and selection of papers in the literature review, based on the framework, did make semantic sense and was successfully used for a literature review. The framework has proven its utility in enabling a faster and more comprehensive exploratory literature review, as compared to competing methods. The framework has increased the speed for analysing a large amount of papers, as well as having increased the reliability in comparison with manual reviews as the same result can be obtained by running the analysis once again. The transparency in the framework is higher than in competing methods, as all steps of the framework are recorded in the code and output files.

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Discussion

This paper presents an approach not often found in academia, by using machine learning to explore papers to identify research directions. Even though the framework has its limitations, the results and ease of use leave a promising future for topic-modelling-based exploratory literature reviews.

The main benefit of the framework is that it provides information about a large number of papers, with little effort on the researcher's part, before time-costly manual work is to be done. It is possible, by the use of the framework, to quickly navigate many different paper corpora and evaluate where the researchers' time and focus should be spent. This is especially valuable for a junior researcher or a researcher with little prior knowledge of a research field. If default parameters and cleaning settings can be found for the steps in the framework, a fully automatic grouping of papers could be enabled, where very little work has to be done to achieve an overview of research directions. From a literature review perspective, the benefit of using the framework is that the decision to include or exclude papers for a literature review will be postponed to a later stage where more information is provided, resulting in a more informed decision-making process. The framework enables reproducibility, as all of the steps in the exploratory review process can be reproduced, and enables a higher degree of transparency than competing methods do, as the entire review process can, in detail, be evaluated by other researchers.

There is practically no limit of the number of papers the framework is able to process, which could enable new practices for exploratory literature reviews. An example is to use the framework to track the development of a research field, by running the topic modelling script frequently or when new papers are published. This is especially potent if new papers are automatically downloaded, enabling a fully automatic exploratory literature review. For example, if an exploratory review was conducted once, the review could be updated constantly whenever new publications are made, grouping the publications into the related topics. For this, the topic model has to be trained properly for the selected collection of papers, where it can be assumed that minor additions of papers would likely not warrant any changes to the selected parameters of the model. However, as time passes and more papers are processed, the model will learn more about the collection of papers and provide a more accurate and updated result. Having an automated process could also enable a faster and more reliable method to do post-processing of the results, reducing the post-analysis cost identified for topic modelling by [30], from moderate to low.

The framework is designed to be easily used by other researchers by designing the framework to require less technical knowledge than a normal topic model usage would entail and by sharing the code used in the case work. The framework is designed as a step-by-step approach, which makes the framework more approachable. However, the framework has yet not been used by other researchers, which would provide valuable lessons for evaluating if the learning curve needs to be lowered even further for researchers to successfully use the framework.

There are, however, considerations that must be addressed when using the smart literature review framework. Finding the optimal number of topics can be quite difficult, and the proposed method of cross-validation based on the perplexity presented a good, but not optimal, solution. An indication of why the number of selected topics is not

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

optimal is the fact that it was not possible to identify a unifying topic label for two of the topics. Namely topics 12 and 20, which were both labelled miscellaneous. The current solution to this issue is to evaluate the relevancy of every single paper of the topics that cannot be labelled. However, in future iterations of the framework, a better identification of the number of topics must be developed. This is a notion also recognised by [6], who requested that researchers should find a way to label and assign papers to a topic other than identifying the most frequent words. An attempt was made by [17] to generate automatic labelling on press releases, but it is uncertain if the method will work in other instances. Overall, the grouping of papers in the presented case into topics generally made semantic sense, where a topic label could be found for the majority of topics.

A consideration when using the framework is that not all steps have been clearly defined, and, e.g., the cleaning step is more of an art than science. If a researcher has no or little experience in coding or executing analytical models, suboptimal results could occur. [11, 25, 27] find that especially the pre-processing steps can have a great impact on the validity of results, which further emphasises the importance of selecting model parameters. However, it is found that the default parameters and cleaning steps set in the code provided a sufficiently valid and useable result for an exploratory literature analysis. Running the code will not take much of the researcher's time, as the execution of code is mainly machine time, and verifying the results takes a limited amount of a researcher time.

Due to the semantic validation method used in the framework, it relies on the availability of a domain expert. The domain expert will not only validate if the grouping of papers into topics makes sense, but it is also their responsibility to label the topics [12]. If a domain expert is not available, it could lead to wrongly labelled topics and a non-valid result.

A key issue with topic modelling is that a paper can be placed in several related topics, depending on the selected seed value. The seed value will change the starting point of the topic modelling, which could result in another grouping of papers. A paper consists of several sub-topics and depending on how the different sub-topics are evaluated, papers can be allocated to different topics. A way to deal with this issue is to investigate papers with topic probabilities close to each other. Potential wrongly assigned papers can be identified and manually moved if deemed necessary. However, this presents a less automatic way of processing the papers, where future research should aim to improve the assignments of papers to topics or create a method to provide an overview of potentially misplaced papers. It should be noted that even though some papers can be misplaced, the framework provides outcome files than can easily be viewed to identify misplaced papers, by a manual review.

As the smart literature review framework heavily relies on topic modelling, improvements to the selected topic model will likely present better results. The results of the LDA method have provided good results, but more accurate results could be achieved if the semantic meaning of the words would be considered. The framework has only been tested on academic papers, but there is no technical reason to not include other types of documents. An example is to use the framework in a business context to analyse meeting minutes notes to analyse the discussion within the different departments in a company. For this to work, the cleaning parameters would likely have to change, and another

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

evaluation method other than a literature review would be applicable. Further, the applicability of the framework has to be assessed on other streams of literature to be certain of its use for exploratory literature reviews at large.

Conclusion

This paper aimed to create a framework to enable researchers to use topic modelling to, do an exploratory literature review, decreasing the need for manually reading papers and, enabling the possibility to analyse a greater, almost unlimited, amount of papers, faster, more transparently and with greater reliability. The framework is based upon the use of the topic model Latent Dirichlet Allocation, which groups related papers into topic groups. The framework provides greater reliability than competing exploratory review methods provide, as the code can be rerun on the same papers, which will provide identical results. The process is highly transparent, as most decisions made by the researcher can be reviewed by other researchers, unlike, e.g., in the creation of coding sheets. The framework consists of three main phases: Pre-processing, Topic Modelling, and Post-Processing. In the pre-processing stage, papers are loaded, cleaned, and cross-validated, where recommendations to parameter settings are provided in the case work, as well as in the accompanied code. The topic modelling step is where the LDA method is executed, using the parameters identified in the pre-processing step. The post-processing step creates outputs from the topic model and addresses how validity can be ensured and how the exploratory literature review can be used for a full literature review. The framework was successfully used in a case with 650 papers, which was processed quickly, with little time investment from the researcher. Less than 2 days was used to process the 650 papers and group them into twenty research areas, with the use of a standard laptop. The results of the case are used in the literature review by [3].

The framework is seen to be especially relevant for junior researchers, as they often need an overview of different research fields, with little pre-existing knowledge, where the framework can enable researchers to review more papers, more frequently.

For an improved framework, two main areas need to be addressed. Firstly, the proposed framework needs to be applied by other researchers on other research fields to gain knowledge about the practicality and gain ideas for further development of the framework. Secondly, research in how to automatically identity model parameters could greatly improve the usability for the use of topic modelling for non-technical researchers, as the selection of model parameters has a great impact on the result of the framework.

Abbreviations
LDA: Latent Dirichlet Allocation; SCM: supply chain management.

Acknowledgements
Not applicable.

Authors' contributions
CBA wrote the paper, developed the framework and executed the case. CM Supervised the research and developed the framework. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
https://github.com/clausba/Smart-Literature-Review (No data).

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

Competing interests
The authors declare that they have no competing interests.

Received: 26 July 2019 Accepted: 2 October 2019
Published online: 19 October 2019

References

1. Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. Int J Adv Comput Sci Appl. 2015;6(1):7. https://doi.org/10.14569/IJACSA.2015.060121.
2. Ansolabehere S, Snowberg EC, Snyder JM. Statistical bias in newspaper reporting on campaign finance. Public Opin Quart. 2003. https://doi.org/10.2139/ssrn.463780.
3. Asmussen CB, Møller C. Enabling supply chain analytics for enterprise information systems: a topic modelling literature review. Enterprise Information Syst. 2019. **(Submitted To)**.
4. Atteveldt W, Welbers K, Jacobi C, Vliegenthart R. LDA models topics… But what are "topics"? In: Big data in the social sciences workshop. 2015. http://vanatteveldt.com/wp-content/uploads/2014_vanatteveldt_glasgowbigdata_topics.pdf.
5. Baum D. Recognising speakers from the topics they talk about. Speech Commun. 2012;54(10):1132–42. https://doi.org/10.1016/j.specom.2012.06.003.
6. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84. https://doi.org/10.1145/2133806.2133826.
7. Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat. 2007;1(1):17–35. https://doi.org/10.1214/07-AOAS114.
8. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res. 2003;3:993–1022. https://doi.org/10.5555/944919.944937.
9. Bonilla T, Grimmer J. Elevated threat levels and decreased expectations: how democracy handles terrorist threats. Poetics. 2013;41(6):650–69. https://doi.org/10.1016/j.poetic.2013.06.003.
10. Brocke JV, Mueller O, Debortoli S. The power of text-mining in business process management. BPTrends.
11. Denny MJ, Spirling A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit Anal. 2018;26(2):168–89. https://doi.org/10.1017/pan.2017.44.
12. DiMaggio P, Nag M, Blei D. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. Poetics. 2013;41(6):570–606. https://doi.org/10.1016/j.poetic.2013.08.004.
13. Elgesem D, Feinerer I, Steskal L. Bloggers' responses to the Snowden affair: combining automated and manual methods in the analysis of news blogging. Computer Supported Cooperative Work: CSCW. Int J. 2016;25(2–3):167–91. https://doi.org/10.1007/s10606-016-9251-z.
14. Elgesem D, Steskal L, Diakopoulos N. Structure and content of the discourse on climate change in the blogosphere: the big picture. Environ Commun. 2015;9(2):169–88. https://doi.org/10.1080/17524032.2014.983536.
15. Evans MS. A computational approach to qualitative analysis in large textual datasets. PLoS ONE. 2014;9(2):1–11. https://doi.org/10.1371/journal.pone.0087908.
16. Ghosh D, Guha R. What are we "tweeting" about obesity? Mapping tweets with topic modeling and geographic information system. Cartogr Geogr Inform Sci. 2013;40(2):90–102. https://doi.org/10.1080/15230406.2013.776210.
17. Grimmer J. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. Polit Anal. 2010;18(1):1–35. https://doi.org/10.1093/pan/mpp034.
18. Grimmer J, Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal. 2013;21(03):267–97. https://doi.org/10.1093/pan/mps028.
19. Guo L, Vargo CJ, Pan Z, Ding W, Ishwar P. Big social data analytics in journalism and mass communication. J Mass Commun Quart. 2016;93(2):332–59. https://doi.org/10.1177/1077699016639231.
20. Jacobi C, Van Atteveldt W, Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit J. 2016;4(1):89–106. https://doi.org/10.1080/21670811.2015.1093271.
21. Jockers ML, Mimno D. Significant themes in 19th-century literature. Poetics. 2013;41(6):750–69. https://doi.org/10.1016/j.poetic.2013.08.005.
22. Jones BD, Baumgartner FR. The politics of attention: how government prioritizes problems. Chicago: University of Chicago Press; 2005.
23. King G, Lowe W. An automated information extraction tool for international conflict data with performance as good as human coders: a rare events evaluation design. Int Org. 2008;57:617–43. https://doi.org/10.1017/s0020818303573064.
24. Koltsova O, Koltcov S. Mapping the public agenda with topic modeling: the case of the Russian LiveJournal. Policy Internet. 2013;5(2):207–27. https://doi.org/10.1002/1944-2866.POI331.
25. Lancichinetti A, Irmak Sirer M, Wang JX, Acuna D, Körding K, Amaral LA. High-reproducibility and high-accuracy method for automated topic classification. Phys Rev X. 2015;5(1):1–11. https://doi.org/10.1103/PhysRevX.5.011007.
26. Mahmood A. Literature survey on topic modeling. Technical Report, Dept. of CIS, University of Delaware Newark, Delaware. http://www.eecis.udel.edu/~vijay/fall13/snlp/lit-survey/TopicModeling-ASM.pdf. 2009.
27. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, Adam S. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun Methods Meas. 2018;12(2–3):93–118. https://doi.org/10.1080/19312458.2018.1430754.
28. Mimno D, Blei DM. Bayesian checking for topic models. In: EMLP 11 proceedings of the conference on empirical methods in natural language processing. 2011. p 227–37. https://doi.org/10.5555/2145432.2145459

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce

29. Parra D, Trattner C, Gómez D, Hurtado M, Wen X, Lin YR. Twitter in academic events: a study of temporal usage, communication, sentimental and topical patterns in 16 Computer Science conferences. Comput Commun. 2016;73:301–14. https://doi.org/10.1016/j.comcom.2015.07.001.
30. Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR. How to analyze political attention. Am J Polit Sci. 2010;54(1):209–28. https://doi.org/10.1111/j.1540-5907.2009.00427.x.
31. Xu Z, Raschid L. Probabilistic financial community models with Latent Dirichlet Allocation for financial supply chains. In: DSMM'16 proceedings of the second international workshop on data science for macro-modeling. 2016. https://doi.org/10.1145/2951894.2951900.
32. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinform. 2015;16(13):S8. https://doi.org/10.1186/1471-2105-16-S13-S8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com

Discussion Publishers edition DPR-2026-1185 https://discussionpublishers.com/research/0ff56493-0bce-419b-9cf3-ed78321802ce